HHS Public Access

Author manuscript

Curr Opin Behav Sci. Author manuscript; available in PMC 2022 November 04.

Published in final edited form as:

Curr Opin Behav Sci. 2021 October ; 41: 114–121. doi:10.1016/j.cobeha.2021.04.020.

Value-free reinforcement learning: policy optimization as a minimal model of operant behavior

Daniel Bennett[a,b], **Yael Niv**[a,c], **Angela J. Langdon**[a]

[a]Princeton Neuroscience Institute, Princeton University, USA

[b]Department of Psychiatry, Monash University, Australia

[c]Department of Psychology, Princeton University, USA

Abstract

Reinforcement learning is a powerful framework for modelling the cognitive and neural substrates of learning and decision making. Contemporary research in cognitive neuroscience and neuroeconomics typically uses value-based reinforcement-learning models, which assume that decision-makers choose by comparing learned values for different actions. However, another possibility is suggested by a simpler family of models, called *policy-gradient reinforcement learning*. Policy-gradient models learn by optimizing a behavioral policy directly, without the intermediate step of value-learning. Here we review recent behavioral and neural findings that are more parsimoniously explained by policy-gradient models than by value-based models. We conclude that, despite the ubiquity of 'value' in reinforcement-learning models of decision making, policy-gradient models provide a lightweight and compelling alternative model of operant behavior.

Keywords

value; decision-making; reinforcement learning; policy gradient; computational modelling

1. Introduction

What is value? In spite of the ubiquity of this term in the field of value-based decision making, there are a number of different ways of defining value, and one's chosen definition has important implications for the kinds of inferences that one is likely to draw about the cognitive and neural processes that subserve operant behavior [1, 2, 3, 4, 5].

In typical usage, 'value' is an explanatory variable that quantifies the degree to which an individual prefers (or is willing to work for) one good or outcome over others [6, 7]. So, for instance, a child who consistently chooses a chocolate bar over fresh fruit would be said to value chocolate more highly. Value—as typically defined—is therefore a latent construct that scaffolds choice behavior by providing a common currency for comparison of different actions (often loosely identified with the similar economic concept of *utility* [8, 9]). In support of the hypothesis that choice behavior is supported by the computation of value,

proponents frequently note (e.g., [10]) that reward-related dopaminergic neural activity is consistent with a class of value-learning algorithms from the field of *reinforcement learning* (RL) [11, 12]. This argument implicitly assumes that different theoretical frameworks each use the term 'value' in the same way. However, there are important differences between typical usage of 'value' and value as defined in RL.

In RL, value is defined as *expected cumulative future reward* [13] (see Box 1). As such, an action's value in RL is not merely a relative quantification of preference, as in the typical usage of 'value'; instead, it exactly quantifies the total amount of future reward[1] that an agent expects to receive if it takes that action. At first glance, this definition of value appears consonant with typical usage, since it seems reasonable that one should prefer actions with greater expected cumulative future reward. Crucially, however, although in RL differences in learned value between two actions do indeed entail differences in choice behavior, the converse is not true. That is, we may observe differences in an individual's preference for different actions despite the agent not having learned different values (in the RL sense) for those actions. This is because there is an entire separate branch of RL algorithms—*policy-gradient* RL—in which agents learn to choose actions by optimizing a behavioral policy directly, without ever taking the intermediate step of learning their values [14].

Here we give an introduction to the distinction between value-based and policy-gradient RL (also termed 'indirect' and 'direct' actors, respectively [15]), and review recent work suggesting that policy-gradient RL provides a good account of neural and behavioral data. We suggest that value (in the RL sense) is not strictly entailed by data from typical laboratory tasks, and propose that value need only be invoked as an explanatory latent construct for phenomena that cannot be accounted for under simpler algorithms like policy-gradient RL. This does not exclude the possibility that the brain does compute value in some circumstances; however, we would argue that policy-gradient RL models are generally favoured by the principle of parsimony, and that 'value' should only be added to these models to explain data that cannot be explained by policy-gradient models alone.

2. Value-based RL versus policy-gradient RL

Value-based RL is a prominent computational model of the cognitive processes subserving simple operant behavior and of the neural substrates of these processes (e.g., [16, 17, 18, 19]). Briefly, in a prototypical value-based RL model (Figure 1A), the agent learns the value of each of a set of discrete actions by trial-and-error, and chooses between actions by mapping these estimated values into a behavioral *policy* using a policy mapping function (see Box 1). As such, using a value-based RL model as a model of a subject's behavior implicitly assumes that, at an algorithmic level of description, the subject learns action-values and makes choices by comparing them.

According to the principle of parsimony, however, in modelling subjects' behavior we should attempt to find the model that best explains the data while minimizing model complexity (i.e., invoking as few latent explanatory constructs as possible) [20]. This raises

[1]After accounting for the temporal discounting of future rewards.

the question: do value-based models provide the simplest RL account of behavior, or can simpler models (such as policy-gradient RL; Figure 1B) account for behavior equally well?

In a policy-gradient algorithm, the agent chooses actions according to a parameterized policy, observes the outcomes of these actions, and then adjusts the parameters of its policy so as to increase the probability of actions associated with more reward and decrease the probability of actions associated with less reward (i.e., it adjusts the parameters of its policy according to the gradient[2] of reward with respect to the parameters of its policy). Rather than the indirect algorithm of value-based RL (learning the values of different actions, and acting by mapping these values into a policy), policy-gradient RL is a conceptually simpler algorithm that directly adjusts the policy without troubling with the intermediate latent construct of value. Consequently, policy-gradient RL is also simpler than value-based RL in implementation as a cognitive model: in its simplest form, policy-gradient RL requires one adjustable parameter per participant (a learning rate for updates to the policy parameters), compared to two adjustable parameters for the simplest form of value-based RL (a learning rate for value updates, plus a policy mapping parameter such as the softmax inverse temperature β)[3].

In the simplest choice setting, where a subject repeatedly chooses between a fixed set of actions in a single environmental state, policy-gradient RL algorithms can, in principle, explain behavior as well as value-based RL algorithms. For instance, the gradient-bandit algorithm described by Sutton and Barto [13] (based on the REINFORCE algorithm of Williams [23]) uses a softmax policy parameterized by a vector of *preferences* for different actions. These preferences are reminiscent of values, but unlike values they are not interpretable as expected discounted future reward; instead, as policy parameters, they directly determine choice probabilities. For instance, consider a state with three actions associated with reward magnitudes of 1 unit, 9 units, and 10 units, respectively. A gradient-bandit agent that learns an optimal policy will show a low preference for the second-best (9-unit) option relative to the best (10-unit) option, in spite of the fact that the *values* of these two actions are relatively similar. As such, value-based and policy-gradient RL have different representations of actions in the environment: whereas value-based algorithms maintain a representation the expected future reward of different actions, policy-gradient algorithms can be thought of as representing actions in terms of which should be taken and which should be avoided. This means that policy-gradient algorithms do not represent the reward amounts associated with different actions. In general, however, this sparser representation does not prevent a policy-gradient RL agent from learning to behave in a manner that maximises its expected future reward.

[2]There exist other policy-optimization algorithms that update the policy without using a gradient, such as trust-region policy optimization [21], but these are beyond the scope of this article.

[3]In fitting value-based RL models with a softmax choice rule, these two parameters are frequently strongly anticorrelated, leading to difficulty in parameter identifiability [22]. We argue that this non-identifiability is a consequence of a deeper conceptual issue: to explain choices using a value-based algorithm, we must posit, in addition to value-learning, an additional cognitive operation by which learned values are mapped into a behavioral policy. However, for any given choice between different actions, there are infinitely many combinations of underlying action-values and policy-mapping functions that will produce identical choice probabilities. This many-to-one correspondence is the root cause of parameter non-identifiability in value-based RL models, and is avoided in policy-gradient RL.

In more complex choice settings, the policy optimized by the RL agent need not be parameterized by preferences. In environments with a continuous one-dimensional action space, for instance, actions might be selected according to a Gaussian policy (e.g., [24]); in such a case, the parameters of the agent's policy would control the mean and variance of the probability distribution over actions that is produced by the policy.

3. Behavioral and neural evidence for policy-gradient RL

The greater parsimony of policy-gradient RL is one reason to prefer it over value-based RL as a model of simple operant behavior. A second reason is that, in some decision-making tasks, policy-gradient methods provide a good account of behavioral and neural data that are more difficult to explain with value-based RL models.

3.1. Behavioral evidence for policy-gradient RL

Two behavioral phenomena better explained by policy-gradient RL than value-based RL are context-dependent preference learning and continuous-action learning.

Context-dependent preference learning.—When given a choice between two options with equivalent reinforcement histories, but where one option was learned in a rich environmental context (high reward availability) and the other was learned in a lean context (low reward availability), humans and other animals display a marked preference for the lean-context option [25, 26, 27, 28]. That is, animals' learned preferences between different options are a function not only of each option's reinforcement history, but also of the environmental context within which each option was experienced. Since action-values in value-based RL are defined cardinally as expected cumulative future reward conditional on taking an action (i.e., action-values are not contextually modulated by the value of their associated environmental state), this phenomenon is difficult to account for using value-based RL models absent additional post-hoc assumptions regarding the reward function (e.g., [28]).

By contrast, context-dependent preference learning emerges straightforwardly from the principles of policy-gradient RL. In a gradient-bandit algorithm, for instance, because the goal of the agent is to update action-preferences to optimize its policy (rather than to learn action-values), learned preferences for each action are solely a function of whether taking an action within its environmental state will lead to maximization of future reward. As such, a gradient-bandit agent will learn positive preferences for actions that are the best available in a state (i.e., actions that should be taken when possible), and learn negative preferences for actions that are the worst available in a state (i.e., actions that should be avoided). If, after learning, the agent is given a choice between the worst action from a rich state versus the best option from a lean state, it will tend to prefer the lean-state action (because of its positive learned preference as the best action in its state) over the rich-state action (because of its negative learned preference as the worst action in its state), even if the reinforcement histories of the two actions are identical.

Continuous-action learning.—In natural environments, behavior frequently involves selecting actions from a *continuous* action space (e.g., controlling a mouse cursor, or driving

a car; see [29]). Standard value-based RL models like *Q*-learning typically perform poorly in such environments, because they operate over a tabular representation of actions in which all actions are equally (dis)similar to one another. This tabular representation leads to inefficient learning in continuous action spaces, since the algorithm cannot generalize between similar actions (e.g., steering a car 10 degrees left is similar to steering 11 degrees left, but dissimilar from steering 40 degrees right). Although there exist continuous extensions of value-based RL involving value-function approximation (e.g., [30]), these approaches have historically focused on continuous state spaces rather than continuous action spaces. Applying value-function approximation to continuous action spaces is not straightforward, because actions need not only be evaluated, but also sampled according to their estimated value. This implies estimation of the whole function (we need evaluate only the current state, but to select an action we need to evaluate all possible actions), which, even for simple choice rules such as those considered here, is computationally intractable for most value-function approximators [31].

By contrast, since policy-gradient methods optimise an overall policy, rather than learning the values of a number of distinct actions, policy-gradient RL algorithms are straightforwardly applicable to continuous action spaces, providing that the functional form of the policy can be sampled from (e.g., a two-parameter gamma distribution for reaction time choices; [32]). In learning a continuous action space, a policy-gradient RL agent will respond to the reinforcement of an action by adjusting the policy such as to increase the choice probability not only of the chosen action, but also of similar actions. This leads to efficient generalization of learning across continuous action spaces.

In line with this proposal, recent research in the field of motor control has suggested that online recalibration processes for motor execution are well-explained by direct updating of an implicit behavioral policy [33, 34]. For instance, a recent study by Hadjiosif et al. [35] used mirror-reversed visual feedback in a sensorimotor adaptation task, and found that subjects' patterns of post-reversal errors were well explained by a model in which an implicit behavioral policy was updated according to a sensory prediction error. As such, policy-gradient RL also represents a point of contact between models of simple operant behavior and models of higher-order motor control. Indeed, even apparently simple operant behaviours such as discrete choice can also be conceptualized as involving continuous actions if we consider that the latency and vigour of these choices constitute a continuously-distributed action space [32].

3.2. Neural evidence for policy-gradient RL

Sensory and action components of midbrain dopamine responses.—Phasic bursts from midbrian dopamine (DA) neurons have been proposed as a neural correlate of the reward prediction error (RPE) signal central to RL theories [36, 11, 37, 38]. This hypothesis regarding the role of dopamine stands for policy-based RL as well. Indeed, the difference between value- and policy-based RL frameworks lies not in the dependence on an RPE, but in what kind of representation is updated by this teaching signal. Decades of evidence in classical conditioning tasks have demonstrated phasic DA responses to sensory stimuli that predict the delivery of a reward, consistent with the learning of

a‘subjective value’ for the predictive stimulus. However, in instrumental settings, DA activity (both phasic and tonic) has also been associated with movement signaling and control: phasic responses in dorsal-striatum-projecting DA neurons can trigger locomotion [39], and movement initiation and response vigor are bidirectionally modulated by transient activity in DA neurons in the substantia nigra pars compacta [40].

Action dependence of DA activity is not restricted to the substantia nigra; ventral tegmental area DA responses are attenuated unless the correct movement is initiated in an operant go/no-go task [41]. More recently, close inspection of DA responses during very early learning in a classical conditioning task show that the timing of the initiation of licking accounts for phasic DA response to rewards at this stage, before more stereotypical sensory components of the RPE signal have emerged [42]. These results suggest that representations of action are, at least in these settings, central to the formation and update of reward predictions. That is, neural reward predictions in dopaminergic circuits appear to be deeply intertwined with neural representations of actions themselves. This stands in contrast with simple value-based RL models, where the learned values of actions are represented separately and only transformed into a policy as needed at the time of choice (top panel, Figure 1).

Neural signatures of policy learning.—The basal ganglia are a central locus in the brain for the initiation and control of movement [43]. Accordingly, much of the evidence for policy learning in the brain hinges on the interpretation of neural activity in the striatum during various instrumental tasks. Representations associated with action-values have been reported in striatal regions [19, 44, 45]; however, unambiguous identification of these activity patterns as value signals rather than action preferences or other related constructs is complicated by the presence of temporal correlations in neural activity recorded from different brain regions [46].

Distinguishing value representations from other arbiters of preference, such as policy, suffers also from the confound that these quantities are correlated in almost all learning models. In fact, specific considerations in task design are necessary to establish the condition in which predictions from value- and policy-based accounts of learning can be disambiguated: policy-based methods predict updates to all (within-context) action probabilities subsequent to the outcome of a single action, leading to a fundamentally relative representation of action preference. Indeed, in a task in which information about the outcome of forgone as well as chosen options is provided, the BOLD signal in the striatum was consistent with a policy-update signal, rather than an action-value update [47]. Neural correlates of context-dependent effects in choice implicate the medial prefrontal cortex and ventral striatum in the relative update of preferences [28], while dopaminergic error-signals related to counterfactual outcomes have been identified in human striatum [48]. These findings speak to interactions between chosen and non-chosen options that are the hallmark of policy-based RL, and suggest such learning recruits the same neural regions and mechanisms that have previously been closely identified with value-based learning in the brain.

4. The intersection of value and policy

In the preceding sections, we have drawn a sharp contrast between value-based and policy-gradient RL. However, this is a false dichotomy, since an entire class of algorithms—*actor-critic* RL—marries value-based and policy-gradient RL. Although we have not discussed these algorithms here, they form the basis of much state-of-the-art work in computer science applications of reinforcement learning [49, 50, 51], alongside other methods that incorporate components of both value-based and policy-gradient RL (e.g., [52, 53]). In the domain of behavioral science, actor-critic RL has shown great promise as a model of diverse phenomena—including conditioned avoidance, matching behavior in response to variable-interval reinforcement, and mood [54, 55, 56]—that are not readily accounted for by either value-based or policy-gradient models alone. In neuroscience, similarly, it has been suggested that dopaminergic neural activity in the basal ganglia is accounted for well by actor-critic models [57, 58]. As such, actor-critic RL represents a potentially fruitful avenue for developing a general model of operant behavior that may resolve the tensions between value-based and policy-gradient RL that we have reviewed here.

More broadly, our advocating for policy-gradient RL as a minimal model of learning in operant settings does not preclude the possibility that more nuanced representations of outcomes—and even value as defined in RL—are recruited in various behavioral tasks. To be clear, there exist a number of phenomena that appear better explained by value-based and model-based RL models than by policy-gradient RL alone. For instance, reward expectations are clearly formed and exploited in classical conditioning [59, 60, 61], outcome devaluation procedures, suggest that a specific expectation of future outcomes is formed and used to guide choice in an adaptive manner [62], and sensory preconditioning studies show that conditioning can result in the formation of associations between different stimuli, and not only between stimuli and responses or stimuli and rewards [63]. Indeed, theories of model-based learning and decision making rely on specifying the precise interaction between states, values and policies to account for flexible behavior in uncertain and dynamic environments [64, 65]. Once again, our argument is not that the brain only ever uses policy-gradient RL; rather, we suggest that, in modelling data, policy-gradient models should typically be favoured in the interests of parsimony, and that the latent construct of value should only be invoked to explain phenomena that are not explicable under this simpler model.

Finally, part of the disjunction between value-based and policy-based RL theories of behavior derives from the historical delineation between operant and classical conditioning paradigms, with the former assumed to pertain to policies for action, and the latter construed as a selective window on value learning. In general, however, the construct of policy is not restricted to discrete choice paradigms; indeed, any engagement in a situation that involves motivationally relevant outcomes will require some targeted regulation of movement, which implies the existence of a policy. More expansive consideration of policies that include continuous action spaces [29, 66], the withholding as well as execution of movements [67], and the inclusion of ‘internal’ actions such as the control of attentional focus [68, 69], naturally complicate the boundary between operant and classical conditioning, merging perspectives from policy- and value-learning into a more integrated whole.

References

[1]. O'Doherty JP, The problem with value, Neuroscience & Biobehavioral Reviews 43 (2014) 259–268. [PubMed: 24726573]

[2]. Miller KJ, Shenhav A, Ludvig EA, Habits without values, Psychological Review 126 (2019) 292–311. [PubMed: 30676040] * This paper presents evidence that habit formation—a process previously linked with model-free learning of action-values—may be instead produced by a value-free process in which choosing an action directly strengthens its future choice probability. Although this is not strictly speaking a policy-gradient model, it nevertheless points to the feasibility of explaining operant behavior in terms of modulation of a policy, without recourse to the explanatory construct of value.

[3]. Juechems K, Summerfield C, Where does value come from?, Trends in Cognitive Sciences 23 (2019) 836–850. [PubMed: 31494042]

[4]. Suri G, Gross JJ, McClelland JL, Value-based decision making: An interactive activation perspective, Psychological Review 127 (2020) 153. [PubMed: 31524426]

[5]. Hayden B, Niv Y, The case against economic values in the brain, 2020. Preprint hosted at PsyArXiv.

[6]. Rolls ET, Emotion Explained, Oxford University Press, 2005.

[7]. Rangel A, Camerer C, Montague PR, A framework for studying the neurobiology of value-based decision making, Nature Reviews Neuroscience 9 (2008) 545–556. [PubMed: 18545266]

[8]. Platt ML, Glimcher PW, Neural correlates of decision variables in parietal cortex, Nature 400 (1999) 233–238. [PubMed: 10421364]

[9]. Levy DJ, Glimcher PW, The root of all value: a neural common currency for choice, Current Opinion in Neurobiology 22 (2012) 1027–1038. [PubMed: 22766486]

[10]. Glimcher PW, Value-based decision making, in: Glimcher PW, Fehr E (Eds.), Neuroeconomics, Elsevier, 2 edition, 2014, pp. 373–391.

[11]. Schultz W, Dayan P, Montague PR, A neural substrate of prediction and reward, Science 275 (1997) 1593–1599. [PubMed: 9054347]

[12]. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ, Temporal difference models and reward-related learning in the human brain, Neuron 38 (2003) 329–337. [PubMed: 12718865]

[13]. Sutton RS, Barto AG, Reinforcement Learning: An Introduction, MIT Press, 2018.

[14]. Mongillo G, Shteingart H, Loewenstein Y, The misbehavior of reinforcement learning, Proceedings of the IEEE 102 (2014) 528–541.* This paper provides an accessible overview of the algorithmic differences between value-based and policy-gradient reinforcement learning, as well as reviewing behavioral and neural evidence in favor of each model family.

[15]. Dayan P, Abbott LF, Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, The MIT Press, 2001.

[16]. Wunderlich K, Rangel A, O'Doherty JP, Neural computations underlying action-based decision making in the human brain 106 (2009) 17199–17204.

[17]. Ito M, Doya K, Validation of decision-making models and analysis of decision variables in the rat basal ganglia 29 (2009) 9861–9874.

[18]. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ, Model-based influences on humans' choices and striatal prediction errors, Neuron 69 (2011) 1204–1215. [PubMed: 21435563]

[19]. Cai X, Kim S, Lee D, Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice, Neuron 69 (2011) 170–182. [PubMed: 21220107]

[20]. Vandekerckhove J, Matzke D, Wagenmakers E-J, et al., Model comparison and the principle of parsimony, in: Busemeyer JR, Wang Z, Townsend JT, Eidels A (Eds.), Oxford Handbook of Computational and Mathematical Psychology, 2015, pp. 300–319.

[21]. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P, Trust region policy optimization, in: Bach F, Blei D (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 1889–1897.

[22]. Ballard IC, McClure SM, Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models, Journal of Neuroscience Methods 317 (2019) 37–44. [PubMed: 30664916]

[23]. Williams RJ, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine Learning 8 (1992) 229–256.

[24]. Degris T, Pilarski PM, Sutton RS, Model-free reinforcement learning with continuous action in practice, in: 2012 American Control Conference (ACC), IEEE, pp. 2177–2182.

[25]. Pompilio L, Kacelnik A, State-dependent learning and suboptimal choice: When starlings prefer long over short delays to food, Animal Behaviour 70 (2005) 571–578.

[26]. Pompilio L, Kacelnik A, Behmer ST, State-dependent learned valuation drives choice in an invertebrate, Science 311 (2006) 1613–1615. [PubMed: 16543461]

[27]. Aw J, Holbrook R, Burt de Perera T, Kacelnik A, State-dependent valuation learning in fish: Banded tetras prefer stimuli associated with greater past deprivation, Behavioural Processes 81 (2009) 333–336. [PubMed: 18834933]

[28]. Palminteri S, Khamassi M, Joffily M, Coricelli G, Contextual modulation of value signals in reward and punishment learning, Nature Communications 6 (2015) 1–14.* This paper uses a carefully controlled task design to show that human participants display behavior consistent with context-sensitive preference learning. Specifically, participants' preferences for an action were modulated by the value of the action's context. One result of this was that avoidance responses accrue positive preference in contexts with a negative expected value.

[29]. Yoo SBM, Hayden BY, Pearson JM, Continuous decisions, Philosophical Transactions of the Royal Society B 376 (2021) 20190664.* This paper points out that naturalistic decision-making frequently involves continuous action spaces and prolonged choice windows, and offers a cogent argument that many common laboratory tasks and models are poorly suited for studying these features of decision making. It suggests that policy-gradient reinforcement learning and control theory provide a solid foundation for studying continuous decisions.

[30]. Doya K, Reinforcement learning in continuous time and space, Neural Computation 12 (2000) 219–245. [PubMed: 10636940]

[31]. Santamaria JC, Sutton RS, Ram A, Experiments with reinforcement learning in problems with continuous state and action spaces, Adaptive Behavior 6 (1997) 163–217.

[32]. Niv Y, The effects of motivation on habitual instrumental behavior, Ph.D. thesis, The Hebrew University of Jerusalem, 2007.

[33]. Haith AM, Krakauer JW, Model-based and model-free mechanisms of human motor learning, in: Progress in motor control, Springer, 2013, pp. 1–21.

[34]. McDougle SD, Ivry RB, Taylor JA, Taking aim at the cognitive side of learning in sensorimotor adaptation tasks, Trends in Cognitive Sciences 20 (2016) 535–544. [PubMed: 27261056]

[35]. Hadjiosif AM, Krakauer JW, Haith AM, Did we get sensorimotor adaptation wrong? implicit adaptation as direct policy updating rather than forward-model-based learning, Journal of Neuroscience 41 (2021) 2747–2761. [PubMed: 33558432] * Using a motor control task with mirror-reversed visual feedback, this paper provides behavioral evidence that implicit motor adaptation is more consistent with the direct learning of a behavioral policy than with the use of a predictive forward model to plan movements.

[36]. Watabe-Uchida M, Eshel N, Uchida N, Neural circuitry of reward prediction error, Annual Review of Neuroscience 40 (2017) 373–394.

[37]. Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N, Arithmetic and local circuitry underlying dopamine prediction errors, Nature 525 (2015) 243–246. [PubMed: 26322583]

[38]. Roesch MR, Calu DJ, Schoenbaum G, Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards, Nature neuroscience 10 (2007) 1615. [PubMed: 18026098]

[39]. Howe MW, Dombeck DA, Rapid signalling in distinct dopaminergic axons during locomotion and reward, Nature 535 (2016) 505–510. [PubMed: 27398617]

[40]. da Silva JA, Tecuapetla F, Paixão V, Costa RM, Dopamine neuron activity before action initiation gates and invigorates future movements, Nature 554 (2018) 244–248. [PubMed: 29420469]

[41]. Syed EC, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME, Action initiation shapes mesolimbic dopamine encoding of future rewards, Nature Neuroscience 19 (2016) 34–36. [PubMed: 26642087] * The authors show that RPE-like changes in DA concentration in the nucleus accumbens (measured using fast-scan cyclic voltammetry) are only evident during the anticipation or execution of overt actions in a go/no-go task, suggesting the roles of dopamine in signaling prediction errors and in motivating purposeful movement are interrelated.

[42]. Coddington LT, Dudman JT, The timing of action determines reward prediction signals in identified midbrain dopamine neurons, Nature neuroscience 21 (2018) 1563. [PubMed: 30323275] ** Demonstrates movement initiation produces phasic responses in midbrain dopamine neurons early in learning during a classical conditioning task in mice, suggesting action-related components of neural prediction error signals both precede and are independent of responses related to learning about reward-predictive sensory cues.

[43]. Klaus A, Alves da Silva J, Costa RM, What, if, and when to move: basal ganglia circuits and self-paced action initiation, Annual review of neuroscience 42 (2019) 459–483.

[44]. Samejima K, Ueda Y, Doya K, Kimura M, Representation of action-specific reward values in the striatum, Science 310 (2005) 1337–1340. [PubMed: 16311337]

[45]. FitzGerald TH, Friston KJ, Dolan RJ, Action-specific value signals in reward-related regions of the human brain, Journal of Neuroscience 32 (2012) 16417–16423. [PubMed: 23152624]

[46]. Elber-Dorozko L, Loewenstein Y, Striatal action-value neurons reconsidered, eLife (2018) 32.* Challenges the popular idea that populations of striatal neurons unambiguously reflect action values, showing that previous analyses that putatively isolate correlates of action value in neural activity cannot dissociate value representations from alternative representations (e.g., policy).

[47]. Li J, Daw ND, Signals in human striatum are appropriate for policy update rather than value prediction, Journal of Neuroscience 31 (2011) 5504–5511. [PubMed: 21471387]

[48]. Kishida KT, Saez I, Lohrenz T, Witcher MR, Laxton AW, Tatter SB, White JP, Ellis TL, Phillips PE, Montague PR, Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward, Proceedings of the National Academy of Sciences 113 (2016) 200–205.

[49]. Schulman J, Moritz P, Levine S, Jordan M, Abbeel P, High-dimensional continuous control using generalized advantage estimation, arXiv preprint arXiv:1506.02438 (2015).

[50]. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, pp. 1928–1937.

[51]. Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, et al. , Grand-master level in StarCraft II using multi-agent reinforcement learning, Nature 575 (2019) 350–354. [PubMed: 31666705]

[52]. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M, Deterministic policy gradient algorithms, in: Proceedings of the 31st International Conference on Machine Learning,, Beijing, China, p. 9.

[53]. Nachum O, Norouzi M, Xu K, Schuurmans D, Bridging the gap between value and policy based reinforcement learning, in: Advances in Neural Information Processing Systems, pp. 2775–2785.

[54]. Sakai Y, Fukai T, The actor-critic learning is behind the matching law: matching versus optimal behaviors, Neural Computation 20 (2008) 227–251. [PubMed: 18045007]

[55]. Maia TV, Two-factor theory, the actor-critic model, and conditioned avoidance, Learning & Behavior 38 (2010) 50–67. [PubMed: 20065349]

[56]. Bennett D, Davidson G, Niv Y, A model of mood as integrated advantage, Psychological Review (in press).

[57]. Barto AG, Adaptive critics and the basal ganglia, in: Models of Information Processing in the Basal Ganglia, The MIT Press, Cambridge, MA, 1994.

[58]. Joel D, Niv Y, Ruppin E, Actor–critic models of the basal ganglia: New anatomical and computational perspectives, Neural Networks 15 (2002) 535–547. [PubMed: 12371510]

[59]. Fanselow MS, Wassum KM, The origins and organization of vertebrate Pavlovian conditioning, Cold Spring Harbor Perspectives in Biology 8 (2016).

[60]. Lichtenberg NT, Pennington ZT, Holley SM, Greenfield VY, Cepeda C, Levine MS, Wassum KM, Basolateral amygdala to orbitofrontal cortex projections enable cue-triggered reward expectations, Journal of Neuroscience 37 (2017) 8374–8384. [PubMed: 28743727]

[61]. Rescorla RA, Wagner AR, A theory of pavlovian conditioning, Classical Conditioning II: Current Theory and Research (1971).

[62]. Balleine BW, Dickinson A, Goal-directed instrumental action: contingency and incentive learning and their cortical substrates, Neuropharmacology 37 (1998) 407–419. [PubMed: 9704982]

[63]. Sharpe MJ, Chang CY, Liu MA, Batchelor HM, Mueller LE, Jones JL, Niv Y, Schoenbaum G, Dopamine transients are sufficient and necessary for acquisition of model-based associations, Nature Neuroscience 20 (2017) 735–742. [PubMed: 28368385]

[64]. Langdon AJ, Sharpe MJ, Schoenbaum G, Niv Y, Model-based predictions for dopamine, Current Opinion in Neurobiology 49 (2018) 1–7. [PubMed: 29096115]

[65]. Dayan P, Berridge KC, Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation, Cognitive, Affective, & Behavioral Neuroscience 14 (2014) 473–492.

[66]. Niv Y, Daw N, Dayan P, How fast to work: Response vigor, motivation and tonic dopamine, in: Weiss Y, Schölkopf B, Platt J (Eds.), Advances in Neural Information Processing Systems, volume 18, MIT Press, 2006.

[67]. Collins AG, Frank MJ, Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive, Psychological Review 121 (2014) 337. [PubMed: 25090423]

[68]. Leong YC, Radulescu A, Daniel R, DeWoskin V, Niv Y, Dynamic interaction between reinforcement learning and attention in multidimensional environments, Neuron 93 (2017) 451–463. [PubMed: 28103483]

[69]. Radulescu A, Niv Y, Ballard I, Holistic reinforcement learning: the role of structure and attention, Trends in Cognitive Sciences 23 (2019) 278–292. [PubMed: 30824227]

Highlights

- Reinforcement learning models can be divided into value-based and policy-gradient
- Policy-gradient RL gives a more parsimonious account of many operant behaviors
- These behaviors can therefore be explained without invoking the notion of ‘value’

Box 1:

Glossary of Reinforcement-Learning Terms

Policy: a function, often denoted π, that specifies a probability distribution over actions given the agent's current state. Actions are sampled from the policy at the time of choice. A policy is an essential component of all RL algorithms—including value-based RL algorithms, which must specify a *policy-mapping function* (see below) that computes a policy given a set of action-values.

Value: expected discounted cumulative future reward. In value-based RL, values can be defined both for states ($V^{\pi}(s)$: the expected future reward associated with being in state s and choosing actions according to the policy π) and actions ($Q^{\pi}(s, a)$: the expected future reward associated with taking action a in state s, and choosing actions according to the policy π thereafter).

Policy-mapping function: a function that specifies choice probabilities for a set of actions given their estimated values, in value-based RL algorithms. Also known as a 'choice rule'. Examples include arg max, ϵ-greedy, and softmax.

- arg max: a policy-mapping function that deterministically selects the action with the highest estimated value
- ϵ-greedy: a policy mapping function that selects either the action with the highest estimated value (with probability $1 - \epsilon$) or a random action (with probability ϵ)
- softmax: a policy-mapping function that stochastically selects actions with a probability that increases based on their estimated value relative to the values of alternative options:

$$\pi(a) = \frac{e^{\beta \cdot Q^{\pi}(s,a)}}{\sum_{\hat{a} \in A} e^{\beta \cdot Q^{\pi}(s,\hat{a})}} \quad (1)$$

  The degree of stochasticity in this mapping is controlled by the inverse temperature parameter β, such that all actions are equally likely when $\beta = 0$ and the softmax function becomes the arg max function as $\beta \to \infty$

Policy gradient: For RL algorithms that use a parameterized policy, the policy gradient is a vector that indicates how much extra reward the agent expects to receive if it made an incremental change to each of the parameters of its policy (technically, the gradient is the vector of partial derivatives of the expected reward function with respect to policy parameters). The gradient of a policy is the key variable estimated (or approximated) by policy-gradient RL algorithms, and is used to adjust the parameters of the policy in the direction in which expected reward is expected to increase most steeply. For this reason, policy-gradient algorithms must use a policy that is everywhere differentiable with respect to its parameters.

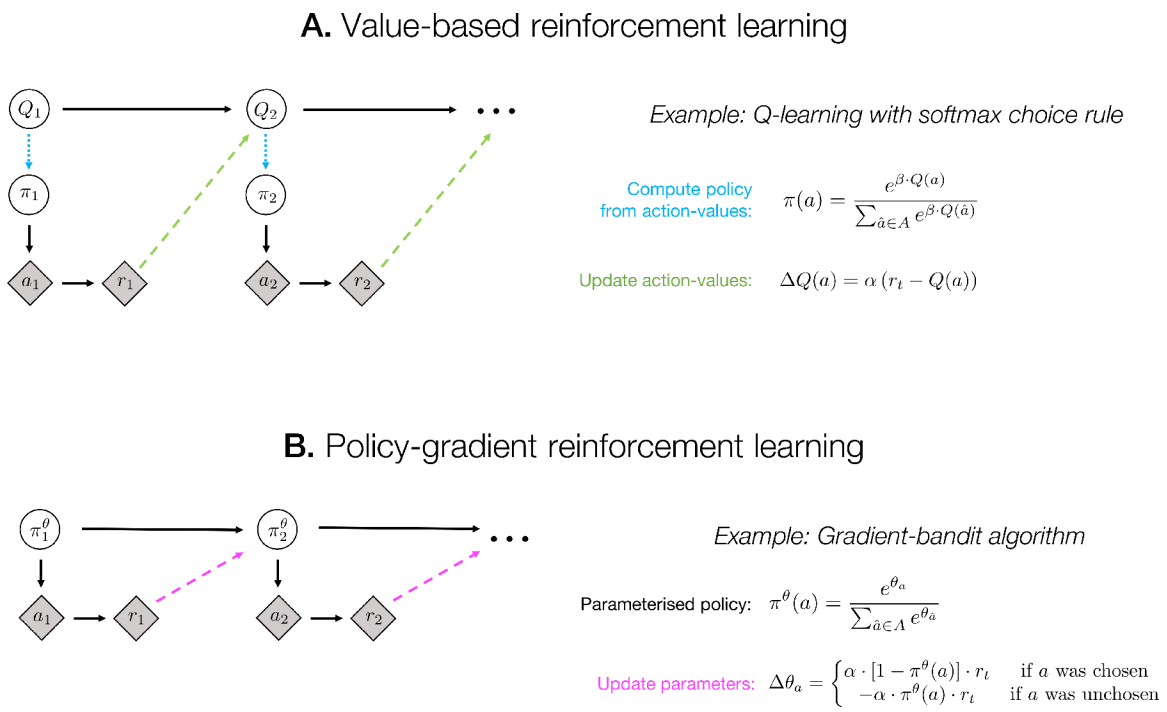

Figure 1:

Update schematics for example value-based and policy-gradient RL algorithms. Shaded diamond nodes denote observable variables, unshaded circular nodes denote latent variables that are internal to the RL agent, and arrows denote dependencies. For simplicity, in these algorithms we do not show the environmental state, which would be an additional (potentially partially) observable variable. **A**: in a value-based RL algorithm (such as the *Q*-learning model presented here), actions (*a*, chosen from a discrete set *A*) are a product of the agent's policy π, which in turn is determined (dotted cyan arrow) by the learned action-values (*Q*). The update rule for action-values (dashed green arrow) depends on the action-values and received reward (*r*) at the previous timestep, and only indirectly on the policy. This algorithm has two adjustable parameters: the learning rate α and the softmax inverse temperature β. **B**: a policy-gradient algorithm (such as the gradient-bandit algorithm presented here; see [13]) selects actions according to a parameterised policy π^{θ}, and updates the parameters θ of this policy directly (dashed magenta arrow; in the gradient-bandit algorithm, θ is a vector of action preferences), without the intermediate step of learning action-values. In the policy-gradient algorithm, by contrast with the value-based algorithm, the size of the update to θ depends more directly on the current policy, since the size of the update to each action preference is scaled by the probability of that action under the policy.