Progress in Neurobiology 91 (2010) 68–76

Contents lists available at ScienceDirect

Progress in Neurobiology

journal homepage: www.elsevier.com/locate/pneurobio

Covert cognition in the persistent vegetative state

P. Nachev [a,*], P.M.S. Hacker [b]

[a] Institute of Neurology, UCL, Queen Square, London WC1N 3BG, UK
[b] St. John's College, Oxford OX1 3JP, UK

ARTICLE INFO

Article history:
Received 19 June 2009
Received in revised form 30 December 2009
Accepted 27 January 2010

Keywords:
Persistent vegetative state
Functional magnetic resonance imaging
Consciousness
Covert cognition

ABSTRACT

Brain damage can sometimes render a patient persistently unresponsive and yet apparently awake, admitting the possibility that the absence of overt voluntary behaviour might conceal a retained capacity for covert cognition. When given instructions to perform a cognitive task, a minority of patients in such a so-called persistent vegetative state (PVS) has recently been found to exhibit patterns of brain activation closely matching those observed in normal subjects obeying the same instructions. These data have been widely interpreted as implying the detection of covert "consciousness". Here we show that this inference is not supported by the extant data because it relies on critical assumptions, obscured by conceptual unclarities, that are either untested or untestable. We set out the proper grounds for ascribing psychological attributes to PVS patients from physiological evidence of any kind, and offer a perspicuous conceptual framework for future empirical studies in the field.

Contents

1. Introduction

Whether patients rendered unresponsive by brain injury have hidden residual cognitive powers[1] has long been the subject of intense speculation. Until recently, an answer has seemed impossible, for it is in the nature of cognition that a capacity for action and communication is required to be able to manifest it, and it is in the nature of unresponsiveness that a capacity for action and communication is absent. With the advent of brain imaging and the remarkable correspondence between the cognitive episodes and processes of a person and the contemporaneous activity of his brain that it has revealed, some researchers have argued that one may be able to identify *neural correlates* of cognition upon which differentiation between different kinds of mental powers and their exercise

Abbreviations: PVS, persistent vegetative state; fMRI, functional magnetic resonance imaging; BOLD signal, blood oxygenation level dependent signal; SMA, supplementary motor area.

* Corresponding author.
E-mail address: p.nachev@imperial.ac.uk (P. Nachev).

[1] Following a convention widely observed in neuroscientific writing we use the term cognition where the broader sense of cogitation is meant.

0301-0082/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pneurobio.2010.01.009

could reliably be made (Boly et al., 2007; Laureys, 2005; Owen and Coleman, 2008a,b).

This theoretical possibility relies on three premises. First, it is assumed that brain injury may affect one power while sparing another. This would seem quite uncontroversial: for example, focal brain damage may affect speech production without affecting comprehension and vice versa. The presence of impairment in one power therefore does not necessarily imply an impairment in any other; whether it does or not is a matter to be settled in each particular case. Thus, that a patient is unable to *respond* does not necessarily mean that he is also unable to *understand*.

Second, it is in the nature of our cognitive powers that their exercise may, at times, be unobservable. It is possible that someone outwardly indistinguishable from a brain-dead body could have the same passive powers of observation, attention, and cognition as someone able to manifest them. This hypothetical case is the asymptote of the "locked-in syndrome", where brain damage has impaired the capacity to produce movement in any part of the body except the eyelids, and other powers are left intact.

Third, although one's cognitive powers normally depend on a body for their expression, an artificial vehicle may, in certain respects, do just as well. For example, a prosthetic hand may be used to gesture in the same way as a real hand, or a speech synthesizer to speak as with a real voice. That a prosthesis is quite unlike real flesh, and an electronic voice quite unlike a real voice need not materially affect the subject's ability to gesture or speak. Indeed, within the framework set by the concept of a living being or indeed of a person, there are no conceptual limits to what "effectors" may be substituted, for none of our cognitive powers depends on features of parts of our bodies that can be neither simulated nor replaced, at least in theory.[2]

If these three premises are straightforward, building a valid conceptual framework from them is not. To the new empirical tools it is therefore essential to add conceptual tools of commensurate power. Without such tools it is impossible to make sense of the new data, just as we cannot make sense of quantum thermodynamics without the exotic mathematics of statistical mechanics, or of general relativity without the geometry of non-Euclidean spaces. Here we show why the conceptual methodology needs radical revision, and give an outline of what seems to us the correct approach. Inevitably, we examine both the conceptual and the empirical, but our focus is principally on the conceptual for that is where the difficulties lie. The consequences of the errors we identify are nonetheless highly practical, potentially distorting the clinical management of unresponsive patients, and – perhaps more importantly – distorting the legal concept of a person itself. Our task is therefore far from merely academic.

[2] It should be noted that contrary to what was suggested by Dainton (2007) the brain is not a limiting case of a mutilated human being. A brain in formaldehyde is not a pickled human being nor a mutilated human corpse. The intelligibility of exhibiting cognitive powers by the exercise of prostheses does not show that the possessor of cognitive capacities is the brain. One can imagine, in science fiction, that a living brain might be linked to prosthetic eyes and ears, mechanical limbs and a computerized voice box. Then, so the story may run, the voice and limbs may exhibit thought and volition. Does this not show that the brain (and not the living human being) is the subject that thinks and wills? No. What it shows is that this imaginary being, which we might dub a cerebroid', does so. We need a brain in order to be able to think (walk and talk), just as an jet aeroplane needs engines in order to fly. But aircraft engines cannot fly any more than brains can think (walk or talk). It is human beings, who have brains, that think – not their brains, which neither have brains (since they are brains) nor minds (there is no such thing as a thought's crossing the brain's mind) or bodies (someone may have a beautiful body, but their brain cannot be said to have a beautiful body). The brain is no more a limiting case of a mutilated human being than an aircraft engine is a limiting case of a damaged aeroplane. For more detailed discussion of Dainton, see Bennett and Hacker (2008).

2. Key concepts and conceptions

To make any kind of empirical assertion about PVS, we must first be clear about the theoretical concepts without which no such assertions can be made, and the fundamental conceptions that are presupposed when they are made. These are the nature and characteristics of the patients in question, the relation between cognition and neural activity, and the fundamental nature of psychological attributes. We take each in turn.

2.1. The persistent vegetative state

Our interest here is in a specific category of brain-damaged patients whose membership is defined not by the aetiology of the injury or the anatomical locus of the damage but solely by their behaviour in the context of brain damage. By definition, such patients show a complete absence of any behaviour that can be unequivocally interpreted as voluntary. This state is differentiable from "brain death" by accompanying evidence of activity for which a functioning brain would seem a necessity such as an apparent sleep/wake cycle. The preservation of such "vegetative functions" and the tendency for such patients to remain stably in such a state has given rise to the term "persistent vegetative state" (Medical Aspects of the Persistent Vegetative State (1), 1994a; 1994a,b; 1996; Jennett and Plum, 1972).

Note that the widely accepted definition makes no reference to consciousness, and therefore does not require a clarification of the nature of consciousness: a controversial topic we have addressed elsewhere (Bennett and Hacker, 2003; Hacker, 2007). PVS is nonetheless widely – and misleadingly thought of as a disturbance of consciousness or awareness. It is true that unconsciousness (in the intransitive sense (Bennett and Hacker, 2003)) has features in common with PVS, but this does not mean that it must play a part in its description. One may be outwardly identical with a PVS patient and yet remain fully aware of one's environment (e.g. asymptotic locked-in syndrome); equally, one may have no awareness and yet exhibit complex behaviour wholly incompatible with a diagnosis of PVS (e.g. complex partial seizures). To label a PVS patient unconscious is rather like labelling an aphonic patient amusic because he is unable to sing, or a Parkinsonian patient indecisive because he is akinetic. The description is neither accurate nor inaccurate – just beside the point. For this reason we do not speak of unconsciousness here, but use the descriptive term "unresponsiveness".

Indeed, it is undesirable to speak of unconsciousness because it clouds the critical distinction we are trying to make. In the presence of complete unresponsiveness, it is outwardly impossible to distinguish between someone unable to manifest what are otherwise normal cognitive powers – i.e. someone in the asymptotic locked-in syndrome – and someone whose cognitive powers are no different from those of a brain-dead patient (Monti et al., 2009). Where along this continuum – and there is no reason to suppose that it is anything but a continuum – a PVS patient might fall is what the clinician or experimenter is trying to establish. As we shall see, many of the difficulties arise from erroneously supposing that PVS patients must be like one or the other with nothing in between.

Whatever label we use, it should be emphasised that the criterion for diagnosing PVS in the context of head injury is behavioural, that is, lack of behavioural response to the environment. Although all PVS patients *must* have brain damage of one kind or another there is no reason why the damage has to be the *same* in each case: PVS may be the common manifestation of a wide array of disparate forms of neural dysfunction. Thus, the physiological disturbance in one case is no guide to any other: the same clinical outcome can have a variety of causes and mechanisms. PVS patients cannot be assumed to be

homogeneous, either aetiologically or physiologically, just as patients with other behaviourally determined syndromes such as hemiparesis or hemianopia cannot. This limits the generalizability of inferences from individual cases.

2.2. Cognition and brain activity

Current technology does not permit us to obtain a comprehensive picture of the relation between behaviour and neural activity in the brain. For such a picture to be comprehensive it needs to capture simultaneously *every* feature of *every* neuron that is material to the behaviour under investigation. Since determining which feature and which neuron are material is precisely what we are trying to establish, a sparse picture cannot be assumed to suffice: we might be missing precisely that which is essential. The markers of neural activity we observe – blood oxygenation level dependent (BOLD) signal, local field potentials, single cell neurophysiology, and others – are therefore mere *correlates* of the behaviour, and will have to remain so until our empirical tools improve.

Now correlations may allow us to make inferences by induction. For example, the exercise of two different kinds of mental powers – e.g. arithmetical calculation or verbal recall – may be paralleled by differences in neural activity that are both stable within a subject and reproducible across subjects. Such a correlation may allow someone inductively to infer from brain activity alone which one of *this* set of mental powers is more likely to be being exercised in *this* particular set of circumstances. The security of the inference here will depend on the tightness of the correlation and the integrity of the assumptions on which the correlation is based. It is, in short, an empirical question, to be established case by case. *Within these limitations*, it is a perfectly valid form of inference.

The presence of a correlation between a pattern of activation and a behaviour does not, however, imply any kind of *causal* relation. The same pattern of activation might occur in the presence of completely different behaviour, or no behaviour at all. Indeed, it is clear that very different kinds of mental activity, on the one hand, or overt behaviour, on the other, may be neurally indistinguishable. For example, the medial motor areas are activated in essentially the same way by any movement irrespective of what, how, and why one moves; only a crude rostrocaudal gradient of the complexity of the movement and its attendant circumstances is discernible (Nachev et al., 2008; Picard and Strick, 1996). Similarly, activation that attends speech is largely indifferent to the content of what is said or meant, and to whether it is spoken or merely mentally rehearsed (Nachev et al., 2008; Picard and Strick, 1996). Even as elementary a contrast as making a movement vs *withholding* a movement – in the context in which a movement may be expected – is barely reflected in the spatiotemporal dynamics of the BOLD signal (Curtis et al., 2005). Importantly, the neural activity may nonetheless be highly characteristic and stereotyped: only it does not distinguish between the kinds of behaviour under investigation. That it should be so is of course profoundly unsurprising: the BOLD response is a crude index of aggregate neural activity, discretized at a resolution whose relation to the underlying neural organisation is unclear (Logothetis, 2008).

The presence of a particular pattern of brain activity therefore does not imply the behaviour with which it has been associated. It may be useful, inductively, only when we are merely choosing from a selection of previously characterized behaviours *one* of which we *already know* is taking place.

3. The nature of psychological attributes

The rules governing the correct application of psychological attributes are more complex than those governing physical attributes. It is difficult to be wrong about the grounds for saying that someone is tall; the grounds for saying that someone is (say) in pain, however, are the subject of a great deal of debate. For our purposes, we need only make two points most neuroscientists do not dispute.

First, like any attribute, a psychological attribute needs to be anchored in observable phenomena if it is possible to refer to it in any kind of discussion: scientific or other. This is not to say that psychological attributes *always* have an external manifestation; only that without *some* link to the outside world, not necessarily a contemporaneous one, they are opaque to discussion, let alone scientific enquiry. Equally, this is not to say that the manifestation of an attribute has to be a conventional one: just as a gagged man can signal his pain with his hands almost as well as with his voice, so any other voluntary behaviour (e.g. changes in breathing rate) may be used to do so. But without *some* manifestation, we can be neither meaningfully contradicted nor supported in our judgements here.

Second, psychological attributes are not the properties of a mind, to be contrasted with the body, but the properties of a living being as a whole. When a being is damaged, and the powers whose exercise psychological attributes reflect are impaired, it may therefore not be possible to ascribe an attribute where it would otherwise have been possible. The pattern of impairments observed in such situations will naturally vary with the location and nature of the damage; this relation is often very complex, particularly when the brain is involved. But we should be clear that there are no *a priori* grounds for believing that when one set of powers is impaired – for example, the ability to speak – another *must* be intact – for example, the ability to understand speech: each case must be assessed on its merits.

The implications of these points are clear. First, the application of psychological predicates depends on observable features of a living being. If an experience has no manifestation in the subject's report, behaviour, physiology or any other discernible aspect, then nothing about it can be asserted or denied. Second, the *vehicle* by which the features are conveyed to us need not be relevant to the attribute in question, although it can make our task harder. For example, one's ability to spell is exhibited no less by handwriting than it is by emailing, although the latter makes the possibility for deception easier. Third, when one ability is affected by injury, we cannot assume that every other is intact. That a mute patient can squeeze our hand to command, for example, does not mean that he can *withhold* a squeeze on command.

4. Empirical data

With these preliminaries in mind, let us examine the key empirical data. We should note that a great deal of speculation rests on very little data, and much of the early data has been extensively criticized by others already (e.g. Owen and Coleman, 2008a). Here we focus on a study of a single PVS patient – reported in highly condensed form (Owen et al., 2006) – which is widely considered to overcome the deficiencies of earlier work, and to settle the matter conclusively in the specific case studied (Monti et al., 2009).

The design of the core experiment is straightforward (Owen et al., 2006). The authors chose different kinds of covert mental activity that could be successfully performed by normal subjects without any kind of physical movement: imagining playing tennis and imagining exploring one's house. By "imagining" the experimenters meant conjuring up a series of mental images of a hypothetical series of real actions, and instructed the (normal) subjects to interpret the command in this way. Although imagining an action need have nothing to do with actually performing one, and is not in itself an action at all (Hacker, 2007; White, 1968), in the specific sense meant in this experiment it is mental activity

that may have a genuine duration and may be differentiable from other kinds of activity at any time. Moreover, since one can choose to engage in it or not, it falls within the spectrum of what is voluntary. Thus a patient who has lost all power to act or communicate (e.g. having been paralysed prior to undergoing an operation and having regained consciousness while still paralysed) may, at least in theory, engage in mental activity in exactly the same way as a normal patient.

To establish an association between the mental activity and a measurable neural correlate the authors collected fMRI data on normal subjects while they alternated between episodes of imagining playing tennis and imagining exploring a house. In keeping with previous imaging studies of similar tasks, the patterns of cortical activation observed in each case were found to differ from each other and from the "rest" condition more than they varied with time and across subjects. Inspection of the contrast between the two patterns of cortical activity was thus sufficient to determine which of the mental activities each of the normal subjects was engaged in.

In the critical experiment, the authors gave the PVS patient the same verbal instructions as the normal subjects and observed the BOLD response across the brain. Fitting an analogous model to the imaging data revealed a similar pattern of activation in each case and a similar difference between them as in the normal subjects.

4.1. Interpretation

From this observation the authors inferred that the PVS patient must have been imagining playing tennis and exploring her house in much the same way as the normal subjects were. Furthermore, since the normal subjects were free to disobey the command to perform the task, they interpreted the subject's apparent "cooperation" as indicating a deliberate *intention*, and therefore a clear capacity for voluntary cognitive activity.

These inferences are logically invalid. First, identity of brain activation on fMRI – even if it could be demonstrated – does not imply identity of mental activity. The authors have shown that two different mental activities are associated with two different patterns of neural activity; the presence of the latter could therefore inductively – and legitimately – be inferred from the former. The converse – that the same neural activity implies the same mental activity (or indeed any mental activity at all) – has not been demonstrated in any shape or form. To argue otherwise is to commit a logical fallacy often referred to as "affirming the consequent", one of the two common errors of reasoning elegantly demonstrated by Wason's four card trick familiar to psychology undergraduates (Wason, 1966). That it is a fallacy requires no proof: it is obvious enough in daily life. If it were not, that gold glitters could be taken to imply that everything that glitters is gold; that the Queen is rich, that everyone who is rich is the Queen; that the Earth is round, that everything that is round is the Earth; and so forth.

Second, one can only speak of voluntary activity where there is evidence of a capacity for making a *choice* (Nachev et al., 2005; Passingham, 1995). Here there seems to be the *possibility* of making a choice – cooperating vs not cooperating with the experimenter – but no possibility of *testing* whether such a capacity is being exercised or not. This is so because not cooperating is indistinguishable from having no capacity to perform the mental activity in the first place. Had the experimenters asked the patient to select x of n mental tasks and shown evidence of performing a task on one occasion and *refraining* from performing it on another, some element of choice could have been demonstrated because *both* alternatives – the criterion for speaking of any kind of choice – would have been tested. But here the question has not been tested at all.

In sum, the patient has not been shown to have engaged in any kind of mental activity: voluntarily or otherwise.

4.2. Inferential validity

We should consider what, if any, circumstances could make the interpretation valid. There are two aspects we need to examine.

4.2.1. Neural activity without mental activity

First, the inference would be correct if it is impossible for the characteristic neural activity to occur without the mental activity with which it is correlated. We should therefore examine whether this is an assumption one could legitimately make. Fortunately, there is evidence both from normal subjects and from patients with brain damage that allows us to answer this question with confidence.

4.2.1.1. Normal subjects. Surveyed at the resolution current neuroimaging offers, essentially identical patterns of neural signalling may be evoked by mental activities that are enormously disparate. For example, activation of the supplementary motor area (SMA) – one of the critical regions in the Owen et al.'s study – is observed not only when imagining playing tennis but also when imagining or performing any kind of action (Nachev et al., 2008; Picard and Strick, 1996; Rushworth et al., 2004). If the patient were to respond to the command to imagine playing tennis by imagining playing chess we would certainly not think of him as having obeyed it. Indeed, far from implying imagining a specific action, SMA activation does not imply imagining or performing any kind of action. Robust neural activity in the region is routinely observed in association with abstaining from actions or passively observing them in others, and with exposure – conscious or subliminal – to a wide range of stimuli incidentally related to action (Nachev et al., 2008). A given pattern of activation could not possibly indicate that any mental activity of imagining action is going on: the relation between the two is far too degenerate. This fact is elegantly illustrated by neuroimaging databases such as BrainMap (http://www.brainmapdbj.org/): even within the limited set of behaviours examined with imaging, the variation in patterns of activation is less than the variation in behaviour.

4.2.1.2. Patients with brain damage. It could be argued that the degeneracy of the relation between brain activity and mental activity observed in normal subjects is immaterial if we are examining a *choice* between a small number of tasks already known to be dissociable, and if we may safely assume the subject will be performing *one* of these tasks if he is performing any task at all. Such an assumption would naturally beg the question, for the presence or absence of a capacity to respond correctly to an instruction to perform a task is precisely what we are trying to establish.

Let us nonetheless examine such a case. Here we cannot use data from normal subjects, for the only way they can fail to engage in a mental activity (at least the kind that is dissociable from another with imaging) is not to try. There are patients with focal brain damage, however, who are able to understand an instruction and perform a task, but only in some circumstances. Here one can reasonably make the assumption that the subject knows what he is meant to be doing – because he does it correctly some of the time – and that therefore finding *no* neural difference between success and failure cannot be caused by the subject's trying to do something else or failing to try to do anything at all. The only possible explanation of such an outcome would be that the neural activity does not inevitably imply performing the task. Patients with disordered attention owing to damage to the posterior parietal lobe offer an elegant example of this (Driver et al., 1999).

Such patients sometimes show the phenomenon of extinction: their ability to perceive a salient event in one side of visual space is impaired by the simultaneous occurrence of another event in the opposite hemifield (Driver, 2001). Thus an ipsilesional event may be used to "gate" the perception of a contralesional event, allowing the neural differences between perceived and unperceived stimuli to be directly compared. The results show these differences to be remarkably small, with robust activation in striate, extrastriate and category-specific visual areas in extinguished trials (Rees et al., 2000, 2002). Had these patients been unresponsive, someone following the logic of Owen et al. (2006) would have wrongly concluded that all of these stimuli were perceived.

In sum, there is enough evidence that a marker of neural activity cannot generally be taken to imply a mental activity: whether it does has to be decided in each specific case, by empirical means. This assumption therefore cannot be used to rescue the inference.

4.2.2. Abductive inference

A second possibility is to appeal to the notion of "best explanation" or so-called abductive inference (Peirce, 1955). The idea here is illustrated by a simple example. Let us suppose we found the grass outside to be wet and wanted to give a causal explanation for this state of affairs. Surely we do not have to observe rain falling on the grass to infer that rain is the best explanation for why the grass is wet? It is of course logically possible for a herd of incontinent cows to be responsible but to insist on excluding this possibility seems merely pedantic. In the present case, if a patient's brain shows a characteristic pattern of neural activity that corresponds to understanding and obeying the command just issued to him, isn't the most obvious conclusion that this is exactly what he is doing? Unfortunately, abductive reasoning does not work here, for three reasons.

First, for the evidence to count in this way there must be independent grounds for foreclosing other possibilities, or at least giving them some kind of probabilistic weighting. But on what basis could we conceivably do this? We are familiar with grass and with the events that usually explain its wetness, and can therefore weight the probabilities appropriately, but we know nothing about the relation between neural activity and psychological attributes in PVS patients: there is no weighting we can do here. Our knowledge of this relation in normal patients is of no help here because it is precisely the difference between the normal and the PVS case that we need to explain.

Secondly, to make any kind of inference we must have a means of knowing whether or not we are right or wrong: here we have none, as a simple example illustrates. To know what X or an X is, is to be able to differentiate things that are X or X-s from things that are not. The two kinds of grounds on which differentiation may be made – logical or empirical – differ in important ways. When we say that gold has atomic number 79 we are citing a logical criterion for something to be gold. Having atomic number 79 is constitutive of what it is for something to be or to consist of gold. So our statement is not an empirical one, but an explicative one. By contrast, when we say that the mineral lump before us is gold we are making an empirical claim that may be correct or erroneous, proved or disproved. Proof here necessarily depends on some kind of *test* which has been validated against some kind of *standard*. In this example, atomic mass spectroscopy may be considered a standard because gold has a unique atomic mass that may be reliably differentiated from *all* other atomic masses.

If one does not have access to a standard test, one may have to rely on some other, for example, whether or not the lump exhibits a golden glitter. If one is to have any kind of confidence in such a test, one would have to explore its adequacy against the "gold standard" across a range of candidate minerals. The measures of adequacy of a test are simple and well-established: we need to know its tendency to produce false positives (specificity) and false negatives (sensitivity). Exhibiting a golden glitter is both insensitive (e.g. sylvanite) and non-specific (e.g. pyrite). Critically, we can only estimate the sensitivity and specificity of our test if we have a standard we can check it against, *and* if we have explored its performance across the range of test cases. Thus, if we have never come across pyrite we might think that exhibiting a golden glitter is a highly specific test for gold.

Let us now translate this example into the domain of the psychological. When we see that someone is in pain, and are asked how we know that he is, we shall cite, as the justification of our judgement, his pain-behaviour in the circumstances. Such justification is logical, not empirical. This kind of behaviour, in such circumstances, is what is called 'pain-behaviour', and in these circumstances it provides logically good evidence (not inductive evidence) for the person's being in pain. Of course, in certain circumstances (but not all) the person may be pretending; so our judgement is defeasible. But a person's *sincere* avowal that he is in pain is not defeasible (although it might be exaggerated). We could not disprove his sincere avowal (and accompanying behaviour and grimaces of pain) by reference to some other standard of correct pain-ascription – for his sincere avowal in these circumstances is our final court of appeal. And this is built in to our very concept of pain.

Now supposing that our patient was unresponsive and we wished to use some physiological test to establish whether he was in pain or not. For the results of this test to be interpretable we have to be able to validate it. Critically, such a validation has to be in the specific context studied: we cannot use normal subjects because it is precisely the difference between normal subjects and those in the pathological state that we wish to capture. But what standard for the application of "pain" can we have here? Without behaviour there is no standard, and without a standard there is no means of validating the inference. Neither the sensitivity nor the specificity of the test can be established: the confusion matrix remains blank.

Third, abductive inference here would in any event argue *against* Owen et al.'s conclusions, not in favour of them. The simplest explanation for the presence of brain activity in the absence of any overt behaviour is that the activity is decoupled from behaviour, just as it is in the patients with extinction we discussed earlier. Indeed, we have excellent grounds for being highly sceptical of the test. If on asking the subject to move his hand we observed the same activation as that seen in someone normal moving his hand the specificity of the test in determining whether or not someone is moving his hand would clearly be shown to be poor. Exactly the same applies to imagining moving one's hand: the only difference is that one cannot show that imagining moving one's hand is not taking place. Using imagined rather than real movements offers no advantage: it merely makes it impossible to prove or disprove what is being inferred.

In sum, none of the potential avenues for escape is available: the inference remains fundamentally invalid. Why one might nonetheless wish to cling to it is an interesting question, but it is a question for the psychology of neuroscience rather than the neuroscience of psychology; we therefore take it up in the Appendix A. Our business now is the grounds of ascribing covert mental powers in PVS.

5. Mental powers in PVS

It should be clear by now that the presence of covert mental activity cannot be inferred from any neurophysiological correlate because the inevitable absence of empirical confirmation or disconfirmation makes it impossible to determine whether or not the inference is correct. Does this mean that one could never establish anything about the mental powers of someone who lacks

the ability to move in any way? No. Neural activity might, in rather special circumstances, provide us with grounds for the ascription of cognitive attributes. In essence, the patient must *use* neural activity in order to communicate via a *code*. To show how this might legitimately be done we need to examine three cases: the locked-in syndrome, the asymptotic locked-in syndrome, and a patient in PVS.

5.1. Communicating in the locked-in syndrome

Consider a patient whose powers of movement are limited to blinking, and whose mental powers are unknown. We can only know the latter as far as they can be revealed by the former. Thus, we can never know the patient's capacity for (say) ironic inflection because there is no code that would allow us to convey tonal inflections by blinking. We can, however, know anything that can theoretically be conveyed by a *binary code*. Importantly, since language can readily be translated into code we can use the medium by which the most distinctively human powers are manifest.

To use a code, however, we must satisfy the requirements for communicating in a code. The source – here the patient – must be able to encode the data, and the recipient – the doctor – must be able to decode it. Communication will be limited by some bandwidth necessarily lower than for unencoded communication of the same level of compression. The processes of encoding and decoding will need to be independent of the data conveyed. Most importantly, we shall need an index of the fidelity of coding.

Each of these things presents serious problems in the locked-in syndrome. Encoding speech in a binary code requires many of the mental powers we are using the code to establish, resulting in an inevitable floor effect. For example, in order to use Morse code, the patient would need to be capable of stably associating each Morse sign with its literal counterpart, and to acquire this set of associations if he does not know them already. The bandwidth of communicating in an uncompressed binary code by blinking is poor. Although compression strategies such as those employed by Dasher (Ward and MacKay, 2002) can help, the limitation is clearly substantial. Furthermore, a blink code cannot be perfect because the subject will occasionally blink by accident. Although in general this should merely add noise, it can cause systematic errors if some irrelevant events, internal or external, make him blink involuntarily.

The major difficulty, however, is establishing the fidelity of the code. In this situation it is logically impossible to distinguish between an error in the encoding and an error in the data being conveyed. This is so because – unlike in the normal state – we have no standard for the correctness of what is conveyed until *after* the encoding process. Thus when someone we know to be in full possession of his mental powers makes an error in sign language (say), we naturally assume that the problem lies not in his ability to conceive the idea but only in relating it in sign language: our grounds for believing this are that he is able to express it perfectly well in ordinary language. In the locked-in case no such assumptions can be made because there is no alternative medium by which we can check where the error arises. For this reason, it is very difficult to establish an index of the fidelity of the code. In the normal case, we can rely on the coherence of what is conveyed to quantify the error; here we cannot, since the coherence of the source is precisely the question at stake.

Our estimate of coding fidelity is made easier if we used an alphabetical code to generate words. Here, since random 10 character strings are effectively unique (MacKay, 2003), we can be reasonably confident that a word so encoded corresponds to what is meant to be conveyed. By contrast, attempts to interrogate the patient via yes/no answers (e.g. "Are you in pain?") require a much longer sequence of questions before we can be confident that the pattern of responding has not occurred by chance. It is very hard to resist the temptation to assume that when a patient blinks in response to a command to blink, he has *understood* it: precisely because in the normal case we require no proof. But one cannot speak of understanding a command if one has no means of distinguishing between understanding, not understanding, and misunderstanding generally, which binary response to a single command does not and cannot provide. For example, we would not say that the patient understood the command if he responded in the same way to the words "do not blink", or the counter-imperative "blink if I say "do not blink" but not if I say "blink" etc.

In summary, the grounds for a psychological attribute – inevitably linguistic or in some other way symbolic – may be conveyed by a locked-in patient if he is able to communicate them via a code. There will inevitably be limits on what is communicated that depend on the patient's abilities to perform the encoding, the limitations of the code itself, and any errors in the code. Encoding errors are impossible to distinguish from errors in the encoded content, for the coded communication is all we have. Critically, the fact that the patient appears able to communicate *something* is no guarantee that he can communicate *anything*: the inferences we may legitimately draw are confined solely to what is communicated.

5.2. Communicating in the asymptotic locked-in syndrome

Now consider a locked-in patient who has lost even the capacity to blink and has become asymptotically locked-in. In theory, one may ask the patient to communicate in exactly the same way as by blinking, except using brain activation. The communication here is also necessarily coded, but the code is *physiological* rather than behavioural. All the foregoing constraints apply, in addition to a few others.

First, deliberately activating one's brain in a specific pattern is not something we naturally *do*. We may naturally choose to engage in cognitive activities that happen to be radiologically dissociable, but we do not naturally *bring about* these activities for an ulterior purpose. The patient will therefore have to translate the message into code and perform the activities on which the code is based in the sequence determined by the encoded message.

Indeed, neural activation can only be used as a means of communication to the extent to which it can be under voluntary control. In general, since neural processes underlie cognition they can never perfectly reflect its outcome. Non-voluntary activation related to the encoding and the bringing about of the cognitive activity will inevitably restrict the range of activation patterns that may be voluntarily obtained. Areas considered to be close to the "effector systems" and remote from areas activated by cognition – for example, primary motor cortex – would seem to be better candidates than those with the converse properties. A region commonly activated during speech or writing would seem to be a poor choice. The SMA – which is activated by language and a substantial array of mental operations – would therefore seem to be especially so.

Since binary encoding here requires a *choice* between voluntarily activating one of two areas, we also have to consider how they differ in their interactions with the message being conveyed. Thus, although hand dominance effects do impinge on the BOLD signal, the hierarchical equivalence between the left and right primary motor cortex means responses where each side is assigned to 1 or 0 will be balanced for complexity much better than in the case of pairings such as SMA and parahippocampal place area. The same applies for the tasks used to activate each area: the greater the mental powers they require and the greater the mismatch between them the higher the floor on complexity and the greater

the interference in what may be conveyed. A task such as imagining waving with one's (right vs left) hand is easily intelligible to someone with very little mental power; playing tennis requires one to understand the concept of tennis, to recall what the game involves, and so on.

Even so, no such code can be perfect because of non-voluntary neural activation. For example, imagine a patient being asked to say something about his right hand: it is obvious that non-voluntary activation of the left primary motor cortex will interfere here. Although this example is easy enough to anticipate, there may be many others where we might not be sure whether or not the primary motor cortex should be activated. Communication may thus fail in context dependent fashion, and just as in the locked-in syndrome case, we will not have any easy means of distinguishing cognitive from communication failure.

The asymptotic locked-in syndrome, then, presents all the foregoing difficulties in addition to others peculiar to communicating via a neural code. These limitations notwithstanding, it is theoretically possible to establish a channel of communication with an asymptotically locked-in patient broad enough to convey grounds for a wide spectrum of psychological attributes.

5.3. Communicating in PVS: the Reverse Turing Test

Let us assume that we succeed in obtaining coded language output from a PVS patient. If we already *knew* that the patient was asymptotically locked-in, that his cognitive powers were otherwise intact, nothing else would be required. But this is precisely what we *do not* know, and what our communication is trying to establish. Does encoded language output allow us to conclude that the patient is covertly "conscious"?

To take the production of a sequence of signs in a binary code in the PVS case as a marker of "consciousness" is to assume – unjustifiably – that the patient is asymptotically locked-in, which is precisely what we are trying to prove. The production of a sequence of signs as such does not in itself require much mental power if any. After all, we do not attribute linguistic powers to tape recorders even though they can produce word-sequences. Nor is it enough for these sequences to be novel, or seemingly responsive to context: simulations of these things are easy enough to achieve with relatively simple computer programmes which no-one would argue confer the status of intelligence on the machines running them. Imagine, for example, that Owen et al.'s patient was an *asymptotic perseverator*, incapable of anything but the repetition of an externally instructed activity until another instruction arrived (which situation would of course perfectly explain the imaging data presented): surely we could no more call such a patient "conscious" than we would a keyboard with sticky keys.

No, the correct way to interpret the language output from a PVS patient is as one would interpret any kind of output from a creature of unknown mental powers: the only powers we can attribute to it are those on display.

Here computer science offers a useful model: the Turing test (Turing, 1950). The simple intuition behind it is that the powers one can legitimately attribute to an unknown entity with which one can communicate only via a terminal are best defined by comparison with a real human being in identical circumstances. Run in reverse, the Turing test can be used to identify communications that cannot be readily explained by any simple algorithm, and therefore imply the exercise of powers justifying the application of such psychological predicates as they normally warrant.

The Turing test is helpful because even very simple programmes can be remarkably good at deceiving an observer. Indeed, there are plenty of algorithms that would do far better than the snippets of communication extracted from comatose patients. For example, the capacity correctly to answer six simple questions requiring recall of major past events (e.g. "Have you ever travelled to country X?") may be taken as evidence of the minimal powers this requires and *no more.* Since it would be easy to write a computer program that does the same thing, one cannot conclude that the patient is therefore asymptotically locked-in. By contrast, no simple algorithm could have generated the book the famous locked-in patient Jean-Dominique Bauby succeeded in "dictating" by blinking: our conclusion that he was in full possession of his mental powers would not have been any different had he been communicating via a neural code rather than by blinking.

The use of the Turing test here is also helpful in illuminating the senselessness of labelling the patient as "covertly conscious" or indeed "unconscious": these terms have a meaning only where the full spectrum of mental powers may be assumed to be broadly intact (Hacker, 2007). But it is precisely the characteristics of these powers, particularly in relation to language – not of any kind of conscious "content" – that we ought to be aiming to characterize. Consequently, the outcome of such an analysis will not be a simple, binary label – "conscious" or "unconscious" – but a complex, continuous description of what the patient can and cannot do, with no assumptions about his mental powers beyond the evidence before us. Thus, just as the ability of a patient with visuo-spatial neglect to detect a single target in his contralesional hemifield is no guide to his ability to detect the *same* target in the presence of competing stimuli in the ipsilesional field, so the ability of a patient to "answer" correctly a specific question on one occasion is no guide to his wider cognitive powers. Moreover, there is no guarantee that the powers we do establish in any particular case – even if they are to do with the manipulation of words – will be any less automatic than the purely vegetative functions we already know to be intact. Someone exhibiting asymptotic perseveration is no less an automaton than someone whose behavioural repertoire is limited to a sleep-wake cycle. Critically, since we have no knowledge of what patterns of dysfunction such patients may have there are no grounds for excluding any possibility: this is uncharted territory.

In summary, obtaining encoded language output from a PVS patient would be the *beginning* – not the end – of establishing his cognitive powers. The nature of encoded communication makes it difficult – but not impossible – to determine the extent of these powers with confidence. A reverse Turing Test is a simple way of conceiving the task and its pitfalls. Critically, the outcome of our assessment should not be a binary label – "conscious" or "unconscious" – but a comprehensive description of the patient's residual powers. Then, and only then, might we be in a position to ascribe consciousness to such a patient.

5.4. Clinical implications

Although the foregoing may seem academic, it has important implications for the clinical management of patients in PVS.

First, if *no* exercise of cognitive powers of any kind can be detected, the possibility will always remain that there are powers the exercise of which current neurophysiological techniques may not reveal, unless of course the damage is so profound that the neural substrate could not conceivably sustain any.

Second, the presence of task-specific brain activation on *imagining* an action is no more proof that imagining is taking place than task-specific brain activation on *failing to perform* an overt action is proof that the observer is blind and the action is actually taking place. The same applies to any other kind of covert cognitive activity.

Third, showing that a patient has the capacity to communicate via brain activity is the *beginning* of the assessment of his powers,

not the end. A direct response to a simple instruction implies no wider cognitive powers than the minimal required for its execution (cf. the "asymptotic perseverator" discussed above).

Fourth, the assessment of the powers of a PVS patient will be constrained by the limitations of the neural code, including the impossibility of distinguishing between errors in the coding and errors in the content the subject is attempting to convey. Inevitably, only what can be communicated via language will be accessible to an observer.

Fifth, the judgment to be made in each case is not *binary* – "conscious" vs "unconscious" – but *continuous* – powers *x*, *y*, *z*... expressed to *m*, *n*, *o*... degree – no discrete "threshold" value can therefore be easily arrived at. The situation is analogous to profound developmental cognitive disability, and the approach to management ought therefore to be similar.

Sixth, although the presence of task-specific activation does not imply a covert capacity to perform the task, it may have prognostic implications for the evolution of the patient's powers, a *marker* of future outcomes. This is something to be determined by longitudinal studies, and is orthogonal to the question of what powers such activity may be legitimately argued to signify.

6. Conclusion

It has been argued that our analysis is unduly strict. That we should observe *any* homology between the neural responses of a PVS patient and those in the normal state would seem at least to open a possibility that the opposite result would have barred. And since existing in the asymptotic locked-in syndrome is so awful to contemplate, even weak evidence of such a possibility merits wide dissemination and further investigation, or so the argument goes.

Three points may be made in reply. First, the published work shows no equivocation in its conclusions: it speaks of "a clear act of intention", "beyond any doubt", and so on. Far from suggesting a possibility, it insists on what it unjustifiably claims to be proven reality.

Second, the clinical decisions on which this evidence is brought to bear are not one sided. For every relative of a living PVS patient given (probably false) hope, another is burdened with the guilt of having acquiesced in the withdrawal of treatment from someone who – he has been led to believe – may have been more alive than it seemed. There are moral costs to false positives as well as to false negatives.

Finally – and ominously – accepting the notion that neurophysiological evidence can *replace* evidential behaviour that warrants ascription of consciousness and associated cognitive attributes is but a short step from accepting the notion that it may sometimes *override* it. Since being a person is defined by possession of appropriate cognitive and volitional powers as manifest in behaviour that exhibits them, such a development would fundamentally distort this key legal and social concept. Indeed, we are already seeing the consequences of this in the questionable use of fMRI in deriving supposedly "objective" measures of pain, sincerity, or belief. The brain – or rather a feeble caricature of the brain – is here taken to be the highest arbiter of what the subject feels, intends, or believes. The narrow question of the relation between brain activity and cognition in near death is thus further clouding the wider and much more important question of their relation in normal life.

Acknowledgements

We are grateful to Professor Max Bennett for his comments on earlier versions of the text, and to the anonymous reviewers whose suggestions have greatly improved the paper.

Appendix A

A.1. Psychological certainty

We have seen that it is not difficult to show that the inferences made in Owen et al. (2006), are fundamentally flawed: the counter-arguments rely on questions of science and logical inference that are well established. How is it that one can be so easily misled?

Here we should draw attention to a point often ignored in cognitive neuroscience. We are familiar and comfortable with the notion that certain *empirical methods* are inherently prone to systematic errors of one kind or another. For example, functional imaging creates arbitrarily discretized pictures of the functional architecture of the brain because of the habit of reporting data in thresholded form. This distortion tends to encourage us to build discrete models of brain function even if the data do not compel them, or even are against it (see Nachev et al., 2008 for a discussion of this in relation to the SMA).

It is also true, however, that our thinking is distorted by errors arising from our *conceptual methods*: the way in which we deploy the concepts on which our models of the brain are based (Bennett and Hacker, 2003). Critically, the nature of the mental is such that conceptual confusion here is an ever-present danger. It is not at all obvious, for example, that despite the superficial similarity, the use and role of the first person singular pronoun "I" is not the same as that of "he" or "she". "He" and "she" admit of misidentification and reference failure, "I" allows no such possibility, although that does not mean that it always involves referential success and correct identification – but rather that it involves no identification at all, and no reference in the sense in which "he" and "she" may refer or fail to refer to someone. Misled by this false analogy, one may try to construe psychological self-ascription on the model of third person ascriptions, and so mistakenly draw parallels between apprehending one's own intentions and coming to know those of others (Lau et al., 2004).

In the present case, the confusion seems to arise from mistakenly treating neurophysiological correlates on the model of psychological attributes. If we see someone writhing in pain in circumstances of injury it is ridiculous to doubt whether he is in pain, since writhing thus is constitutive evidence for being in pain: the relation between the grounds for our ascription and what we ascribe is logical, not empirical. The truth of this is easy to see when one considers the absurdities of a question such as: "He has broken his leg and is screaming and writhing – I wonder whether he is in pain". Here the grounds for pain-ascription are *logically* good evidence for his being in pain, and in the absence of defeating conditions doubt is out of place. Even more obviously, in the first person case, such sentences as "I wonder whether I am in pain", "I think it hurts, but I am not sure", "I believe I am in pain, but I may be wrong" are patently absurd. Self-ascription of pain is groundless – an acculturated extension of natural pain-behaviour. So the question of doubt in one's own case cannot arise (unless it is a matter of a borderline instance of pain), and neither therefore can that of certainty.

It therefore makes no sense to speak of the sensitivity or specificity of a form of behaviour – e.g. crying out, 'It hurts!' – in determining whether or not someone is in pain, for if the subject is sincere neither false negatives ('I cried out in pleasure by mistake') nor false positives ('I cried out in pain but there was no pain there') are possible. (Although, of course, one may cry out in mistaken anticipation of pain.) Thus, if one makes the relatively obscure error of treating neurophysiological correlates as if they were logically on the same level as psychological attributes, the otherwise obvious error of ignoring the uncertainty of the empirical correlation of neurophysiological events and mental activity is concealed.

This confusion is interesting because a much commoner error in neuroscience is the converse: the assumption that the application of psychological predicates is ordinarily based on inductive inference. For example, it is mistakenly argued that someone's report of (say) pain is a contingent response to an internal neural state of affairs that some other means – perhaps functional imaging (Cruccu et al., 2004) – could allow us to apprehend better. We have already seen that this cannot make sense: there are no neurophysiological grounds on which a sincere self-report may be over-ruled.

References

Bennett, M.R., Hacker, P.M.S., 2003. Philosophical Foundations of Neuroscience. Blackwell Publishers.

Bennett, M.R., Hacker, P.M.S., 2008. History of Cognitive Neuroscience. Wiley-Blackwell, Oxford, pp. 242–243.

Boly, M., Coleman, M.R., Davis, M.H., Hampshire, A., Bor, D., Moonen, G., Maquet, P.A., Pickard, J.D., Laureys, S., Owen, A.M., 2007. When thoughts become action: an fMRI paradigm to study volitional brain activity in non-communicative brain injured patients. NeuroImage 36, 979–992.

Cruccu, G., Anand, P., Attal, N., Garcia-Larrea, L., Haanpaa, M., Jorum, E., Serra, J., Jensen, T.S., 2004. EFNS guidelines on neuropathic pain assessment. European Journal of Neurology 11, 153–162.

Curtis, C.E., Cole, M.W., Rao, V.Y., D'Esposito, M., 2005. Canceling planned action: an fMRI study of countermanding saccades. Cerebral Cortex 15, 1281–1289.

Dainton, B., 2007. Wittgenstein and the brain. Science 317, 901.

Driver, J., 2001. Perceptual awareness and its loss in unilateral neglect and extinction. Cognition 79, 39.

Driver, J., Vuilleumier, P., Husain, M., 1999. Spatial neglect and extinction. In: Gazzaniga, M. (Ed.), The New Cognitive Neurosciences.. MIT Press, Cambridge, USA, ISBN: 0262071959.

Hacker, P.M.S., 2007. Human Nature: The Categorial Framework. Blackwell Publishing.

Jennett, B., Plum, F., 1972. Persistent vegetative state after brain damage A syndrome in search of a name. Lancet 1, 734–737.

Lau, H.C., Rogers, R.D., Haggard, P., Passingham, R.E., 2004. Attention to Intention. American Association for the Advancement of Science, pp. 1208–1210.

Laureys, S., 2005. The neural correlate of (un)awareness: lessons from the vegetative state. Trends in Cognitive Sciences 9, 556–559.

Logothetis, N.K., 2008. What we can do and what we cannot do with fMRI. Nature 453, 869.

MacKay, D.J.C., 2003. Information Theory Inference and Learning Algorithms. Cambridge University Press.

Medical Aspects of the Persistent Vegetative State (1), 1994a. The multi-society task force on PVS. The New England Journal of Medicine 330, 1499–1508.

Medical Aspects of the Persistent Vegetative State (2), 1994b. The multi-society task force on PVS. The New England Journal of Medicine 330, 1572–1579.

Monti, M.M., Coleman, M.R., Owen, A.M., 2009. Neuroimaging and the vegetative state. New York Academy Sciences Annals 1157, 81–89.

Nachev, P., Kennard, C., Husain, M., 2008. Functional role of the supplementary and pre-supplementary motor areas. Nature Reviews 9, 856–869.

Nachev, P., Rees, G., Parton, A., Kennard, C., Husain, M., 2005. Volition and conflict in human medial frontal cortex. Current Biology 15, 122–128.

Owen, A.M., Coleman, M.R., 2008a. Detecting awareness in the vegetative state. Annals of the New York Academy of Sciences 1129, 130–138.

Owen, A.M., Coleman, M.R., 2008b. Functional neuroimaging of the vegetative state. Nature Reviews 9, 235–243.

Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., Pickard, J.D., 2006. Detecting awareness in the vegetative state. Science (New York, N.Y.) 313, 1402.

Passingham, R.E., 1995. The Frontal Lobes and Voluntary Action. Oxford University Press, USA.

Peirce, C.S., 1955. Philosophical Writings of Peirce. Courier Dover Publications.

Picard, N., Strick, P.L., 1996. Motor areas of the medial wall: a review of their location and functional activation. Cerebral Cortex 6, 342–353.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., Driver, J., 2000. Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. Brain 123, 1624–1633.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., Driver, J., 2002. Neural correlates of conscious and unconscious vision in parietal extinction. Neurocase 8, 387–393.

Rushworth, M.F.S., Walton, M.E., Kennerley, S.W., Bannerman, D.M., 2004. Action sets and decisions in the medial frontal cortex. Trends in Cognitive Sciences 8, 410–417.

The Permanent Vegetative State, 1996. Review by a working group convened by the Royal College of Physicians and endorsed by the Conference of Medical Royal Colleges and their faculties of the United Kingdom. Journal of the Royal College of Physicians of London 30, 119–121.

Turing, A.M., 1950. Computing machinery and intelligence. Mind 59, 433–460.

Ward, D.J., MacKay, D.J.C., 2002. Fast Hands-free Writing by Gaze Direction. Arxiv preprint cs.HC/ 0204030.

Wason, P.C., 1966. Reasoning. New Horizons in Psychology 1, 135–151.

White, A.R., 1968. The Philosophy of Action. Oxford University Press, USA.