Neural Mechanisms of Decision Making and the Personal Level

Nicholas Shea

For the *Oxford Handbook of Philosophy and Psychiatry*, KWM Fulford, M Davies, G Graham, J Sadler, G Stanghellini and T Thornton (eds.), OUP.
January 2012
Word count 10,920 (including abstract)

Abstract

Can findings from psychology and cognitive neuroscience about the neural mechanisms involved in decision-making can tell us anything useful about the commonly-understood mental phenomenon of making voluntary choices? Two philosophical objections are considered. First, that the neural data is subpersonal, and so cannot enter into illuminating explanations of personal level phenomena like voluntary action. Secondly, that mental properties are multiply realized in the brain in such a way as to make them insusceptible to neuroscientific study. The paper argues that both objections would be weakened by the discovery of empirical generalisations connecting subpersonal properties with the personal level. It gives three case studies that furnish evidence to that effect. It argues that the existence of such interrelations are consistent with a plausible construal of the personal-subpersonal distinction. Furthermore, there is no reason to suppose that the notion subpersonal representation relied on in cognitive neuroscience illicitly imports personal-level phenomena like consciousness or normativity, or is otherwise explanatorily problematic.

Keywords

decision making, reward, prediction error, desire, subpersonal level, representation, information processing, reduction, mind-brain identity, reasons as causes

1 Introduction

How is the brain relevant to understanding the mind? A venerable body of opinion says that it isn't. Many in psychiatry and psychology, as well as in philosophy, still hold that it is a mistake to study the brain if you are trying to understand the mind. It would be like trying to figure out what an abacus does by plotting the physical dynamics of beads on wires. That might give you a reasonable grip on the trajectories of the beads but would entirely lose sight of the calculations being performed. Very few are explicit dualists, of course, but many think that because a straightforward reduction of the mental to the neural is untenable, we should study mental phenomena in isolation from goings on in the brain.

This paper examines some of the obstacles to relying on neural information in understanding the mind. A major objection is formulated in terms of the distinction between personal and subpersonal properties (section 2), objecting in particular to the idea of subpersonal representations (section 3). Multiple realisability presents a separate challenge (section 4). Section 5 argues that the well worked-out body of research on reward-guided decision making has overcome those challenges. Section 6 shows how that work can be used to explain some real practical cases: inter-individual differences in decision making (§6.1), choice behaviour in addiction (§6.2), and positive symptoms in schizophrenia (§6.3).

An overarching theme is that this philosophical account suggests a particular perspective on patients, as people, that is relevant to the clinical situation. Neural evidence can explain the personal level phenomenon of voluntary decision making not because it shows how patients are being caused to act by their brain – the 'brain made me do it' approach to the problem, which treats patients like mechanisms – but because it allows us to see why these personal level processes should unfold in unusual ways.

2 Personal and Subpersonal

Cognitive neuroscience purports to explain behaviour by reference to recognisably psychological properties: perceiving a state of affairs, representing an outcome, valuing a reward. But it attributes those properties to parts of the brain. That presupposes that the mind can be unified with neuroscience relatively straightforwardly.

The unification is achieved via the idea of internal representations. A paradigmatic example of a representation is a written sentence, e.g. "Snow is white". The sentence is a collection of marks on the page that also has semantic properties: it concerns some other things in the world (snow, whiteness) and can be true or false, depending on how the world is. Indeed, giving the conditions under which a written sentence is true does a lot to capture its meaning. The internal representations postulated by psychology and cognitive neuroscience share these features of paradigmatic representations. They are physical particulars, proper parts of people (wholly or substantially in the brain), that have semantic properties. These internal physical particulars enter into causal processes, and they way those processes unfold depends upon their physical properties. However, the system's dispositions to move between these physical particulars can be set up so as to be faithful to their semantic properties. That is the crucial insight behind the success of computers, and it is a critical assumption of information-processing psychology: intelligent behaviour is a result of information processing over internal representations, which are internal particulars with semantic properties.

Cognitive neuroscience often goes further. Not only are there representations in the brain that are as physically real as words on the page, but a person may instantiate mental properties (e.g. perceiving a tomato or desiring orange juice) in virtue of having appropriate internal representations in the brain (neural representations of, e.g., features of the tomato or the subjective value of orange juice). The idea is that, when neural representations play the right functional role in the rest of the neural architecture, then they can be the basis on which a whole person has mental properties like perceiving and desiring.

A standard objection opposes the first move: the very idea of neural representations. Properties of people (and their brains and bodies) can be divided into those at the *personal level*[1] and those at the *subpersonal level* (Dennett 1969). The commonsense understanding of the mind, *folk psychology*, is at the personal level. Personal level properties are those which are familiar from everyday descriptions of people and their mental lives: perceptual states like seeing, hearing and tasting; cognitive states like believing, desiring and remembering; emotional states and moods; feelings like being in pain; and so on. Explanations at the personal level rationalise actions in terms of mental states.

Even if the distinction cannot in the end be rigidly delineated, two characteristic features of the mind do allow us to pick out paradigmatic instances of personal level properties. The first is consciousness. Conscious experiences are at the personal level, as are mental states that can be brought to consciousness. For example, episodic memories are at the personal level, even when not being experienced occurrently, since they can come to be consciously experienced. So are unattended features of the visual field that can draw our attention and become conscious. The second characteristic feature is having contents that are suitable for explaining and justifying why a person acts as she does. It is not just a brute causal fact that perceiving chocolate makes a person reach out and take it. Combined with the near-universal desire for chocolate, the perceptual experience rationalises the action.

By contrast, properties at the subpersonal level are not conscious and may not be suitable for rationalising explanations. At the subpersonal level we can observe that a person's chocolate-seeking behaviour correlates with (and is perhaps caused by) the firing of neurons in the orbitofrontal cortex, and that as the person eats more chocolate, the firing of those neurons attenuates and their chocolate-seeking behaviour reduces. But the fact that a particular neuron in orbitofrontal cortex, call it neuron 42375, fires does not rationalise any of the actions it causes. We can describe various properties of neuron 42375: its neural type (e.g. pyramidal cell), anatomical location, network connections and firing rate. But none of these throws any rational light on the action of eating a chocolate. Even if the person's reaching out to take the chocolate is caused in part by neuron 42375, neither its firing pattern, nor any other of its purely neural properties, rationalises the action. Nor is neuron 42375 conscious. No one thinks that the activity of a single neuron is sufficient for consciousness, even if, integrated in the appropriate neural networks, this neuron were the basis for the neural difference between desiring chocolate and being satiated. So the firing pattern of neuron 42375 is a subpersonal property of the agent.

According to one influential view, explanations that deal in personal level properties are of a distinctive kind, importantly different from the kinds of explanations offered in science, which deals in nomological generalisations about how the world happens to work (Hornsby 1997). John McDowell argues that it is a mistake to mix together properties from the two explanatory schemes:

> concepts of the propositional attitudes have their proper home in explanations of a special sort: explanations in which things are made intelligible by being revealed to be,

[1] I follow the convention of using italics when introducing a term by using, rather than mentioning it.

> or to approximate to being, as they rationally ought to be. This is to be contrasted with a style of explanation in which one makes things intelligible by representing their coming into being as a particular instance of how things generally tend to happen. (McDowell 1985: 389)

Others think that there is no deep difference between personal and subpersonal level properties and that many acceptable explanations appeal to both (Rey 2001), in particular that it can be appropriate to explain a personal level property like an action in subpersonal terms. Whether that is legitimate is one of the topics for this paper. However, even those who refuse to dichotomise can still accept a rough-and-ready distinction, with the conscious, contentful states appealed to by folk psychology being paradigmatic of the personal level.

The personal-subpersonal distinction forms the basis of an attack of the very idea of subpersonal representation. The idea of representing is found at the personal level. If it thereby belongs to a different explanatory scheme, inapplicable to things like neural structures (which are not persons), then it would be a mistake to invoke the idea of neural representation at all. The acceptability of the idea of subpersonal representation therefore depends in part on how deep the divide is between the personal and subpersonal level. Those who rely on subpersonal representations need not subscribe to the view that they are of the same kind as personal level representations like beliefs and desires. But they do have to assert the existence of a subpersonal variety of representation and reject the view that the personal-subpersonal distinction marks an insuperable explanatory divide.

This is not the place to mount a philosophical argument for the interaction between personal and subpersonal levels (see Davies 2000). However, its plausibility partly turns on whether there are actual cases where subpersonal level facts furnish useful generalisations about personal level phenomena. The aim of this paper is to offer some worked examples of interaction between personal- and subpersonal-level properties.

3 Subpersonal Representation

3.1 Making space for subpersonal representation

Some have argued that scientific psychology can talk "as if" subpersonal physical particulars are representations, provided the representational talk is ultimately discharged in favour of unproblematic subpersonal level properties (Hornsby 2000). However, scientists are not restricted to choosing between using the personal level notion of representation metaphorically or not using it at all. Psychology and cognitive neuroscience can define and work with their own technical terms, including a subpersonal concept of representation (Davies 2000). Their use of subpersonal representation need not inherit all the properties of personal level varieties of representation (beliefs, desires, perceptions, etc.); in particular the normativity and consciousness. Their commitment to the existence of subpersonal mental representations is unproblematic provided the notion of internal representation introduced above – that of subpersonal particulars interacting causally in virtue of physical properties in ways that respect associated semantic properties like their correctness or satisfaction conditions – does not introduce properties that are part of an explanatorily isolated realm.

Cognitive neuroscientists work out what a bit of the brain is representing by seeing how its response profile connects to the behaviour of the organism and to features of the outside world. For example, some patterns of neural firing are said to represent the probability of reward (Yang and Shadlen 2007). They are thought to have that content because their firing rate correlates with the probability that choosing a particular stimulus will lead to a reward, in the context of evidence that the neural firing is causally important for the choice behaviour. So the key question is whether there is a subpersonal variety of representation that comports with cognitive neuroscientific practice of this kind. Given scientists' use, various straightforwardly causal properties seem appropriate to characterising these internal states, like carrying correlational information, and having various natural functions (Davies 2005).

The project of giving a fully-satisfying account of the contents that figure in subpersonal information-processing psychology remains an open philosophical question, and an important one. But there is no reason to doubt that the scientists' concept or concepts of subpersonal representation can be explicated without appeal to problematic properties like normativity and consciousness. As it is used in subpersonal contexts, representation is tied closely to other clearly subpersonal properties like correlation, natural information, isomorphism, natural teleology, and so on; not to people, their social relations and associated norms. That is so whether or not subpersonal representation turns out to be reducible to other natural properties. A broadly reductive approach to subpersonal representation remains promising, although it is far from fully worked-out. But even if a reduction is not available, subpersonal representation may still be a perfectly acceptable physical property, on a par with other non-reducible properties in special sciences like biology, chemistry and geology. It can be integrated with other scientific properties though generalisations and *ceteris paribus* laws without being reducible to them.

Either way, there is nothing constitutively normative about subpersonal representation. That is to say, it is not constitutive of an information processing state having the content it does that certain norms apply to that state. The difference between a correct and an incorrect representation is just another descriptive distinction. Norms may apply to that distinction, as they do to many other descriptive distinctions (e.g. whether a person is cycling on a road or on the footpath). For example, if a piece of information processing in the brain relies on a subpersonal representation in calculating how to satisfy your desires, it turns out to be false, and things go badly for you as a result, then we might indeed say there was something wrong, normatively, with the subpersonal representation. But the normativity is not built into the property of subpersonally-representing. In this domain correct versus incorrect is a descriptive distinction to which norms may attach. Even if personal level representations were constitutively normative (which remains open to controversy), subpersonal representations do not share that feature. What they do share with personal level representations is that something like correctness conditions or satisfaction conditions play a role in explaining behaviour.

3.2 Interaction of subpersonal representations with the personal level

Having made space for subpersonal representations, whether reductive or non-reductive, the question arises as to their relation to personal level representations like beliefs, desires, perceptions, and so on. On the most optimistic view personal level representation will reduce to one or another variety of subpersonal representation. At the moment that looks unlikely.

Even without reduction, however, there are many ways of integrating personal level representation with the subpersonal that would not create an unbridgeable divide. The nature of personal level representation may lie in its constitutive connections with other personal level properties: various normative connections between representations inter se, or between states and actions, may be constitutive of their identity. But the personal and subpersonal levels may interact, for example where empirical findings at the subpersonal level show that a given personal level conception does not correspond to a property that humans actually instantiate (Davies 2000). There may also be non-strict generalisations (admitting of exceptions or *ceteris paribus* clauses) or other explanatory relations connecting personal level representations with subpersonal information processing.

Those who argue that personal level properties fall on the far side of an unbridgeable gulf need to do more than to appeal to non-reducibility or anomalousness of the mental (in the sense of Davidson 1970), since anomalous monism is compatible with there being robust *ceteris paribus* laws linking the mental and the physical (Shea 2003). They need to show that the personal level resists any kind of explanatory integration with the subpersonal. The force of those arguments depends in part on the success of the cognitive neuroscientific enterprise. If scientific psychology discovers lots of robust generalisations linking the personal and subpersonal, then premises about the explanatorily hermetic status of personal level properties start to look rather less secure.

The examples below show that cognitive neuroscience has indeed had some success in integrating phenomena across the personal-subpersonal divide. Alive to the distinction, we can assess these examples without being beguiled by the familiarity of personal level talk, aware of the possibility that personal level properties could be being smuggled in and predicated of parts of the brain in inappropriate ways. Humans have a tendency to over-attribute intentionality, for example when we see intentions at work in floods and storms (Bloom 2004). So we should be careful that representation talk in cognitive neuroscience is not metaphorical or merely instrumental (eliminable "as if" talk), or straightforwardly false.

A second merit of marking the distinction is that it forces us to accept that different representational properties are in play. At the very least there is personal level representation on the one hand and subpersonal representation on the other; more likely, there are several varieties of each. For the reasons discussed above, subpersonal representational content is likely to be rather different in kind than personal level representational content. We don't have to subscribe to the thesis that brains and persons fall under incommensurable schemes of explanation to think that, as a matter of empirical fact, the firing of a single neuron just cannot have belief-type content. Most neuroscientists are extremely cautious about their claims, but there are still plenty of papers that overreach, especially amongst those that are picked up outside the field. So it is as well to be alive to inappropriate deployment of personal level properties.

Even if the personal-subpersonal distinction is not a huge gulf that undermines the very idea of subpersonal representation, as some claim, paying attention to the distinction does alert us to the important difference between attributing content to some aspect of neural activity as part of an information processing account of the performance of a task, on the one hand, and uncovering the constitutive basis of personal level phenomena like believing, desiring and perceiving, on the other.

4 The Challenge from Multiple Realisability

A second reason for caution about cognitive neuroscientific practice derives from a mainstream part of naturalistic philosophy of mind. The philosophical orthodoxy of the last 40 years has been that mental states are multiply realizable in real physical systems, and are likely to be multiply realized in the brains of actual people. The prospect of finding a neural property shared by all and only those organisms that are in pain, say, which was the hope of the central state materialists of the 1950s and 1960s, has been displaced. However, cognitive neuroscience seems to have ignored the philosophical lesson about multiple realisation and dived straight in to look for information processing mechanisms in the brain that are shared by many or all human subjects. What is the status of those findings, and of the philosophical motivations that they seem to contradict?

Standard practice in cognitive psychology is to proceed in a way that is independent of knowledge of brain mechanisms. From observing subjects' patterns of behaviour in experimental settings, especially reaction times and patterns of error, but also dissociations between abilities in pathological cases, we are able to build up a picture of how information is being processed to drive behaviour. In principle these information processing steps could be realised differently in different people's brains, and could even be realised differently in a given person's brain on different occasions.

The information processing account divides the mechanism leading to behaviour into a series of boxes, each of which performs its own step or computation and then communicates the results to other stages of processing. The stages of processing are functionally defined. What makes it the case that a particular part of the brain is performing a particular part of the information-processing is the functional role of that part of the brain. And those functional roles can be re-assigned dynamically. Analogously, a standard personal computer stores many representations in random access memory (RAM). There is no systematic correlation between different parts of a RAM chip and different kinds of stored information (hence 'random'). Indeed, there is no strict mapping between the pieces of information appreciated by the user and the information stored at particular locations. The words that appear serially in a sentence on the screen need not be stored serially in RAM locations. So the computer analogy supports the idea that there may be only a very loose connection between stages of the true information processing account of the system's operation and the neural areas that are involved.

Before the rise of cognitive neuroscience, the practice of psychology tended to respect this assumption. Neuroscience might study the wiring of the brain and some of its basic mechanisms, like the synaptic plasticity involved in long term potentiation and long term depression, but the psychological level of description was taken to be relatively autonomous from these details, and multiply realized in them. With the rise of spatially-detailed brain imaging techniques like fMRI and PET that assumption has been displaced by the assumption that some psychological functions are relatively consistently realized. Cognitive neuroscientists now look for repeatable, generalisable events in the brain that correspond directly to particular stages of the information processing story.

For example, it is now well-established that there are fairly stable mappings between some basic psychological functions and particular brain areas. That is particularly true of

functions that are proximal to input and output: the processing of basic perceptual features, and the motor programs that drive action sequences. Many of these commonalities apply, not only to all normal humans, but also to other primates. For example, the perception of movement depends upon visual area V5/MT and it is now very plausible that neural firing in this area is the basis of the content difference between different perceived directions of motion in both monkeys and humans (Rees, Friston, and Koch 2000). Similar results have been established for many other contents of perception. There is also a mapping between motor cortex and effectors that is generalisable across individuals. Such broad-level generalisations leave space for a lot of variable realisation at the level of neural firing. In very simple brains like in the sea slug or nematode it is possible to reidentify individual neurons across individuals, with particular neurons being dedicated to the same range of functions in all members of the species (White et al. 1986). Nothing like that is true in primate cortex. The distributed patterns of neural firing that are the basis of representing a particular perceptual content vary between individuals.

Multiple realisability leaves it open that the same psychological state could be realised in different ways in the same individuals at different times. Here too neural evidence suggests that, even if mental properties are multiply realizable, actual variation in their realisation is less disparate that it might be. fMRI data can be analysed by powerful pattern classifiers to identify the detailed spatial patterns that are associated with different contents. For example, pattern classifiers can learn the distributed pattern of voxel activation in early visual areas that is characteristic of the orientation of edges being viewed (Kamitani and Tong 2005). At higher levels, pattern classifiers have been trained to predict, with reasonable accuracy, whether a subject intends to add or subtract a pair of numbers that are yet to be presented (Haynes et al. 2007). Although there are very many neurons in each voxel, these results show that representing the same content on different occasions has a relatively consistent effect on the local distribution of oxygenated haemoglobin, which is in turn coupled to local differences in firing rates (Mukamel et al. 2005). So these results demonstrate the relative intra-personal stability of the distributed patterns of neural activation responsible for perceiving particular contents. They show that there can be a tight correspondence between a subject's instantiating a psychological property and what is going on in their brain. That suggests that there are cases where the mind-brain relation is intimate enough that neural data can be illuminating about personal level phenomena. The relation appears to lie within the family of options in the metaphysics of mind that make neural data a good basis for inferences about personal level phenomena.

There are three broad options in this family. The first is identity theory: that each psychological property is identical to some non-psychologically-specifiable, non-functionalist, property of a person's body and brain. The relevant non-psychological properties are likely to be complex, and the identities are likely to concern fine-grained properties – the state of feeling a sharp stabbing pain in the top right hand corner of the left big toe, rather than the more general property of feeling pain. The second option is functionalism: that psychological properties are constituted by causal relations to physically-specifiable inputs, outputs and other functionally-specifiable states. However, as we've seen above, even if such functional properties are multiply realisable, there is enough commonality of realisation that we are able to make reasonable inferences about a person's psychological state from observations of physical

properties of the brain, at least for some states. The third option is some form of mere token physicalism, albeit one where there are quite robust *ceteris paribus* generalisations linking the psychological and the physical.

According to some philosophical views, for example a property dualism that takes the neural properties to be merely nomologically connected to mental properties, the deliverances of cognitive neuroscience are at best merely evidence of what mental state a person is in. However, according to all three of the physicalist views just mentioned, cognitive neuroscience is delivering more than mere evidence. The configuration of the person's brain is part of what makes it the case that they are in that psychological state: because the brain state is identical to, realises, or provides the supervenience base for being in the psychological state.

It is tempting to view the neural evidence quite a different way: that it concerns a brain state that causes me, the person, to be in a given personal level psychological state. In non-specialist discussions of neuroscientific results, it is often said that the neural state is a cause of the mental state – that neuroscientists are discovering ways in which brains make people do things. However, if that is assumed to be true in all cases, it betrays an intuitive dualism about the mental. Although it is unsurprising that commonsense views subscribe to a kind of dualism about the mind, the philosophical arguments in favour of the physicalist positions listed above should lead us, on reflection, to reject this aspect of commonsense. Often the brain state is not just a regular cause, and therefore evidence of the person's psychological state. It is part of the constitutive basis of their psychological state. In this way neuroscience can deliver evidence of what makes it the case that a person instantiates various personal level psychological properties.

Although not intended as a rigorous philosophical argument, the last two sections do indicate how philosophical concerns about the idea that neural properties have a role to play in explaining personal level phenomena may be overcome. The next section turns to a worked-out example.

5 Representational Models and Brain Mechanisms

One of the best cases of convergence between a psychological information processing theory and an account of what is going on in the brain is offered by work on the neural basis of reward-guided decision making. Subjects are asked to make a series of choices, for example choosing between pairs of stimuli. Neither stimulus delivers a sure-fire reward, but there is a certain probability that each will be rewarded when chosen (e.g. 70% for A, 30% for B). Those probabilities may change during the experiment. Subjects are typically asked to make a long series of rapid choices and typically have little awareness of why they choose as they do. Nevertheless, they often perform much better than chance, sometimes nearly optimally.

Mathematical models have been developed showing what the optimal choices are, given a certain string of choices and feedback. For example, if option A has mostly been rewarded in recent trials, it makes sense to distribute a high proportion of your choices to option A; but not exclusively so, otherwise you will miss out on finding out when the reward contingencies switch and B becomes the higher-probability option.

Experiments measure the neural activity occurring in people and other animals when they perform these tasks. There turns out to be a surprising convergence between the way

some of the mathematical models suggest that optimal actions ought to be calculated and the quantities that seem to be processed in parts of the brain in subjects carrying out such tasks.

There are many ways that a subject could decide what to do next, given a history of reward. For example, she could try to work out the causal structure of the system with which she is interacting. Or she could even try to second-guess the intentions of the experimenter. A simpler approach is to use an algorithm that keeps a running estimate of how much reward is delivered, on average, by each option. Reinforcement learning models take that approach. They use the feedback about the magnitude of reward received from taking an option in order to update a representation of the expected value of each option available in the situation.

A subclass of reinforcement learning models that have proven to be particularly successful are those that employ temporal difference learning (Sutton and Barto 1998). The temporal difference approach overcomes the problem that the reward received from an action may not be immediate. Some actions may be useful because they take the agent one step further towards being able to take an action that receives a payoff. The system keeps track of expected long-run rewards, but generates predictions about the reward expected at a time by taking the difference between long run rewards across that time step: $V_t - V_{t+1}$. In actor-critic models using temporal difference learning, an 'actor' selects an action as a function of the relative long-run values of the available options. The critic generates a prediction about the amount of reward that should be delivered at the current time step, given that choice, and compares the prediction with the feedback actually received. It subtracts the expected reward from the reward actually received to generate a prediction error δ_t.

Prediction error at t	=	reward received at t	– reward expected at t
δ_t	=	r_t	$- (V_t - V_{t+1})$

The prediction error is used to update reward expectations to be closer to that actually experienced. The rate at which expected values are revised up or down in the direction of the most recent feedback is fixed by the learning rate α. The new expected value is the old expected value plus a fraction α of the prediction error δ_t.

The key finding in this literature is that something very like the prediction error signal posited by temporal difference models is found in the brain of subjects performing the task. So in single unit recording in macaques, dopamine neurons in ventral tegmental area (VTA) and substantia nigra pars compacta have been shown to have the response profile expected of a prediction error signal (Schultz, Dayan, and Montague 1997; Schultz 1998; Bayer and Glimcher 2005). When an action that has not previously been rewarded leads to a reward, the neurons fire. As that action continues to be rewarded, their firing decreases, in line with the increasing expectation of a reward for that action that would be observed if the subject was updating its expectations based on reinforcement learning from the history of reward. Instead, it is now the predictive stimulus that elicits a prediction error signal. The predictive stimulus indicates a change of state, since it signals that a reward will be coming, so changes the agent's expectations about the long-run value of the current state. When the reward predicted by the stimulus is subsequently delivered on cue, it elicits no response, because it is fully predicted. In a further manipulation a predictive cue that has been learned is presented and the expected

reward is omitted. In that case reduced firing is observed at the time of expected reward, consistent with the negative prediction error postulated by the temporal difference learning model.

Similar results are obtained in humans by looking at fMRI data about activity in the brain. fMRI measures the flow of oxygenated blood in parts of the brain (the "BOLD" signal), which reflects neural firing rates. Subjects are asked to perform the kind of probabilistic reward-based task described above. Parameters in the temporal difference model, principally the learning rate α, can then be estimated by fitting the model to subjects' choice behaviour. The goodness of fit of the temporal difference model can also be assessed from examining the subject's behaviour, and compared to rival models. Once parameters have been set, the model itself makes predictions about the prediction error that the agent should be generating on each trial, if it is calculating the quantities posited by the model. Trial-by-trial estimates of the prediction errors that the subject should calculate are then compared to the trial-by-trial variation in the BOLD signal in order to identify neural areas in which the BOLD signal covaries with prediction error (if there are any). This method consistently finds a reward prediction error signal in the VTA (D'Ardenne et al. 2008), and in areas in the ventral striatum to which dopamine neurons in the VTA project (McClure, Berns, and Montague 2003; O'Doherty et al. 2003; Haruno and Kawato 2006).

This impressive body of work is sometimes interpreted as showing directly that the brain implements the temporal difference learning algorithm in an actor-critic architecture, but that is too quick. The ubiquitous problem with imaging methods that the brain activity being recorded may just be a side effect of, rather than the constitutive basis of, the information processing which gives rise to the behaviour in question has been partly addressed by obtaining converging evidence from a variety of sources (neurophysiology, fMRI, EEG, TMS, etc.). A more important problem concerns the validity of model-based analysis of fMRI data (Corrado et al. 2009).

It is likely that a whole family of algorithmic models would show a reasonable match to the empirical data (Mars et al. 2010). It is hard to differentiate the particular temporal difference learning model that is used to account for trial-by-trial variations in neural activity from other reinforcement learning models in which prediction error signals play a role. So we should reach a more tentative conclusion: that representations of expected values and reward prediction errors posited by temporal difference learning accounts, *or some closely-related quantities*, are being processed in the brain and probably have a causal role in generating choice behaviour. To put this another way, we can conclude that the temporal difference learning model captures something important about the target phenomenon – the neural information processing underpinning reward-guided decision making – but that the exact relation between model (temporal difference algorithm) and target (neural information processing) remains unclear at this stage. The philosopher of science Peter Godfrey-Smith has argued that it is a distinct advantage of model-based science in general that the relation between model and target can remain loose and unspecified as the enquiry proceeds (Godfrey-Smith 2006).

Even with this caveat, the cognitive neuroscience of reward-guided decision making offers a worked-out example of our understanding of the mind-brain relation where relations between the psychological and neural levels turn out to be more tractable than some philosophical arguments suggest they might be. If psychological properties are functional kinds,

it turns out in this domain that there is less multiple realisation, both within and between subjects, than there might have been. Non-reductive physicalism remains a viable option as well, because there is no suggestion here of there being exceptionless laws linking the psychological with the neural. But the kind of radical anomalousness of the mental that some have expected is not observed in this field of activity.

What does this work say about whether there is an unbridgeable divide between personal and subpersonal levels? A deflationary answer is to argue that all of the data is subpersonal. Although conscious awareness has not been studied extensively in these paradigms, it does not seem that subjects are consciously aware of their reward expectations, of the reasons they choose as they do, or of the reasons that they revise their choice behaviour after feedback (Pessiglione et al. 2008). But it is unsatisfactory to assimilate the whole pattern of behaviour to the subpersonal level. The subjects are fully conscious normal adults, behaving as they do in the experiment because they have understood and are following instructions. In almost all experiments the stimuli are also consciously perceived. Subjects are motivated by the cash rewards available in the experiment and their behaviour shows sensitivity to the structure of those rewards. It is hard to deny that they are acting voluntarily when they select one stimulus over another or push one button rather than another. Even if the psychological processes leading up to the behaviour are forever unavailable to consciousness, therefore subpersonal, the thing to be explained – the subject's behaviour – is a voluntary action at the personal level. So the temporal difference model together with its brain basis provides a putative subpersonal level information-processing explanation of a personal level phenomenon.

Although it is far from clear that this scheme of explanation will extend to cover all thought and action, it does offer a template for the relation between the personal and subpersonal levels that shows that they are not always separate and incommensurate schemes of description. When a subject makes a choice, neural data will allow us to differentiate between several alternative explanations of why they decide as they do. Perhaps a large prediction error was generated on the last trial and they have changed their valuations as a result, leading them to choose differently. Or it could be that there have been no prediction errors over a series of trials and the subject is continuing to choose as she was before. A third possibility is that the subject has changed choice, not because of a prediction error, but because this is one of the occasions when their stochastic decision rule has selected a low-value option. These are competing explanations of why the subject acts as she does on each occasion.

Neural evidence allows us to differentiate between these options, and hence explain something about why the subject acted as she did. The subpersonal processes behind these action choices are not like an external cause that compels the agent to behave one way rather than another. On the contrary, they are the very mechanisms that allow the agent's choices to reflect her overall plans (to comply with the experimental instructions) and desires (to take home as much money as possible from the experiment). They allow the agent's behaviour to be responsive to incentives and feedback. So there is good reason for thinking that these mechanisms are part of what makes it the case that the agent's behaviour is voluntary.

Two features of this framework are important for what follows. First is the existence of robust explanatory connections between personal level phenomena (voluntary actions) and subpersonal mechanisms in which they are implemented. Second is the observation that the subpersonal properties are not acting as causes external to the agent, compelling her actions or

moving her around like a mechanism. The connection is much more intimate than that (although consistent with several different approaches to the metaphysics of mind, as discussed above). If the subpersonal mechanisms captured by reinforcement learning models of reward-guided decision making are part of what makes it the case that agents take voluntary decisions as they do, then malfunctions of those mechanisms can help explain how the personal level mental life of patients goes awry in some of the cases studied by psychiatry. The model here is not of powerless agents being compelled by external forces outside of their control. Instead, the picture is of personal level processes operating in an anomalous way because of systematic and explicable malfunctions in the way they unfold.

6 Applications

6.1 Variability in learning from experience

In this section we will see three examples of the way that the kind of work discussed above can explain inter-individual variations and pathologies in voluntary action.

The first example deals with how swiftly we update our expectations in the light of unexpected feedback. When a piece of feedback fails to match expectations, a reward prediction error is generated. How much should expectations then be adjusted? The reward prediction error only reflects how far the expectation was awry. In a very static environment it would make sense to discount this information and rely instead on a long history of experience suggesting that the choice is generally rewarding and that this particular skipped reward is just a fluke. In a very changeable environment, on the other hand, an unexpectedly absent reward is much more likely to carry useful information, indicating that the probabilities of different actions being rewarded have changed, in which case it makes sense to update your expectations to better reflect the new reward contingencies.

This difference has been probed in an experiment that manipulated the variability of the environment experienced by subjects in a pared down reward-based scenario (Behrens et al. 2007). During periods of stability the contingencies between stimuli and outcomes varied little and any failure to match expectations was more likely to be due to chance than to a change in the reward contingencies. During other periods in the experiment the reward contingencies changed rapidly, so reward prediction errors were much more likely to reflect the fact that the reward contingencies really had changed. Fitting a model of a Bayesian optimal learner to the behavioural data suggested that subjects altered their setting for the learning rate parameter (corresponding to α in the model discussed above) depending upon whether they were in a volatile or stable phase of the task. Neural activity in the anterior cingulate cortex (ACC) in subjects performing the task correlated with the estimates of volatility derived from the optimality model, suggesting that subjects were keeping track of the volatility of the environment and using it to weight the value of the information carried by a prediction error signal, thereby altering the rate at which they learnt from prediction errors.

Most interestingly for our purposes, the study found that variations in the size of this ACC signal across subjects predicted variations in their learning rates. Some subjects altered their estimates of expected value more than others when feedback failed to match expectations. If it is true that a learning rate is represented in the brain in a way that is reflected in ACC

activity, and is sensitive to the volatility of the environment, then the variations in that represented value reflected in the BOLD signal generated by the ACC can explain why some subjects are more sensitive to unexpected feedback than others.

One criticism of reliance on brain imaging to explain behaviour is that the brain data adds nothing to an observed difference in behaviour. If subjects behave differently in two different experimental conditions, then we should expect there to be some differences in brain activation that reflect the contrast between those two types of behaviour. Claiming that such differential brain activity explains the behavioural difference is indeed a bit thin because, taken alone, it doesn't tell us anything more about computations or psychological processes. We knew there would be some neural differences between two conditions and we find that there are. However, the explanatory framework above goes much further. It explains changes in behaviour (choosing one action rather than another) in terms of a series of computations carried out in the brain, including calculations of the reward prediction error. And it explains inter-individual differences in behaviour in terms of differences in another quantity, the learning rate, which is involved in the computations that constitute the making of a choice. So the neuroimaging data is not just telling us that there is some neural difference between conditions (which happens to be reflected in activity in the ACC). It gives us a window into the calculations being performed within the subject in the process of choosing an action. The evidence then allows us to explain subjects' behaviour in terms of the mental representations which are being processed; and differences in behaviour in terms of differences in the values being represented.

These representations are 'hidden variables', which help choose between competing hypotheses consistent with the same range of behavioural data. They do not simply reflect a direct contrast between subjects or conditions. Instead, the brain imaging data is allowing us to choose between different hypotheses about the multi-stage internal computational processes that lead to behaviour. It is precisely because such processes do not correlate with simple experimental parameters (stimulus type, behaviour type, feedback type, etc.) that model-based brain imaging is so useful as an independent source of data about such internal processing.

In a separate series of experiments, Krugel et al. discovered that some inter-individual differences in how quickly subjects adapt to changed reward contingencies can be explained by genetic differences in a gene involved in dopamine metabolism (recall that dopamine mediates the transmission of the prediction error signals discussed above) (Krugel et al. 2009). That is a second kind of case where features of the brain mechanisms that instantiate various calculations leading to behaviour can be identified in a way that is relatively independent of the patterns of behaviour. That in turn means that such data can form part of an explanation of why subjects are behaving as they are.

So here we have a personal level phenomenon – voluntary actions that are sensitive to rewards – an aspect of which is susceptible to subpersonal explanation. When people differ in how much they react to unexpected outcomes, that can be explained both by variations in how heavily they weigh feedback information computationally (reflected in the neural signal from ACC) and in genetic variations in the underlying mechanisms. Neither of these is an external cause of the agent's behaviour. They are more likely to be constituents of the mechanisms that

instantiate or realize their behaviour.[2] So they give us an explanatory window onto the question of why individual subjects are the kind of (personal level) decision makers that they are.

The observation that genetic differences and differences in neural activation may predict and explain inter-individual differences in personal level phenomena like voluntary action may well be unsurprising. It is, however, more controversial when this framework is applied to pathologies in personal level phenomena. Two examples are discussed below.

6.2 Addiction

The well-confirmed framework above for explaining reward-guided decision making can also play a role in explaining pathological behaviour. With the framework in place it should be immediately clear that errors in the prediction error signal could lead to radically false representations of expected value and thereby motivate behaviour that is decoupled from actual rewards. The neurons originating in the VTA / substantia nigra signal prediction errors by releasing dopamine at their terminals in the striatum and medial prefrontal cortex. This dopamine release in turns drives the short term plasticity by which representations of expected value are modified. A straightforward consequence is that any direct interference with the neuropharmacology of dopamine will stop this mechanism operating as it should.

That prediction is widely confirmed by data from rats, other primates and humans. To take just one case, there is good evidence that at least some forms of drug addiction occur because of the abnormal operation of the dopamine system. Cocaine acts directly in the brain to block dopamine reuptake in the postsynaptic terminal, thereby increasing the concentration of extracellular dopamine (Koob and Bloom 1988), and other addictive drugs have similar effects on dopamine concentration in synapses particularly in the ventral striatum (Johnson and North 1992). Recall that when the prediction error system operates normally, a large prediction error is generated when an unexpected reward is delivered. But as the same stimulus leads repeatedly to reward, the rewards become predicted and the prediction error signal declines. It seems that the direct pharmacological action of cocaine has the effect of perpetuating the production of a misplaced prediction error signal. That effectively tells the circuit calculating expected value that the experienced value of the action just performed was higher than expected so that the expected value for the next occasion should be revised upward. The result is that the represented expected value continues to increase even though the feedback is not experienced as hedonically pleasurable and is not increasingly intrinsically rewarding (Wyvell and Berridge 2000; Berridge and Robinson 1998). There is not yet consensus on the details of this explanation, but it is widely thought that the direct action of cocaine on the dopamine-mediated system for reward-guided decision making is part of the reason that some addicts continue to be motivated to get their drug even when the feedback they experience as a result of taking the drug stops being hedonically positive (Hyman 2005).

There are two ways to understand this pathology. On one understanding, because of these effects of the drug on the brain the addict is behaving in a way that is beyond her control.

---

[2] In a similar vein, structural differences in white matter tracts between subjects can explain inter-individual variability in behaviour (Buch et al.).

It is as if she were being compelled by an external force. Her drug seeking behaviour is being driven by a system that is external to her 'self', or will or her capacity for voluntary self-control. The alternative picture sees the drug as acting on and altering the mechanisms that constitute personal level voluntary behaviour. If so, it's not that the addict's personal-level psychology is losing a battle with a strong external cause; it's rather that the constitution of the addict's personal-level psychology has been changed, via a non-standard route.

The fact that addicts' drug-taking behaviour is sensitive to incentives is sometimes taken to be evidence addicts are ordinary decision makers who prioritise the pleasure of taking their drug over other priorities (Foddy and Savulescu 2010). But the model above shows how a mechanism that implements decision-making can go wrong in ways that leave agents still showing sensitivity to incentives, but having represented pathologically high expected values for some outcomes. The fact that addicts' drug taking behaviour is still sensitive to rewards is some evidence against the first picture above, which treats the addict's behaviour as coming from outside of and overpowering the personal level phenomena of voluntary action guidance. So it gives us reason to prefer the second picture according to which the drug has interfered directly with the mechanisms that generate representations of value that are partly constitutive of the personal level phenomenon of acting voluntarily.

Which picture is correct has implications for clinical engagement with people suffering from a drug addiction. The balance of evidence to date seems to favour the second picture according to which at least some drug addictions are due to a pathology of the personal level phenomenon of voluntary behaviour. That would have the advantage from the clinical perspective of encouraging clinicians to engage with addicts as rational agents, rather than as mechanisms being pushed around by forces beyond their control (Pearce and Pickard 2010). The person with a drug addiction really is strongly motivated to take their drug, and that motivation feeds into the mechanisms of voluntary action control in many of the ordinary ways, allowing for complicated reasoning and forward planning. But they are suffering because the usual mechanisms that keep expected values aligned with feedback have been interfered with. Their motivational system is being driven by artificially-inflated expected values, induced by the persistence of a prediction error signal long after the experienced reward associated with drug taking is fully predicted, or has even abated entirely. Furthermore, this direct action of the drug on the brain also seems to block the normal routes by which our higher order desires – our conscious decisions about what to want or value – act on our first order desires or motivations (Holton 2009).

In this way the dopamine system gives us a useful subpersonal explanation of the personal level phenomenon of abnormal voluntary action. Exactly how that explanation will go, and what it means for treatment, will depend upon the details of the story, which are still being worked out. But the outlines of the account so far suggest that this is another case where a subpersonal level understanding of the mechanisms that constitute a personal level phenomenon will help us to explain and intervene on the personal level problem. Rather than the subpersonal being a disempowering external cause – 'my brain made me do it' – the subpersonal is coming in as part of what makes it the case that the patient is as she is: that the way that she chooses voluntarily works as it does. This body of work suggests that it would be wrong to treat those suffering from drug addiction as automata deprived of the capacity for rational agency. But nor are they just ordinary people with selfishly hedonistic values. The

cognitive neuroscience gives us good reason to think that the values that go into their rational decision making are produced in a non-standard, pathological way.

6.3 Delusions in Schizophrenia

It is worth giving a final brief example, because of its tight connection with prediction errors and the dopamine system. A prominent theory of the positive symptoms of schizophrenia (e.g. hearing voices, loss of sense of agency) traces them to 'dysconnection': abnormal regulation of NMDA-receptor-mediated synaptic plasticity by neuromodulators including dopamine (Stephan, Friston, and Frith 2009). The idea is that there is an underlying pathology in the way representations of the world are updated by error signals, caused by abnormal dopamine neurotransmission (Fletcher and Frith 2009).

To illustrate the approach, consider the phenomenon of hearing voices. A difference between inner dialogue and hearing the speech of others is that inner speech is predictable in a fine-grained way from the agent's occurrent beliefs and other mental states, in a way that the speech of others is not. The hypothesis is that there is a subpersonal mechanism that keeps track of this distinction in order to tell whether a voice is one's own or the voice of another. Now consider what happens if there is a problem with dopamine neuromodulation so that a false prediction error signal[3] is generated when the patient is engaging in inner speech. The speech which should generate no prediction error (matching prediction) falsely generates a large prediction error, indicating non-match, which is taken to indicate that the voice is the speech of another.

This explanation of the positive symptoms of schizophrenia is still controversial (Gallagher 2004). But the hypothesis is nevertheless interesting for our purposes because of what it illustrates about the potential relation between personal level phenomena (the experienced life of a person suffering from schizophrenia) and subpersonal mechanisms. If the very mechanisms that are involved in weighing evidence and drawing conclusions in a roughly Bayesian-rational way are impaired, then the phenomenon itself becomes an interesting mix of the personal and subpersonal. The positive symptoms of schizophrenia are part of the experienced conscious life of the patient, and thereby belong to the personal level. But the kinds of rational connections that are the other paradigmatic feature of the personal level may have broken down very systematically. (Two factor accounts of delusions are designed to do justice to that fact: (Davies et al. 2001).) So here we have a subpersonal explanation of why the personal level phenomenon should fail to exhibit all the signs of being at the personal level.

If a patient's experience of hearing voices can be explained by a pathology in the subpersonal mechanisms that constitute the normal capacity for responding to evidence in rational ways, then we again have a subpersonal factor that is more than just an external cause of the personal level phenomenon. It is part of the constitutive basis of that phenomenon. But because it is operating abnormally, the nature of the personal level phenomenon itself changes.

If all of that is right, it suggests that a very particular mindset may be called for in the clinical encounter. It would not just be a case of the patient being pushed around and

[3] Or, on some views, a false estimate of the precision or confidence to be attached to the prediction error signal, and hence how strongly it should be weighted in subsequent processing.

compelled by some external cause. Nor would it be right to treat him as a mere mechanism. Rather, to engage with the problem is to engage with the person. But the dysconnection hypothesis suggests that the person himself is different. The clinician will have to be alive to the fact that the usual ways that a very central feature of interactions between persons – the way we evaluate and weigh evidence – may be impaired. The patient's personhood or capacity for agency may be altered in respects that normally underpin fluid interpersonal interactions. Of course, that will be no kind of news to the experienced clinician. What is worth remarking, though, is that appealing to subpersonal properties to explain the personal level pathology in no way detracts from the need to treat the patient as a person, not a mechanism: a person whose psychological processes depart in a central way from the paradigm of the personal level, albeit in constrained and predictable ways.

7 Conclusion

Although there is not scope here to mount a full-scale rebuttal of the claim that there is an unbridgeable personal-subpersonal divide, the discussion above shows that there are respectable philosophical positions according to which neural data can help explain personal level phenomena. The cognitive neuroscience of reward-guided decision making offers a well worked-out example where there is sufficiently little variable realisation that neurally-based generalisations are tractable, and where there are indeed robust explanatory connections between the personal and the subpersonal. To illustrate the point we saw how subpersonal neural properties can explain particular phenomena: inter-individual variability in weighing the value of new evidence about rewards, motivation in drug addiction, and delusions in schizophrenia. None of this requires us to treat people as mechanisms that are being pushed around by their brains. The framework strongly suggests instead that the neural properties appealed to in these explanations are part of what makes it the case that people are voluntary decision-makers, guided by their goals and sensitive to feedback. If so, this philosophical approach to issues in the metaphysics of mind is in sympathy with a clinical approach that treats patients as people rather than mere mechanisms.

Acknowledgements

The author would like to thank Martin Davies and Neil Levy for discussion of the issues canvassed in the paper; and Tim Bayne and Richard Gipps for comments on an earlier draft. The support of the John Fell OUP Research Fund, the Oxford Martin School and the Wellcome Trust (grant 086041 to the Oxford Centre for Neuroethics) is gratefully acknowledged.

References

Bayer, H. M., and P. W. Glimcher. 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47 (1): 129-141.

Behrens, T. E. J., M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth. 2007. Learning the value of information in an uncertain world. *Nature Neuroscience* 10 (9): 1214-1221.

Berridge, K. C., and T. E. Robinson. 1998. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews* 28 (3): 309-369.

Bloom, P. 2004. *Descartes' baby: How the science of child development explains what makes us human*: Basic Books.

Buch, E. R., R. B. Mars, E. D. Boorman, and M. F. S. Rushworth. A network centered on ventral premotor cortex exerts both facilitatory and inhibitory control over primary motor cortex during action reprogramming. *Journal of Neuroscience* 30 (4): 1395.

Corrado, G. S., L. P. Sugrue, J. R. Brown, and W. T. Newsome. 2009. The trouble with choice: studying decision variables in the brain. In *Neuroeconomics: Decision making and the brain*, eds. P. W. Glimcher, C. F. Camerer, E. Fehr and R. A. Poldrack, 463-480. Amsterdam: Elsevier.

D'Ardenne, K., S. M. McClure, L. E. Nystrom, and J. D. Cohen. 2008. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* 319 (5867): 1264.

Davidson, Donald. 1970. Mental Events. In *Experience and Theory*, eds. L. Foster and J. W. Swanson. Amherst, Mass.

Davies, M. 2000. Interaction without Reduction: The Relationship between Personal and Sub-personal Levels of Description. *Mind & Society,* 2 (1): 87-105.

Davies, M. 2005. Cognitive science. In *The Oxford Handbook of Contemporary Philosophy*, eds. F. Jackson and M. Smith, 358-394. Oxford: OUP.

Davies, M., N. Breen, M. Coltheart, and R. Langdon. 2001. Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology* 8 (2): 133-158.

Dennett, D. C. 1969. *Content and consciousness*. London: Routledge.

Fletcher, P. C., and C. D. Frith. 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience* 10 (1): 48-58.

Foddy, B., and J. Savulescu. 2010. A liberal account of addiction. *Philosophy, Psychiatry, & Psychology* 17 (1): 1-22.

Gallagher, S. 2004. Neurocognitive models of schizophrenia: A neurophenomenological critique. *Psychopathology* 37 (1): 8-19.

Godfrey-Smith, Peter. 2006. The strategy of model-based science. *Biology and Philosophy* 21: 725–740.

Haruno, M., and M. Kawato. 2006. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of Neurophysiology* 95 (2): 948.

Haynes, J. D., K. Sakai, G. Rees, S. Gilbert, C. Frith, and R. E. Passingham. 2007. Reading hidden intentions in the human brain. *Current Biology* 17 (4): 323-328.

Holton, Richard. 2009. *Willing, wanting, waiting*. Oxford: Oxford University Press.

Hornsby, Jennifer. 1997. *Simple Mindedness: A Defence of Naïve Naturalism in the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

———. 2000. Personal and sub-personal: A defence of Dennett's early distinction. *Philosophical Explorations* 3: 6-24.

Hyman, Steven E. 2005. Addiction: A Disease of Learning and Memory. *American Journal of Psychiatry* 162: 1414-1422.

Johnson, S. W., and R. A. North. 1992. Opioids excite dopamine neurons by hyperpolarization of local interneurons. *Journal of Neuroscience* 12 (2): 483.

Kamitani, Y., and F. Tong. 2005. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8 (5): 679-685.

Koob, G. F., and F. E. Bloom. 1988. Cellular and molecular mechanisms of drug dependence. *Science* 242: 715-723.

Krugel, L. K., G. Biele, P. N. C. Mohr, S. C. Li, and H. R. Heekeren. 2009. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences* 106 (42): 17951.

Mars, R. B., N. J. Shea, N. Kolling, and M. F. S. Rushworth. 2010. Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *The Quarterly Journal of Experimental Psychology* (1): 1-16.

McClure, S. M., G. S. Berns, and P. R. Montague. 2003. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38 (2): 339-346.

McDowell, John. 1985. Functionalism and anomalous monism. In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, eds. E. LePore and B. P. McLaughlin. Oxford: Blackwell.

Mukamel, R., H. Gelbard, A. Arieli, U. Hasson, I. Fried, and R. Malach. 2005. Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* 309 (5736): 951.

O'Doherty, J. P., P. Dayan, K. Friston, H. Critchley, and R. J. Dolan. 2003. Temporal difference models and reward-related learning in the human brain. *Neuron* 38 (2): 329-337.

Pearce, S., and H. Pickard. 2010. Finding the will to recover: philosophical perspectives on agency and the sick role. *Journal of Medical Ethics*.

Pessiglione, M., P. Petrovic, J. Daunizeau, S. Palminteri, R. J. Dolan, and C. D. Frith. 2008. Subliminal instrumental conditioning demonstrated in the human brain. *Neuron* 59 (4): 561-567.

Rees, G., K. Friston, and C. Koch. 2000. A direct quantitative relationship between the functional properties of human and macaque V5. *nature neuroscience* 3 (7): 716-723.

Rey, Georges. 2001. Physicalism and psychology: towards a substantive philosophy of mind. In *Physicalism and Its Discontents*, eds. C. Gillet and B. Loewer. Cambridge: Cambridge University Press.

Schultz, W. 1998. Predictive reward signal of dopamine neurons. *Journal of neurophysiology* 80 (1): 1.

Schultz, W., P. Dayan, and P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275 (5306): 1593.

Shea, Nicholas. 2003. Does Externalism Entail the Anomalism of the Mental? *The Philosophical Quarterly* 53 (211): 201-213.

Stephan, K. E., K. J. Friston, and C. D. Frith. 2009. Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. *Schizophrenia Bulletin* 35 (3): 509.

Sutton, R. S., and A. G. Barto. 1998. *Reinforcement learning: An introduction*: The MIT press.

White, J. G., E. Southgate, J. N. Thomson, and S. Brenner. 1986. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 314 (1165): 1.

Wyvell, C. L., and K. C. Berridge. 2000. Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: enhancement of reward" wanting" without enhanced" liking" or response reinforcement. *Journal of Neuroscience* 20 (21): 8122.

Yang, T., and M. N. Shadlen. 2007. Probabilistic reasoning by neurons. *Nature* 447 (7148): 1075-1080.