Chapter 4

How to read minds

Tim Bayne

Introduction

Most animals have mental states of one sort or another, but few species share our capacity for self-awareness. We are aware of our own mental states via introspection, and we are aware of the mental states of our fellow human beings on the basis of what they do and say. This chapter is not concerned with these traditional forms of mindreading—forms whose origins predate the beginnings of recorded history—but with the prospects of a rather different and significantly more recent form of 'mindreading': the capacity to ascribe mental states to a creature on the basis of information derived from neuroimaging.

The thought that we might be able to read minds by inspecting brains has stimulated philosophical interest for decades (Dennett 1978), but with recent advances in neuroimaging this idea has now passed from science fiction and into science: mindreading—or 'brain decoding', as it is also known—is now a burgeoning industry.[1] Here are three examples of mindreading—or at least attempted mindreading. In one study, Haynes and colleagues asked subjects to decide either to add or subtract two numbers that had been presented to them (Haynes et al. 2007; see also Haynes, this volume). On the basis of fMRI data, the experimenters were able to determine with up to 70% accuracy whether the subjects would sum the presented numbers or whether they would subtract one number from the other. In another study, Spence and colleagues suggested, on the basis of neuroimaging evidence, that a woman who had been convicted of intentionally inducing illness in a child may have been innocent (Spence et al. 2008). In a third study, Owen and colleagues concluded that a vegetative state patient was conscious on the grounds that she showed neural activity in brain areas implicated in motor imagery and spatial navigation when instructed to either imagine herself playing tennis or visiting the rooms of her home (see also Boly et al. 2007; Monti/Vanhaudenhuyse et al. 2010).

These studies are of great interest in their own right, but they also raise more general questions about the nature and scope of brain-based mindreading. One set of questions concerns methodology. How might one justify the ascription of a mental state to a creature on the basis of neuroimaging data? A second set of questions concerns the scope of mindreading. Under what conditions, and with respect to which kinds of mental states, might mindreading be possible? A third set of questions concerns the interaction between

[1] For other examples of mindreading see Chadwick et al. (2010); Dehaene et al. (1998); Haynes and Rees (2005, 2006); Kamitani and Tong (2005); Polyn et al. (2005); Richiardi et al. (2011); and Shirer (2011).

brain-based mindreading and the more familiar forms of mindreading mentioned earlier, namely, those that involve introspection and behaviour. How might these three forms of mindreading be related to each other?

Rather than attempt to provide definitive answers to these questions, I will provide a framework in which such answers might be usefully pursued. With this goal in mind, I will avoid engaging with the questions raised by the limitations of current neuroimaging technologies (see Haynes, this volume), but will focus instead on the foundational issues that are likely to confront the use of any neuroimaging technology to read minds, no matter how sophisticated it may be.

The methodology of mindreading

In principle there are two ways in which one might attempt to ascribe mental states to a creature on the basis of neuroimaging data. One way would be to use neuroimaging to determine what brain states a person is in, and then employ an explanatory model of how brain states give rise to mental states in order to determine what mental states the person is in. The idea behind this approach is that one should be able to infer a creature's mental states from its neural states in much the same way in which one can infer a substance's gross behavioural properties from its molecular structure. For obvious reasons we might call this the *chemical model* of mindreading.

Alas, we do not have a chemical model of the mind. Moreover, there are good (although far from incontrovertible) reasons to think that we may never have such a model. One reason for pessimism concerns the semantic or contentful aspects of the mind. Since the cognitive revolution in psychology and the rise of functionalism in philosophy, it has become commonplace to view the mind as the software of the brain (Block 1995). On this picture, although meaning is realized by neural states, there is no necessary connection between the identity of a neural state and the particular content that it carries, and in principle thoughts with the same content can be 'carried' by any one of a number of different neural state types. Just as there are various ways in which public languages can represent tigers, so too there are various ways in which the brain can represent tigers. The way in which a creature's tiger-related thoughts are neurally realized may depend on the evolutionary history of the species to which it belongs and its individual learning history. Even if, as a matter of fact, each of our tiger-related thoughts is realized by a single type of brain state, we can no more identify which brain state that is by investigating the brain than we can identify the meaning of words in an unknown language from investigating the shape of the script in which it is written. Instead, in each case we need a translation manual or 'Rosetta stone' in order to move from syntax to semantics.

A second reason for pessimism concerning the 'chemical' model of mindreading concerns the experiential aspects of mentality. Broadly put, the problem is that we lack an explanatory grip on the relationship between neural states and experiential states (Levine 1983). We do not know why some neural states are associated with experiential states whilst others are not, nor do we know why those neural states that are associated with experiential states are associated with the particular experiential states that they are (say, the taste of strawberries) rather than others (say, the smell of sardines). Some theorists hold that our ignorance in this matter is merely temporary and that it will be ameliorated by advances in science; others argue that features of our cognitive architecture will prevent us from ever grasping the explanatory nexus between neural states and experiential states; and still other theorists hold that there is no explanatory relationship between neural states and experiential states to be grasped. Whatever the truth of this matter, the explanatory gap is unlikely to be closed any time soon.

Between them, the twin challenges just outlined suggest that the chemical model will not provide us with a viable account of mindreading. There is, however, another way in which mindreading might proceed. Rather than attempting to identify mental states from neural states on the basis of first principles (as the chemical model does), one might employ *independently established* correlations between neuroimaging data and mental states. Indeed, this is precisely the methodology adopted by the three mindreading studies mentioned above. In each case, researchers employed independently established correlations from a certain type of neuroimaging state N_1 to a certain type of mental state (M_1) in a population P to argue that a particular member of P was probably in mental state M_1 on the grounds that he or she was in neuroimaging state N_1. I will refer to this approach to mindreading as the *correlational method*.[2]

We will explore the correlational method in some detail below, but let us first note that the method avoids the problems that undermine the chemical approach. The correlational method avoids the problem of the explanatory gap, for it is possible to identify and employ a correlation without making any assumptions whatsoever about what underlies that correlation. Perhaps more surprisingly, the correlational approach also avoids the problems posed by the multiple realizability of mental states. To see this, suppose that there is a certain type of mental state—pain, for example—that is realized by neural state N_1 in some members of P, by N_2 in other members of P, and by N_3 in still other members of P. That this is so does not prevent us from ascribing pain to any member of P on the grounds that he or she is in (say) neural state N_1. What matters from the point of view of the correlational method is not the mapping from mental states to neural states, but rather the mapping from neural states to mental states. In other words, the challenge facing the correlational method is not that particular mental states might be associated with multiple kinds of neural states, but rather that particular neural states might be associated with multiple kinds of mental states. We will return to this point.

The correlational method is not undermined by the possibility of multiple realization, but perhaps it faces challenges from other quarters. It is sometimes suggested that mindreading is possible only if there is a language of thought (see e.g. Haynes, this volume). Let us understand the language of thought hypothesis to be the claim that thoughts have combinatorial structure, such that the semantic structure of a thought is roughly mirrored

[2] The correlational method involves what Poldrack (2006) calls a *reverse inference*—'reverse' because cognitive neuroscientists are typically interested in inferences from mental states to neural states rather than from neural states to mental states.

by its syntactic structure (Davies 1998; Fodor 1975, 2008; Maloney 1989; Rey 1995). The idea, in other words, is that thoughts are built up out of symbols, where a symbol makes the same contribution to the semantic properties of whatever thought it occurs in. Just as tokens of the symbol 'tiger' make the same semantic contribution to the sentence 'The gardener chased the tiger' as they do to the sentence 'The tiger bit the butler', so, too, advocates of the language of thought hold that there is a mental symbol which refers to tigers, tokens of which occur in such thoughts as <The gardener chased the tiger> and <The tiger bit the butler>.

Thus understood, it should be clear that the correlational method does not assume the existence of a language of thought. Indeed, it would be possible to employ the method without assuming that thought has any syntactic structure at all, let alone a syntactic structure that is roughly isomorphic to its semantic structure (as advocates of the language of thought claim). In principle, all that the correlational method requires is that there be some reasonably robust mapping from neural states to mental states—it does not require that there also be a robust mapping from neural states to the *constituents* of thoughts.

That said, the prospects of the language of thought hypothesis do have a bearing on the practice of mindreading. For one thing, many mindreading experiments are concerned with the constituents of thought. In order to ascribe tiger-related thoughts to subjects, theorists might look for the neuroimaging response that is specific to thoughts about tigers as such. However, this search might be doomed to failure if there is no language of thought. The brain state that the subject is in when thinking <The gardener chased the tiger> might have nothing in common with that which he or she is in when thinking <The tiger bit the butler>.[3] Moreover, the absence of a language of thought would restrict the potential interest of mindreading. Suppose that there is no language of thought—or at least, that there is no language of thought that we might have any chance of deciphering. In that case, the mindreader would be in the position of a tourist who speaks only a guidebook version of the local language. She would be able to attribute thoughts that figure in the correlations to which she has access, but she would not be able to attribute to individuals *novel* thoughts. If, on the other hand, our would-be mindreader has deciphered the language of thought, then she would—at least in principle—be able to attribute thoughts that do not figure in the correlations that are listed in her database ('her guidebook'). For example, if she knows the 'Mentalese' (language of thought) words for <tiger>, <butler>, and <bit>, then she might be able to attribute the thought <The tiger bit the butler> even if this thought does not appear anywhere in her list of correlations.

It is, of course, controversial whether there is a language of thought (see e.g. Dennett 1981; Matthews 2007). Even if there is a language of thought, it is a further question whether any two thinkers share a common language of thought, or whether the language of thought is 'solipsistic', such that no mental symbol in any one thinker's lexicon can be type-identified with that which occurs in the lexicon of another thinker. If the Mentalese were solipsistic in this way, then one would need to learn a new version of Mentalese for

[3] I am grateful to Nicholas Shea here.

every potential target of mindreading. Although this would not put one back in a position of the guidebook speaker—for, after all, one could ascribe to that thinker thoughts that one had not already come across—it would radically undermine one's ability to generalize from one group of thinkers to another. Unless we share a version of Mentalese, the lexicon derived from the study of one cohort of thinkers could not be used to unlock the thoughts of another cohort. In short, although the correlational method does not as such require a language of thought, debates about the language of thought do have implications for the scope of mindreading.[4]

The scope of mindreading

Let us turn now to the correlational method itself. At the heart of the method are correlations from neuroimaging states to mental states of the following form:

> *Neuroimaging Correlations (NC)*: For any arbitrary member S of a population P, if S is in neuroimaging state N_1 then there is a high probability that S is in M_1.

Although I have been discussing correlations from neural states to mental states, NC itself refers to correlations from *neuroimaging* states to mental states. Neuroimaging data is, of course, grounded in brain-based activity of some kind, but there are debates about precisely what kind of neural activity is being measured by neuroimaging techniques. By couching the correlations employed in mindreading as correlations from neuroimaging states to mental states, we can avoid taking a position on what precisely it is that neuroimaging techniques are tapping.[5]

We should also note that the correlational method does not require that there be a strict inference from the neuroimaging state to a particular mental state—that is, it does not require that the probability of the mental state conditional on the neuroimaging state is 1—but only that the neuroimaging evidence raises the antecedent probability that the target is in a particular mental state. Of course, if the neuroimaging data raises the probability that the target is in the relevant mental states only slightly, then it might not be accurately to describe it as facilitating an act of 'mindreading'; we might want to reserve that label for contexts in which the neuroimaging data raises the probability of a certain mental state above a certain threshold.

In some cases neuroimaging data may indicate that the target is in one of a number of independent mental states, rather than in any particular mental state. For example, it could be that there is a strong correlation from a neuroimaging state to a particular set of

[4] Of course, there may be relevant neural generalizations across subjects even if thought is solipsistic. For example, dog thoughts may have features that are shared across people even if their type-identity—the thing that makes them the particular mental symbol they are—is not shared. The central point is that although solipsism allows for such generalizations it does not guarantee them. Thanks to Nicholas Shea for this and a number of other points.

[5] However, we might need to determine what kinds of neurofunctional states are responsible for our neuroimaging data if we want to integrate it with brain-based data of some other kind (say, lesion data) or indeed with another kind of neuroimaging data.

mental states {M_1, M_2, and M_3}, but only a very weak correlation between N_1 and any individual member of this set. In such a case, the neuroimaging data give one good reason to believe that the target is in either M_1, M_2, or M_3, without giving one any clue as to which of these three states it is in.

A further feature of the correlational method that deserves comments concerns the fact that the correlations from neuroimaging states to mental states are relativized to particular populations. We can assume that the 'standard' mindreading population will be neurologically unimpaired adult human beings. Although the correlational method can in principle be applied to many different types of individuals—including human neonates, humans who suffered some form of severe neurological insult, and even the members of non-human species—there will often be severe obstacles in the application of applying mindreading techniques to such 'non-standard' populations. This is because it is typically much easier to identify the NCs that characterize neurologically normal adult humans than it is to identify the NCs that characterize other populations. With respect to neurologically normal adult humans, not only are we able to avail ourselves of introspective reports, we also have a reasonably robust capacity to use an individual's behaviour to constrain attributions of mental states to it. Neither of these things is true—at least not to the same extent—when it comes to the very young or severely brain-damaged members of our own species or the members of other species.

There are two ways in which one might attempt to get around the challenges posed by 'non-standard populations'. On the one hand, one might attempt to extend the NCs derived from the study of normal adult humans to non-standard mindreading targets. An example of this approach is provided by the work of Owen and colleagues, who used correlations drawn from neurologically unimpaired individuals as the basis for their ascription of conscious imagery to a vegetative state patient (Owen et al. 2006). It is arguable that this extension of a standard NC is legitimate, for it seems unlikely that the brain damage that this patient had suffered would have disrupted the specificity of these neural responses. However, there are many contexts in which it will be quite unclear whether the application of standard NCs to a non-standard mindreading target is justified.

A second approach to the challenge posed by non-standard population involves looking for NCs that are specifically tailored to that population. For example, we know that in congenitally blind individuals who have learned to read Braille, activity in visual cortex is correlated with tactile experience rather than visual experience (Merabet and Pascual-Leone 2010; Sadato et al. 1996). Thus, any attempt to read the mind of a Braille reader will need to use NCs that are specifically tailored to the members of this population rather than those that are derived from the study of the sighted. Identifying NCs that are tailored to the congenitally blind is relatively straightforward, for such individuals can report their experiences. However, when dealing with 'non-standard' populations whose members are not able to produce introspective reports it may be extremely difficult to identify such specifically-tailored NCs.

Let us turn from the challenges posed by 'non-standard' cases to those posed by neurologically normal adult human beings. How selective are 'our' neural states? The answer to this question will depend on the kinds of neural states and the kinds of mental states that we employ in our analysis.

Consider first the issues raised by neural kinds. It is often thought that many neural areas are highly selective for specific kinds of mental states. There is some truth to this, especially when it comes to low-level sensory areas, but recent neuroscience suggests that many neural areas that have been traditionally thought to be content-specific are in fact implicated in a wide variety of mental states and processes. Indeed, it is not uncommon for theorists to describe the brain as 'essentially multisensory' (Driver and Noesselt 2008; Ghazanfar and Schroeder 2006; Macaluso 2006; Pascual-Leone and Hamilton 2001). Take the *pars opercularis* (Brodmann Area 44), for example. This region has been implicated in the production and comprehension of phonetic structure; auditory imagery; automatic imitation and 'mirror' activity; the manipulation of a musical sequences; deductive and inductive reasoning; the evaluation of causal relations, and a number of other domains. Moreover, there is no reason to think that the *pars opercularis* is any less selective than many other neural areas. In an important meta-analysis of 1,469 subtraction-based fMRI experiments, Anderson (2010) found that the typical cortical region is activated by tasks drawn from any one of nine out of 11 task domains.

Although the non-selectivity of neural states represents something of an obstacle to the correlational method, it is not an insurmountable obstacle. For one thing, the subject's environment can be structured so as to 'screen off' certain interpretations of the neural activity. Suppose that neural state N_1 has been implicated in mental states M_1, M_2, and M_3. If we knew nothing about the subject (S) under consideration other than that they were in N_1 then we would not be justified in ascribing of any one of these three mental states to S. However, information about S's environment might count against the ascription of (say) M_1 and M_2 to S, and count in favour of the ascription of M_3. It is important to note that, in order to usefully contribute to the task of mindreading, information about a subject's neural states need not determine a unique ascription of mentality but need only shift our prior probabilities concerning the matter. In addition, new methods of mindreading are being developed which focus not on the activity of particular neural areas but on the functional connectivity between disparate areas (Haynes and Rees 2006; Norman et al. 2006; Richiardi et al. 2010; Shirer et al. 2011). These techniques have the potential to identify spatio-temporally complex states that may be significantly more selective than those that form the mainstay of current mindreading research.

Let us turn now to questions of mental taxonomy. There are a number of dimensions along which mental states can be distinguished from each other. Firstly, we can distinguish coarse-grained mental states, such as the state of being conscious, from fine-grained mental states, such as the state of hearing a bell ringing. Cutting across this distinction is a distinction between mental episodes or events (also known as 'occurrent mental states') on the one hand, and dispositional mental states on the other. Attempting to add two numbers together, visually identifying a word, or being in pain are mental episodes—they characterize one's psychological life for discrete periods of time. By contrast, being depressed, having prosopagnosia, intending to retire to the south of France, and believing that tigers are dangerous are dispositions, capacities, or long-term states rather than episodes or events. One can believe that tigers are dangerous without that state manifesting itself in one's behaviour or stream of consciousness. Yet a third distinction contrasts those mental states that are primarily sensory, perceptual, or affective in nature with those that are primarily cognitive. In this regard, we can draw a rough distinction between (say) states of bodily pain and visual experiences of motion on the one hand, and (say) the judgement that justice is more important than peace on the other. How might these three distinctions bear on the prospects of mindreading?

Let us begin with the question of grain. *Prima facie*, one might think that mindreading will be most straightforward with respect to very coarse-grained mental states, for it seems reasonable to assume that the neural states with which coarse-grained mental states are correlated will themselves be coarse-grained and thus relatively easy to identify. However, although there may be some kind of correlation between the 'grain' of neural states and that of mental states, it is far from obvious that coarse-grained mental states will always be correlated with coarse-grained neural states. Instead, coarse-grained mental states may be correlated with the disjunction of various fine-grained neural states. Because of this, in order to ascribe a coarse-grained mental state to a creature on the basis of neuroimaging data one may often have to go 'via' the ascription of a fine-grained state. Rather than looking for a neural correlate of consciousness as such, it may often be easier—and, depending on the neural basis of consciousness, perhaps even necessary—to look for a neural correlate of a particular kind of conscious state, and infer consciousness on that basis.

What implications might the distinction between mental episodes and dispositional mental states have for mindreading? There is good reason to think that neuroimaging will need to take quite different approaches to mindreading depending on whether the feature in question is episodic or dispositional. Episodic features will need to be detected by looking at dynamic neural activity, whereas the direct detection of dispositional states may require the identification of more stable forms of neural structure. However, given the close connections that hold between episodic and dispositional mental states, it will also be possible to indirectly ascribe dispositional mental states to an individual by ascribing episodic mental states to them. We will shortly encounter an example of this.

Finally, let us consider the distinction between those mental states that are primarily sensory, affective, or motoric from those that are primarily cognitive. One's views about how this distinction plays out in the context of mindreading will depend to some degree on one's views of cognitive architecture. According to an influential view, whereas perception involves a number of separate modules that process information in relative autonomy from each other and from the agent's background beliefs and desires, cognition is essentially non-modular in nature (Fodor 1983, 2000). Although modularity is primarily a matter of informational encapsulation, it is typical for theorists to associate modularity with neural localization and the lack of modularity with the absence of neural localization. Should this view of cognitive architecture be correct, then (roughly speaking) the closer a mental state is to the sensory periphery the more likely it is that it will have a dedicated neural basis and the easier it will be to identify by means of neuroimaging (Anderson 2010).

An opposing conception of cognitive architecture holds that some degree of modularity (and hence, perhaps, neural specificity) applies not just to perception but also to cognition. Versions of this view are defended by the advocates of massive modularity, such as Carruthers (2006) and Sperber (2001). In contrast with the advocates of the Fodorian view sketched above, proponents of massive modularity are likely to argue that, in general, it will be no harder to identify cognitive states on the basis of neural information than it will be to identify perceptual states.

By way of putting some flesh on these rather abstract bones, let us consider how these points might apply to the three examples of mindreading introduced earlier. Consider the study conducted by Haynes and colleagues (2007), who used fMRI to determine whether subjects were adding or subtracting numbers. The first point to note here is that the NCs used in this study were derived from the very individuals that were the targets of mindreading, and hence the NCs employed were ideal. Furthermore, because the subjects of this study were neurologically unimpaired adults whose veracity was not in question, the mental ascription produced by the decoding algorithm could be checked against the reports of their own mental states. (For obvious reasons, this kind of independent checking was not available in either of the other two examples of mind-reading.) In addition, the experimental context in which this study was conducted was highly constrained, and the experiment made critical use of the fact that the subjects had been instructed to perform one or other of two specific tasks. Clearly the experimenters would not have achieved the high levels of predictive accuracy that they did had their subjects been operating in a relatively unconstrained naturalistic environment.

Let us turn now to the study conducted by Spence and colleagues of a woman who had been convicted of intentionally causing illness in her child (Spence et al. 2008). This study did not attempt to directly determine the subject's beliefs or what her intentions had been. Instead, the experimenters attempted to determine whether or not the woman had been telling the truth by requiring her to agree or disagree with a series of statements, some of which endorsed the version of events that she had publicly defended and some of which contradicted that narrative. In other words, although this study only directly probed the subject's occurrent mental episode, the environmental context was such that this event was diagnostic of the subject's belief—a long-term dispositional state.

In suggesting that this woman was not lying when she asserted her innocence, the researchers relied on previous research indicating that deceptive responses activate ventrolateral prefrontal and anterior cingulate cortices when contrasted with sincere responses (see e.g. Abe et al. 2006; Kozel et al. 2004, 2005; Langleben 2002; Nunez et al. 2005; Spence 2001). These studies involve acts of deception that differed in a number of ways from the kind of deception of which this woman had been accused. Most obviously, they required subjects to engage in novel acts of deception, whereas this woman had repeated her account of the events so often that its representation was by now highly automatic (Spence 2008). However, this difference does not undermine the interpretation of the neural data given by the authors of this study. The reason for this is as follows. If the subject's account of events was a highly-routinized act of deception that required little deliberative control on her part, then one would expect to see no significant difference between the 'truth-telling' and the 'lying' conditions. In other words, this objection fails to explain why the experimenters found a significant difference between the two conditions.

The third of our three mindreading studies is perhaps the most problematic. As you will recall, the subject of this study was a 23-year-old female victim of a car accident who had been in a vegetative state for five months and was scanned whilst she was played a pre-recorded instruction to engage in a specific act of imagery—either to play tennis or to walk around each of the rooms in her house. In these two conditions, the BOLD (blood oxygenation level dependent) signal from those brain areas preferentially involved in motor imagery and spatial navigation—that is, the supplementary motor area (SMA) and the parahippocampal place area (PPA) respectively—was indistinguishable from that seen in healthy controls. The authors of this study concluded on this basis that the patient was indeed conscious. How plausible is this conclusion?

We can begin with an objection voiced by Nachev and Hacker (2009). They argue that the ascription of conscious motor imagery to this patient is undermined by the fact that SMA activation is seen in subjects who observe someone perform an action, and indeed in subjects who are merely exposed to action-related stimuli (Nachev et al. 2008; Rushworth et al. 2004). This objection might be worrying if we had no information about this patient other than the fact that she had, on certain occasions, shown SMA activation, but this is not the situation in which we find ourselves. Indeed, we have a great deal of information about the temporal parameters of the patient's neural responses and the environmental context in which it occurred. We know that the SMA and PPA activity was time-locked to the instructions 'imagine playing tennis' and 'imagine visiting the rooms in your home' respectively—that is, it commenced immediately after the relevant imagery instruction was given and ceased immediately after the instruction to stop engaging in the relevant form of imagery was given. This fact enables us to 'screen off' alternative interpretations of the patient's neural activity in favour of that provided by Owen and his collaborators. Although it is *possible* that this patient's SMA activity might have subserved (say) imagery of someone else performing an action or representations of an action-related stimulus, the fact that it was time-locked to an instruction to engage in motor imagery surely raises the probability that this is precisely what the patient was doing.

There is, however, an objection to the interpretation of this experiment given by its authors that cannot be straightforwardly met by appealing to the role of the patient's environment. The worry concerns the legitimacy of applying a NC that has been derived from the study of neurologically unimpaired individuals to individuals, such as this woman, who have suffered massive brain damage. Even if (say) SMA activity is robustly correlated with conscious motor imagery in normal human beings—indeed, even if SMA activity was robustly correlated with conscious motor imagery in this particular patient prior to brain damage—it is a further question whether it is robustly correlated with conscious motor imagery in individuals with massive brain damage.

One might argue that this question can be met by invoking the response just made to the previous objection: if SMA and PPA activity in this patient was not correlated with motor imagery and spatial navigation imagery respectively, then why was it time-locked to the instructions that the patient was given? This response is fine as far as it goes, but in and of itself it doesn't provide any reason to rule out the possibility that the patient was engaged in acts of *unconscious*, stimulus-driven imagery (Levy 2008).

In order to see what lies behind this worry, it is useful to distinguish between two components of a conscious mental state's total neural correlate, what we might call its 'differentiating correlate' and its 'non-differentiating correlate' (Bayne 2010; see also Chalmers 2000; Block 2005). A differentiating neural correlate is a neural state that is specifically correlated with the presence of a certain kind of content consciousness. For example, SMA activity is a differentiating correlate for experiences of motor imagery. A non-differentiating correlate, by contrast, is a neural state that is implicated in all conscious states, irrespective of their content. Although SMA activity is correlated with the presence of conscious motor imagery, it is very unlikely that it represents a total correlate of such states. Instead, it is far more plausible to suppose that SMA gives rise to such experiences only when it is suitably integrated with various kinds of 'non-differentiating' neural activity.

Non-differentiating correlates are not always of central importance to discussions of the neural correlates of consciousness, but they are clearly of vital relevance in the present context, for the central question in which we are interested is whether this patient was conscious at all, rather than whether she was conscious in a particular manner. Unfortunately, we don't really know whether the non-differentiating correlates of consciousness were active in this patient. For one thing, we don't know exactly what the non-differentiating correlates of consciousness are. Moreover, to the extent there are plausible hypotheses about the locus of the non-differentiating correlates of consciousness, those hypotheses were not investigated in this experiment. The upshot is that this study falls some way short of vindicating the claim to have 'demonstrated' that this patient was conscious, although the evidence that it provides is certainly suggestive.[6]

Mindreading, behaviour, and introspection

How might brain-based mindreading of the kind with which we have been concerned interact with the more familiar forms of mindreading that involve behaviour and introspection? As we have already seen, certain aspect of this relationship are broadly 'supportive'. Because available NCs will often fail to determine a unique mental ascription to a subject, theorists will often have reason to appeal to the subject's behaviour and

[6] Note, however, that there is another way in which the ascription of consciousness to this patient might be justified. Briefly put, one might use the correlational method to ascribe mental imagery to the patient, and then use assumptions about the nature of mental imagery—such as the fact that it was sustained for 30 seconds—to argue that it was likely to have been conscious (Shea and Bayne 2010).

introspective reports in order to adjudicate between competing mentalistic interpretations. Available NCs might fail to adjudicate between competing mentalistic hypotheses, but one or more of these hypotheses might be either undermined—or, alternatively, confirmed—by the target's introspective reports and/or behaviour. In this way, introspection, behaviour, and neural data may be thought of as simply different sources of evidence about a person's mental state.

This line of thought also raises a possibility that is decidedly less rosy. By recognizing the possibility of brain-based mindreading are we not also undermining the authority that a person has over the contents of their own mind? Let us consider two manifestations of this concern, one introspective and one behavioural.

Imagine that we have strong neuroimaging evidence for thinking that a certain subject, S, is in pain. We have found strong correlations between N_1 and the presence of pain in the population to which S belongs, and we know that S is in state N_1. But suppose that S denies being in pain, and that we have no reason to doubt the sincerity of S's denial. (In fact, it is possible to finesse the issue of sincerity by supposing that *you* are S.) On the face of things, it is not implausible to suppose that the introspective judgement that one is not in pain is infallible (that is, could not be false); at the very least, we tend to assume that such judgements are incorrigible (that is, could not be rationally corrected by information derived from other sources). The same might be said, incidentally, of the introspective judgement that one *is* in pain. And yet if we are prepared to allow evidence derived from neuroimaging to carry some weight with respect to the ascription of mental states in general, it seems that we ought to allow neuroimaging evidence to lower one's credence in the proposition that one is not in pain. But this conclusion flies in the face of highly plausible views about the kind of epistemic warrant that introspectively-based ascriptions of pain enjoy. Intuitively, the authority that they possess cannot be undermined by third-person data of the kind provided by neuroimaging.

A parallel form of conflict appears to be possible between neuroimaging data and behaviour, where the notion of behaviour is to be understood broadly. To modify a case introduced into the literature by Dennett (1978), imagine that one has neuroimaging evidence for the claim that S believes that he has a brother in Cleveland. However, S does not reason in the ways that someone who had this belief would reason, nor does he act in the ways in which we would expect someone with this belief to act. For example, he denies—with apparent sincerity—that he has a brother in Cleveland. Again, the notion that we should allow our neuroimaging data to trump S's behaviour appears to threaten the authority that we typically accord to behaviour in such contexts.

One might attempt to respond to these challenges by questioning whether they are really coherent. After all, one might argue, given that NCs are grounded in introspection and behaviour, is there not a methodological guarantee that the mindreading data derived from neuroimaging *cannot* dissociate from that which is provided by introspection and behaviour? Although tempting, this line of thought should be resisted. Arguably the correlational method does guarantee that introspective and behavioural data will not *in general* dissociate from neuroimaging data, but it does not guarantee that they cannot dissociate in particular cases. And if they were to dissociate, we would then be faced with the question of how to weigh the evidence drawn from neuroimaging against that derived from behaviour and introspection.

But perhaps we shouldn't put this issue in terms of weighing competing lines of evidence. Let us contrast two conceptions of the relationship between mental states themselves and our introspective and behavioural 'access' to them. According to the first view, mental states are only contingently related to their introspective and behavioural manifestations. This view allows that introspection and/or behavioural dispositions might provide extremely good *evidence* of an individual's mental states, but it denies that they are constitutively related to them. A rival view holds that the relationship between mental states on the one hand and our introspective and behavioural 'access' to them is, or at least can be, constitutive of their possession. If this view were correct, then there might be situations in which the evidence provided by neuroimaging would simply be irrelevant to the question of what mental states the individual was in, for such questions would already have been decided on the basis of introspective and/or behavioural considerations.

The debate between these two views is one of the central questions in the philosophy of mind, and it would be foolish to attempt to engage with it in any serious fashion here. However, it may be useful to consider, albeit in outline sketch, the motivation for each of the two conceptions. Before proceeding to that sketch, we should note that these two views are not straightforwardly exclusive and various hybrid accounts are possible. For example, certain types of mental states might be constitutively related to introspective judgements; others might be constitutively related to behavioural dispositions; and still others might have no constitutive connection to either introspection or behaviour.

Let us begin with the question of whether introspection might be constitutively connected to certain kinds of mental states, such that introspective judgements to the effect that one is (or is not) currently in the state in question are incorrigible. It is clear that there are many kinds of mental states for which such a claim would be highly implausible. For example, our introspective judgements concerning our reasons for action, our character traits, and our behavioural dispositions often involve significant amounts of confabulation and post-hoc rationalization (Wilson 2002). Indeed, there is reason to think that introspection can lead one astray even with respect to one's current conscious states (Bayne and Spener 2010; Haybron 2007; Schwitzgebel 2008; Spener MS). Although many people think that they enjoy visual experiences of the world that are rich in detail, there is good reason to think that such judgements are false and that visual experience is typically sparse in content. These points notwithstanding, there is something to be said for the thought that certain kinds of introspective judgements may be incorrigible. Suppose that you are looking at a tree on a normal summer's day, and you suddenly become aware that this is what you are doing. Arguably, there is a constitutive connection between your introspective judgement and the visual experience towards which it is directed, such that this introspective judgement ('I am now having a visual experience like *this*') could not be false (or at least could not be corrected) (Chalmers 2003; Gertler 2001; Horgan and Kriegel 2007). And if that is the case, then neuroimaging evidence indicating that one was not visually conscious in this manner would need to be explained away rather than accommodated.

The case for ascribing this kind of authority to introspection is not restricted to perceptual experiences but extends—although perhaps for different reasons—to certain types of thoughts. Suppose that I ask myself what I am currently thinking about, and it occurs to me that I am (or have just been) thinking about the prospects of discovering intelligent life elsewhere in the universe. Arguably, this introspective judgement does result from an attempt to *identify* an independent state that I am in—an attempt that might or might not be successful—but is rather a feature of the very fact that I am conscious thinking this thought. In other words, with respect to certain types of conscious thoughts, there may be no gap between introspectively judging that one is thinking that such-and-such and actually thinking that such-and-such. And if that is right, then one's introspective judgements about the contents of one's current thoughts would be immune to corrections by neuroimaging data.

What about constitutive relations between behaviour and mentality? Although there is little to be said for the idea that mental states are constitutively tied to any *particular* behavioural response, there is a great deal to be said for the thesis that certain kinds of mental states involve what Ryle called multi-track dispositions. For example, being angry involves the disposition to produce one or more of a certain range of behaviours in particular contexts, such that someone who was not disposed to produce at least some of these behaviours in relevant contexts simply would not qualify as angry. (Anger, of course, might also involve a particular kind of phenomenal state.) This picture is also attractive as an account of belief. Arguably the notion of belief has essential behavioural elements, such that someone who fails to act in certain ways simply lacks the belief in question. Since we do not typically have privileged access to information about our behavioural dispositions, we can be wrong about our own beliefs (in a way that perhaps we cannot be wrong about our occurrent thoughts). Those who know us well might have insights into our beliefs that we ourselves lack, for they might be better at tracking our behavioural dispositions than we ourselves are. (Perhaps we have been blinded to the true nature of our beliefs by self-deception.) Given the behavioural element to belief, neuroimaging evidence that an individual is (or is not) in a certain belief state might simply be irrelevant. This isn't to say that neuroimaging is incapable of grounding attributions of belief to a subject, but it is to say that in so doing it must respect whatever constitutive connections there are between belief and behaviour.

Finally, some types of mental states are likely to have no constitutive links to either introspection or behaviour. Consider the fact that some people are inclined to report that their dream phenomenology is black and white; others are inclined to report that they dream in colour; and still others are quite unsure of just what their dream phenomenology is like (Schwitzgebel 2011). Perhaps there is a great deal of inter-subjective variation with respect to whether people dream in black and white or in colour. Whatever the facts of the matter, there is reason to doubt whether there are constitutive connections between the nature of our dream experiences and either our introspective judgements or our behavioural dispositions. In cases such as this, we should attempt to integrate the neuroimaging data that we have with whatever we can glean from introspective and behaviour.

Author query: sense

Conclusion

Although I have referred to the practice of ascribing mental states on the basis of neural data as 'mindreading' we have seen that the term is somewhat misleading, for identifying someone's mental states on the basis of information about their neural states is far from direct or unproblematic. 'Mindreading' is possible, but it is a risky business, for it requires a host of assumptions, many of which will be controversial.

This chapter has had a rather narrow focus, for I have restricted my attention to the question of whether and under what conditions neuroimaging might be used to ascribe the kinds of mental states that are already recognized by 'folk psychology'—the intuitive, pre-theoretical framework that we use for understanding the mind. This question can be contrasted with a number of other—and in some ways more radical—questions that might be addressed in connection with mindreading. For example, one might ask whether neuroimaging might be able to reveal personal-level mental states that folk psychology does not yet recognize. One might ask what light neuroimaging might be able to shed on the nature of mental processes. And one might ask what capacity neuroimaging has for revealing the sub-personal architecture of the mind.[7] Although these questions are continuous in certain respects with the question on which I have focused, it is far from clear whether the correlational method that I have articulated here might be able to answer them. That, however, is a topic for another occasion.[8]

References

Abe, N., Suzuki, M., Tsukiura, T., Mori, E., Yamaguchi, K., Itoh, M., and Fujii, T. (2006) Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex* 16: 192–9.

Anderson, M. (2010) Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33(4): 245–313.

Bayne, T. (2010) *The Unity of Consciousness*. Oxford: Oxford University Press.

Bayne, T. and Spener, M. (2010) Introspective humility. *Philosophical Issues* 20: 1–22.

Berlucchi, G. and Buchtel, H.A. (2008) Neuronal plasticity: historical roots and evolution of meaning. *Experimental Brain Research* 192: 307–19.

Block, N. (1995) The mind as the software of the brain. In *An Invitation to Cognitive Science, Vol. 3*, eds D.N. Osherson, L.Gleitman, S.M. Kosslyn, S.Smith, and S. Sternberg, 377–426. Cambridge (MA): MIT Press.

Block, N. (2005) Two neural correlates of consciousness. *Trends in Cognitive Sciences* 9(2): 46–52.

[7] For discussion of this issue see (e.g.) Coltheart (2004), Coltheart (2010), Harley (2004), Henson (2005), Loosemore and Harley (2010), Roskies (2009), and Poldrack and Wagner (2004).

[8] I am grateful to Nicholas Shea and Sarah Richmond for their very helpful comments on a previous draft of this chapter.

Boly, M., Coleman, M.R., Davis, M.H., Hampshire, A., Bor, D., Moonen, G., Maquet, P.A., Pickard, J.D., Laureys, S., and Owen, A.M. (2007) When thoughts become action: an fMRI paradigm to study volitional brain activity in noncommunicative brain injured patients. *Neuroimage* 36: 979–92.

Carruthers, P. (2006) *The Architecture of the* Mind. New York: Oxford University Press.

Chadwick, M.J., Hassabis, D., Weiskopf, N., and Maguire, E.A. (2010) Decoding individual episodic memory traces in the human hippocampus. *Current Biology* 20: 544–7.

Chalmers, D. (2000) What is a neural correlate of consciousness? In *The Neural Correlates of Consciousness*, ed. T. Metzinger, 17–39. Cambridge (MA): MIT Press.

Chalmers, D. (2003) The content and epistemology of phenomenal belief. In *Consciousness: New Philosophical Perspectives*, eds Q. Smith and A. Jokic, 220–72. Oxford: Oxford University Press.

Coltheart, M. (2004) What has functional neuroimaging told us about the mind (so far)? *Cortex* 42: 323–31.

Coltheart, M. (2010) What is functional neuroimaging for? In *Foundational Issues in Human Brain Mapping*, eds S.J. Hanson and M. Bunzl, 263–72. Cambridge (MA): MIT Press.

Davies, M. (1998) Language, thought, and the language of thought (Aunty's own argument revisited). In *Language and Thought*, eds P. Carruther and J. Boucher, 226–47. Cambridge: Cambridge University Press.

Dehaene, S., Le Clec'H, G., Cohen, L., Poline, J.B., van de Moortele, P.F., and Le Bilan, D. (1998) Inferring behavior from functional brain images. *Nature Neuroscience* 1: 549–50.

Dennett, D. (1978) Brain writing and mind reading. In *Brainstorms*, 39–50. Cambridge (MA): MIT Press.

Dennett, D. (1981) A cure for the common code. In *Brainstorms*, 90–108. Cambridge (MA): MIT Press.

Driver, J. and Noesselt, T. (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron* 57: 11–23.

Fodor, J. (1975) *The Language of Thought*. Cambridge (MA): Harvard University Press.

Fodor, J. (1983) *The Modularity of Mind*. Cambridge (MA): MIT Press.

Fodor, J. (2000) *The Mind Doesn't Work that Way*. Cambridge (MA): MIT Press.

Fodor, J. (2008) *LOT 2: The Language of Thought Revisited*. Oxford: Oxford University Press.

Gertler, B. (2001) Introspecting phenomenal states. *Philosophy and Phenomenological Research* 63: 305–28.

Ghazanfar, A.A. and Schroeder, C.E. (2006) Is neocortex essentially multisensory? *Trends in Cognitive Sciences* 10: 278–85.

Harley, T.A. (2004) Does cognitive neuropsychology have a future? *Cognitive Neuropsychology* 21: 3–16.

Haybron, D. (2007) Do we know how happy we are? On some limits of affective introspection and recall. *Noûs* 41(3): 394–428.

Haynes, J.-D. and Rees, G. (2005) Predicting the stream of consciousness from activity in human visual cortex. *Current Biology* 15: 1301–7.

Haynes, J.-D. and Rees, G. (2006) Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* 7(7), 523–34.

Haynes, J-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R.E. (2007) Reading hidden intentions in the human brain. *Current Biology* 17: 323–8.

Henson, R.N.A. (2005) What can functional imaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology* A 58: 193–233.

Horgan, T. and Kriegel, U. (2007) Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues* 17(1): 123–44.

Hurley, S. and Noë, A. (2003) Neural plasticity and consciousness. *Biology and Philosophy* 18: 131–68.

Kamitani, Y. and Tong, F. (2005) Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8: 679–85.

Kozel, F.A., Johnson, K.A., Mu, Q., Grenesko, E.L., Laken, S.J., and George, M.S. (2005) Detecting deception using functional magnetic resonance imaging. *Biological Psychiatry* 58: 605–13.

Kozel, F.A., Padgett, T.M., and George, M.S. (2004) A replication study of the neural correlates of deception. *Behavioural Neuroscience* 118: 852–6.

Langleben, D.D., Schroeder, L., Maldijan, J.A., Gur, R.C., McDonald, S., Ragland, J.D., O'Brien, C.P., and Childress, A.R. (2002) Brain activity during simulated deception: an event-related functional magnetic resonance study. *Neuroimage* 15: 727–32.

Levine, J. (1983) Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64(4): 354–61.

Levy, N. (2008) Going beyond the evidence. *American Journal of Bioethics* 8(9): 19–21.

Loosemore, R. and Harley, T. (2010) Brains and minds: On the usefulness of localization data to cognitive psychology. In *Foundational Issues in Human Brain Mapping*, eds S.J. Hanson and M. Bunzl, 217–40. Cambridge (MA): MIT Press.

Macaluso, E. (2006) Multisensory processing in sensory-specific cortical areas. *Neuroscientist* 12(4): 327–38.

Maloney, C. (1989) *The Mundane Matter of the Mental Language*. Cambridge: Cambridge University Press.

Matthews, R. (1987) *The Measure of Mind*. Oxford: Oxford University Press.

Matthews, R. (2007) *The Measure of Mind: Propositional Attitudes and Their Ascription*. Oxford: Oxford University Press.

Merabet, L.B. and Pascual-Leone, A. (2010) Neural reorganization following sensory loss: the opportunity of change. *Nature Reviews Neuroscience* 11(1): 44–52.

Monti, M.M., Vanhaudenhuyse, A., Coleman, M.R., Boly, M., Pickard, J.D., Tshibanda, J-F.L., Owen, A.M., and Laureys, S. (2010) Willful modulation of brain activity and communication in disorders of consciousness. *New England Journal of Medicine* 362(7): 579–89.

Moran, R. (2001) *Authority and Estrangement: An Essay on Self-Knowledge.* Princeton (NJ): Princeton University Press.

Nachev, P. and Hacker, P.M.S. (2010) Covert cognition in the persistent vegetative state. *Progress in Neurobiology* 91: 68–76.

Nachev, P., Kennard, C., and Husain, M. (2008) Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience* 9: 856–69.

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences* 10(9): 424–30.

Nunez, J.M., Casey, B.J., Egner, T., Hare, T., and Hirsch, J. (2005) Intentional false responding shares neural substrates with response conflict and cognitive control. *Neuroimage* 25(1): 267–77.

Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., and Pickard, J.D. (2006) Detecting awareness in the vegetative state. *Science* 313: 1402.

Pascual-Leone, A. and Hamilton, R. (2001) The metamodal organization of the brain. In *Progress in Brain Research* 134: 427–45, eds C. Casanova and M. Ptito. Amsterdam: Elsevier.

Poldrack, R.A. (2006) Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences* 10(2): 59–63.

Poldrack, R.A. and Wagner, A.D. (2004) What can neuroimaging tell us about the mind? Insights from pre-frontal cortex. *Current Directions in Psychological Science* 13(5): 177–81.

Polyn, S.M., Natu, V.S., Cohen, J.D., and Norman, K.A. (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310: 1963–6.

Quartz, S.R. and Sejnowski, T.J. (1997) The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences* 20: 537–56.

Rey, G. (1995) A not 'merely empirical' argument for a language of thought. *Philosophical Perspectives* 9: 201–22.

Richiardi, J., Eryilmaz, H., Schwartz, W., Vuilleumier, P., and Van De Ville, D. (2011) Decoding brain states from fMRI connectivity graphs. *NeuroImage* 56: 616–26.

Roskies, A. (2009) Brain-mind and structure-function relationships: A methodological response to Coltheart. *Philosophy of Science* 76(5): 927–39.

Rushworth, M.F.S., Walton, M.E., Kennerley, S.W., and Bannerman, D.M. (2004) Actions sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences* 8: 410–17.

Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M.P., Dold, G., and Hallett, M. (1996) Activation of the primary visual cortex by braille reading in blind subjects. *Nature* 380(6574): 526–8.

Schwitzgebel, E. (2008) The unreliability of naïve introspection. *The Philosophical Review* 117(2): 245–73.

Schwitzgebel, E. (2011) *Perplexities of Consciousness*. Cambridge (MA): MIT Press.

Shea, N. and Bayne, T. (2010) The vegetative state and the science of consciousness. *British Journal for the Philosophy of Science* 61: 459–84.

Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M.D. (2011) Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex* <doi: 10.1093/cercor/bhr099>.

Spence, S.A., Farrow, T.F.D., Herford, A.E., Wilkinson, I.D., Zheng, Y., and Woodruff, P.W. (2001) Behavioural and functional anatomical correlates of deception in humans. *Neuroreport* 12: 2849–53.

Spence, S.A., Kaylor-Hughes, C.J., Brook, M.L., Lankappa, S.T., and Wilkinson, I.D. (2008) ‘Munchausen’s syndrome by proxy’ or a ‘miscarriage of justice’? An initial application of functional neuroimaging to the question of guilt versus innocence. *European Psychiatry* 23: 309–14.

Spener, M. MS. *Phenomenal adequacy and introspective evidence*. University of Oxford.

Sperber, D. (2001) In defense of massive modularity. In *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*, ed. E. Dupoux, 47–57. Cambridge (MA): MIT Press.

Wilson, T.D. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge (MA): Harvard University Press.