REWARD IS EVIDENCE OF VALUE

Paul de Font-Reaulx

University of Michigan, Ann Arbor

Abstract

What does the reward function in reinforcement learning correspond to in natural agents, such as humans? Previous responses to this question have been shaped by the assumption that we are optimizing for reward itself. In this paper, I argue that this is false. I present and defend evidentialism about reward: the view that biological reward signals provide pre-reflective evidence about a functional quantity of basic value that our minds represent and optimize for. This is broadly analogous to how perceptual signals provide pre-reflective evidence about other properties. I also show that this is consistent with computational reinforcement learning models.

1 Introduction

Reinforcement learning (RL) is a formal framework for modelling agents as expected value maximizers that learn from experience. Recently, it has emerged as a prominent paradigm for explaining the motivational system of natural agents (Haas, 2022; Sripada, 2025). The reason is that it seems to capture important aspects of the mental processes that generate our motivational states. A central component in many RL algorithms is a reward function that specifies the immediate payoff associated with a state or transition that RL agents estimate and strive to maximize in the long run. What, if anything, does the reward function correspond to in us?

This question has generated substantial debate (Aronowitz, n.d.; Haas, 2022; Juechems & Summerfield, 2019; Molinaro & Collins, 2023a, 2023b; Singh et al., 2009; Weber et al., 2025). We might assume that it corresponds to biological reward or punishment signals innately generated in response to select stimuli like food, sex, and bodily harm. But this implies a reductive picture of human motivation in which all our desires are ultimately aimed at the activation of such signals in our brains. Others have proposed that it corresponds to

the satisfaction of acquired goals or desires. But this explanation risks circularity, since RL is supposed to explain how we acquired those goals in the first place. This leaves us without appealing options. Call this *the puzzle of reward*.

In this paper, I defend a new proposal to resolve the puzzle that I call *evidentialism* about reward. In standard RL algorithms, the reward function provides a signal that both teaches the agent what to do and constitutes the scalar the agent strives to maximize. I argue that this is a simplifying modelling assumption and that in natural agents these functions come apart. Biological reward signals provide our minds with simple pre-reflective evidence that updates internal representations of a simple functional quantity of basic value that our minds strive to optimize. This is broadly analogous to how innately determined perceptual signals update our representations of other quantities, such as distance, while remaining agnostic about the referential content. Evidentialism implies that the reward signal is not the optimization target; its content of basic value is. This avoids the issues of the alternative proposals and resolves the puzzle of reward.

Here is an overview of the rest of the paper. In Section 2, I present some background on reinforcement learning and its relation to cognitive science. In Section 3, I present and critically evaluate proposals for how to interpret the reward function as applied to natural agents. In Section 4, I present evidentialism about reward. In Section 5, I consider the objections that evidentialism is in tension with the axioms of reinforcement learning and that it relies on strong assumptions about the content of value representations, and argue that neither constitutes a problem. Finally, in an Appendix, I present an analogue of evidentialism in RL formalism and demonstrate that evidentialism is consistent with learning algorithms such as temporal-difference learning.

2 Background

Reinforcement learning comprises both a problem specification for modelling an ideal agent as maximizing expected value, and a particular set of algorithms for how the agent learns to do so by interacting with its environment (Sutton & Barto, 2018). An ideal RL agent is defined as maximizing the expected discounted sum of a scalar quantity, typically called the *reward*. To know what to do in any given moment, an RL agent typically uses a representation of expected future reward. This is called the *value function*. By interacting with the world through trial and error, the agent learns and updates its value function towards a more accurate representation of the rewards associated with different states or transitions in ways

that can be precisely modelled with RL learning algorithms. This allows it to take actions that lead to more reward.[1]

As it turns out, RL is effective not only for modelling abstract agents, but also for explaining the functioning of natural ones. In recent decades, the cognitive science of motivation has converged on the idea that our minds, and those of many other animals, carry simple non-conceptual representations of expected value (Daw & Shohamy, 2008; Levy & Glimcher, 2011; O'Doherty et al., 2017; Padoa-Schioppa & Assad, 2006). These representations function as a mental index of what our minds register as valuable to pursue and causally steer our behavior towards the options with the highest score (Ballesta et al., 2020). This idea is supported by a large and convergent body of neural and behavioral evidence showing that value representations are encoded in specific areas of the brain and explain a range of mental functions, including choice (Levy & Glimcher, 2012), attention (Bisley & Goldberg, 2010), and cognitive control (Shenhav et al., 2016). It has also inspired a range of philosophical work (Burnston, 2025; Haas, 2023; Hendrickx, 2026; Sripada, 2025; Wu, 2023).

Like other representations, value representations get updated in response to new experiences, and RL provides the leading computational framework for modelling this learning process. The most famous example is provided by the temporal-difference (TD) learning algorithm. The intuitive idea of TD learning is that an agent updates its value function in response to whether things went better or worse than expected. This difference in expected value between prediction and outcome is called the temporal-difference error (TDE). Here is a simplified version of the algorithm, where $V(s)$ is the agent's value function representing the expected value of a state s and δ is the TDE:[2]

$$V_{\text{new}}(s) = V_{\text{old}}(s) + \delta. \tag{1}$$

To calculate the TDE, we compare the uninformed estimate of the expected value of s with the informed estimate we get once we have seen how the next state s' turns out. This informed estimate can be decomposed into the immediate reward as determined by a reward function, r, and the new estimate of future reward, $V(s')$:

$$\delta = \left[r_{s \to s'} + V(s')\right] - V(s). \tag{2}$$

[1] An RL agent can be defined as the solution to a Markov decision process, which is defined as a tuple $(\mathcal{S}, \mathcal{A}, T, r, \gamma)$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ a set of actions, T a transition function, $r\colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ a reward function, and γ a discount factor. A policy π assigns to each state a probability distribution over actions. An agent learns to approximate a true value function V that represents the actual expected cumulative discounted reward, given some policy: $V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s\right]$.

[2] Specifically, this version lacks discount and learning rates.

Consider an example to illustrate. Suppose that a mouse hears a sound (s) and then shortly thereafter receives a piece of cheese (s'). Suppose further that $r_{s \to s'} = 1$ and that $V(s) = 0$, meaning that the transition to cheese carries a positive reward and that the mouse expected no reward. In this case, s turned out to be better than expected. Assuming $V(s') = 0$, we can calculate the TDE:

$$\delta = 1 = \big[1 + 0\big] - 0. \tag{3}$$

Following the TD-learning algorithm, we can then use the TDE to update the mouse's $V(s)$:

$$V_{\text{new}}(s) = 1 = 0 + 1. \tag{4}$$

The mouse has now updated its value representation of the sound, and will be more motivated to bring it about (for example, by pulling a lever that activates the sound). This case is an illustration of simple motivational learning in natural agents through a process of conditioning. In one of the most famous experiments in neuroscience, Schultz et al. (1997) showed that dopamine activity, known to be involved in the modulation of motivation, seems to encode the TDE. In addition, the resulting behavior in both humans and other animals is what TD learning would predict. This strongly suggests that TD learning closely models real mental processes in our minds that update our value representations and causally regulate our motivation.[3]

There is a strong relation between value functions in RL and the value representations in our minds. However, RL algorithms also contain another crucial component: the reward function, which conveys how directly rewarding a given state or transition is. In artificial agents, reward can be treated as an aspect of the environment and handcrafted by an engineer. But what is rewarding for natural agents depends on what kind of agent we are—what is rewarding to me will not be to a dung beetle—meaning that something in our minds must determine this too. As it turns out, identifying what that is is difficult.

3 Theories of Reward

There is a burgeoning literature aiming to explain what the reward function corresponds to in natural agents. It has clustered around two broad strategies. One identifies the reward

[3]More recently, other RL algorithms have been fruitfully deployed to explain other aspects of human motivation, including model-based algorithms (Daw et al., 2011) and the successor representation (Momennejad et al., 2017).

function with innately determined reward signals, and the other identifies it with acquired attitudes such as desires or goals. These should be seen as extremes on a spectrum of how much the reward function depends on ontogeny, but considering them will reveal issues that most proposals will face some combination of.

Starting on the side of simple nativism, we can imagine that the reward function corresponds to simple signals generated in response to stimuli in innately determined ways. Such stimuli are called primary reinforcers by psychologists, and paradigmatic examples include high-calorie food, sexual activity, and bodily harm. For example, in the conditioning example above, we assume that the cheese constitutes a primary reinforcer for the mouse. According to this view, value representations estimate what states will cause the firing of these signals, guiding our actions towards those that are expected to generate the most activity. Many neuroscientists seem to implicitly support a view close to this (Kringelbach & Berridge, 2009; Rolls, 2013).

There are two problems with this view. The first is that it seems, intuitively, that we optimize for a richer set of outcomes than those we were selected to find innately attractive. We may learn to pursue science, love, and greatness as highly valued outcomes in their own right. Yet from the perspective of the nativist signal view, all such pursuits would be in the service of maximizing the achievement of primary reinforcers in expectation. If we could achieve those reinforcers directly, we could in principle skip everything else. This problem is familiar from issues with psychological hedonism and the implausibility that humans only strive to maximize pleasure.

The second problem is one of *wireheading*, a term for when a process optimizes for a signal rather than the state causing that signal. Strictly speaking, the view implies that value representations represent the internal activity of the reward system, rather than anything in the world. This suggests that, if we had the option to press a button to activate reward signals at the expense of all other pursuits, humans would be acting optimally by doing so (barring occasional pauses for maintenance, allowing for more button-pressing in the long run). This seems implausible. Although humans are susceptible to obsessive behavior induced by reward signals—for example, when we struggle to stop doomscrolling on our phones—these are typically seen as pathological behaviors. Crucially, even though it can be difficult, we can recognize them as such and become less motivated to pursue them by reflection and effort. It is difficult to explain how our motivational system would permit this if our minds did in fact optimize only for the signals themselves.

A view that is more popular among philosophers and cognitive scientists is that the reward function corresponds to some of our largely acquired conative attitudes, such as terminal desires or goals, which represent outcomes we want for their own sake (Juechems

& Summerfield, 2019; Molinaro & Collins, 2023b).[4] Such goals might include proving the Goldbach conjecture, ensuring that our children do well, and performing the morally right action. On this view, the computational notion of reward corresponds to the achievement of these goals, or progress towards them. This has the added benefit of establishing a strong relation between RL and more familiar representations of motivation, such as desire-belief psychology.

This view raises a natural question: how do we acquire goals, including terminal goals, in the first place? This seems like the kind of question that RL is meant to answer. However, if terminal goals constitute the analogue of the reward function, then we need to have them in place to explain how we acquire terminal goals. This seems potentially circular. One response is that we bootstrap our goals by gradually integrating learned goals into our reward function. For example, Schroeder (2004) has argued that we can acquire new terminal desires via a process of conditioning in which a neutral outcome is associated with a rewarding outcome and then gradually becomes rewarding itself. This process is the conditioning we saw modelled by TD learning before.

The problem is that this is not a description of how something becomes integrated into the reward function, and the impression that it is stems from an ambiguity in the word "reward". What TD learning shows is that states can become "rewarding" merely by acquiring expected value, because the TDE is a function of both direct and expected reward. For example, when the mouse becomes conditioned to value the sound, it can induce other behavior (for example, lead it to learn to pull a lever to activate the sound). But this is because it predicts *future* reward in the RL sense, not because it provides reward itself.

We might think that this is a simplifying artifact of TD learning, and that value representations eventually become rigidified by some other process into terminal desires that can play the role of a reward function. This is dubious. The stability of the reward function in conditioning is what allows us to explain how we both acquire and *lose* desires, in what is known as extinction. Specifically, if the sound were no longer followed by the cheese, then it would lose its value, and so would pulling the lever. If we were to "merge" the value of the sound into the reward function, we would no longer be able to explain how such extinction occurs.[5] In other words, appealing to conditioning as a source of a changing reward function requires abandoning our best explanation of the acquisition and loss of desires.

[4] I treat desires and goals as interchangeable here.

[5] Sometimes model-free learning like TD learning is identified with habit formation, which tends to be permanent. More recent work demonstrates that these are separable (Miller et al., 2019).

There are plausibly more proposals that can be provided, and several intermediate positions have been defended (Haas, 2022; Keramati & Gutkin, 2014; Weber et al., 2025).[6] The point of this discussion is to show that any such proposal that treats reward as an optimization target will need to navigate a seemingly impossible strait: it needs to make the reward function sophisticated enough to capture what humans value while avoiding wireheading, and yet stable enough to explain the empirical success of models like TD learning that rely on a stable signal. As I argue next, the key to solving this puzzle is to reject the premise: reward is not the optimization target.

4 Evidentialism about Reward

I propose a new solution to the puzzle, which I call *evidentialism about reward*. According to evidentialism, innate reward signals update our value representations, but they do not *constitute* the value that our value representations aim to estimate accurately. Instead, they provide simple pre-reflective evidence about that value, and our minds use that evidence to update value representations.

In standard RL, the reward function plays two distinct roles. First, it provides a signal for any given state that the agent can use to update the value function (for example, in TD learning). Second, that signal constitutes the optimization target for the agent: the scalar value that value functions estimate and that the agent acts to maximize in the long run. Evidentialism implies that, in humans and other animals, these two functions come apart. For clarity, let us call the optimization target *basic value* and reserve *reward* for the signals that update value representations. The assumption that reward is basic value simplifies the formal models and allows us to explain many dynamics of human motivation. However, I argue that it is a simplifying assumption that generates the tension seen in the previous section.[7]

To understand the idea of evidentialism, it is helpful to consider perception as an analogy. Our perceptual system is innately sensitive to a range of features in the environment, which generate signals that are subsequently consumed to update our internal representation of the environment. Consider depth perception: when I look at a stop sign that is ten feet away from me, light hits my retinas and is transduced into simple representations which, after being propagated through my visual system, are eventually interpreted by my mind as evidence that the stop sign is ten feet away (Marr, 1982).

[6] Aronowitz (n.d.) arguably defends a view that falls outside this spectrum.

[7] Evidentialism is similar to the thesis defended by Carruthers (2018, 2024, 2025) that valence represents value, though my focus is the relation between reward and value in particular.

Evidentialism suggests that biological reward signals function in a structurally similar way. They are triggered by aspects of the world to which we are sensitive and generate information that is innately interpreted by our minds as evidence that whatever caused the signal has basic value. The crucial observation is that, just as we do not need to take our perception of distance to *constitute* distance, we do not need to take the reward signal to constitute basic value. In both the case of distance and value representations, the content of the representation is determined by the content of the signals that inform it, rather than by the vehicles of that content.

Evidentialism resolves the issues identified above. Like the simple nativist view, it accommodates the apparent need for a stable reward signal in algorithms such as TD learning, thereby avoiding the problems that arise for goal-based views. This can be seen by analogy with perception: even major shifts in belief do not typically involve major changes in our basic perceptual capacities, and plausibly for good functional reasons. Evidentialism also avoids the wireheading and reductionist issues that afflict the simple nativist view. Since evidentialism does not take reward signals themselves to be the basic value that is optimized, there is no pressure for an agent to optimize the achievement of simple reward signals as such. In fact, doing so would amount to deliberately generating misleading evidence, since one would know that the signals are not indicative of something in the environment having basic value.

Finally, evidentialism allows that reward signals are one of many sources of evidence that bear on our value representations. The analogy to perception is again helpful. When we look at an optical illusion—for example, one where objects look large because our mind interprets them as being far away—we receive evidence that our minds innately interpret in misleading ways. However, we can also correct our representations, for example by drawing inferences from our other representations (such as a belief that objects of that kind tend to be smaller).

Similarly, we humans can draw on reflection to influence our value representations (Atlas et al., 2016; Bromberg-Martin et al., 2010; Hare et al., 2009). For example, if I see a donut, my mind might associate high expected value with eating it. However, when I read the calorie content, I infer that it would be bad for me, which—admittedly, not always—decreases my value representation and makes me want it less. By modelling basic value as an independent quantity about which reward signals provide one important source of noisy evidence, we can make sense of our ability to critically evaluate desires formed through rewarding experiences, especially when we know those reward signals were designed to induce obsessive behavior, as with social-media feeds.

This provides the basic idea of evidentialism. However, it also raises substantial questions. In the next section, I respond to what seem to me the most pressing ones.

5 Questions and Objections

An initial objection to evidentialism is that it appears to contradict the axioms of RL. It assumes that we can separate basic value from reward. But, the critic might say, in RL the optimization target just *is* reward. If so, evidentialism becomes nonsensical.

Although identifying reward with the optimization target is ubiquitous in RL, it is not a necessary component of the framework. Dissociating them requires more sophisticated problem formulations than those commonly used, however. In the Appendix, I present a model that distinguishes basic value from observations of reward. I show that the correct value function for basic value can be learned using Bayesian conditionalization, and also approximated using TD learning as in standard RL, assuming that reward signals are not systematically biased.[8]

Another question concerns what basic value is supposed to be. In the case of perception, perceptual signals carry information about a property in the world that they represent, such as distance (Neander, 2017; Shea, 2018). The analogous suggestion for value representations is that reward signals track some real property of basic value. However, this would seem to commit evidentialism to value realism, which seems implausible.[9]

There are deep questions about the content of value representations that are largely beyond the scope of this discussion. However, they will plausibly not undermine the basic descriptive picture defended here, because a mental representation need not successfully refer to anything in order to play a causal-functional role in cognition (Egan, 2025; Ramsey, 2007). For example, imagine Kurt, whose only desire in life is to maximize holiness. He travels the world visiting the biggest cathedrals and collects items thought to have belonged to prophets. It seems that we can accurately explain Kurt's motivation and predict what he will do, even if there is no property of holiness instantiated in the world.

Similarly, the content of value representations does not depend on a particular referent. We could explain why agents' motivational systems work the way they do even if value representations failed to refer to anything in the world. Instead, what makes the scalar represented by value representations *value* is its function in our mental economy: it is a quantity that causally regulates motivation in proportion to its magnitude. The same is true for the reward signals that inform these representations. What makes them provide evidence

[8] Notably, prominent papers in RL have made use of a similar distinction between reward observations and an optimization target. Everitt et al. (2017) present a model where an agent receives reward signals through a corruption function that distorts them, and they demonstrate conditions under which the agent can learn about the distribution of "true" reward. Hadfield-Menell et al. (2016) present a model in which an agent learns about and optimizes for the reward function of a principal.

[9] Carruthers (2024) faces a similar issue and identifies value with reproductive success via teleosemantics, which is an option I set aside here.

is not that they successfully inform about a property in the world—Kurt can receive evidence that something is holy, even if an error theory of holiness is correct[10]—but that the mind interprets their content as evidence for updating a value representation.[11]

This means that we can articulate the descriptive picture implied by evidentialism while remaining quiet on difficult questions about intentionality. However, we might additionally ask the normative question whether value representations and reward signals can be accurate. What we can plausibly say is that our value representations behave *as if* they were aiming to be accurate when they get updated by reward signals. As we have seen, this process is well modelled by TD learning, and TD learning is a method for converging on a true representation of the expected value of the environment (Dayan, 1992). Whether there is such a thing as an accurate value representation to converge on, however, must remain a topic for future work.

6 Conclusion

Reward signals in natural agents are best understood not as the quantity the motivational system ultimately values, but as one important source of evidence about a functional quantity of basic value estimated and optimized for by that system. This preserves the explanatory role of stable reward signals in RL models while avoiding both the reductionism of simple nativist views and the circularity that threatens goal-based views. It also helps explain why reflection and other cognitive inputs can revise what we are motivated to pursue.

Appendix: Formalizing Evidentialism

In this Appendix, I show that RL is consistent with evidentialism. I present an RL model that separates reward observations from basic value, and then show that the resulting value function can be learned by Bayesian conditionalization. I also show that TD learning converges on the same value function if reward signals are unbiased.

For comparison, consider a standard simple Markov decision process (MDP) with an infinite discounted horizon and a fixed policy π:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \gamma, \pi). \tag{A1}$$

[10] On a Bayesian understanding of evidence, p is evidence for q just if conditionalizing on p increases our subjective probability in q, given our priors.

[11] Value representations can minimally be understood as what Ramsey (2007) calls an implicit representation that can do causal work without a referent.

Here $\mathcal{S}$ and $\mathcal{A}$ are finite state and action sets, $T(s' \mid s, a)$ is the transition function specifying the probability of arriving at s' from s when choosing a, $r(s, a, s')$ is the reward, and $\gamma \in [0, 1)$ is the discount factor. Its value function is

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k}\, r(s_{t+k}, a_{t+k}, s_{t+k+1}) \;\middle|\; s_t = s\right]. \tag{A2}$$

In this MDP, the same scalar r is both what is optimized and what teaches the agent, for example via TD learning.

Compare this to another MDP. Let θ be a latent parameter determining the basic value of a transition:

$$u_{\theta}(s, a, s'). \tag{A3}$$

The correct value function, relative to the true latent parameter θ^{*}, is

$$V^{\pi}_{\theta^{*}}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k}\, u_{\theta^{*}}(s_{t+k}, a_{t+k}, s_{t+k+1}) \;\middle|\; s_t = s\right]. \tag{A4}$$

$V^{\pi}_{\theta^{*}}$ is the ideal value function the agent is trying to track. However, the agent does not observe θ^{*} directly, and it might believe that some other θ is true. A Bayesian learner can update this belief in response to evidence. Suppose that after the transition (s_t, a_t, s_{t+1}), it receives an observation o_{t+1} generated by an observation model

$$P(o_{t+1} \mid s_t, a_t, s_{t+1}, \theta). \tag{A5}$$

Let

$$h_t = (s_0, a_0, o_1, s_1, \ldots, a_{t-1}, o_t, s_t) \tag{A6}$$

be the history up to time t, and let

$$b_t(\theta) = P(\theta \mid h_t) \tag{A7}$$

be the agent's posterior probabilistic belief over basic-value hypotheses. Bayes' rule then gives

$$b_{t+1}(\theta) = \frac{b_t(\theta)\, P(o_{t+1} \mid s_t, a_t, s_{t+1}, \theta)}{\sum_{\widetilde{\theta}} b_t(\widetilde{\theta})\, P(o_{t+1} \mid s_t, a_t, s_{t+1}, \widetilde{\theta})}. \tag{A8}$$

Given this, we can define the agent's estimate of the value function as

$$\widehat{V}_t^{\pi}(s) = \sum_{\theta} b_t(\theta)\, V_{\theta}^{\pi}(s), \tag{A9}$$

where

$$V_{\theta}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k\, u_{\theta}(s_{t+k}, a_{t+k}, s_{t+k+1}) \;\middle|\; s_t = s\right]. \tag{A10}$$

If the model is well specified and the evidence is not systematically misleading, then accumulating observations will make the posterior concentrate on θ^*, in which case $\widehat{V}_t^{\pi}$ approaches $V_{\theta^*}^{\pi}$.

Can TD learning be used to learn this as well? Yes. Suppose that the scalar signal fed into TD learning is this same observation variable o_{t+1}, now treated as the input to the update rule. A simple sufficient condition is that this observational signal be conditionally unbiased for one-step basic value:

$$\mathbb{E}\left[o_{t+1} \mid s_t, a_t, s_{t+1}, \theta^*\right] = u_{\theta^*}(s_t, a_t, s_{t+1}). \tag{A11}$$

In that case, the observation signal is noisy evidence about basic value. Let $\widehat{V}_t$ be the agent's current value estimate at time t, and let α_t be the learning rate. The TDE is

$$\delta_t = o_{t+1} + \gamma \widehat{V}_t(s_{t+1}) - \widehat{V}_t(s_t), \tag{A12}$$

and the TD update is

$$\widehat{V}_{t+1}(s_t) = \widehat{V}_t(s_t) + \alpha_t\, \delta_t. \tag{A13}$$

If the agent observed basic value directly, the corresponding one-step TD target would be

$$u_{\theta^*}(s_t, a_t, s_{t+1}) + \gamma \widehat{V}_t(s_{t+1}). \tag{A14}$$

Since $\widehat{V}_t(s_t)$ and $\widehat{V}_t(s_{t+1})$ are fixed once we condition on $s_t, a_t, s_{t+1}, \theta^*$, the conditional expectation applies only to o_{t+1}. So

$$\mathbb{E}\left[\delta_t \mid s_t, a_t, s_{t+1}, \theta^*\right] = \mathbb{E}\left[o_{t+1} \mid s_t, a_t, s_{t+1}, \theta^*\right] + \gamma \widehat{V}_t(s_{t+1}) - \widehat{V}_t(s_t). \tag{A15}$$

So by (A11),

$$\mathbb{E}\left[\delta_t \mid s_t, a_t, s_{t+1}, \theta^*\right] = u_{\theta^*}(s_t, a_t, s_{t+1}) + \gamma \widehat{V}_t(s_{t+1}) - \widehat{V}_t(s_t). \tag{A16}$$

Thus TD learning can treat reward as noisy evidence about basic value, and under the familiar tabular assumptions it will still converge to $V^{\pi}_{\theta^*}$, given the unbiasedness assumption above (Dayan, 1992). By contrast, if o_{t+1} is systematically misleading, then we should not expect convergence (for example, if θ^* is healthy eating and o_{t+1} is sugar).

References

Aronowitz, S. (n.d.). Locating Values in the Space of Possibilities. *Philosophy of Science*.

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016). Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala (T. E. Behrens, Ed.). *eLife*, *5*, e15192. https://doi.org/10.7554/eLife.15192

Ballesta, S., Shi, W., Conen, K. E., & Padoa-Schioppa, C. (2020). Values encoded in orbitofrontal cortex are causally related to economic choices. *Nature*, *588*(7838), 450–453. https://doi.org/10.1038/s41586-020-2880-x

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21. https://doi.org/10.1146/annurev-neuro-060909-152823

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, *104*(2), 1068–1076. https://doi.org/10.1152/jn.00158.2010

Burnston, D. C. (2025). Epistemic reduction of the concept of 'decision'. *Synthese*, *205*(2), 74. https://doi.org/10.1007/s11229-025-04917-8

Carruthers, P. (2018). Valence and Value. *Philosophy and Phenomenological Research*, *97*(3), 658–680. https://doi.org/10.1111/phpr.12395

Carruthers, P. (2024, January). *Human Motives: Hedonism, Altruism, and the Science of Affect*. Oxford University Press.

Carruthers, P. (2025). *Explaining our Actions: A Critique of Common-Sense Theorizing*. Cambridge University Press. https://doi.org/10.1017/9781009585781

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Daw, N. D., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning [Place: US]. *Social Cognition*, *26*(5), 593–620. https://doi.org/10.1521/soco.2008.26.5.593

Dayan, P. (1992). The convergence of TD(
lambda) for general
lambda. *Machine Learning*, *8*(3), 341–362. https://doi.org/10.1023/A:1022632907294

Egan, F. (2025, March). *Deflating Mental Representation* (F. Recanati, Ed.). MIT Press.

Everitt, T., Krakovna, V., Orseau, L., Hutter, M., & Legg, S. (2017, August). Reinforcement Learning with a Corrupted Reward Channel [arXiv:1705.08417 [cs]]. https://doi.org/10.48550/arXiv.1705.08417

Haas, J. (2022). Reinforcement Learning: A Brief Guide for Philosophers of Mind. *Philosophy Compass*, *17*(9), e12865. https://doi.org/10.1111/phc3.12865

Haas, J. (2023). The Evaluative Mind. In J. Haugeland, C. Craver, & C. Klein (Eds.), *Mind Design III: Philosophy, Psychology, and Artifical Intelligence* (pp. 295–314). MIT Press.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative Inverse Reinforcement Learning [arXiv:1606.03137 [cs]]. https://doi.org/10.48550/arXiv.1606.03137

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-Control in Decision-Making Involves Modulation of the vmPFC Valuation System. *Science*, *324*(5927), 646–648. https://doi.org/10.1126/science.1168450

Hendrickx, M. (2026). Difficulty. *Mind*, fzaf080. https://doi.org/10.1093/mind/fzaf080

Juechems, K., & Summerfield, C. (2019). Where Does Value Come From? *Trends in Cognitive Sciences*, *23*(10), 836–850. https://doi.org/10.1016/j.tics.2019.07.012

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability (E. Marder, Ed.). *eLife*, *3*, e04811. https://doi.org/10.7554/eLife.04811

Kringelbach, M. L., & Berridge, K. C. (Eds.). (2009). *Pleasures of the brain*. Oxford University Press.

Levy, D. J., & Glimcher, P. W. (2011). Comparing Apples and Oranges: Using Reward-Specific and Reward-General Subjective Value Representation in the Brain. *Journal of Neuroscience*, *31*(41), 14693–14707. https://doi.org/10.1523/JNEUROSCI.2218-11.2011

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Marr, D. (1982, January). *Vision* (3rd edition). W.H. Freeman.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*(2), 292–311. https://doi.org/10.1037/rev0000120

Molinaro, G., & Collins, A. G. E. (2023a). Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLoS biology*, *21*(7), e3002201. https://doi.org/10.1371/journal.pbio.3002201

Molinaro, G., & Collins, A. G. E. (2023b). A goal-centric outlook on learning. *Trends in Cognitive Sciences*, *27*(12), 1150–1164. https://doi.org/10.1016/j.tics.2023.08.011

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680–692. https://doi.org/10.1038/s41562-017-0180-8

Neander, K. (2017). *A Mark of the Mental: A Defence of Informational Teleosemantics*. MIT Press.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making [_eprint: https://doi.org/10.1146/annurev-psych-010416-044216]. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value [Number: 7090]. *Nature*, *441*(7090), 223–226. https://doi.org/10.1038/nature04676

Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge University Press.

Rolls, E. T. (2013, December). *Emotion and decision making explained*. Oxford University Press.

Schroeder, T. (2004). *Three Faces of Desire*. Oxford University Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control [Number: 10]. *Nature Neuroscience*, *19*(10), 1286–1291. https://doi.org/10.1038/nn.4384

Singh, S., Lewis, R., & Barto, A. (2009). Where do rewards come from? [Paper presented at the 31st Annual Meeting of the Cognitive Science Society]. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Sripada, C. (2025). The valuationist model of human agent architecture [_eprint: https://doi.org/10.1080/09515089 *Philosophical Psychology*, *0*(0), 1–30. https://doi.org/10.1080/09515089.2025.2485323

Sutton, R. S., & Barto, A. G. (2018, November). *Reinforcement Learning: An Introduction* (F. Bach, Ed.; 2nd ed.). Bradford Books.

Weber, L. A., Yee, D. M., Small, D. M., & Petzschner, F. H. (2025). The interoceptive origin of reinforcement learning. *Trends in Cognitive Sciences*, *29*(9), 840–854. https://doi.org/10.1016/j.tics.2025.05.008

Wu, W. (2023). *Movements of the Mind: A Theory of Attention, Intention and Action*. Oxford University Press.