Reward Prediction Error Signals are Meta-Representational

Nicholas Shea
Faculty of Philosophy, University of Oxford

1. Introduction

It is often thought that metarepresentation is a particularly sophisticated cognitive achievement. There is substantial evidence that dolphins and some primates can represent some of their own psychological states, but the existence of 'metacognition' in any other species remains highly controversial (Carruthers 2009; Hampton 2009; Smith 2009). Research on the extent to which humans metacognize their own psychological states is expanding in parallel (Fleming et al., 2012). Representing the mental states of others is believed to be even rarer or non-existent in non-human animals (Hare et al. 2001; Heyes 1998). That capacity, under the label 'theory of mind', was long thought to be a crucial and relatively late developmental transition even in humans (Leslie 1987; Perner et al. 1989; Wimmer and Perner 1983). Recent looking-time experiments suggesting infants have an ability to track others' perceptions and beliefs at a very young age raise the possibility that infants have some lower-level capacity for mentalizing (Kovács et al. 2010; Onishi and Baillargeon 2005; Surian et al. 2007), in which case they may be able to metarepresent before they have a fully-fledged concept of belief, desire, or any other psychological state (Apperly and Butterfill 2009). However, macaques do not show the same behaviour (Marticorena et al. 2011) and there has been no suggestion that the capacity for non-conceptual metarepresentation extends more widely than that.

This paper argues that non-conceptual metarepresentation does extend much more widely, but based on a different set of considerations, located in a field where the issue of metarepresentation has been entirely overlooked: the literature on reinforcement learning in reward-guided decision-making tasks. Research on humans and other animals has produced an impressive body of converging evidence that midbrain dopamine neurons produce a reward prediction error signal (RPE) that is causally involved in choice behaviour (Rushworth et al. 2009; Schultz et al. 1997; Schultz 1998). RPEs are found in humans, primates, rodents and perhaps even insects (Claridge-Chang et al. 2009). This paper argues that RPEs carry metarepresentational contents.

A metarepresentation is a representation whose content concerns the content of another representation. For the purposes of this paper, a *non-conceptual* representation is a representation without semantically-significant constituent structure. It follows that the use of a non-conceptual representation does not require the possession of any concepts. RPEs are non-conceptual representations. So there is no suggestion that deploying RPEs involves having a theory of mind or having concepts of mental states. RPEs are a more low-level

form of representation, probably non-conscious,[1] and quite different from the kind of thinking about the mental states of ourselves and others that is familiar from everyday experience. It is clear that the brain does implement many forms of low-level information processing, beyond the personal level representations that occur in the familiar stream of conscious thought. For example, low-level non-conceptual information processing has been found in some systems for perception and motor control.

Could a non-conceptual representation ever represent another representation as a representation? In one sense, no. It is natural to understand 'representation-as' so as to require deployment of a concept. A thought represents Jane as a professor only if the thought includes the concept PROFESSOR and predicates it of Jane. For a putative metarepresentation M to represent another representation R as a representation (in this sense), M must have a constituent that refers to the property of being a representation (or some other representational property of R, like being a *belief*, having a certain *content* or *truth condition, referring* to some particular, etc.). That is clearly ruled out since non-conceptual representations lack semantically-significant constituent structure.

However, there is nothing in the idea of non-conceptual content that excludes representational properties from figuring in the correctness condition or satisfaction condition of a non-conceptual representation. Just as for all non-conceptual contents, such properties can figure in the correctness condition without there being a corresponding semantically-significant constituent of the representation.[2] Nevertheless, we can still discern a sense in which a non-conceptual representation M might fail to represent another representation R as a representation: M might represent only vehicle properties of R. For example, the correctness condition of M might be: such-and-such neural network is firing more strongly than normal; or the variance of the firing rate of such-and-such neural assembly is high. If the neural assembly figuring in the content is a representation R, then M is indeed about another representation; but M does not represent R as a representation, since no representational property of R figures in the content of M. On the other hand, the correctness condition of M could concern the content of R, for example: the current visual representation of the location of the light is likely to be false. Then M would indeed represent R as a representation—in the sense in which non-conceptual content admits of the distinction. Accordingly, M's having a correctness condition or satisfaction condition that concerns the content of another representation is taken to be a sufficient condition for M to be a metarepresentation. That is a reasonably stringent test. It is not enough that M concerns another representation. A representational property must figure in M's correctness condition or satisfaction condition.

This paper argues that non-conceptual information processing is responsible for the reward-guided decision-making that is elicited in certain simple experimental paradigms; and that one element of this information processing, the RPE signal, happens to have meta-level non-

---

[1]Shea and Heyes (2010) discussed whether there is a form of metarepresentation that is plausibly sufficient for consciousness. The RPE is not a candidate for that role.

[2]I use 'correctness', 'veridicality' and 'accuracy' conditions interchangeably as shorthand for the conditions that give the content of an indicative representation. (Some use 'correctness conditions', or more commonly 'truth conditions' more narrowly, for instance reserving the term as only applicable to conceptual contents at the personal level.)

conceptual content. Since the mechanism which deploys a phasic dopaminergic signal for reinforcement learning of reward-driven behaviour is widespread, it follows that a form of non-conceptual metarepresentation is relatively common in the animal kingdom. But the argument does not suggest that metarepresentation is ubiquitous. It is only because of particular features of the way RPEs are generated and processed that an argument for metarepresentation can be sustained. The argument is that the modelling and empirical findings combine to produce good evidence for metarepresentation—the considerations are evidential, not constitutive. No sufficient condition is proposed which entails that a system contains metarepresentations. Nevertheless, these evidential considerations may also apply to other systems in which the difference between a prediction and feedback is used to update the prediction for the future (Friston 2010; Wolpert et al. 2011). However, the point does not generalise to all comparator circuits, nor to all mechanisms for combining two sources of information (Ernst and Banks 2002). Nor is the model obviously applicable to the data on the seemingly metarepresentational looking behaviour in infants mentioned above.

Section 2 below summarises the evidence that RPEs are involved in reward-guided decision making. Section 3 shows that widely-accepted models of the information processing responsible for subjects' choice behaviour in these settings presuppose that RPEs carry metarepresentational contents. That furnishes a prima facie reason to think that RPEs are metarepresentational. Although the question of metarepresentation has not been canvassed in the RPE literature, it has been much discussed in the literature on 'metacognition' in non-human animals. Section 4 examines strategies deployed in the metacognition literature to displace a metarepresentational reading. Those strategies can be used to test the claim that RPEs are metarepresentational. Section 5 argues that the prima facie case that RPEs are metarepresentational is not undermined by the arguments found in the metacognition literature. Instead, the kinds of considerations advanced there, together with a plausible framework for content attribution, add up to a positive argument that RPEs have metarepresentational contents. They have both indicative and imperative contents (they are so-called pushmi-pullyus). The indicative content is that the content of another representation—the agent's (first-order) representation of the reward that will be delivered on average for performing a given action—differs from the current feedback, and by how much. The imperative content instructs that it be revised upwards or downwards proportionately.

2. Reward-Guided Decision Making

Experiments on reward-guided decision making ask subjects to choose between two or more options in order to receive probabilistic rewards. Subjects typically choose between stimuli (e.g. one fractal pattern vs. another), or between actions directly (e.g. left vs. right button press), or between actions in the context of a stimulus. The rewards offered for each choice are not wholly predictable, and the chance of each option being rewarded can vary. For example, a monkey might have a 0.3 chance of receiving 0.7ml of juice if it selects stimulus A, but a 0.7 chance of receiving 0.2ml of juice if it selects stimulus B. The chances and payoffs might then change during the course of the experiment. Human subjects are typically rewarded in money at the end of the experiment.

How should subjects distribute their choices? Clearly the optimal strategy is always to pick the option with the highest objective expected reward (= chance × payoff). But subjects are not told what the chances are. Chances must be inferred from feedback. Furthermore, the chances and payoffs associated with each option may change over time. So when a subject receives an unexpected payoff, that may just be due to chance, it may be because they have misrepresented the expected reward delivered by that option, or it may be that the probabilistic reward schedule has changed. A policy of just choosing the option represented as having the highest expected reward would be suboptimal, because it would not allow the subject to gather data about the changing payoffs of other options.

Normative accounts have been developed that show how an agent ought to distribute their choices in the light of a given history of reward. The term 'model-based' is used for solutions which work out the causal structure of the system of interest and use it to predict what is going to happen next. A simpler approach is just to keep a register of how valuable each available option is on average, and to update that register in the light of the feedback received on each trial. Solutions in that family, which do not require any grasp of causal structure, are called 'model-free'. A model-free decision-making system estimates how much reward each available option will deliver on average and updates those estimates through reinforcement learning based on feedback. Many animals can use reinforcement learning to perform such tasks successfully. Although humans could do it in other ways, the set up we have just described, requiring a series of rapid responses for relatively short-term rewards, encourages human subjects also to solve the task using model-free reinforcement learning.

The problem of calculating the optimal way to behave in these settings has long been studied by mathematical psychologists and computational modellers (e.g. Bush and Mosteller 1951). A major breakthrough was the discovery of the class of *temporal difference* algorithms. TD learning can deal with cases where rewards only occur at the end of a series of choices, but nevertheless distribute credit for the eventual reward to choices taken earlier in the series, solving the so-called credit assignment problem (Sutton and Barto 1998). While TD algorithms are undoubtedly fundamental to the conspicuous success of reinforcement learning models of decision-making and understanding its neural basis, the philosophical questions about metarepresentation arise in just the same way in cases where there is immediate feedback for each choice. To keep the metarepresentation issue centre-stage we will work with a pared-down model with immediate feedback. But the argument for metarepresentation in no way turns on this simplification, as is made clear in the Appendix.

TD learning models, including our simplified version, give a central role to a reward prediction error that is calculated at each time-step. The RPE is the difference between the average payoff expected at a time-step, given the current stimulus / context / behavioural choice, and the reward received at that time-step. Crucially, it is this prediction error, rather than the absolute value of the feedback directly, that drives learning. The agent's expectation about the average reward that will be received in a context is revised upwards or downwards in accordance with this RPE.

RPEs started life as a feature of computational models designed for optimal decision-making, without reference to how the problem is actually solved by humans or other animals. But then Wolfram Schultz and colleagues discovered that midbrain dopamine neurons broadcast an RPE signal (Schultz et al. 1997; Schultz 1998). Relative to a background 'tonic' level of firing, there is a transitory 'phasic' increase when a unpredicted reward is delivered, and a phasic decrease in firing when a predicted reward is not delivered. This finding brought the computational modelling rapidly back into contact with real psychology, galvanised the cognitive neuroscience of decision-making and launched the science of neuroeconomics.

The strategy that has been so successful in this area is to use behavioural and neural data in tandem to pin down the information processing involved. Rival models are fitted to the behavioural data (the pattern of choices made and rewards received) and compared in terms of how well they fit the data. Models from the TD learning class often perform well. To find which brain areas are responsible for implementing the different steps of the algorithm, trial-by-trial variations in the quantities posited by the model are compared with data reflecting neural activity, most commonly the fMRI BOLD signal. The trick is to look for regions of the brain whose activity varies parametrically in line with some quantity calculated over according to the model (e.g. that shows a close quantitative match with the RPE posited by the model). Such trial-by-trial correlations point to brain areas likely to be involved in performing various steps of the algorithm.

Similarly, the patterns of neural firing obtained through single unit recording can be correlated with quantities in the model like the RPE. Dopaminergic neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta have been found to have a firing profile corresponding to the RPEs posited by TD learning models of appetitive conditioning (Bayer and Glimcher 2005; Schultz et al. 1997; Schultz 1998). Applied to fMRI data, the approach has found correlates of the RPE in the VTA (D'Ardenne et al. 2008), and in areas of the ventral striatum that receive dopaminergic inputs (Haruno and Kawato 2006; McClure et al. 2003; O'Doherty et al. 2003).

This kind of data, together with a large range of converging evidence that the measured neural activity is causally relevant to the observed behaviour, has produced a substantial consensus about the mechanism that is responsible, how it is neurally realised, and how to describe its quantitative properties mathematically. Issues remain unresolved, of course, for example over whether prediction errors concerning actions are calculated in addition to or instead of reward prediction errors (Li et al. 2011). Dissenting voices also remain that say the phasic dopamine signal does not function as any kind of prediction error (Redgrave and Gurney 2006). But the current state of the art is as strong a scientific consensus as a philosopher could possibly hope for. So we can take it that TD learning models capture something true and important about a system of low-level representations found in real brains, and that the phasic dopamine response is a prediction error signal.

3. Content in the Model

For the sake of concreteness, we will work with the popular actor-critic version of TD learning (Sutton and Barto 1998). The actor-critic model separates the algorithm into two parts, which map well onto separable components of neural activity. This section sets out a simplified version of the actor-critic model. We will see that, in common with other TD learning algorithms, the actor-critic model presupposes that the RPE has meta-level content: it represents (non-conceptually) that the current feedback differs from the system's representation of expected value, and by how much, and directs the value representation to be adjusted accordingly.

In the actor-critic model one system, the actor, follows a 'policy' that selects actions based on the average payoff associated with each available option. The policy makes a choice probabilistically, giving proportionately higher probability to options represented as proportionately more valuable. Another system, the critic, makes use of the prediction about the chosen action. Because of the element of chance in the action selection policy, the expected value of the chosen action may be high or low. The critic compares the expected value with the outcome actually received from the chosen action. The difference between these values is the RPE. The RPE signal is used by the critic to update the prediction associated with the chosen action for the future. Then we return to the first step: that updated value is used by the actor, together with predicted values for the other available options, as the basis for selecting the next action.

The RPE signal is generated by midbrain neurons in the VTA and substantia nigra pars compacta. Phasic changes in the dopamine released at their terminal projections in the ventral striatum modulate synaptic plasticity in such a way that the predicted rewards represented in the striatum are modulated, increasing or decreasing the predicted reward associated with the just-chosen option in accordance with the RPE signal. The adjacent dorsal striatum is thought to implement the actor, selecting actions on the basis of the ventral striatum's predictions (Daw et al. 2006; O'Doherty et al. 2004).

We will work with simplified version of the actor-critic model, assuming rewards are immediate. We can take it that the neuroscience has established how the various components are connected together, how the activity of each is dependent on the activity of other components, and how the system is connected to actions in the world and feedback from the world. So we will assume that the wiring diagram in figure 1 is a correct description of the neural system. That system can be described in terms of its non-contentful properties, saying how firing rates in each component depend on the others, lead to action, and depend on feedback; and saying how those connections are modulated. When the system is embedded in a problem space, receiving feedback from the external world for real actions performed on the world, then we can also describe it as performing computations over representations about the world. Parameters in the computational model describe how components in the system relate to items in the world (actions, rewards). These are further real, relational properties of components of the neural circuit (e.g. how neural firing in one internal component covaries with the volume of juice delivered). The system performs an algorithm over components with these relational properties. As an implementation of a computational

model the various components of the system have putative correctness conditions or satisfaction conditions when they are in certain states. Those are the contents presupposed by the model.

The computational model plus details of its putative neural implementation constitutes a hypothesis about how quantities described by the model are realised in distributed patterns of neural firing it the brain, and how learning is realised by synaptic plasticity. What do these putative contents add to a purely non-contentful, neural description of how the system operates? They allow us to explain the operation of the system in terms of properties that connect, explanatorily, with aspects of the environment to which the system is receptive and on which it acts. Those contents are more than merely instrumentally justified if they do indeed capture real relational properties of the system (content in these kinds of systems being a certain kind of complex relational property). In this section I will simply set out the contents presupposed by the actor-critic model. The predictive success of the model is a prima facie reason to accept the putative contents it attributes. The rest of the paper aims to test that prima facie case.

Consider just two actions, action 1 and action 2 (e.g. left and right button presses), where probabilistic rewards depend only on which action is chosen and are not conditional on any other stimulus or contextual feature. The critic keeps a model-free representation of the reward that will be received on average for performing each action: V1 and V2. The actor uses these values as input to a policy π that chooses an action to perform probabilistically, with the probability that a given action is selected increasing as its expected value increases relative to other available actions.[3] Once the action is executed the critic receives feedback in the form of a reward of some magnitude r (including zero), which it uses to calculate an RPE and then update the value for the chosen action. The value is moved in the direction of the feedback actually received by adding a proportion of the RPE. The proportion is given by the learning rate α. α accordingly determines how quickly the predicted reward is adjusted in the direction of the most recent feedback. So the representations involved are:

Expected value of action	1V1
Expected value of action	V2
Chosen action	Ai (either A1 or A2)
Reward received	r
Prediction error (having chosen Ai)	$\delta = r - Vi$
Learning rate	α
Updated expected values:	
Chosen action	$AiVi \rightarrow Vi + \alpha\delta$
Unchosen action	$AjVj \rightarrow Vj$

These putative representations are realised by patterns of neural firing in the brain areas described above. According to the computational model these vehicles have the following

[3]E.g. using a softmax rule or drift diffusion mechanism.

contents. r represents the reward actually received in the current time step. The chosen action is represented by A1 or A2, respectively. These representations are tightly connected to outputs: representation A1 reliably causes action 1 to be performed and that is part of its function. So the model treats the Ai as having directive or imperative content: *do action i*. It may also be that they are relied on in updating expected values, to tell the system which action was chosen. If so, they also have indicative content: *action i was chosen*.

V1 and V2 predict the reward that would be obtained on average if action i were repeatedly selected in the current environment. (Without loss of generality we focus on V1.) That is, the content is an expectation in the probabilistic sense of expectation (probability × magnitude): *on average the reward for selecting action 1 will be V1*.[4] The claim that the system represents an expectation can sound metarepresentational in its own right, before we even reach the RPE. But that is a play on the word 'expectation'. Here the expectation value is a represented objective quantity (objective probability of reward times its magnitude). V1 is the agent's current best estimate of that value. We can put this in terms of the veridicality conditions for V1. According to the model, V1 is accurate iff the average reward payoff that would be achieved by repeatedly choosing action 1 in the current environment is V1. The pattern of behaviour driven by V1 will work best for the system only if V1 is accurate in that sense. If the average reward that would be obtained by repeatedly choosing action 1 is actually higher than V1, then the agent's behaviour will be suboptimal in that the system will chose action 1 less than it should; conversely if V1 overestimates the average reward that would be obtained the system will choose action 1 too often.

The RPE, δ, is used to update the reward prediction for the type of action that was just chosen. It indicates that the reward received was higher or lower than the predicted value and directs that the predicted value be adjusted accordingly. So the model presupposes that the indicative content of δ is: *the reward received for the last action was δ higher / lower than currently-represented expectation for that action*. δ then has the function of updating the corresponding representation of expected value. So the model also presupposes an imperative content for δ: *increase / decrease the predicted value Vi in proportion to the magnitude of δ*. These indicative and imperative clauses are, at this stage, simply descriptions of the content that is presupposed by the model. The indicative content describes the condition under which the RPE signal would be accurate, according to the model. The imperative content describes the condition under which downstream processing would have responded appropriately to the RPE signal; that is, it states a satisfaction condition that the model presupposes δ to have. Although it is not remarked upon at all in the reinforcement learning literature, representational properties figure in both the correctness condition and the satisfaction condition attributed to δ. That makes δ metarepresentational.

The presupposition that the RPE has metarepresentational content makes intuitive sense in the context of the model. δ closely correlates with the difference between expected value and reward (because that is how it is calculated). And it is relied on in subsequent processing for

---

[4]I follow the convention of using letters both to refer to the representations involved and, where appropriate, to pick out the quantities variably represented by those representations.

that correlation, not for any further information it carries about things in the world. By way of analogy, consider two ways that I might update my beliefs about the temperature outside. Suppose I read 16°C on the slightly unreliable thermometer outside the window. If a friend then tells me it is 20°C outside, I will revise my estimate based on both sources of evidence, perhaps to an average of 18°C. If instead of telling me about the weather, she tells me about the thermometer—that it under-reads by 2°C—then I would also revise my estimate to 18°C. In the first case I am relying on my friend for some evidence she has about the world and forming a conclusion based on both sources of evidence. In the second case I am relying on her for some evidence about the accuracy of my first estimate, and revising that estimate accordingly. The RPE δ is used in the second way—for the content it carries about the degree of inaccuracy of the previous estimate.

4. How to Deflate a Metarepresentational Reading

Although the metarepresentational status of RPEs is very little discussed, there is a large literature on 'metacognition' in other animals—the central issue being whether representations with meta-level contents are involved in performing various tasks. That makes it a useful place to look for arguments that test the prima facie case that RPEs are metarepresentational. Without language it is much harder to tell if the animal has any representational states with meta-level contents. I follow the practice of using the term 'metacognition' to cover all tasks that involve some kind of self-monitoring, irrespective of whether metarepresentation is involved.

Two distinctions should be noted at the outset. First, meta-'cognitive' could suggest a contrast between the cognitive and the perceptual, affective or motoric, so perhaps also a limitation to conceptual representations with constituent structure, built up out of concepts. In fact the literature is not restricted in that way. Many of the representations at issue, with putatively meta-level contents, are non-conceptual representations with no constituent structure. Nor need they be cognitive in any other narrow sense. They may include perceptual representations, motor programs and low-level representations in subpersonal systems. Any representations at the psychological level (personal or subpersonal)—i.e. that figure in a realist information processing explanation of behaviour—are candidates.

The second distinction is between the content and control senses of metacognition. Some call all representations or processes which supervise or monitor some other mental process 'metacognitive' (Carruthers 2009, p. 170). But it is of course a substantive question whether, in giving an information processing account of some process that depends upon monitoring other processes, we should appeal to metarepresentational contents. For example, a robot is capable of monitoring whether it is continuing to make forward progress and, if not, initiating a movement in another direction (Anderson and Perlis 2005). That kind of self-monitoring clearly does not require metarepresentation, but could be considered to fall within the 'control' sense of metacognition (Carruthers 2009: 129). Our focus is on metacognition in the content, not just the control sense: on representations whose content concerns the content of another representation.

Optimal performance in a typical metacognition experiment requires the animal to keep track of its own chance of success in some object-level task. For example, in (Hampton 2001) macaque monkeys were trained to remember a visual stimulus, and then tested on whether they could pick out that stimulus from an array of four different pictures some minutes later, with a food reward for getting it right. The monkeys were sometimes given the option, shortly before taking the memory test, of opting out for a sure-fire but lesser reward. If they opted to take the memory test they would get a peanut (favoured) for correct answers, otherwise nothing. If they opted out they were sure to get a reward, but only an un-exciting pellet of monkey chow. The optimal behaviour would be to opt out of the test on those occasions where, because its memory was poor, the monkey would be unlikely to get the right answer if took the test.

Naturally the memory test becomes more difficult as the delay between initial stimulus and test increases (up to 4 minutes). Both of Hampton's monkeys opted out more at longer delays. However, that could reflect learning a general rule that the probability of getting it right tends to be lower at longer delays. The crucial question is whether, at a given delay duration, the animal is sensitive to trial-by-trial variations in the strength of its memory of the initial stimulus. One of the two monkeys did distinguish, at a given delay, between trials in which it was likely to succeed and those in which it was likely to fail, evidenced by the fact that it was more likely to be correct in trials where it was given the choice and opted in, than in those trials (33%) where it was forced to take the test. Of course, there could be a general accuracy cost to being forced to take the test (e.g. because these trials are rarer, or because the forced-choice situation is more stressful). But that cost should be the same at each delay duration, whereas Hampton's first monkey derived an increasing benefit from opting out as the delay increased. Its chance of succeeding in the opt-in condition remained relatively constant at increased delays while its chance of succeeding in the forced choice condition grew progressively worse. This was an important result because it provided good evidence that a macaque monkey could make an opt-out decision based on a purely internal indicator of its own chance of success: the strength of its memory of the stimulus, or some correlate thereof.

Understandably, comparative psychology has focused on establishing whether these kinds of tasks are being solved in reliance on an internal or merely an external cue. But interestingly, that issue does not determine the metarepresentation question, in either direction. An internal cue may be relied on in processing simply for the information it carries about the world (e.g. internal cues about where the edges are in a visual scene are relied upon for information about where the objects are). And an external cue may be relied upon for information that it carries about the accuracy of the animal's own representational states. For example, the animal's own behaviour in vacillating between two options is an externally observable cue. But if the animal reacts to this information, not simply by acting to resolve the response conflict, but by doing something epistemic like gathering more information before deciding, then it may be relying on the externally-observable cue for information it carries about the animal's own representational states.

Conflating metarepresentation with reliance on an inner resource will make it seem as if any hierarchical information processing system involves metarepresentations. Consider a

feedforward set up in which layer 1 of a system represents the spatial distribution of visual contrast and layer 2 uses that information to calculate the location of edges. Does the fact that layer 2 is wholly reliant on an internal resource, namely the output of layer 1, thereby make activation in layer 2 metarepresentational? No, because layer 2 can rely on layer 1 for the information it carries about the world. Speaking metaphorically, layer 2 need not be interested in the processing of layer 1 for its own sake—for information layer 1 carries about the system itself. By analogy, I might ask a friend to describe an event in order to find out what went on, or I might know what occurred and just be interested in assessing his veracity. Only in the latter case need I metarepresent the information he conveys. In the former case I could simply take him to be an instrument that carries information about the event. Typically models that involve hierarchical layers of information-processing filters (Marr 1982) are not metarepresentational because, although each layer after the first is wholly reliant on an internal resource, those internal resources are relied on for the information they carry about external features.

In an experiment similar to Hampton (2001), Kiani & Shadlen recorded directly from neurons in the macaque brain that seemed to be causally relevant to the animal's decision to opt out of a two-alternative perceptual decision test. They interpreted these neurons as representing the monkey's confidence in the accuracy of its first-order perceptual judgement (Kiani and Shadlen 2009). But a rival first-order explanation is available, which nicely illustrates the thorniest issue in metacognition experiments. Kiani and Shadlen found neurons in area LIP for which a high firing rate predicted making a perceptual judgement, and an intermediate firing rate predicted choosing the opt-out option. The trouble with their interpretation is that these neurons could equally well be representing the expected value of the target option (reward magnitude × probability) relative to the other options. Rather than representing its own confidence in a particular perceptual judgement, the system could simply be keeping track of the expected value of all three options, two of which deliver high rewards probabilistically and the third of which delivers a low reward with certainty. The monkey can learn by reinforcement that when the noise in its representation of the perceptual stimulus is high, it is less likely to be rewarded for its subsequent perceptual judgement. In those circumstances the opt-out option will be have the highest expected value, so the monkey will choose that one. In fact, for much of the paper Kiani and Shadlen treat this neural population as representing the 'accumulation of evidence in favour of one or the other option' (p. 761) or the 'expected chance of success' (p. 762). It is only in the discussion section that they describe this as a representation of certainty, but without offering a good reason to distinguish certainty from expected value (p. 763).

That line of reasoning exemplifies the basic challenge to meta-level explanations of self-monitoring experiments. Peter Carruthers makes similar claims about all the different experiments which purport to show metarepresentation in other animals. Whether the argument successfully undercuts a metarepresentational conclusion varies from experiment to experiment, but our focus is its basic structure. The tactic is to argue that keeping track of the probability of various rewards, together with subtle calculations over first-order expected values (reward value × probability), is adequate to explain behaviour.[5]

For example, Carruthers takes on an argument made by Davidson that surprise is necessarily metarepresentational (Carruthers 2008: 61-63; Davidson 2001: 104-05). Carruthers says that a feeling of surprise can be generated simply by a clash between first order representations, without thereby being second order. You may believe not-p and then see that p, the contradiction occasioning surprise, without thinking about your own mental states or metarepresenting their contents in any way. As a matter of exegesis, Carruthers' argument does not succeed as a refutation of Davidson, since Davidson was drawing a distinction between being surprised and being startled, indicating that he meant surprise to cover only some of the cases: 'Surprise involves a further step. … Surprise requires that I be aware of a contrast between what I did believe and what I come to believe' (p. 104). It is surprise in that sense that Davidson claims entails having beliefs about beliefs—which is plausible, given the way he sets up his terms. But, Davidson interpretation aside, Carruthers is right that the basic phenomenon of reacting to a contradiction between representational states need not involve metarepresentation. Carruthers relies on these kinds of considerations to argue against Joëlle Proust's claim that some 'epistemic feelings' are metacognitive.

Proust and Carruthers on Metacognitive feelings

Proust has developed a sophisticated account of 'metacognitive feelings' and associated phenomena (Proust 2003, 2007, 2008, 2009a, 2009b, 2010). These states are more low-level in several respects than thoughts or cognitions. Centrally for our purposes they are non-conceptual representations. Proust points to bodily changes and reactions occasioned by uncertainty, response conflict or surprise. An animal might use these reactions as cues to act so as to improve its information, or to opt out of the current situation so as to make new choices available. We humans make similar uses of these 'epistemic feelings'. Proust argues that they are non-conceptual representations but are not meta-representational, having instead a special kind of functional role intermediate between object-level and meta-level representations, involving the control of object-level representational states.

Carruthers claims that animals making use of such feelings as cues might thereby be capable of passing any of the tests of metacognition yet devised in comparative psychology, but would be 'wholly incapable of metacognition in the metarepresentational sense' (2009, p. 171). Carruthers accepts that such states carry information about other internal states, in the purely correlational sense (Shannon 1949), but rightly objects that mere correlations are insufficient for representation. For example, the feeling of fear may be a reliable consequence of the thought that something is dangerous, so carries the information that such a thought has occurred (Carruthers 2008, p. 62). But if correlational connections between representations were sufficient, metarepresentation would be absolutely ubiquitous. General reasons for rejecting purely information-based accounts of content (Fodor 1987; Millikan 1990) apply with equal force here.

Several authors argue that, amongst all the profusion of correlational information carried by any representation, the way to home in on content is to look at the way that the representation is acted on or 'consumed' in downstream processing (Godfrey-Smith 2006;

---

[5]Another strand in Carruthers' work is to posit additional mechanisms for dealing with response conflict—for making a choice when there is a near tie in first order value.

Millikan 1984; Papineau 1987; Shea 2007). Carruthers offers that kind of criterion for deciding whether a putative representation, which carries correlational information about another representation, is in fact metarepresentational:

> 'purely informational accounts of intentional content face notorious difficulties … . And then the question for us becomes: Does the animal *make use of* the epistemic feelings in question in such a way that the feeling is thereby constituted as a nonconceptual representation of a cognitive state?' (2009, p. 171; emphasis in original)

> 'It seems that a nonconceptual representation of some property of the world represents what it does partly in virtue of its role in guiding thought and action that is focused on that aspect of the world.' (2009, p. 171)

We can apply Carruthers' test to the LIP signal recorded by Kiani and Shadlen. It carries correlational information about the world (how likely it is that the monkey will be rewarded if it chooses an option) and about the monkey's representations (how accurate its representation of the stimulus is). How does the system make use of this resource? If the signal ramps up only slowly, the animal chooses the opt-out option, thereby increasing the average reward delivered. This way of using the LIP signal seems to be focused on the world—the average reward payoff (probability × magnitude) associated with various options —and not the accuracy of the monkey's object-level representations as such.

Carruthers suggests that this type of deflationary approach applies across the board, with the consequence that epistemic feelings never have meta-level contents. In response Proust makes a strong case that some epistemic feelings play a genuinely metacognitive role that cannot be fully captured in terms of ordinary first-order contents (they 'concern', but do not metarepresent, the content of other thoughts, according to Proust). Our focus is not on resolving that debate, but on whether the deflationary tactic identified by Carruthers applies to the RPEs involved in reward-guided decision making.

5. A Deflationary Treatment of RPEs?

5.1 Dispensing with prediction errors

This section sets out the evidence that RPEs have genuinely metarepresentational content. 5.1 gives the evidence against an account of reward-guided behaviour that dispenses with prediction errors entirely. 5.2 applies the Carruthers test to the RPE signal and argues that the way it is made use of is 'focused on' the information it carries about another of the system's representational states. 5.3 shows that alternative first-order contents are not compatible with the way correctness / incorrectness of the RPE signal is used to explain the system's behaviour. And 5.4 briefly assess how far these kinds of considerations apply to other kinds of system.

Here is how the task could be solved without calculating a prediction error. Recall that reinforcement learning models posit a system that keeps a tally of the expected reward for each action (the chance of success multiplied by the value of the reward), adjusts that tally on the basis of feedback, and uses it as input to a stochastic decision rule in which the action

with the highest relative value is more likely to be selected. The updating algorithm is as follows (having chosen action 1):

$$V1 \rightarrow V1 + \alpha\delta = V1 + \alpha\,(r - V1)\,[\delta = r - V1]$$

Rearranging:

$$V1 \rightarrow (1 - \alpha)\,V1 + \alpha r\;(1)$$

So the expected value is reset to a weighted average of the current feedback and the old cumulative expected value. The relative weight to be attached to the latest feedback is determined by the learning rate parameter α—the lower α, the more the next choice will be affected by the rewards received deep in the reward history of that action.[6] That algorithm could be implemented without relying on a prediction error signal. The new value of the expected reward is calculated in an operation that takes as input the expected value, reward feedback and learning rate, and outputs a new expected value as a result (a linear operator algorithm: Bush and Mosteller 1951). The RPE quantity, which was our candidate for meta-level content, has dropped out of the calculation entirely.[7]

That way of solving the task would be behaviourally equivalent to the actor-critic algorithm in the tasks I have considered so far. Therefore this task can be solved first order, using an algorithm that does not depend on an RPE. However, suspending the simplification introduced in section 2, in some tasks agents have to make a number of choices in series before they achieve an eventual reward. We saw above that temporal difference learning algorithms, which do rely on an RPE, provide an optimal way of solving this credit-assignment problem. A linear operator algorithm that simply re-weights expected values in proportion to feedback is not suited to such tasks (see Appendix).

Furthermore, we are not restricted to behavioural evidence about which algorithm subjects are using. Neural evidence can identify putative representations in terms of their functional profile: which internal and external parameters they correlate with most closely. The question of which model best fits the pattern of neural evidence can be assessed without presupposing that quantities calculated over in the model carry any particular content. And the evidence is very strong that a signal carried by dopamine neurons projecting from the VTA and substantia nigra pars compacta to the ventra striatum is causally involved in choice behaviour. Its functional profile corresponds closely to the RPE quantity in the model. Interestingly, this is a case where, in addition to neurophysiological recording in animals, the strategy of model-based analysis of fMRI data, recently on the rise in cognitive neuroscience, is able to take us beyond the kinds of inferences about internal variables that could be made using the standard methods of cognitive psychology (Corrado et al. 2009;

---

[6]In fact in many models there is an additional 'decay rate' parameter γ applied to V which introduces a second degree of freedom, the effect of which is to allow the weightings that are applied to V1 and r in the sum above to vary independently.
[7]Eliasmith & Anderson make an analogous point about Kalman filters: although standard versions involve a prediction error term, a mathematically equivalent formulation is available that produces the same input-output behaviour and dispenses with the prediction error term, instead updating via a weighted average of two sources of evidence (Eliasmith and Anderson 2003: 288-93).

Mars et al. 2010). In short, many different lines of evidence now lend strong weight to the conclusion that an RPE signal is causally involved in generating the patterns of behaviour observed in these experiments.

Proust points out that externally-available information about the pattern of choices is sufficient to calculate choice behaviour. An objective observer can keep track of the agent's success rate just as the agent himself does, on the basis of actions performed and feedback delivered (Proust 2007, p. 283). That is right, but we saw above that information being externally available does not determine the question of what is being represented using that information. The evidence from neuroscience and computational modelling amounts to a compelling case that the decision-maker uses that externally-available evidence in a series of internally-implemented calculations, one stage of which involves an RPE signal (that is, its non-contentful correlational profile corresponds closely to the RPE quantity in the model). The fact that the standard interpretation of the algorithm presupposes that the RPE signal carries meta-level content completes the prima facie case for metarepresentation.

5.2 What is use of the RPE focused on?

So there is strong evidence that model-free reward-guided decision making is achieved in many species using an algorithm which, like our simplified version, relies on an RPE. Here I assess the presupposition that the RPE signal meta-represents, adopting Carruthers' suggestion that we look at how it is made use of in downstream processing, in particular asking whether those uses are focused on an aspect of the world.

When a state correlates with one property it usually correlates with many. For example, the needle on a car fuel gauge covaries with the volume fuel in the tank; but also with the mass of fuel in the tank, the height of fuel in the tank, the pressure exerted on the electromagnetic sensor in the tank, the current flowing in the wire leading to the gauge, and so on. The RPE signal too carries correlational information about many things. Precisely what it correlates with will depend upon the details of the experimental set up. For example, in a task in which the environment is very unstable, with frequent changes in the payoff probabilities attached to each option, rewards will be unpredictable, keeping predicted rewards Vi low. So an RPE signal will tend to be produced by every delivery of reward. That is, the RPE will correlate with the reward just delivered, and to some extent with the expected reward (magnitude $\times$ probability) for the just-chosen option.

In a stable environment in which some options are consistently rewarded with high probability, most feedback will be fully predicted, so most RPE signals will be small. The few large negative RPE signals will be caused by occasional omitted rewards. Large positive RPE signals will be even less frequent, requiring that on one of the rare occasions when the system selects the low value option, it is also by chance unexpectedly rewarded. So in these circumstances the RPE is inversely correlated with the system's stable reward expectation for the chosen option.

In the midst of these and all the other types of correlational information carried by the RPE signal, which information is being made use of in downstream processing? How do we answer Carruthers' question: is the way the RPE signal is used to guide thought and action

directed on some aspect of the world, or is the use directed on the creature's own representational states?

Let's pin down the question a bit. Godfrey-Smith (2006) argues that the use of a representation by some other system (the 'consumer') is an essential part of the 'basic representationalist model'. The consumer takes the representation to stand in for some state of the world and reacts to the representation as it would if it were able to react to that state of the world directly. In this way, the consumer system is making use of some relation between the representation and the world. Carrying information (correlating) is just one of several relations between representation and world that could be exploited by the consumer in this way.

Fred Dretske says more about consumers exploiting correlations. In chapter 4 of *Explaining Behaviour* (Dretske 1988) he advances a theory of content that differs from the purely informational treatment in *Knowledge and the Flow of Information* (Dretske 1981). Content is based not just on information, but on information that is used to guide learning for action. Dretske's theory is an account of what makes it the case that certain representations have the content they do. Here I appeal to it only as offering plausible evidential considerations, rather than as a true constitutive account. Adapting Dretske's framework slightly for our purposes, consider a system that learns by instrumental conditioning. Suppose an animal can detect an external stimulus by means of tokening an internal state B. Suppose, too, that the animal has an action-selection system that initially responds at random to B, producing internal state R that drives a reach to the right. If that action is rewarded, the connection between B and R is strengthened—that is, the chance increases that R will be produced when B is tokened (see Fig. 2).

Internal state B carries correlational information about many things: the identity of the stimulus, properties of the stimulus like colour and shape, and all the other worldly properties with which the stimulus correlates—a whole series of properties F, F', … .[8] These are not restricted to properties of the stimulus. For example, and crucially for our purposes, internal state B correlates with there being a peanut on the right (because the external stimulus driving B is so-correlated).

Now our question: is the way B is used to drive thought and action directed on F, or F', or …? In producing response R, which aspect of the correlational information carried by B is being made use of? Dretske's tactic is to focus on the condition in virtue of which response R is rewarded. Some consequence of response R was registered by the system as rewarding (the animal's ingesting a peanut), and the B→R connection was strengthened as a result. Assuming that R's being rewarded is somewhat specific to the circumstances when the stimulus is present, one of the features with which the stimulus correlates, and hence B correlates, is responsible for the response R being rewarding; namely there being a peanut

---

[8]These are not different ways of coding information, but different pieces of information carried by the same code. For example, the brain may use a rate code and/or a phase code. The different codings are different properties of the system's internal states (the rate or phase of neural firing), each of which correlates with a whole range of external properties. Suppose firing rate makes a difference to downstream processing whereas information carried by phase is discarded. Still, firing rate correlates with a range of properties F, F', … .

on the right. Call that fact G. B's correlating with G explains B's being connected with motor state R, hence with reaching to the right. G, then, is the piece of B's correlational information that is also representational—use of B can be said to be directed on G.

For example, in Fig. 2, B correlates both with *there is a dark grey stimulus* and *there is a peanut on the right*. When the animal responds to B by looking right and finds the result rewarding, it is there being a peanut on the right, rather than the colour of the stimulus, which acts as the primary reinforcer. So it is in virtue of the correlation between B and peanuts on the right (G) that the B→R connection was established. In sum, this is the evidential test I take from Dretske: which aspect of the correlational information carried by a putative representation explains the fact that it is wired up to behaviour in the way it is?

We can apply this reasoning to the RPE signal δ. It is not wired up directly to behaviour as in Dretske's model, but we can still ask which correlational information explains the fact that it is processed as it is to issue in behaviour. We saw above that δ correlates with many external-world properties, and also with an internal property: the difference between expected value and feedback, as relied on by the actor-critic model. How is the RPE signal used to guide subsequent information processing and action? It is used to adjust the expected value of the just-chosen action, the updated value of which is then used in selecting the next behaviour. Applying the Dretske test (inspired by the Carruthers quote), we ask: which correlation explains the fact that the system is set up to process and then act on δ in that way? The computational modelling results show that the overall system is an optimal way for an agent to harvest maximal rewards in a certain range of environments (Sutton and Barto 1998). So we can suppose that the system has been set up by evolution and/or learning to maximise the overall delivery of reward to the agent.

The represented expected value V1 of the just-chosen action (supposing it was action 1) is increased if action 1 was more rewarding than represented, and V1 is decreased if the feedback was less than represented; in both cases by a quantity proportional to the difference between represented expectation and reward. The way the Vi feed into the decision policy means that, when V1 is set higher, the actor will choose action 1 more often than before. According to the optimality conditions proved in the reinforcement learning literature (Sutton and Barto 1998), that change will only be beneficial on average for the overall system if the reward received really was more than the previously-represented expected reward from action 1. Otherwise the adjustment would tend to reduce long run payoffs. Correlatively, when V1 is adjusted to be lower, the actor will choose action 1 less often than before. That change will only be beneficial to the system on average if the reward received really was lower than the previously-represented expected reward from action 1. So the way that δ is acted upon—to revise the represented expected values Vi—is beneficial to the system in virtue of the tight correlation between δ, on the one hand, and the difference between represented expected value and feedback, on the other. Of course δ also correlates with the difference between the current feedback and the (worldly) objective expected reward for that option (i.e. its objective chance of being rewarded times the reward magnitude). But the fact that the current feedback differs from the chance × magnitude for option i does not make sense of changing the estimate of expected reward Vi or of the resultant change to the action policy, to select action i more frequently. The feedback on a

particular trial will never match the actual expected reward (objective chance × magnitude) even if the expected value is correctly represented, and unless the system is representing that the current estimate of expected reward is actually incorrect, there is no rationale for changing the estimate or the action policy.

We are supposing that the system has been set up the way that it is, by evolution or learning, in order to maximise overall average payoffs to the agent. It is the correlational information carried by δ about difference between represented expectation and feedback that contributes to achieving this overall outcome. So that correlation explains why δ is wired up to be processed in the way that it is. Applying the Dretske-inspired test, that is evidence that δ is representing that the reward was more/less than the represented expected value and telling downstream processing to revise expected values accordingly—which is the content presupposed by the cognitive neuroscientific explanation (section 3). That content partly concerns the content of another of the system's representations (Vi), and so is metarepresentational.

Carruthers considers the possibility that the forms of behaviour elicited in experiments on animal metacognition might be generated by *non-conceptual* metarepresentations; and rejects it (2009, p. 170-1). Although elsewhere he suggests that metarepresentation requires concepts of mental states (2008, p. 58; and indeed 2009, p. 171, last para.)—which would entirely rule out non-conceptual metarepresentations in animals lacking concepts of representational properties—in the passage at (2009, pp. 170-1) he rejects non-conceptual representation for a different reason. He applies the test we relied on above and concludes that the use made of the non-conceptual representations in question in these animals is 'aimed at answering a first-order question about the object—"Is it edible?", "Is it safe?", "What comes next?", and so on—rather than being aimed at changing the animal's own degrees of belief' (2009, p. 171). Applying that test to δ, the use made of it *is* aimed at changing the agent's own representational states. So δ has non-conceptual meta-level content of the kind that Carruthers explicitly rejects in the case of animal metacognition. (Given his other commitments, it is likely that Carruthers would want to resist the conclusion that a representation with meta-level content, even of a non-conceptual variety, is involved in low-level reward-guided decision-making.)

Proust argues that behaviour in animal metacognition tasks is driven by non-conceptual representations, but does not think they are metarepresentational (Proust 2007, 2009b; pace Carruthers 2009, p. 170). Proust develops a subtle view according to which metacognition is neither first-order nor second-order (Proust 2007, p. 287). The non-conceptual representations involved are driven by other representations, but rely on properties of the vehicles of those representations like the strength of a trace (Proust 2003: 352). The states they are about are indeed representations, but their contents only concern non-representational properties of those states (Proust 2010: 996). So Proust would reject the kind of metarepresentation at issue here, in which a representational property of another representation figures in its non-conceptual content.

If the theory Proust has developed to account for animal metacognition were applied to the RPE signal, then the content of δ would concern only non-representational properties of the

vehicles of the other representations to which it is connected (expected values Vi and reward feedback r). Such vehicle properties are indeed further properties with which δ correlates. If Vi and r are coded by firing rates, then δ correlates with the difference between those firing rates. However, applying the Dretske-inspired test relied on above, the correlation with firing rates is not the most direct explanation of why δ is acted on in downstream processing as it is. Relying on such a correlation would only make sense because there happens to be a further correlation between it and a relevant contentful property. The correlation with a representational property of Vi—that its content fails to match the current feedback—offers of more direct explanation of the fact that δ is processed as it is.

In short, the considerations relied on here suggest that the RPE has content that would count as metarepresentational on Proust's view, and also on Carruthers view when he is not tying metarepresentation constitutively to the possession of concepts; and both reject the idea that metarepresentations of this kind are involved in animal metacognition.

5.3 Alternative explanations—worldly correlates

We noted above that δ also correlates with many external properties. Could it be argued that it is by making use of the RPE's correlation with some external property that the system manages to harvest rewards optimally? Carruthers' deflationary tactic was to replace (meta −) representations of confidence with first-order representations of probability of reward. In this section we will consider that potential worldly correlate of δ and others, and see that none satisfactorily explains how δ contributes to the system's performing its task.

Consider whether δ represents the likely average payoff (probability × magnitude) for the previously selected action. We saw above that δ does indeed carry some correlational information about the expected value of the last action (action 1, say), especially in an unstable environment. If that was the correlation being used, then there should be nothing wrong with producing a large positive RPE signal in response to receiving a large reward for action 1, even if that reward were fully predicted by V1. However, we know that to do so would lead to adverse consequences for the agent. The system would then select action 1 even more often, which would be a suboptimal response. It would stop the agent exploring action 2 often enough to learn that the reward contingencies have changed when they do. A similar argument undermines the idea that δ is made use of for its correlation directly with the reward that is received.

In fact, a prominent theory of drug addiction is that this is precisely what is going wrong in addicts' reward system (Hyman 2005). Even though the rewards from an action are fully predicted, a false RPE signal is generated by the direct action of the drug on the dopamine system, leading reward expectations to be revised ever upwards, far beyond the actual levels of reward received. If that is right, making δ a more direct correlate of the reward value of the chosen option, rather than reward prediction error, leads to pathological behaviour.

What about an imperative content for δ: *perform the most recent action*, or for negative values: *do not perform the most recent action*. The trouble with this putative content is that the connection between δ and the next action is not very tight, nor should it be. If action 1 has a low expected value V1, is chosen (because of the role of chance in the decision rule),

and then rewarded, a large positive δ will result. But the expected value V2 of the other action may still be much higher, making it much more likely that action 2 will be chosen next (as it should be). Similarly for large negative δ—the action may still be likely to be chosen again, if starting from a high expected rewarded relative to the other option. The RPE does not directly dictate the chances of choosing an option, but only increases or decreases those chances, relative to their previous level. (Furthermore, in both cases the extent of the change is also modulated by the learning rate α). If the strength of the connection between δ and action selection were to be increased so that the chance of performing the next action was directly dictated by the magnitude of δ, then the agent would perform less well because the system would no longer be able to take account of reward history, but would only be able to take the most recent feedback into account when calculating what to do. That is a suboptimal strategy in all but the most radically stochastically variable environments. So direct object-level imperative contents about action are not appropriate contents to capture the information processing role of δ.

This discussion suggests an intuitive test as to which correlation is being made use of by the system: is it a correlation the strengthening of which would lead to greater overall benefits for the system? Would its weakening reduce the benefits obtained by the system? If a stimulus is being consumed as a probabilistic sign of a predator, and that probabilistic correlation is strengthened, then the agent will benefit.[9] My suggestion is that we hold fixed how all the other components of the system are disposed to interact with each other and with the outside world (at input and output). Then we select one of the correlations into which δ enters and consider what would happen if that correlation were strengthened or weakened. If strengthening that correlation would increase the kind of benefit for the agent which the system has been designed to achieve (by evolution or learning), then that is evidence that δ represents that content (*mutatis mutandis* for weakening a correlation). Given the use that is in fact made of δ in downstream information processing, and the optimality results in the literature on reinforcement / temporal-difference learning, the correlation whose strengthening will be most beneficial to the agent is the correlation between δ on the one hand, and the difference between feedback and expected value on the other.

The other broad tactic for explaining apparent metacognitive performance first order is to see it as the result of some form of response conflict (Hampton 2009). There need be no response conflict involved in these reward-guided decision making tasks. One option may be consistently more likely to be rewarded than the other options. Nevertheless, rewards that are unexpectedly delivered or omitted continue to produce RPE signals, which are in turn used to update reward expectations in a way that affects choice behaviour. So an explanation in terms of response conflict is unlikely to succeed.

In sum, if the midbrain dopamine RPE signal is part of the implementation of an actor-critic reinforcement learning algorithm, and we use the Carruthers / Dretske approach as an evidential test for what the signal represents, then we have good evidence that the RPE does indeed have metarepresentational content. Leading alternative accounts of the information

---

[9]Trade offs between misses and false positives also have to be considered (Godfrey-Smith 1991).

processing involved in these tasks, including those that are plausibly first-order, are less well supported by behavioural data, or neural data, or both.

5.4 Contrast Cases

How much further does this argument extend? Do the kinds of evidential considerations we have been relying on in the case of RPEs also apply to other kinds of information processing system? Does it follow that metarepresentation arises in every comparator circuit? Here I offer, briefly and more tentatively, some thoughts on the potential for the argument above to generalise to other cases.

The argument does not apply to the well-known Ernst & Banks comparator mechanism. This mechanism combines two sources of information about a stimulus (visual and aural, say), weighting them by the variance on the respective channels (Ernst and Banks 2002). We argued above, in relation to Kiani and Shadlen (2009), that the noise in a signal could be relied on as a sign of the probability of some external world fact. The same applies to two sources of information about a stimulus. Weighting these estimates by variance effectively takes the variance of each to be an indicator of the probability that the world is as that channel says it is. So it is not a case of metarepresentation.

Nor does the argument obviously apply to a control system that compares a represented target state with feedback from the world about the agent's progress towards reaching that state. That comparator signal may simply represent the fact of how the agent's action lies in relation to a particular worldly state. If that process is taken offline, however, the case is quite different. In some models of motor control, when a motor program is executed it is used in parallel to produce a prediction of the outcome that will be produced (Wolpert et al. 1995). That prediction is compared to the target state and a prediction error signal is used to adjust the details of the motor program being run. Such adjustments can be made before there has been time for any sensory feedback about the result of executing the motor program. That architecture would seem to mirror the role of the RPE. So it is plausible that such offline motor prediction errors do metarepresent—representing the discrepancy between a target state or imperative representation and a prediction about the state that is currently likely to be achieved, and being used to adjust the motor program being run as a result.

This brief overview suggests that, although the considerations we have been pointing to might extend to some other forms of low-level information processing, they do place substantive constraints on how widely metarepresentation arises. However, on some views predictive coding is an absolutely ubiquitous phenomenon in the brain (Friston 2010), in which case our form of metarepresentation might be very widespread indeed. However, that radical hypothesis is by no means empirically established. If it did turn out to be true, then the consequence that metarepresentation is ubiquitous would fairly capture what is so surprising about the proposal.

6. Conclusion

Cognitive neuroscience promises to bring together behavioural data with neural evidence and computational modelling so as to understand the information processing carried out by parts of the brain. Research on probabilistic reward-guided decision making has gone a long way towards fulfilling that promise. Many lines of neural evidence converge on the conclusion that the brain is making the calculations captured by one of a small family of algorithms, all of which rely ineliminably on a signed reward prediction error signal. On the face of it, actor-critic models take the RPE to have meta-level contents, which is intuitive, given the way the RPE signal is used in downstream processing to modify representations of expected value. Deflationary approaches to metarepresentation, found in the discussion of metacognition in other animals, do not succeed in providing a first-order reading of the RPE signal. Indeed, one approach to representational content proposed in the metacognition literature furnishes a positive argument that RPEs are metarepresentational. If so, low-level non-conceptual metarepresentation is much more widespread than previously thought.

Acknowledgments

Appendix: Temporal Difference Learning Algorithms

The model above makes two major simplifying assumptions. First, it assumes that each trial consists of just one time-step, so it does not raise the problem of assigning credit to actions in a sequential task requiring a series of correct choices to obtain a reward. Secondly, it does not make the reward for an action conditional on a stimulus; that is, it contains no *conditioned stimulus* (CS). Prediction error signals in the brain were initially discovered by Schultz and others in a Pavlovian conditioning paradigm in which a CS predicts an unconditioned stimulus / reward, not requiring any action to be performed.

In the Pavlovian task each trial consists of multiple time steps. For example, a CS might be presented at step 1 and rewarded at step 3 of a 6 step trial (t = 1 to 6). The temporal difference learning algorithm generates, at each time step, a prediction V(t) of the total reward that will be delivered (on average) over the course of the rest of the trial from that time step onwards. V(t) is conditional on the stimulus (calculated as the dot product between a learnt weight vector and a vector representing the stimulus). The reward expected to be delivered *during* a given time step t is V(t) minus V(t+1) (except at the last time step, when it is just V(t)). The difference between this expected value and the reward actually received during that time step constitutes the prediction error:

$$\delta\left(t\right) = r\left(t\right) - \left(V\left(t\right) - V\left(t+1\right)\right)$$

As before, $\delta(t)$ is used to update the expected values $V(t)$ that are carried forward to the next trial. That is achieved by adjusting the weights by which the stimulus vector is multiplied to arrive at $V(t)$. For simplicity, we can treat this as if it were a direct impact of $\delta(t)$ on $V(t)$: $V(t)$ is increased or decreased in line with the value of $\delta(t)$, scaled by the learning rate α:

$$V\left(t\right) \to V\left(t\right) + \alpha\delta$$
$$= V\left(t\right) + \alpha\left\{r\left(t\right) - \left(V\left(t\right) - V\left(t+1\right)\right)\right\}$$

Rearranging:

$$V\left(t\right) \to \left(1 - \alpha\right)V\left(t\right) + \alpha r\left(t\right) + \alpha V\left(t+1\right) \quad (2)$$

[Compare (1), the non-TD version from section 5.1 : $V \to (1 - \alpha)V + \alpha r$]

As in our non-TD version (section 5.1), the expected value is adjusted by giving some weight to the previous prediction and some weight to the most recent feedback, the relative weightings determined by the learning rate α. But in the temporal-difference version here there is an additional term: $V(t)$ is also adjusted by a proportion of the reward expected at the next time step. The effect of this is to make reward expectations propagate backwards to the earliest time step at which they are fully predicted. If an initially unpredicted reward is delivered at step 3 then a large positive prediction error $\delta(3)$ will be generated then, leading $V(3)$ to be increased. If the same reward is delivered next time round, the prediction error $\delta(3)$ will be less, since the reward is now partly predicted by the revised $V(3)$. But this will generate a prediction error at step 2, $\delta(2)$, because of the effect of $V(3)$ (the third term in formula (2) above), which moves $V(2)$ in the direction of the total reward to be delivered by the end of the trial, even though no reward is delivered at step 2. By this means all the reward expectations before the reward is delivered are gradually increased, starting from the time step at which the CS predicts the reward, so that they all come to fully predict the total reward that will be delivered, on average, by the end of a trial.

This accounts for the classic RPE signal observed in Pavlovian experiments (Schultz et al. 1997; Schultz 1998). Initially a positive RPE is recorded at the time of the reward. This gradually shifts backwards in time to the time of the stimulus that predicts the reward, with no signal observed at the time of reward delivery. If a predicted reward is then omitted, there is a transitory reduction in the dopamine response at the time of the anticipated reward.

In an instrumental conditioning task, payoffs depend upon which action is performed as well as the stimulus context. In a multi-stage instrumental task we need to supplement the model with predictions about the value of available actions at each time step. The credit assignment problem is overcome because the expected value of eventual rewards propagates backwards, as we saw above, to earlier states that predict those rewards. Those state values can then be used to select an action that leads to a more favourable state at the next step. A particular

merit of the actor-critic model is that the same RPE signal δ is used both (i) to update reward expectations (critic); and (ii) to update the policy by which actions are selected at each time step based on state values (actor). That level of complexity is omitted from the simple model set out above, but is essential to the way the TD learning algorithm overcomes the knotty credit assignment problem.

We saw in section 5 that our simple rule for updating expected values based on the RPE could be reformulated to eliminate prediction errors, becoming just a weighted average of the previous expected value and the most recent feedback (formula (1)). The RPE can also be eliminated from the updating of value in the TD algorithm (formula (2)), albeit with an extra term reflecting the expected value at the next time step. We saw in section 5 that, despite the availability of a first-order calculation, there is strong neural evidence that the RPE is in fact calculated as a separate step and relied on in choice behaviour. Now we can see why that should be. It is because, as well as updating reward expectations, the RPE is simultaneously used by the actor to update the policy by which actions are chosen. The computational efficiency of this actor-critic algorithm, with its dual role for an RPE signal δ(t), explains why the task is not in fact solved by the first-order calculation set out in formulas (1) and (2) above.

References

Anderson ML, Perlis D. Logic, self-awareness and self-improvement: the metacognitive loop and the problem of brittleness. Journal of Logic and Computation. 2005; 15:21–40.

Apperly IA, Butterfill SA. Do humans have two systems to track beliefs and belief-like states. Psychological review. 2009; 116(4):953. [PubMed: 19839692]

Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron. 2005; 47(1):129–41. [PubMed: 15996553]

Bush, Robert R.; Mosteller, Frederick. A mathematical model for simple learning. Psychological Review. 1951; 58(5):313–23. [PubMed: 14883244]

Carruthers, Peter. Meta-cognition in animals: a skeptical look. Mind and Language. 2008; 23(1):58–89.

Carruthers, Peter. Mindreading underlies metacognition. Behavioral and Brain Sciences. 2009; 32(02):164–82.

Claridge-Chang A, Roorda RD, Vrontou E, Sjulson L, Li H, Hirsh J, Miesenböck G. Writing memories with light-addressable reinforcement circuitry. Cell. 2009; 139(2):405–15. [PubMed: 19837039]

Corrado, GS.; Sugrue, LP.; Brown, JR.; Newsome, WT. The trouble with choice: studying decision variables in the brain. In: Glimcher, PW., et al., editors. Neuroeconomics: Decision making and the brain. Elsevier; Amsterdam: 2009. p. 463-80.

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD. BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science. 2008; 319(5867):1264. [PubMed: 18309087]

Davidson, D. Subjective, intersubjective, objective. Oxford University Press; USA: 2001.

Daw, ND.; Niv, Y.; Dayan, P. Actions, policies, values and the basal ganglia. In: Bezard, E., editor. Recent breakthroughs in basal ganglia research. Nova Science Publishers; New York: 2006.

Dretske, Fred. Knowledge and the Flow of Information. MIT Press; Cambridge, M.A.: 1981.

Dretske, Fred. Explaining Behaviour: reasons in a world of causes. MIT Press; Cambridge, MA: 1988.

Eliasmith, Chris; Anderson, Charles H. Neural Engineering: computation, representation, and dynamics in neurobiological systems. MIT Press; London / Cambridge MA: 2003.

Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002; 415(6870):429–33. [PubMed: 11807554]

Fleming SM, Dolan RJ, Frith CD. Metacognition: computation, biology and function. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2012; 367:1280–1286.

Fodor, Jerry. Psychosemantics. MIT Press; Cambridge, MA.: 1987.

Friston, Karl. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010; 11(2):127–38.

Godfrey-Smith, Peter. Signal, decision, action. Journal of Philosophy. 1991; 88:709–22.

Godfrey-Smith, Peter. Mental representation, naturalism and teleosemantics. In: Papineau, David; Macdonald, Graham, editors. New Essays on Teleosemantics. Oxford: OUP; 2006.

Hampton RR. Rhesus monkeys know when they remember. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(9):5359–62. [PubMed: 11274360]

Hampton RR. Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? Comparative cognition and behavior reviews. 2009; 4:17. [PubMed: 20046911]

Hare B, Call J, Tomasello M. Do chimpanzees know what conspecifics know? Animal Behaviour. 2001; 61(1):139–51. [PubMed: 11170704]

Haruno M, Kawato M. 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. 2006; 95(2):948. [PubMed: 16192338]

Heyes CM. Theory of mind in nonhuman primates. Behavioral and Brain Sciences. 1998; 21(01):101–14. [PubMed: 10097012]

Hyman, Steven E. Addiction: A Disease of Learning and Memory. American Journal of Psychiatry. 2005; 162:1414–22. [PubMed: 16055762]

Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. science. 2009; 324(5928):759. [PubMed: 19423820]

Kovács ÁM, Téglás E, Endress AD. The social sense: Susceptibility to others' beliefs in human infants and adults. Science. 2010; 330(6012):1830. [PubMed: 21205671]

Leslie AM. Pretense and representation: The origins of "theory of mind.". Psychological review. 1987; 94(4):412–26.

Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND. Differential roles of human striatum and amygdala in associative learning. Nature Neuroscience. 2011; 10(14):1250–52.

Marr, D. Vision. W H Freeman & Co.; New York: 1982.

Mars RB, Shea Nicholas, Kolling N, Rushworth MFS. Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. The Quarterly Journal of Experimental Psychology. 2010; (1):1–16.

Marticorena, Drew C. W.; Ruiz, April M.; Mukerji, Cora; Goddu, Anna; Santos, Laurie R. Monkeys represent others' knowledge but not their beliefs. Developmental Science. 2011 DOI: 10.1111/j. 467-7687.2011.01085.x.

McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003; 38(2):339–46. [PubMed: 12718866]

Millikan, Ruth Garrett. Language, Thought and Other Biological Categories. MIT Press; Cambridge, MA: 1984.

Millikan, Ruth Garrett. Compare and Constrast Dretske, Fodor and Millikan on Teleosemantics. Philosophical Topics. 1990; 18:151–61.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003; 38(2):329–37. [PubMed: 12718865]

O'Doherty JP, Dayan P, Schultz P, Deischmann J, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304:425–54.

Onishi KH, Baillargeon R. Do 15-month-old infants understand false beliefs? Science. 2005; 308(5719):255. [PubMed: 15821091]

Papineau, David. Reality and Representation. Blackwell; Oxford: 1987.

Perner J, Frith U, Leslie AM, Leekam SR. Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. Child Development. 1989; 60(3):689–700.

Proust J. Does metacognition necessarily involve metarepresentation? Behavioral and Brain Sciences. 2003; 26(3):352.

Proust J. Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? Synthese. 2007; 159(2):271–95.

Proust J. Epistemic agency and metacognition: an externalist view. Proceedings of the Aristotelian Society. 2008; 108:241–68.

Proust J. Overlooking metacognitive experience. Behavioral and Brain Sciences. 2009a; 32(02):158–59.

Proust, J. The representational basis of brute metacognition: a proposal. In: Lurz, R., editor. The Philosophy of Animal Minds: New Essays on Animal Thought and Consciousness. C.U.P.; Cambridge: 2009b.

Proust J. Metacognition. Philosophy Compass. 2010; 5(11):989–98.

Redgrave P, Gurney K. The short-latency dopamine signal: a role in discovering novel actions? Nature Reviews Neuroscience. 2006; 7:967–75.

Rushworth MFS, Mars RB, Summerfield C. General mechanisms for making decisions? Current opinion in neurobiology. 2009; 19(1):75–83. [PubMed: 19349160]

Schultz W. Predictive reward signal of dopamine neurons. Journal of neurophysiology. 1998; 80(1):1. [PubMed: 9658025]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275(5306):1593. [PubMed: 9054347]

Shannon, Claude E. The mathematical theory of communication. In: Claude, E Shannon; Warren, Weaver, editors. The Mathematical Theory of Communication. University of Illinois Press; Urbana: 1949.

Shea, Nicholas. Consumers Need Information: supplementing teleosemantics with an input condition. Philosophy and Phenomenological Research. 2007; 75(2):404–35.

Shea, Nicholas; Heyes, Cecelia M. Metamemory as evidence of animal consciousness: the type that does the trick. Biology and Philosophy. 2010; 25:95–110. [PubMed: 20234826]

Smith JD. The study of animal metacognition. Trends in cognitive sciences. 2009; 13(9):389–96. [PubMed: 19726218]

Surian L, Caldi S, Sperber D. Attribution of beliefs by 13-month-old infants. Psychological Science. 2007; 18(7):580. [PubMed: 17614865]

Sutton, RS.; Barto, AG. Reinforcement learning: An introduction. The MIT press; 1998.

Wimmer H, Perner J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition. 1983; 13(1):103–28. [PubMed: 6681741]

Wolpert DM, Ghahramani Zoubin, Jordan Michael I. An internal model for sensorimotor integration. Science. 1995; 269:1880–2. [PubMed: 7569931]

Wolpert DM, Diedrichsen J, Flanagan JR. Principles of sensorimotor learning. Nature Reviews Neuroscience. 2011

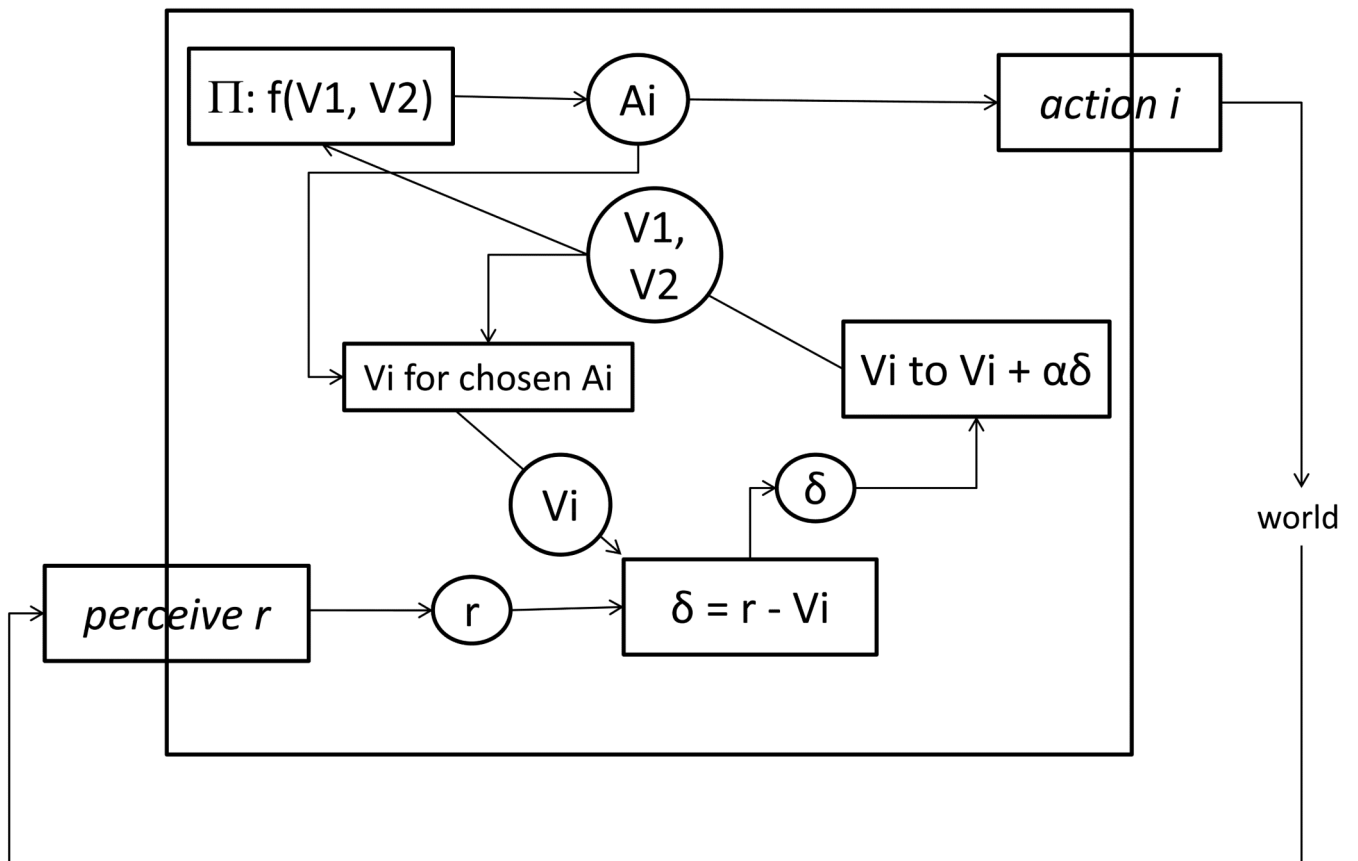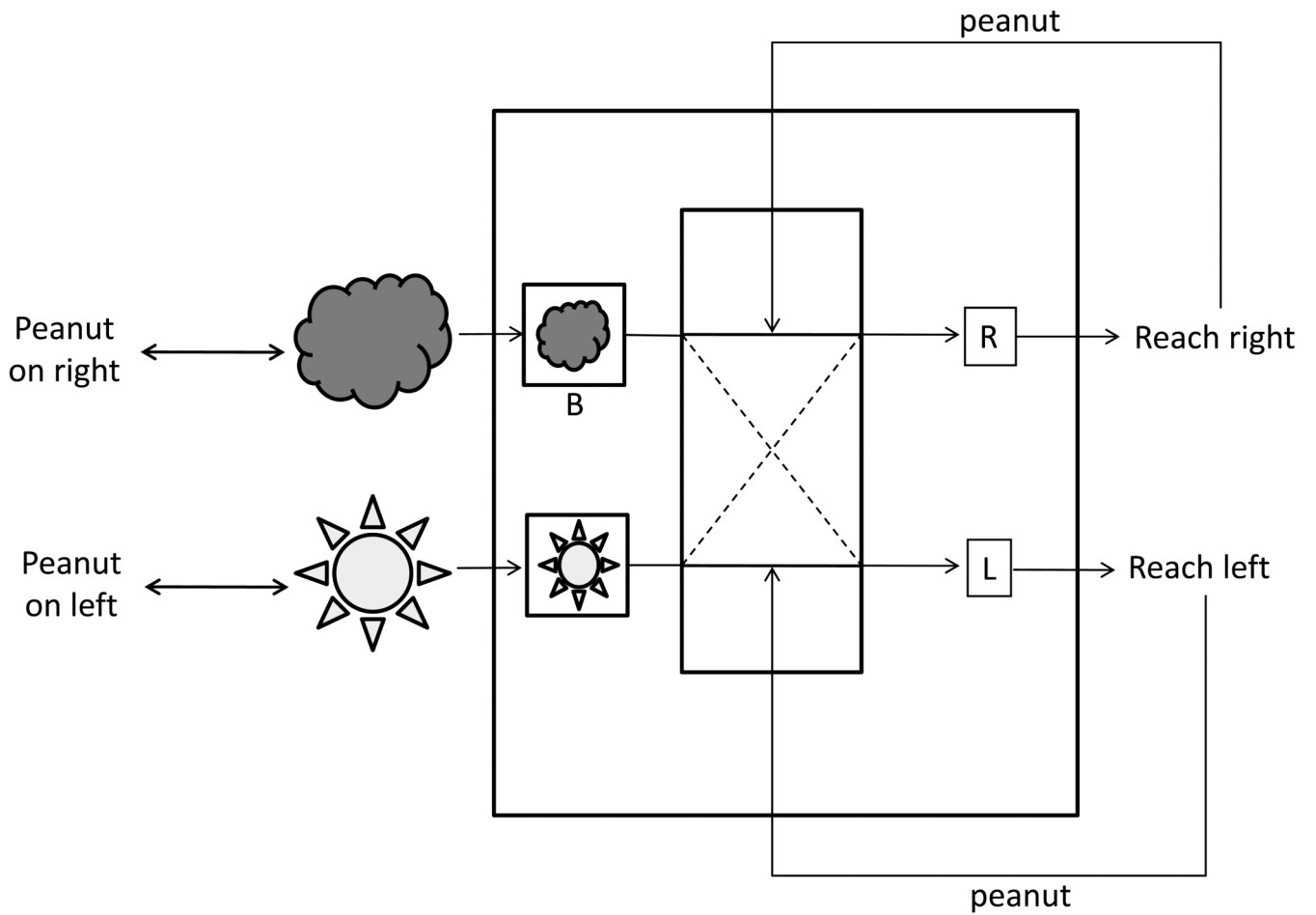**Fig. 1. Simplified actor-critic model**

Fig. 2. Content from instrumental conditioning (following Dretske 1988)