Evidentialism about Reward

Paul de Font-Reaulx
University of Michigan, Ann Arbor · Sophron Research
First Philosophy and Reinforcement Learning Symposium · UMass Amherst · May 1, 2026

The question

Reinforcement learning has emerged as a leading framework for modelling natural agency. Its central component is a reward function that provides immediate payoff.

> *What—if anything—does the reward function correspond to in us?*

Singh, Lewis & Barto 2009 · Juechems & Summerfield 2019 · Haas 2022 · Aronowitz n.d. · Sripada 2025

The puzzle of reward

Two opposing demands on what reward must be

FUNCTIONAL DEMAND

Provide update signals

What plays the role of reward in us should be simple enough to correspond to innate reward signals.

NORMATIVE DEMAND

Be an optimization target

It should also be something that provides *basic value* for an agent: what it tries to achieve in expectation.

Can anything capture both in us? This is the *puzzle of reward*.

PLAN

Thesis & plan

THESIS: EVIDENTIALISM

Separation of roles: Reward signals update by providing *evidence* of a functional quantity of *basic value* that minds represent as their optimization target.

1. **Background** *RL, value representations, TD learning*
2. **Main proposals** *Nativism and the goal-based view — and their problems*
3. **Evidentialism** *Reward signals as evidence about a latent quantity*
4. **Objections** *Is this still RL? What is basic value?*

(Note: 'Basic Value' ≠ Value in RL. It's what the value function is the long-term expectation of. Alternative: "True reward")

Reinforcement learning and motivation

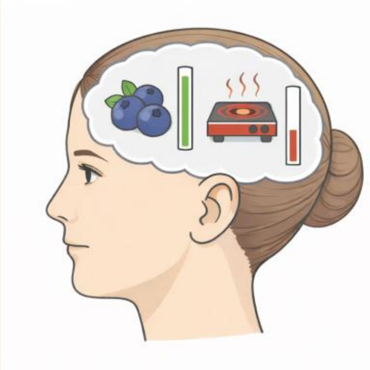
What is RL?

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*.

What's the connection to us?

Convergent evidence for ***valuationism*** *(Sripada 2025)*: minds carry simple non-conceptual **value representations** that index expected value and steer behavior.



Daw & Shohamy 2008 · Levy & Glimcher 2011 · Padoa-Schioppa & Assad 2006 · Sripada 2025

Temporal-difference learning

Value representations get updates in ways modelled by RL learning algorithms – e.g. TD learning

THE ALGORITHM

$$\boldsymbol{\delta} = [r + V(s')] - V(s)$$

$$V_{new}(s) = V_{old}(s) + \boldsymbol{\delta}$$

TD error δ encodes the difference between expectation and discovery.

Example

A mouse hears a sound (s), then receives cheese (s′). Cheese is rewarding: r = 1. δ = 1, so V(s) is updated upward. The sound now predicts value.

Schultz, Dayan & Montague 1997 — dopamine activity encodes the TD error.

What does reward correspond to?

Two broad strategies

Nativism

Innate, simple

Reward = signals from primary reinforcers (food, sex, harm).

Goal-based view

Acquired, sophisticated

Reward = achievement of acquired terminal goals (truth, fairness, the well-being of one's children).

Strategy 1 · Nativism

Reward = innately determined signals from primary reinforcers

Primary reinforcers are stimuli that innately trigger reward signals — paradigmatically: high-calorie food, sexual activity, bodily harm.

Best fit for the role of reward in standard TD learning: a stable scalar that updates value representations in response to fixed, identifiable stimuli.

Secondary reinforcers (cheese-predicting sound, $20 bill) acquire *value* through expected reward.

Implicit in much neuroscience — e.g., Kringelbach & Berridge 2009; Rolls 2013.

Problems with nativism

REDUCTIONISM

All motivation as expected reward signal

Pursuits like science, love, and greatness become merely instrumental to maximizing expected primary-reinforcer activation. Familiar objection from psychological hedonism.

WIREHEADING

Optimizing the signal, not its cause

If reward is the signal itself, the optimal life is pressing a button to maximize it. But we treat such behavior (e.g., compulsive doomscrolling) as pathological — and can resist it on reflection.

Strategy 2 · The goal-based view

Reward = the satisfaction of acquired terminal desires or goals

On this view, the reward function is constituted by what we want for its own sake — sometimes including sophisticated, learned goals.

> **Examples:** *proving the Goldbach conjecture · one's children flourishing · doing the morally right thing.*

Establishes a strong tie between RL and folk-psychological desire-belief explanation.

Juechems & Summerfield 2019 · Molinaro & Collins 2023b

Problems with the goal-based view

REGRESS

Where do the goals come from?

Acquiring new terminal goals is naturally explained as a learning process. But this requires an optimization target. That's what a terminal goal is supposed to do!

FUNCTIONAL MISMATCH

Sophisticated vs. simple

Reflective goals (e.g. proving the Goldbach conjecture) and simple primary reinforcers (calories) play very different roles.

Evidentialism about reward

The two roles come apart in natural agents–assumed that reward does both in standard RL

Biological reward signals provide pre-reflective ***evidence*** about a latent quantity of ***basic value*** that the mind represents and optimizes for.

Reward signals — innately generated, simple, scalar; the input that updates value representations (cf. TD learning).

Basic value — the latent functional quantity our minds aim to estimate accurately. The optimization target, not the signal.

An analogy with perception

Reward signals do for value what perceptual signals do for distance

PERCEPTION

object at distance → retinal signal → depth representation

REWARD

stimulus / reinforcer → reward signal → value representation

In both cases, the signal updates the representation but is not what the representation is about.

How evidentialism resolves the puzzle

Stable target

The optimization target — basic value — never changes itself; value representations shift in light of new evidence.

No wireheading

Manufactured r is misleading evidence — like an optical illusion we can screen off once we know its source.

Top-down reflection

Inference can revise value representations directly. (The donut: I want it; I read “600 calories”; I want it less.)

Objection 1 · Is this still RL?

Yes. *Standard MDPs are a special case of one where the optimization target is determined by a latent parameter — and reward signals are observations.*

Value in Standard MDP

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

r teaches and is the target.

Value in Evidentialist MDP

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

Latent θ fixes basic value u_θ; the target is V^*_θ.

Cf. Hadfield-Menell et al. 2016 (CIRL); Everitt et al. 2017 (corrupted reward channels).

Objection 2 · What is basic value?

"Aren't you committed to value realism?"

No. Basic value is a functional quantity defined by its role in cognition: a scalar that causally regulates motivation in proportion to its magnitude.

> **Imagine Kurt**, who optimizes for *holiness*. We can fully explain his psychology even if there is no property of holiness in the world.

Whether anything in the world makes value representations accurate is a further question

Three upshots

01 The puzzle of reward dissolves

Stop trying to identify the signal with the target. Read reward signals as informing about basic value, not constituting it.

02 A different picture of agency

Nothing has terminal or instrumental value full stop. Every motivational state aims at the same thing: tracking expected basic value.

03 Implications for AI safety

A rational evidentialist agent doesn't wirehead — that would be deliberately generating misleading evidence about its own optimization target.

Thank you.

The evidentialist MDP

Setup. *Let θ be a latent parameter that fixes basic value $u\theta(s, a, s')$. The agent does not observe $\theta*$ directly, but receives observations through $P(o_{t+1} \mid s_t, a_t, s_{t+1}, \theta)$.*

Standard MDP

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

$M = \langle S, A, T, r, \gamma, \pi \rangle$. Same scalar r teaches and is the target.

Evidentialist MDP

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

Latent θ fixes basic value u_θ; the target is V^*_θ.

Cf. Hadfield-Menell et al. 2016 (CIRL); Everitt et al. 2017 (corrupted reward channels).

Bayesian conditionalization

Learning θ. *After each transition, the agent receives an observation o_{t+1} and updates its posterior belief $b_t(\theta) = P(\theta \mid h_t)$ over basic-value hypotheses.*

Posterior update

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

Bayes' rule for the new evidence o_{t+1}.

Value estimate

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

Expectation over basic-value hypotheses.

Given a well-specified model and evidence that is not systematically misleading, the posterior concentrates on $\theta*$ and the value estimate converges to $V^*\theta*$.

TD learning still converges

No Bayes required. *Treat the observation o_{t+1} as the scalar fed to the standard TD update rule. One sufficient condition delivers convergence.*

Sufficient condition

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

Signal is conditionally unbiased for one-step basic value.

TD update with this signal

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

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

Same shape as standard TD; the signal is now noisy evidence.

Under tabular assumptions, $\hat{V} \to V^*\theta*$ (Dayan 1992). Failure: if the signal is systematically misleading (e.g., $\theta*$ = healthy eating, signal = sugar), no convergence.