Evolution, Consciousness, and the Internality of the Mind

I think we can cast light on the problems of mind -- and in particular the problem of consciousness -- by taking them in the perspective of evolution. In this we consider the mind, and also our conception of the mind, as adaptations produced by natural selection. We can do so even in the absence of detailed hypotheses about their mode of evolution; and this heightens our sense of the distinction between having a mind and, in addition, having a conception of the mind.

It seems that many animals have minds, at least in the sense that they experience or perceive how the world is, seek to attain their goals in light of what they learn, are subject to emotions comparable to anger, lust, and fear, and so on. Simplifying, let us say that to have a mind is in part to have a system of *internal representations* which makes it possible to attain goals *intelligently*, that is, in light of a variety of information which bears on their attainment. To have a *conception* of this sort of mind, by contrast, is thus to represent not only the attainment of goals in the environment, but also the inner representations which make this possible, and which govern the intelligent behavior of creatures with minds.

It is unclear how far animals besides ourselves have such a conception, and so for example can consider a range of goals to choose among them, or seek to know or influence the minds of others. This, in turn, seems related to a difference in adaptive function. The advantages of possessing a mind would seem to derive from representing the self in relation to the environment, including other creatures and their behavior. Those of possessing a conception of mind, by contrast, apparently flow from representing these behavior-governing representations themselves, and so further anticipating and controlling the behavior which they regulate.

In this perspective the problems of the mental and the physical are not those of relating two distinct worlds or realms of being; rather they are those of reconciling two different ways of thinking, both natural to us, and both serving to represent significant and connected aspects of the natural world which we inhabit. And since we are seeking to understand both ways of thinking as representing linked aspects of the same world, we can expect to find that they cohere with one another, and that their relations are not ultimately mysterious.

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

We take the behavior-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 is witnessed by the fact that we with all our science are scarcely beginning to fathom them.) So, as we can say, our species -- or our nervous system -- has had to evolve the capacity to represent behavior-governing states and events in nervous systems *blindly*, that is, without making use of perceptual information about neurons as disposed in space and having features perceptibly related to the computational and controlling operations which they perform.

This seems part of the reason why in employing our concept mind we do not represent the neural causes of behavior *as such*, but rather *by relation to observable behavior and the perceivable environment*. In thinking of persons as having minds we construe their behavior 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 behavior, built up through evolution, and programmed, in Chomsky's phrase, to grow in the mind (or brain).

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

The causal role played by the representation of a goal in intelligent behavior is partly straightforward. The representation should function so to govern behavior as to bring about (cause)

the attainment of the goal, and in addition the creature should register 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 behavior by '-[causes]->', we have overall:

A has the goal that A stops that flow -[causes]-> A stops that flow

And when the creature registers the attainment of this goal, and this causes the goal-representation to cease to operate, we have:

A registers that A has stopped that flow-[causes] -> A ceases to have the 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 by what they represent; so the underlined sentence 'A stops [+ tense] that flow serves to specify both the situation or event in the environment which is the animal's goal, and the representation of this which we take to be realized within. The same holds for the use of the sentence to describe the internal registration of the 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 in such a way that the latter ceases to operate.

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

G: A's 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 also partly covers the veridical registration of information about the environment. For as part of this 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:

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

This appears to schematize the animal analogue of perception-based veridical belief; and of course animal goal-seeking is characteristically informed by perception and memory of the environment in something like this 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:

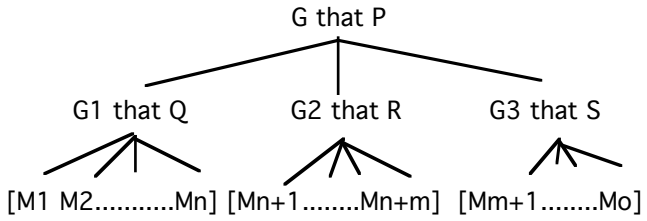
Initial Goal: A has G that P (that it stops that flow)
Information: A registers that if Q then P (if it moves that branch then it stops that flow.)
Derived Goal: A has G that Q (that it moves that branch).

This process is clearly 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

Although we apply them to non-speaking animals, the schema **G** and **PR** are obviously related to the human case. Thus substitution of 'desires' for 'has goal', and 'believes' for 'registers' yields schema which apply to human desire, belief, and practical reason.

Animal action is commonly driven by numerous goals related by complex information. Thus even in a very simple case a creature may have that 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 too 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 predictive and constrained. 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.

3. A possible advantage of representing inner causes as opposed to regularities in behavior.

This provides a minimal first sketch of the kind of ordering and modification of goals which we find in intelligent creatures. 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. As noted we make use of information about such trees, and frame goals in relation to them, when we choose among alternative courses of action, or again when we seek to anticipate or influence the motives of others.

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) tried to convince behaviorists to take account of causes concealed in the brain. Suppose we are studying the behavior 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.

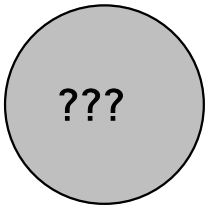


Figure 1

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

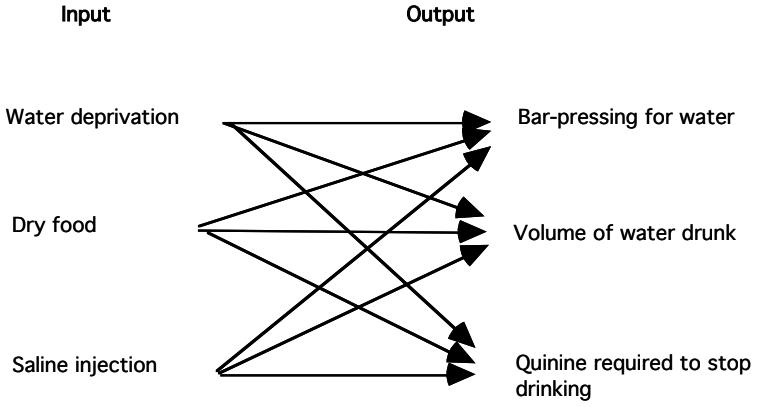


Figure 2

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:

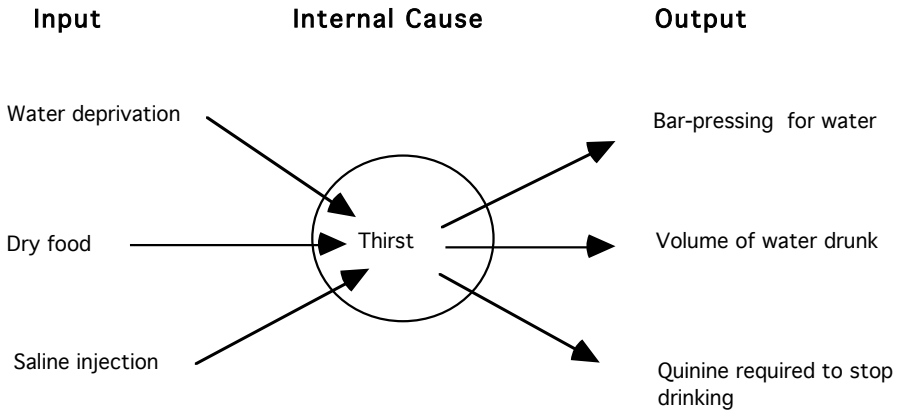


Figure 3

Now 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 our representation (the intermediate inner cause) in effect enables us to drop three others (three correlation-marking arrows). In general it seems that even simple explanatory theories which postulate inner causes may enable us to *compress* (in the computational sense) the correlational data which they cover (compare Dennett 1991). But then it seems clear that there may be cognitive and hence reproductive advantage not only in representing correlations which hold for our own behavior or that of creatures with which we interact, but also in representing these correlations more efficiently, as Figure 3 does by comparison with Figure 2.

4. How we represent the inner causes of behavior: sentential descriptions of goals and information as an empirical theory which operates via norms of language.

So far we have 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.; and each of these admits of 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. In general, and as a matter of basic linguistic understanding, we map our sentences to the perceptible objects and situations which constitute their conditions of truth. Then in articulating our conception of the mind we map the same sentences again to our unperceived insides -- to the internal mental (or neural) causes of behavior -- by taking their conditions of truth as those of the satisfaction of desires, the fulfillment of intentions, the realization of hopes and fears, and so on. In this we pre-reflectively bond the representing mind/brain to the represented environment, by conceiving the inner causes of behavior in terms of the sententially described outer situations which it is their adaptive function to produce (as in the case of desires, cf. Millikan 1984) or to reflect (as in the case of beliefs or thoughts). This makes use of our capacity to relate sentences and situations via the normative concept of truth, to yield a potential infinity of semantic/causal descriptions of otherwise imperceptible motives; and this implements our conception of the mental as having *intentionality*, that is, as engaging both causally and normatively with the world.

This sentential mode of description enables us to trace the role of otherwise inscrutable inner causes of behavior via our linguistic understanding of the sentences we use to describe them. We see this semantic specification of causal role in some detail in **G** and **PR**, taken in terms of belief and desire. Thus consider the schema relating to perception and belief

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

This uses 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 situation described by 'P' as applied to a situation in the environment, through the perception described by 'P' as applied to information in the visual apparatus, to the belief described by 'P' as applied to the mind/brain; and this semantic specification also marks perception as veridical and the belief as justified and true. The specifications in **PR** and **D** likewise use sentences to mark stages in internal processing and motor output, this time as instances of reasoning which is logical, and action which succeeds in satisfying and pacifying desire. The mappings which inform our commonsense concept of mind thus enable us to represent the brain as a *virtual semantic engine*, that is, one whose causal workings are specified via sentences and the situations they describe, and so in terms of truth, reason, and the satisfaction of desire.

As in Whiten's conception, this representation as a virtual engine also serves to compress sentential correlations between perceptual input and behavioral output; and as the engine is sensitive to syntax and deductively driven, the compression is potentially powerful. We can suggest this by a diagram of inputs, internal causes, and outputs, as follows:

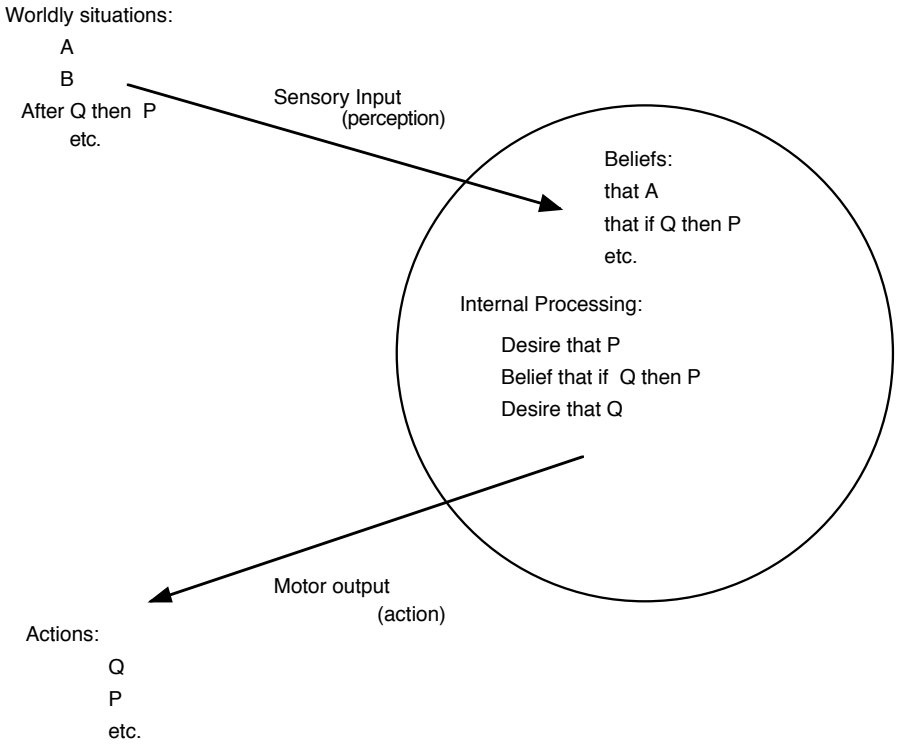


Figure 4: The sentential image of the mind or brain.

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

We have described interpretive understanding 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 motives. This is an important component of our notion of consciousness, which, as we take it, 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 we understand the inward and external mappings by which our sentences describe our minds and the world, and the former as relating to the latter, coherently and correctly. That we do so 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 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 can understand with greater precision and certainty than our own language, and hence those we take to share it with us.

5. 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 it is surely something which should admit of explanation. We can sketch a part of the required explanation by attending to the contrasting roles of speech 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 behavior which we could *not* understand in isolation from the rest of the behavioral 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 interpret these sounds or marks, or regard them as language at all. (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 behavior without relying on language, at least up to a point. We can generally see the purposive patterns in people's behavior in terms of their performance of commonplace intentional actions, as in accord with **G** 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.

We thus have a general claim about interpretive understanding. Words with no relation to deeds are unintelligible, and deeds with no relation to words are inarticulate. It follows that the understanding of people we actually attain, in which we take their deeds to spring from motives with determinate and precisely conceptualized content, requires us to integrate our understanding of verbal and non-verbal behavior, and hence to correlate and co-ordinate the two. It is some such integration which enables us to tie 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 interpret the rest of behavior as informed by experience and thought which, like that expressed in language, has fully articulate content.

I think that the particular mode of integration which we use involves what we can regard as a process of interpretive triangulation. In this we systematically relate the motives which we take to be expressed in speech -- including desires, beliefs, and experiences -- to those upon which we take speakers non-verbally to act. We thus triangulate from speech and non-verbal action to focus upon their common causes, that is, motives which can be specified by relation to both verbal and non-verbal behavior. 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 behavior 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 behavior 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 on trust. My confidence in my interpretations would be owed 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 favor 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 drive confirmation of the hypothesis that each understood the idiolect of the other steadily upwards. So by this means, it seems, we could in favorable circumstances come to regard our possession of mutual linguistic understanding as confirmed to the highest degree. (This, it will be recalled, is one of the things we are seeking to explain; and the principles illustrated here apply to more other and more complex cases, as discussed in Hopkins 1998.)

Triangulation of this kind presupposes an interpreter with a capacity to think in an effective hypothetical way about the motives and experiences which explain both verbal and non-verbal actions, and an interpretee who can provide both non-linguistic and linguistic behavior, 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 so continually explore and confirm the presuppositions of successful interpretation of this kind. Of course in practice we cannot always interpret accurately, and our first-person authority may fail. But an interpreter can still correct bad interpretations in light of the evidence which the interpretee provides, and also check and, if relevant, try to correct or make allowance for failures in the interpretee's first-person authority. All these processes admit continual repetition and refinement. So the fact that each of us is *both* a potentially accurate interpreter *and* a potentially authoritative interpretee would appear to allow us to calibrate our interpretations of verbal and non-verbal behavior continuously and cumulatively, and so as to give both something like the degree of precision and accuracy which we observe them to enjoy.

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 interwoven. Taken this way, moreover, first-person authority does not seem solely or primarily self-directed. Rather it appears as a social complement to the ability to interpret: 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, moreover, also seems such as to have been shaped by evolution. Roughly, and other things being equal, we should expect that an increase in the ability to understand and anticipate the behavior of others should be an advantage to members of a species who possess it; and the same should hold for an increase in the ability to influence the way in which one is so 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 could get its hands on behavior which expresses motive to cull and save in favor of both these abilities. There seems reason to hold that such a process has been accelerated among the social primates, and particularly in our own species; and that this has resulted in the interpretive and expressive abilities that underlie our conceptions of language and mind. (See, e.g., Deacon 1997). So it seems that we might begin to sketch an evolutionary account of the process by which our capacities for expressing and describing motives have become welded to the relevant behavior-governing states and events in our brains with the tightness and precision which we observe in our exercise of first-person authority.

6. The internality of the mind.

Now it is also a striking feature of our conception of the mind that we regard mental items, events, or states 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. Thus as Colin McGinn (1989) writes

> Our acquaintance with consciousness could hardly be more direct...'Introspection' is the name of the faculty through which we catch consciousness in all its vivid nakedness. By virtue of possessing this cognitive faculty we ascribe concepts of consciousness to ourselves; we thus have 'immediate access' to the properties of consciousness.

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' or 'in the head', where only one person can see it (Wellman 1993). Nonetheless it remains a puzzling matter. 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. 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 this private introspectible space therefore seems somehow distinct from the public physical space inside the tooth and body. And as is familiar, the introspectible quality manifest in this internal space seems to be the defining or essential feature of the sensation of pain.

7. 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, and thus with dualism more generally. 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 by any physical thing. This raises the question as to how the qualities which we take ourselves to know in introspection can be related to the physical qualities we encounter in perception and scientific understanding, and particularly to those of the brain.

As the consideration of pain above indicates, this opposition, which many philosophers regard as *the* problem of consciousness, seems one member of a series, all of which are related to the internality of the mind. Thus qualities which are introspectible and so phenomenal also seem *subjective*, in that there seems no clear distinction between how they appear in introspection 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*. Again, qualities which are introspectible also seem *private*, since 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.

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

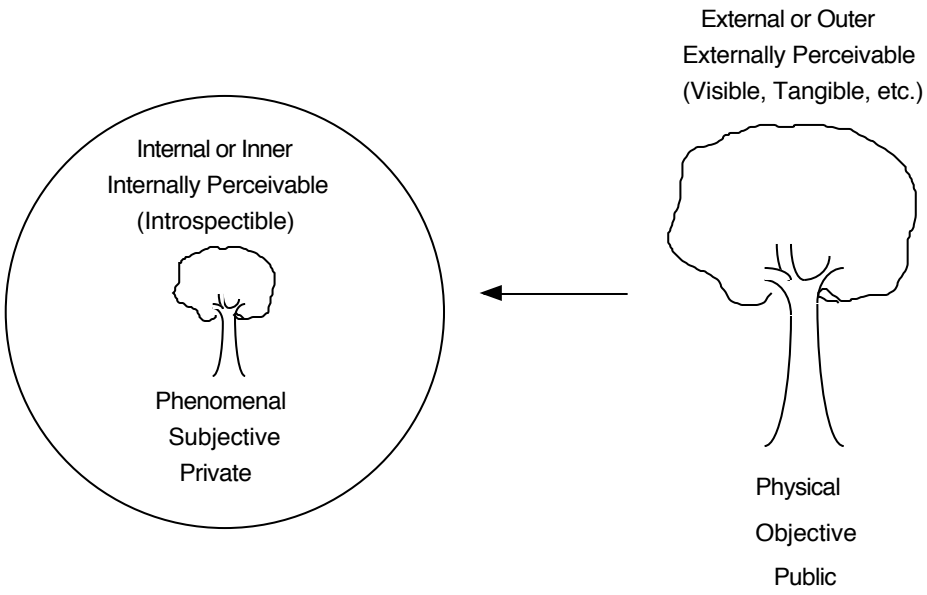


Figure 5: An image of the internality of the mind.

Now if we consider these oppositions together I think we begin to see a *prima facie* case for taking them to be systematically related. We can set this out as follows:

(i) The terms of opposition seem naturally to fall into an *internal* and an *external* series. Qualities which are external and visible (or tangible, etc.) are physical, and these are also objective and public; whereas qualities which are internal and introspectible are phenomenal, and these are also subjective and private.

(ii) Each series seems to have an underlying unity, in that each term seems closely (perhaps essentially) related to the others. It seems essential to what is external and so externally perceivable to be physical, and essential to what is physical to be objective and public. Likewise it seems essential to what is internal (in this mental sense) and so introspectible to be phenomenal, and essential to what is phenomenal to be also subjective and private.

(iii) While both series seem to be organized around the opposition internal/external, it seems that the notion of internality may at least partly explain some other terms in the internal series. For example the subjectivity of phenomenal qualities seems to derive from the idea that their whole nature is internal and so manifest in introspection; and the privacy of these qualities seems owed to the idea that each of us can introspect only what is internal to a single mind.

(iv) Each term seems utterly incompatible with its opposite number; and this incompatibility is like that between different colors, in seeming at once experiential and logical. Just as it seems (experientially and logically) impossible that anything phenomenal should be physical, so it seems impossible that anything which exists internally to the mind should also exist externally to it; that an object of introspection should also be visible to the eyes; that anything subjective should also be objective; or that anything private should also be public. Thus each pair of oppositions could be said to present a problem in understanding consciousness, or a particular part of an encompassing (four- or five-fold) problem of consciousness. Again, each opposition could be used to argue -- either intuitively, or via modal or epistemic considerations -- for various distinctions between the mental and the physical, or again various types of dualism. (And indeed other oppositions of the series are regularly used to invoke the problem of consciousness, even by authors whose main purpose is to contrast the phenomenal with the physical).

Now (ii), the apparent underlying unity of the series, suggests that they may have an underlying explanation. Their organization around the opposition internal/external noted in (i),

together with the apparent explanatory dominance of internality noted in (iii), suggests further that the key underlying conception may be that of the internality of the mind. (It is as though in representing the mind as internal we somehow split off the internal from the rest of the world, so that the series of internal terms coheres as a kind of negation of that from which the internal is split off, yielding the incompatibility noted in (iv)). If anything like this is true, then understanding our conception of internality may be the key to understanding the rest of the inner series. But then since phenomenal qualities simply are those which are internal and introspectible, it seems possible that their ostensibly non-physical nature should be explicable by reference to their internality as well.

8. Representing internality: the example of conceptual metaphor.

No doubt our conception of the internality of the mind reflects the fact that the behavior-governing events which we describe in mental terms actually occur inside the body. But the mere physical location of these events by itself does nothing explain how we manage to *represent* them as internal, much less how we have come to represent them as internal (as well as phenomenal, subjective, and private) in the particular and remarkable way we do. So how might we suppose that our particular mode of representation of mental events as internal or inner has arisen?

Here it seems worth noting that we already make use of another environmental mapping, by which we represent the internality of some mental events. Recently George Lakoff, Mark Turner, and a number of others have argued that we frequently represent things via *cognitive metaphor.* In this we systematically map one domain of objects and properties (called the *source domain*) to another (called 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 conceptualization 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 opposition in the properties of the metaphorical fluids to which we map them. This is one of many examples of representation of the mind as an inner space or container, and indicates something of the tacit systematic nature of thinking of this kind.

9. Evolution, metaphoric representation, and the concept of mind.

Such thinking bears comparison with the mapping of motives to sentences discussed above. In both, as it seems, we represent the causal role of neural states or processes by mapping them to something external to the body. In this case, however, the mapping 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 takes place *inside* the mind/body container, as a fluid acts *inside* a vessel.

Now we said at the outset that evolution posed nervous systems the problem of representing the behavior-governing states of nervous systems blindly, that is, without recourse to perceptual information bearing directly upon neural operation. In previous sections we speculated

that this achievement was facilitated by mapping the neural states in question to things which could be perceived -- in particular, by linking them, via sentences, to the environmental situations which they were adapted to produce or reflect. Such mapping provides a representation of the causal role of the items to which it is applied, but not of their internal status itself, nor to specifically internal aspects of their causal role. This lack we can envisage addressed by the kind of mapping exemplified by the metaphor of the mind/body container. In this we represent events which are (i) *perceptually and causally inscrutable* and *(ii) hidden inside the body* by mapping them to others which are (i) *perceptually intelligible* but which may be (ii) *hidden in containers in the external environment*. We thus use our information about *containment in space in the external environment* to create an image of *containment in an inner space*, which we apply to the neural events which we characterize as both mental and having an important internal role.

I think it is plausible to speculate that the process of evolution (together, perhaps, with other sources of cognitive development) has built upon our information about relations of containment in the space about us in a comparable way, to enable us to frame our basic image of the mind as internal. In particular, we can see this image as formed by a mapping of *the external process of perception* into *the space internal to the body* and thence to *the internal neural events which realize experience*. To indicate the general nature of this hypothesis let us put an instance into a cartoon related to the illustrations above. Thus suppose we had no way of representing the internality of visual experience, as in Figure 1 above

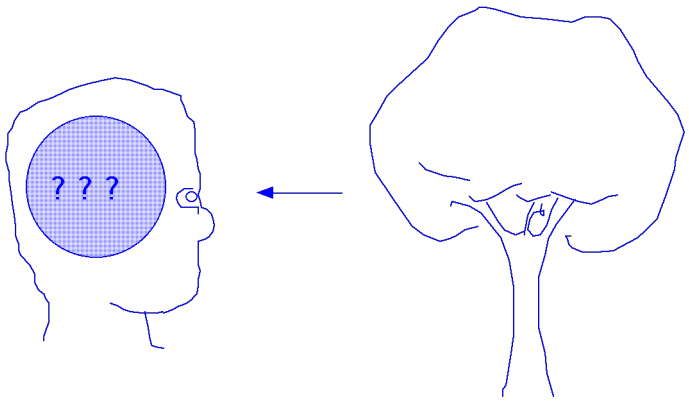


Figure 6a

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, which we could illustrate as follows.

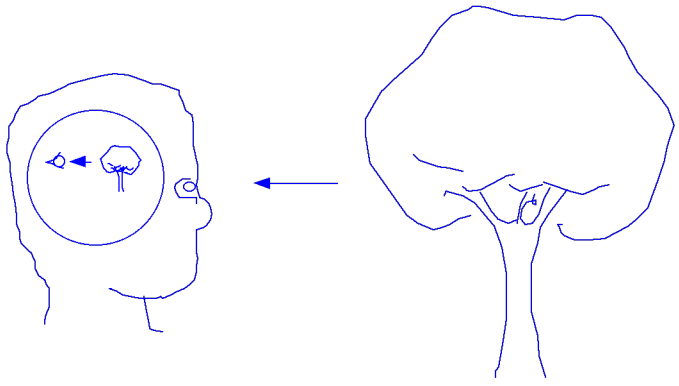


Figure 6b

This mapping, however, seems essentially that involved in Figure 5 above. Now of course this is only a cartoon, and related to just one kind of experience, that of a visual image, or the visual field, as 'in the mind or head'. But the idea that our image of experience involves some such

internalization of external perception can also be applied to a range of other cases, in all of which we regard experience as presented in something like *internal perception* (introspection), 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, that of kinesthetic 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 conceive of visual space as having to do with the eyes, and so, perhaps, as somewhere behind them; auditory experiences as having to do with the ears, and so, perhaps, as somewhere inside them; and so on. Hence we 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 notion of space involved had been derived from our experience of that external to the body and transposed to the space within.

As illustrated by comparison between figures 5 and 6b, such mapping seems a possible constituent of our image of the mind as inner; and this would complement the sentential image discussed above, by representing internality as well as causal role. So overall we have a rough hypothesis about the origin and nature of our conception of mind: our concept of the *mind* arises as part of the activity of the *brain* (or the nervous system) in forming a representation of itself by mapping inwards from the perceptible environment. The mapping inwards of sentences yields an image of the mind/brain as a *virtual semantic engine,* and the mapping inwards of external relations of spatial containment yields an image of neural space as a *virtual inner space*. So the brain's image of itself -- and in particular its image *as from an inside* -- yields our image of the inner space, and the inner life, which we take as that of the mind. (My brain's image of itself, in short, is my from-inside image of me.)

10. Virtual inner space and the apparent non-physicality of the mental.

The idea that the brain has evolved to represent itself as a virtual semantic engine in a virtual inner space bears on a number of problematic features of the concept of mind. First, it contradicts the idea that introspection gives us some sort of direct access to the mind (cf. McGinn's 'consciousness in all its vivid nakedness' above). Although introspection inherits the first-person authority which goes with understanding language, it is less direct than visual perception itself; for in introspection, on this account, we represent events in a second-hand way, by re-using a neural image derived from perceiving space outside the body. Further, introspection radically misleads us about the nature of inner events themselves; for this second-hand representation creates the deep but illusory appearance of non-physicality which marks our naive conception of the mind.

Empirical observation indicates that in using mappings of the kind we are considering we automatically and unconsciously 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 see this principle operating in our use of the metaphor of the mind as a 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 it may 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 we also think of as not fully physical. Still, this de-physicalization retains a physical ontology. It flows from requirements of coherent representation upon a mapping which has both

physical sources (physical fluids and containers) and physical targets (changes inside the body involved in emotion), so that nothing both real and non-physical actually comes into question. So it seems that a comparable process, also consistent with a physical ontology, might account for the more radical de-physicalization which we find in the virtual internal world of our image of the internality of the mind.

11. Virtual inner space and the problems of consciousness.

This suggests the possibility of an account of the problematic features of consciousness, and as dependent on our notion of the internality of the mind, as sketched in section 7 above. Roughly, in conceiving what we introspect in terms of a virtual internal space, we tacitly and automatically conceive this space and the things in it as distinguishable from what is physical, and so as phenomenal. Further, since items thus represented as internal have no existence apart from this space their *esse* seems *percipi*, so that they are also subjective; and since they have their existence in the virtual space of a single mind, they are also private.

Clearly this account requires further elaboration and defense. Still, the sketch here suggests that 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 that between physical causes and mental reasons. Just as our inwardly mapped sentential specification of the causal role of motives can wrongly suggest that this role is semantic as opposed to physical, so an inwardly mapped spatial representation of the innerness 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 weigh the rainbow. We come to understand such phenomena not by altering our science of nature, but by recognizing that their place in nature is not as it appears, and so by studying them as forms of illusion.

_____

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. (1998) 'Patterns of Interpretation: Speech, Action, and Dream', in L Marcus, ed, *Cultural Documents: The Interpretation of Dreams*, Manchester: Manchester University Press.

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.

McGinn, C. (1989) 'Can we solve the mind-body problem?' *Mind*, Vol 98, July 1989.

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.

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 behavior-reading become mind-reading' in P. Carruthers and P. K. Smith, eds, *Theories of theories of mind*, Cambridge: Cambridge University Press, 1996.