# Reward Prediction Error Signals are Metarepresentational

**Nicholas Shea (nicholas.shea@philosophy.ox.ac.uk)**
Faculty of Philosophy, University of Oxford
10 Merton Street, Oxford, OX1 4JJ, UK

**Abstract**

Although there has been considerable debate about the existence of metarepresentational capacities in non-human animals and their scope in humans, the well-confirmed temporal difference reinforcement learning models of reward-guided decision making have been largely overlooked. This paper argues that the reward prediction error signals which are postulated by temporal difference models and have been discovered empirically through single unit recording and neuroimaging do have metarepresentational contents.

**Keywords:** metarepresentation; metacognition; reward-guided decision making; temporal difference learning

## Introduction

It is often argued that the capacity for metarepresentation is a particularly sophisticated cognitive achievement (Carruthers, 2008). In the animal literature authors debate whether success on tasks that seem to require self-monitoring can be achieved without metarepresentation (Carruthers, 2009; Hampton, 2001; Smith, 2009). The same question is debated about tasks that seem to require keeping track of the mental states of others (Hare, Call, & Tomasello, 2001; Heyes, 1998). It is assumed that evidence that non-human animals are processing metarepresentations is a sign of considerable psychological sophistication, even consciousness (Cowey & Stoerig, 1995; Smith, Shields, & Washburn, 2003; Stoerig, Zontanou, & Cowey, 2002); although some have argued that some forms of metarepresentation can be achieved more easily (Shea & Heyes, 2010). In developmental psychology the capacity to have beliefs about others' belief states is seen as a particularly important developmental transition (Leslie, 1987; Perner, Frith, Leslie, & Leekam, 1989; Wimmer & Perner, 1983), although here too there is increasing evidence that some forms of very early behaviour depend upon representing or keeping track of others' representations (Apperly & Butterfill, 2009; Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007).

This paper argues that there is already strong evidence of metarepresentation in a different literature – one in which issues about metarepresentation have seldom been canvassed. Research on reward-guided decision making has produced an impressive body of converging evidence that midbrain dopamine neurons generate a reward prediction error signal (RPE) that is causally involved in choice behaviour (Rushworth, Mars, & Summerfield, 2009). I argue that such RPEs carry a metarepresentational content. The system is conserved across primates and rodents, and perhaps more widely (Claridge-Chang et al., 2009). Some animals doubtless make more sophisticated use of metarepresentations than this. But this result does show that there is at least one variety of metarepresentation that is found very widely in the animal kingdom.

Metarepresentations are representations whose content in part concerns the content of another representation. The sentence: 'The main headline in the *Post* today is in huge letters' is not metarepresentational. It concerns another representation, but not its content. The sentence: 'The main headline in the *Post* today is about Gaza' is metarepresentational.

To assess whether reward prediction error signals are metarepresentational I examine the standard information-processing account of their role in generating behaviour and ask what content RPEs would have to have for that account to be vindicated.

## Reward Prediction Errors

The prediction error signal postulated by temporal difference learning models of reward-guided decision making (Sutton & Barto, 1998) was discovered empirically through single unit recording in the awake behaving macaque (Schultz, Dayan, & Montague, 1997). The central idea is that the brain keeps track of the expected value of various possible actions. When the animal performs an action, it computes an expected value of the current behaviour. When feedback does not match that expected value a prediction error signal is generated. The signal is used to update the stored representation of the value associated with that behaviour, by an amount given by the learning rate. For example if an animal pulls a lever for the first time and obtains a reward, that will generate a prediction error signal. The actual reward will have exceeded any expectation of reward. (If the animal has some general expectation of there being some rewards in this environment, then it will have a mild general expectation of reward.) So the unexpected reward will generate a prediction error signal.

Normative models of reinforcement learning attempt to capture the best way of calculating what to do given a history of rewarded and unrewarded actions (under various computational constraints). The popular temporal difference models suggest that reward prediction error signals will be used to update the expected value of the chosen action. As a result, on future occasions the animal will expect slightly more from pressing the lever. How much more depends upon the learning rate.

After enough learning, the animal will come to expect reward when it presses the lever. If it presses the lever and receives no reward, that will again create a RPE, but in the opposite direction. The effect will be to reduce the animal's

expectation of obtaining a reward from that action in the future.

The fact that an animal's behaviour in experimental situations is well-described by a temporal difference learning model is not enough to show that it is really processing over internal representations that represent the quantities found in the model. On an instrumentalist approach to representation it would be enough to show that the model is adequate to the data and predictively accurate. But that fact also gives us *some* evidence that the animal really is processing over real internal variables that correspond to the quantities in the model: expected values and prediction errors. We get stronger evidence by investigating brains directly.

Of course, there could be real internal representations that are coded in a very non-obvious format. So if the search for evidence of internal representations in the brain were to deliver a negative result, that would be far from conclusive evidence against the existence of internal representations. Fortunately in the case of RPEs, it looks as if there are internal representations with a fairly stable, tractable neural basis. There are midbrain dopamine neurons whose firing patterns correspond to the quantities found in the model.

In single unit recording in monkeys, 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 for appetitive conditioning (Bayer & Glimcher, 2005; Schultz, 1998; Schultz et al., 1997). Functional magnetic resonance imaging (fMRI) in humans shows a similar pattern of effects. By fitting temporal difference learning models to the behavioural data, trial-by-trial estimates of a subject's representations of value and RPE are generated and correlated with the fMRI response. These find a BOLD response consistent with RPEs both in the VTA (D'Ardenne, McClure, Nystrom, & Cohen, 2008) and in areas of the ventral striatum receiving dopaminergic inputs (Haruno & Kawato, 2006; McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003).

## Content in the Model

What do these prediction error signals represent? To answer that question we need to examine the way they are produced and how they enter into subsequent processing. In order to have a fixed target, we shall presuppose that the current state of the evidence supports the conclusion that the calculations hypothesized by temporal difference learning models are being performed in the brains of humans and other animals when they make rapid simple decisions for probabilistic rewards, and that these calculations are responsible for the observed patterns of choice behaviour.

At the outset of a trial, when a number of behavioural options are presented, sometimes following a cue, the system activates an expected value for each option. A decision rule makes use of these values to choose an option. For example a softmax decision rule increases the probability of choosing one option as its value relative to other options increases.

When the agent has made his choice and feedback has been received, the system calculates a prediction error: the (signed) difference between the expected reward and the actual reward. For example, if a moderate reward was expected with only low probability, a large positive RPE will be generated if the reward is delivered. The same level of reward would produce a much smaller RPE if it were anticipated. The omission of an expected reward generates a large negative prediction error.

The RPE is then used to update the expected reward for that action, which in turn is used to make the next decision. The updated expected reward is moved in the direction of the reward received. The extent to which it is moved is moderated by the learning rate. If the learning rate is low, the expected value is adjusted only slightly in the direction of the reward just delivered. If the learning rate is high, the adjustment is more substantial. At the limit, were the learning rate equal to one, the expected value would be reset to the value of the last reward.

So the putative representations of interest that figure in the information processing story are as follows.[1]

| | |
|---|---|
| Expected value at t of option 1 | $V1_t$ |
| Expected value at t of option 2 | $V2_t$ |
| Chosen behaviour at t | Bi (B1, B2) |
| Actual reward at t | $r_t$ |
| Prediction error (having chosen i) | $\delta_t = r_t - Vi_t$ |
| Learning rate | $\alpha$ |
| Updated expected values: | |
| Chosen behaviour i | $Vi_{t+1} = Vi_t + \alpha\delta_t$ |
| Unchosen behaviour j | $Vj_{t+1} = Vj_t$ |

What should we think of these values as representing if the information processing story is to make sense? We have to use words to capture these contents, but the with the caveat that the words are not aiming to capture either (a) what the system or the agent understands the contents of the states to be; or (b) constituent structure – the states whose contents we are describing have none of the constituent structure that is found in the sentences we use to describe them.

### Reward and Value Representations

Quantity $r_t$ is straightforward: it represents the value of the reward actually received t (the value of so many ml of juice, for example): *$r_t$ was received*. B1 and B2 are also straightforward. Bi has a directive content: *do action i*, or *choose the action that will select option i.*

$V1_t$ and $V2_t$ seem to be stating facts about causal conditionals. However, they do not simply predict the value of the next chosen action. Rather, they predict the reward that will be obtained on average if an option is repeatedly

[1] The symbols are used both to refer to the representations involved, and to pick out the quantities variably represented by those representations.

chosen in the current environment. That is, they represent expected values in the probabilistic sense of expectation (summing probability x magnitude): *if option 1 is chosen, then the expected reward will be $V1_t$.* Here expected reward is an objective quantity of which $V1_t$ is the agent's current estimate. Expected rewards should not be confused with the agent's (subjective, represented) expectations about rewards. So the success condition for behaviour driven by representation $V1_t$ looks to be something like this: the average reward payoff that would be achieved by repeatedly choosing option 1 in the current environment would be $V1_t$.[2] If the actual expected gain from option 1 is higher than this, then the agent's behaviour will be suboptimal in that it will choose option 1 less than it should. Conversely, if the actual expected gain from option 1 is lower than that represented by $V1_t$ then the agent's behaviour will be suboptimal in that it could increase its chances of receiving higher overall rewards by selecting option 1 less frequently.

When an actual reward is received when there is a relatively low $V1_t$ that could be because the estimate of expected value is wrong, or it could be that this is one of those low-probability occasions where option 1 is rewarded. Temporal difference learning models finesse this information gap by re-jigging the value representation in every case, in effect treating it as possible that this bit of feedback is a sign that the current estimate is wrong (either because of insufficient learning or because the environment has changed). This leads the estimate of expected reward to be altered for future trials. That recalculation is mediated by the magnitude and sign of the difference between represented expected value and feedback: the RPE $\delta_t$.

## Reward Prediction Error

Characterizing the content of $\delta_t$ is more tricky. An input-driven approach to content looks at the parameters with which a representation covaries and uses them to ascribe content (Dretske, 1981). The notorious difficulty is that a given representation that correlates with some inputs will thereby correlate with very many others too (Fodor, 1987). Considered informationally $\delta_t$ will carry some information about the actual reward, some information about the expected reward, and even more reliable information about the difference between them. There are good reasons to be suspicious of the idea that the content of a representation is that feature with which it correlates most strongly (Millikan, 1984). For example, consider a predator-detector set up to produce lots of false positives. Its strongest correlation may be with shadows, rather than predators.

This suggests that we should also look at how a representation is used (Godfrey-Smith, 2006). The firing of a predator detector leads to avoidance behaviour whether or not the stimulus was just a shadow. Thus, the way a representation enters into downstream processing helps us to focus in on its content. Downstream, $\delta_t$ is used to update the expected value of the reward for future trials. The best way of describing how $\delta_t$ is being acted on cannot help but advert to the fact that it is used, not directly to select an action, but to update a second internal register – to update another internal representation.

Sometimes hierarchical information processing involves a series of steps (e.g. filters) or the combination of information from different sources to form a new representation (e.g. conjunctive feature detectors). These are also cases where downstream representations are changed in reliance on upstream representations. But in those cases the upstream representations are relied upon for information they carry about some external fact of the matter. RPEs, by contrast, are acted upon to update value representations not directly because of information they carry about reward, but because they carry information directly about the accuracy of previous estimates. Whether a reward has just been received or not, the job of $\delta_t$ is to reset the expected value $Vi_{t+1}$ by a lot if the feedback on the current occasion was a long way from the average that was expected over repeated trials, $Vi_t$, and by only a little if $Vi_t$ closely matches the current feedback.

Consider the kinds of things that could go wrong in neural processing and why, according to the temporal difference learning model, these would constitute errors. Suppose that because of some glitch a large positive prediction error were generated on an occasion where the chosen option B1 was not expected to be very rewarding and was not rewarded. We can't understand this error straightforwardly in terms of some mistaken behaviour on the next trial, because the decision rule might well lead B2 to be chosen on the next trial. The error is not in how the system acts on the next trial, but in how it changes its expectations on the next trial, because it will have mistakenly increased V1, its expected reward for option 1. Correlatively, suppose $\delta_t$ is produced in the regular way so it does reflect the difference between $r_t$ and $V1_t$, but is then ignored in downstream processing. Here we would say that it correctly represented the possibility that the previous prediction $V1_t$ was mistaken, but that it wasn't acted on correctly to update $V1_{t+1}$ for the next trial.

While these commonsense considerations do not amount to an unassailable argument, they do give us good reason to take the assumptions of the normative model at face value. Surprisingly, it has been little-remarked that temporal difference learning models attribute metarepresentational content to $\delta_t$. Its content can be roughly described as having both descriptive and directive aspects (Millikan, 1996) as follows:

> *The reward for the current chosen option is higher/lower than the predicted expected value $Vi_t$ by an amount $\delta_t$; increase/decrease $Vi_{t+1}$ in proportion to the magnitude $\delta_t$.*

[2] Note that we can talk sensibly about this quantity in counterfactual terms even if the environment is changing so that the agent does not have the opportunity repeatedly to sample option 1 in the current environment.

Notice that both the descriptive and directive aspects of the content make reference to the content of another representation: $Vi_t$ / $Vi_{t+1}$. The reward prediction error signal does not just describe some aspect of the agent's environment. Nor does it just direct a particular action on the part of the agent. Instead we should take seriously the assumption in the temporal-difference learning literature that the RPE's content partly concerns the content of another representation. That is to say, it is a metarepresentation.

## A Competing First-Order Interpretation

In a series of papers Proust has elucidated a form of what we have been calling metarepresentation that differs from the kind of explicit conceptual-level attribution of mental states to oneself and others that is often the focus of the literature on metacognition (J. Proust, 2007, 2008; 2009). She identifies metacognitive 'feelings', like the feeling you know a list of names, as a locus of non-conceptual but meta-level cognition. That is a complementary body of work, which supports the direction taken here by showing how meta-level cognitive phenomena arise within non-conceptual thought, well before the level of explicit, conceptual re-representation of representational contents.

In the course of one of her discussions Proust considers an argument that the signals processed according to temporal difference learning models are first-order and do not involve metarepresentations (Proust 2007, pp. 282-285). This is one of very few existing discussions of whether RPEs are metarepresentational, so merits investigation. Proust's response to the argument is that, in the kinds of self-monitoring paradigms she is interested in, it is not possible to explain performance in terms of the agent keeping track of its objective chance of success. In experiments such as Hampton (2001) the animal seems to be drawing on information beyond that delivered by the problem situation, but that depends upon keeping track of trial-by-trial variation in the agent's own informational resources. That is, the animal's performance (one of the two animals, in the Hampton experiment) seems to depend upon procedural self-knowledge.

Proust's own response leaves the original objection, as it applies to the ordinary cases of reward-guided decision making captured by temporal difference learning models, standing — namely that subjects' behaviour in these experiments can be fully captured in first order terms. The argument is that there is no substantive difference between keeping track of the reliability of one's estimates of expected value (second order) and keeping track of one's chances of succeeding when performing particular behaviours (first order) (Proust 2007. p. 283). That argument does indeed apply to the agent's representation of expected reward (the $Vi_t$ above). Although we could describe these as measuring how well the agent knows that a given option will be rewarding, we have seen above that a first order explanation is preferable. The content to be attributed to the $Vi_t$ is rather subtle, involving a subjective estimate of an objective probabilistic expectation, but the temptation to think of this as metarepresentational is just a mistake. It probably derives from the ambiguity of 'expectation'. In $Vi_t$ the agent is keeping track of an expectation, but 'expectation' here is not what the agent (or anyone else) expects, but an expected value in sense of probability theory: the average of the magnitudes of the available options weighted by their objective probabilities. $Vi_t$ is keeping track of this quantity, which is fixed by external parameters of the problem space, rather than anything about what the agent itself expects.

However, the fact that the expected values Vi should be ascribed first-order contents is not the end of the matter. The argument above was only that the RPE signal was second-order. The objection Proust considers, when levelled against RPE, would then be that $\delta_t$ can be understood in terms of the agent's chances of succeeding, rather than keeping track of any kind of internal state. But it cannot. A very small RPE is compatible with there being a very high chance of succeeding, for example if reward expectations were already high and matched the reward actually received on the current trial. But a very small RPE is also compatible with there being a very low chance of succeeding, for example if reward expectations were low and no reward was delivered. Conversely, a large RPE is compatible both with a high and a low chance of succeeding. What the RPE is telling the agent is not well captured by its connection to the chance of succeeding in future behaviour. If the temporal difference models are anything like on track, what the RPE signal is doing is telling the agent something about how well or badly its representations of expected value for an option match the current feedback. What it does with that information, namely to re-jig its reward expectations proportionately, also makes much more sense in the light of meta-level contents. In short, there is no easy way to replace the meta-level contents inherent in temporal difference models of reward-guided decision making with a first-order reinterpretation.

## Conclusion

The conclusion that non-conceptual metarepresentations are processed during reward-guided decision making in many animals opens up several questions for further research. What distinguishes these representations from the more sophisticated forms of metarepresentation involved in keeping track of the mental states of others, or of the agent's own mental states? To what extent does the temporal difference model connect with decision making at the personal level, or does it just describe a subpersonal system? How inferentially promiscuous are the representations involved in model-free reward guided decision making? Are they conscious or do they have some impact on consciousness?

All these questions are interesting and important. A less obvious question also merits attention. In temporal difference learning models of model-free reward-guided

decision making we have a well-understood, normatively-based model of behaviour with a well-confirmed neural basis. The whole amounts to one of the strongest results of the project of cognitive neuroscience: of finding psychological-level information-processing accounts of behaviour that can be mapped onto neural processes. Once we have a good grip on the kinds of content ascriptions that are supported by these theories, including the metarepresentational contents discussed here, they provide us with an excellent arena against which to test philosophical theories of content. That is, they provide another test case, quite different from the usual repertoire from perception and the cognitive psychology of concepts, of which we can ask: in virtue of what do these representations have the content they do? It will be an important constraint on that theorizing that metarepresentational contents can already be realized these relatively low-level systems.

## Acknowledgments

Work on this paper has been supported by the Oxford University Press John Fell Research Fund, the Oxford Martin School and the Wellcome Trust (grant 086041 to the Oxford Centre for Neuroethics).

## References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states. *Psychological review, 116*(4), 953.

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

Carruthers, P. (2008). Meta-cognition in animals: a skeptical look. *Mind & Language, 23*(1), 58-89.

Carruthers, P. (2009). Mindreading underlies metacognition. *Behavioral and Brain Sciences, 32*(02), 164-182.

Claridge-Chang, A., Roorda, R. D., Vrontou, E., Sjulson, L., Li, H., Hirsh, J., et al. (2009). Writing memories with light-addressable reinforcement circuitry. *Cell, 139*(2), 405-415.

Cowey, A., & Stoerig, P. (1995). Blindsight in monkeys. *Nature, 373*(6511), 247-249.

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

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, M.A.: MIT Press.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA.: MIT Press.

Godfrey-Smith, P. (2006). Mental representation, naturalism and teleosemantics. In D. Papineau & G. Macdonald (Eds.), *New Essays on Teleosemantics*. Oxford: OUP.

Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences of the United States of America, 98*(9), 5359-5362.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour, 61*(1), 139-151.

Haruno, M., & Kawato, M. (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.

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*(01), 101-114.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological review, 94*(4), 412-426.

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

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Millikan, R. G. (1996). Pushmi-pullyu Representations. In J. Tomberlin (Ed.), *Philosophical Perspectives, vol. 9* (pp. 185-200). Atascadero, CA: Ridgeview Publishing.

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

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development, 60*(3), 689-700.

Proust, J. (2007). Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese, 159*(2), 271-295.

Proust, J. (2008). Epistemic agency and metacognition: an externalist view. *Proceedings of the Aristotelian Society, 108*, 241–268.

Proust, J. (2009). The representational basis of brute metacognition: a proposal. In R. Lurz (Ed.), *The philosophy of animal minds*. Cambridge: C.U.P.

Rushworth, M. F. S., Mars, R. B., & Summerfield, C. (2009). General mechanisms for making decisions? *Current opinion in neurobiology, 19*(1), 75-83.

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

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

Shea, N., & Heyes, C. (2010). Metamemory as evidence of animal consciousness: the type that does the trick. *Biology and Philosophy, 25*, 95-110.

Smith, J. D. (2009). The study of animal metacognition. *Trends in cognitive sciences, 13*(9), 389-396.

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences, 26*(03), 317-339.

Stoerig, P., Zontanou, A., & Cowey, A. (2002). Aware or unaware: assessment of cortical blindness in four men and a monkey. *Cerebral Cortex, 12*(6), 565.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580.

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128.