Reinforcement learning:
A brief guide for philosophers of mind

Julia Haas | DeepMind

Draft prepared for *Philosophy Compass*

As a learning problem, reinforcement learning asks: how can an agent optimize its behavior by learning from interactions with its environment? For example, how does a baby plover learn to survive in its environment, simply by hopping around in it? Or again, how does a newcomer to London find her way around, just by using a map and a bit of trial-and-error? As a research program, reinforcement learning refers to a branch of computer science, together with associated interdisciplinary approaches, that analyzes formal versions of this question and develops computational solutions to it (Dayan & Abbott, 2001; Montague, 2007; Glimcher and Fehr, 2013). Finally, reinforcement learning methods are the suites of computational algorithms that aim to solve the learning problem (Sutton and Barto 1998, 2018).

Over the past fifty years, the interrelated problem, program and methods have had a significant impact on the development of artificial intelligence and empirical scientific inquiry. This is at least in part because, while reinforcement learning methods are roughly normative in nature - that is, they characterize optimal methods for learning from the environment - they also successfully predict and explain a wide range of empirical findings and applications across the natural sciences, most notably in neuroscience, psychology, psychiatry, and economics, or the so-called 'the decision sciences.'[1] Increasingly, non-computational or conceptual approaches have also taken up the framework to play an explanatory role in theorizing, including in some cases in the philosophical literature (Schroeder 2004; Huebner, 2012, 2015; Arpaly & Schroeder 2014; Colombo 2014; Colombo & Wright 2017; Railton 2017a, 2017b, [xxxx] 2018, 2020, 2021).

This plurality of analyses, findings, and theories is, at a minimum, a sign of reinforcement learning's theoretical productivity (Kitcher 1982, 35-48). However, philosophical engagement with it remains limited, especially as compared with, say, the predictive processing research program (Hohwy, 2013; Clark, 2013, 2015; Metzinger & Wiese, 2017), other areas of machine learning (e.g., Buckner, 2019), and computer science more broadly construed (Haugeland, 1997; Cummins, 2000; Copeland & Proudfoot, 2006). This lack of engagement is regrettable. There are key theoretical points of convergence between reinforcement learning and philosophy, especially in the philosophy of mind, as well as real prospects for productive, bidirectional collaborations between the two fields.

In this opinionated review, I aim to draw attention to some of the contributions reinforcement learning can make to philosophical inquiry, and particularly to questions in the philosophy of mind.[2] I specifically highlight

[1] These approaches are both effective and tractable means of generating good behavior, which is highly desirable in the design of artificial intelligence.

[2] Given this focus on the philosophy of mind, it is worth highlighting some significant approaches that the review does not cover. These include the relationship between reinforcement learning and the philosophy of artificial intelligence,

reinforcement learning's foundational emphasis on the role of reward in agent learning; I appeal to some of the contributions reinforcement learning has made to the decision sciences as a kind of stepping stone between the basic machine learning framework and more traditional, philosophical folk psychological views on the other. In particular, I canvass two ways in which the proposed nature and workings of reward may help advance our understanding of perception and motivation, respectively.

For precision, I make several assumptions about the nature of reinforcement learning and its instantiation in minds like ours. I sketch these assumptions, together with their relationship to other versions of reinforcement learning, in Section 2. In Section 3, I review contributions of reinforcement learning methods to advancements across the decision sciences. In Section 4, I show how principles from reinforcement learning can shape philosophical debates regarding the nature of perception and characterizations of desire. I briefly conclude in Section 5.

2. Foundations and assumptions

As a general approach, the reinforcement learning framework makes certain foundational and technical assumptions, with specific versions of the framework commiting to some assumptions while suspending or relaxing others.[3] Here, I sketch what I call the 'reinforcement learning and decision-making' (RLDM) framework, which draws on assumptions made in both machine learning and computational neuroscience.[4] Specifically, in addition to assuming many of the somewhat more basic features of the framework, this version assumes that reinforcement learning is, to some degree, meaningfully instantiated in the minds of biological organisms. It also takes a particular - if minimal - view regarding the problem of 'where rewards come from' in biological systems. Throughout, it will be useful to remember that this is just one - though perhaps particularly philosophically useful - variant of the framework among many.

Let's start with the basics. In a reinforcement learning framework, we have an agent and an environment. The agent is the learner or decision-maker in question, and selects different actions in its environment, where actions can be understood as "any decisions we want to learn how to make" (Sutton & Barto, 2018, 50). The environment refers to everything 'outside' of the agent, which the agent cannot arbitrarily change but rather that the agent interacts with.[5] The agent and the environment interact in the sense that the agent is presented with sensory information from the environment, and the agent chooses different actions within the environment, where the environment is affected by these actions. Here, choosing a given action within a given state is sometimes called a 'state-action pair.' Notably, the agent may not be able to observe the complete environment and may not have prior knowledge of the environment dynamics. In addition, the agent may (but by no means

understood as a topic roughly independent from the philosophy of mind (e.g., Bostrom, 2017), as well as historical and foundational (philosophy of science) approaches to reinforcement learning (Colombo, 2014, *in prep*).

[3] This section is indebted to Sutton & Barto (2018) and, especially, to Neil Rabinowitz.

[4] Name adapted from Gęsiarz & Crockett (2015).

[5] For example, in many cases, even parts of the agent's body are considered to be a part of the environment.

needs to) build a model of the environment in order to choose actions in and learn from it (for further discussion of models, see Section 3, below).

Crucially, one of the distinguishing features of the reinforcement learning framework is the role of reward. Roughly, in reinforcement learning, the agent's objective in the environment is to maximize the cumulative reward it receives over time, where rewards are passed from the environment to the agent (for more on this, see below). In their influential text, Sutton and Barto call this framing the *reward hypothesis*, specifying, "all of what we mean by [an agent's] goals and purposes can be well thought of as the maximization of expected value of the cumulative sum of a received scalar signal (called reward)" (2018, 53). That is, the agent's objective is to maximize its yield of reward as it acts in the world, and this objective is characterized by assigning a quantity of intrinsic desirability to each state (or to taking each action in each state), known as the reward.[6]

This intrinsic desirability assigned to each state (or taking each action in each state), or *reward*, can be contrasted with the notion of *expected value*, which captures the expected, discounted, future sum of reward associated with each state (or each action in each state), conditional on a certain policy of action. Here, 'policy' refers to the function that determines what actions to perform given a set of observations and history. We can elucidate the distinction between reward and expected value further using an example from Silver (2015). Imagine an agent in an environment with a wall. Upon arriving at the wall for the first time, the agent receives a *reward* from the environment. But this reward can also be used to assess how relatively good (valuable) individual states are expected to be to the extent that, conditional on a certain action policy, they *lead to* the wall and hence the reward. Hence, an agent's ongoing interactions with its environment enable it to continually revise the expected value it attributes to a given state or state-action pair conditional on a certain policy, upgrading or downgrading as needed. And this enables the agent to *learn* the most appropriate actions in the most appropriate states in order to maximize cumulative reward over time, conditional on a certain policy, in spite of the fact that states (or state-action pairs) can be of high *expected value* without being intrinsically worthwhile (rewarding).[7] This partly helps explain why not every state in an environment needs to be directly rewarding in order for an agent to act appropriately within it.

As a branch of machine learning, reinforcement learning represents the foregoing conceptual features in computational terms. There are countless reinforcement learning algorithms, or processes for learning effective policies, each with a distinctive computational profile. For example, the *temporal-difference learning algorithm* represents a computationally efficient way of making predictions about reward in the future. One (non-TD) way to improve predictions over time is to make a prediction about an actual outcome, compare the difference

[6] Thanks to Neil Rabinowitz for this formulation.

[7] We can take the example of making and having coffee to help illustrate the difference between a state of high expected value that is nonetheless not technically rewarding. Although only drinking a cup of coffee itself may be intrinsically worthwhile (rewarding), and the grinding of the beans almost certainly is not, the state-action pair of grinding the coffee is nonetheless associated with *expected value*, as it is, conditional on a certain policy, a necessary step or state-action pair on the way to having the coffee.

(or error) between the two, and then update that prediction. To borrow an example from Sutton (1988, 10), one can make a prediction on Monday about the weather on Saturday, wait until Saturday, and then update Monday's prediction based on the difference between Monday's prediction and Saturday's actual weather. The temporal difference approach does something a little neater by updating its predictions throughout. That is, it is more akin to one making a prediction on Monday about the weather on Saturday, but then comparing Monday's prediction to *Tuesday's* prediction about Saturday and adjusting accordingly, and so on. For instance, if Monday's prediction for Saturday is a 90% chance of rain, but Tuesday's prediction for Saturday is only a 60% chance, then the temporal difference approach is to lower Monday's prediction.[8]

Notably, given that different problem settings present different challenges, there are myriad different RL algorithms in use today. These trade off factors such as memory consumption, computation cost, data efficiency, and stability; some are useful for very small environments and others for very large; some for discrete action spaces, and others for continuous ones.[9] Thus, 'reinforcement learning' refers to a general learning problem and a suite of computational algorithms, as well as to the branch of computer science devoted to studying them, rather than to any token solution to the problem.[10]

The RLDM version of reinforcement learning adds two assumptions to the basic reinforcement learning framework. First, it assumes a relationship between reinforcement learning and the minds of biological creatures like us. This assumption is by no means universally held: machine learners can pursue decades of research and remain entirely agnostic regarding the role of reinforcement learning in biological agents. Similarly, cognitive and comparative psychologists can study the nature of learning and behavior without any appeals to the reinforcement learning framework. However, RLDM follows computational neuroscientists and other decision scientists who suspect that reinforcement learning does, in fact, capture something special about minds like ours. As Dayan and Niv (2008, p. 1) put it, reinforcement learning appears to offer "more than just a computational, 'approximate ideal learner' theory for affective decision-making. [Reinforcement learning] algorithms, such as the temporal difference (TD) learning rule, appear to be directly instantiated in neural

[8] For a more detailed discussion, see Sutton and Barto (2018, Chapter 6, and especially Example 6.1.).

[9] Thanks to Neil Rabinowitz for this formulation.

[10] Very briefly, deep reinforcement learning combines the foregoing group of methods with a second, general computational framework known as deep learning (for an excellent overview of the integration between the two systems and its applications in neuroscience research, to which this subsection is indebted, see Botvnick, Wang, Dabney, Miller, & Kurth-Nelson, 2020). Deep learning systems are often characterized in architectural terms, namely, as groups of artificial neurons connected with artificial synapses (for a philosophical introduction to these networks, see Buckner, 2019). Like neurons, these units transmit scalar values based on their inputs, weighted by the relative strength of their connection (Goodfellow et al., 2016). Following Botvinick et al. (2020, 605), then, the resulting combination of deep reinforcement learning can be defined as "any system that solves an RL problem (i.e., maximizes long-term reward), using representations that are themselves learned by a deep neural network (rather than stipulated by a designer.)" For example, a deep reinforcement learning system was able to develop superhuman play in the game Go by combining supervised learning of human games of Go and reinforcement learning that was used to improve the foregoing policy by having the system play itself (Silver et al., 2016; see also Silver et al., 2017; Silver et al., 2018).

mechanisms, such as the phasic activity of dopamine neurons. That [reinforcement learning] appears to be so transparently embedded has made it possible to use it in a much more immediate way to make hypotheses about, and retrodictive and predictive interpretations of, a wealth of behavioral and neural data collected in a huge range of paradigms and systems." However, we are free to relax the condition that reinforcement learning is directly *instantiated* in the workings of the brain. It is sufficient to say that reinforcement learning provides remarkably useful frameworks for thinking about decision-making and selection in the mind.

RLDM's second assumption has to do with the nature of reward. As noted above, in the basic reinforcement learning framework, rewards are passed from the environment to the agent when an agent enters certain states of the environment, or when the agent takes certain appropriate actions in certain appropriate states. This external nature of reward is unproblematic in the context of machine learning, because the reward is simply designed by the researcher as a means of communicating what the researcher wants the artificial agent to achieve. But things get thornier when we get to biological organisms, since it's not clear where rewards would then come from. This question regarding the origin of reward in biology generates what Juechems and Summerfield call the *paradox of reward*. The issue is paradoxical, the authors contend, because

> No external entity exists that can directly quantify the consequences of each action, like the points that are awarded in a video game for completing levels or shooting monsters. Nor is it obvious that biological systems have a dedicated channel for receipt of external rewards that is distinct from the classical senses. Rather, rewards and punishments are sensory observations – the taste of an apple, the warmth of an embrace – and so stimulus value must be inferred by the agent, not conferred by the world. In other words, rewards must be intrinsic, not extrinsic" (2019, 837-838).

Exactly how this conversion between sensory observations and assignments of intrinsic rewards occurs - assuming that it occurs at all - remains the subject of lively theoretical debate. One possible explanation is that minds like ours have evolved specific mechanisms that convert sensory observations into hedonic signals (e.g., see Schultz, 2015). Another, complementary possibility is that, in addition to the evolved mechanisms for basic rewards (e.g., food and water), human beings develop cognitive setpoints, akin to homeostatic setpoints, on which reward amounts to a by-product of computing the distance to self-defined goals (e.g., such as getting married or going to graduate school) (Juechems & Summerfield, 2019). Here, RLDM again takes a minimal approach, and merely assumes that minds like ours subpersonally assign subjective rewards to sensory observations; it remains provisionally agnostic about how this assignment takes place.

3. Contributions to the decision sciences

As noted at the outset, there are important points of contact between formal, normative reinforcement learning algorithms and the decision sciences, which, when considered, could in turn have appreciable implications for core issues in the philosophy of mind. We can now turn to what Dayan and Niv above called the "wealth of behavioral and neural data" understood and interpreted against the backdrop of RLDM, with an emphasis on cognitive neuroscientific approaches.

As gestured at above, arguably the most significant connection is between RLDM and the reward system in the mammalian brain. In the mid-1990s, theoretical and empirical work showed that the firing of dopamine neurons is closely described by the temporal difference learning algorithm (for narrative accounts of the discovery, see Montague, 2007; Redish, 2013; see also, Colombo, 2014). That is, dopamine neurons fire when an organism experiences a higher- or lower-than-expected value in association with a given state (Schultz, Dayan, and Montague 1997). This seminal finding in turn led to the use of reinforcement learning methods to study the neuroscience of vision (Hayhoe and Ballard, 2005; Hikosaka 2000; Hickey et al. 2010), attention (Della Libera & Chelazzi 2009; Chelazzi et al. 2014; Anderson & Kim 2018), memory (Patil et al., 2017; Ergo, De Loof, & Verguts, 2020), prospective memory (Krishnan & Shapiro, 1999; Katai et al., 2003; Kliegel et al., 2005; Walter & Meier, 2014), cognitive control (Savine & Braver, 2010; Botvinick & Braver, 2014; Chiew & Braver, 2014; Cubillo, Makwana, & Hare, 2019) , and above all, decision-making (Sutton and Barto, 2018; Dayan and Niv 2008, Rangel et al. 2008, Dayan 2011, Glimcher & Fehr 2013).

For example, a systematic body of evidence now indicates that the reward system guides visual fixation and saccadic eye movement, i.e., what we look at, when, and in what order (Liao & Anderson, 2020). Similarly, reward guides what we do or don't attend to - more precisely than do either location or salience (Anderson and Kim, 2018). Conversely, deficits and disruptions (e.g., by addictive substances) to the reward system are not only implicated in diseases such as Parkinson's and Tourette's, but also in a range of psychiatric disorders, including depression (Huys, Daw, & Dayan, 2015) and addiction (Hyman, 2005; Redish, Jensen, & Johnson, 2008; Redish, 2013). Arguably, methods from reinforcement learning thus represent an important - and, to date, under-utilized - framework for elucidating the nature and mechanisms underlying selection between competing states of affairs across a range of 'low'- as well as 'high-level' kinds of cognitive processing.

Equally rich cross-pollination exists between methods in RLDM and the field of neuroeconomics (for an engaging discussion on the relationship between the two, especially regarding the slightly orthogonal uses of key concepts as 'reward' and 'value,' see Padoa-Schioppa & Schoenbaum, 2015). Unlike in neo-classical economics, where value is mainly taken to be a theoretical construct, (that is, where economic choice is accounted for *as if* the choosing subject maximized an internal value function), the core idea in neuroeconomics is that the brain actually assigns subjective values to various choice alternatives, such that value acts as a 'common currency' for choosing between prima facie incommensurable alternatives. Hence, the study of neuroeconomics can roughly be described as efforts to understand the calculation of value across competing alternatives. By extension, a key theoretical virtue of neuroeconomic approaches lies in their ability to measure the subjective values associated with various decision factors, notably factors such as delay, risk, loss aversion, cognitive effort, social considerations, and normative considerations, together with their underlying neural mechanisms (see Westbrook & Braver 2015 for a review).

Specifically, a standard strategy asks participants to choose between alternative offers, uses these choices to infer the subjective value of each alternative (consistent with standard economic principles), and then measures these values against participants' neural signals (for a review, see Padoa- Schioppa 2011). For example, the subjective value of larger economic offers is systematically reduced when paired with increasingly demanding tasks, such that subjective effort can be quantified in terms of cost (Westbrook, Kestner, & Braver, 2013). Findings such as these have, in turn, enabled us to localize subjective valuation in the brain, and even to show that the subjective costs of cognitive effort are represented by domain-general valuation mechanisms in the brain (Westbrooke, Lamichhane, & Braver, 2019). Reinforcement learning methods and, by extension, neuroeconomic paradigms thus provide a significant avenue for philosophers interested in moving beyond introspection and self-report to more systematic and quantifiable methods for studying certain capacities in the brain (significant issues concerning our cognitive taxonomies notwithstanding).

There are also substantial interactions between RLDM and animal learning (for an excellent extended review, see Sutton and Barto, 2018, Chapter 14). Foundationally, a relatively small number of algorithms offer principled explanations of learning phenomena such as classical and instrumental conditioning, delayed reinforcement, Kamin blocking, extinction, and cognitive maps. The differing contours of several key reinforcement learning algorithms have also directly and indirectly been used to distinguish between decision systems, most notably between the Pavlovian, habit-based, and goal-directed decision systems, and their interplay in decision-making, in human as well as non-human animal decision-making (Kable & Glimcher, 2009; Dolan & Dayan, 2013).

Roughly, the *Pavlovian* system produces basic, stimulus-driven behavioral responses.[11] These behaviors are not learned, and they do not appear to be controlled by the animal at all. Rather, they are most likely the products of a lengthy evolutionary history, which has selected for a range of automatic, appropriate responses in the face of appetitive or aversive stimuli (Macintosh 1983). These unconditioned responses include both outcome-specific responses, such as inflexibly licking water, and more open-ended, valence-dependent responses, such as generally approaching something rewarding. Both classes of response are characteristically recalcitrant to changes in outcome, as when chickens will continue to peck at a feeder that will not dispense any seeds over hundreds of trials (Macintosh, 1983; Huys et al. 2011, 2012).

The habit-based decision system, often associated with model-free learning, akin to trial-and-error learning (e.g., Daw *et al.*, 2005), produces instrumental responses by choosing actions based on their previously learned expected values in different contexts. For example, a button-press that has previously resulted in a reward is a good state-action pair, while a button-press that has previously resulted in a punishment is a bad state-action

[11] Notably, this sense of the word 'Pavlovian' should not be confused with its everyday usage, i.e., in reference to the conditioned response. In psychology, attention is paid to Pavlov's discovery of the bell as the conditioned stimulus and its learned association with the unconditioned response, the salivating. But for the purposes of RLDM, it is really the relationship between the *unconditioned* stimulus (i.e., the food) and the *unconditioned* response (i.e., the salivating) that is of interest (for an interesting discussion of how these two interpretive traditions view Pavlovian learning differently, see Rescorla 1988).

pair. However, unlike the foregoing Pavlovian system, the model free system is not "hardwired," and does gradually update.

The goal-directed decision system, often associated with model-based learning, or methods that explicitly use models and planning (e.g., Daw *et al.*, 2005), uses a forward-looking model to represent possible actions, outcomes, and associated expected values. Decision-making using such a model is often achieved by traversing a decision tree. Each node in the tree represents a possible choice; the model-based system can then "search" through the decision tree to find the branch with the highest total expected value. For example, a chess player may represent three upcoming moves in a game of chess, with each possible move further branching into a wide range of subsequent moves. To win, the player seeks to find the best possible sequence of moves overall. Such modeling enables the agent to both plan future sequences of actions, as well as to generalise to new situations (although for failures in future representations, see, e.g., Redish, Jensen, & Johnson, 2008).

Of course, the connections between reinforcement learning, neuroscience, neuroeconomics, and psychology extend well beyond what is described here. The aim here is not to catalogue these intersecting bodies of literature here in a comprehensive way. Rather, my goal is to highlight the two rich, intersecting frameworks. To complement this, the next section offers three substantive illustrations of how principles from reinforcement learning are already being taken up in philosophically-oriented inquiry, specifically, in discussions of perception and the nature of desire, respectively. In both cases, the notion of reward - as introduced in reinforcement learning, but taken up and in some cases adapted in the decision sciences - plays a guiding, explanatory role.

4. Reward in perception and desire

4.1. Perception

In addition to its contributions to the decision sciences, RLDM has recently begun to inform philosophical approaches to the study of perception and, in particular, philosophical approaches to the study of binocular rivalry. Binocular rivalry occurs when one stimulus is shown to one eye at the same time as a different stimulus is shown to the other. The resulting experience is of the two images alternating back and forth. For example, if one eye is shown an image of a *face* and the other eye is shown an image of a *house*, then, rather than seeing the face and the house superimposed over one another, the experience is of seeing a face, then a house, and so on. *Perceptual dominance* in binocular rivalry refers to one of the two images appearing first, or for a longer period of time during the overall duration of the experience of alternation. Since binocular rivalry is a case of perception clearly not representing 'what's going on out there,' it constitutes an important test case for theories that aim to explain why we have particular perceptual states at particular times.

To try and explain the puzzling nature of binocular rivalry, Jakob Hohwy and colleagues (2008) propose a predictive processing approach that recast the phenomenon in terms of Bayesian inference. Briefly, Hohwy and colleagues propose to explain binocular rivalry in terms of priors, likelihoods, and prediction errors. For example, they suggest that the perceiver experiences a *house* or a *face*, say, because *face-and-house* combined has a much lower prior than do either *face* or *house*: it is a priori improbable that what is being seen is really a *face-and-house*, and interactions with the environment that could have induced a prior for such a hypothesis are unlikely. Thus, as long as the low prior offsets the likelihood advantage for *face-and-house* over *face* or *house*, *face-and-house* will not be selected over either *face* or *house*. Similarly, they propose that alternation between the percepts occurs because the expectation for either *face* or *house* only explains half of the stimuli. As a result, *face*, say, results in a strong prediction error signaling *house*, and so on. Since no single expectation combines a high prior and high likelihood, alternation between the two hypotheses results. The approach is widely cited across the cognitive sciences as an illustration of predictive coding's ubiquity in the brain, as well as its explanatory power as a general theory of the mind (e.g., for a review, see Clark (2013) for a prominent discussion; see also Metzinger & Wiese, 2017).

However, RLDM provides a hint that something may be missing from the predictive coding explanation ([xxxx], 2021). This is because, if reward plays a significant role in saccadic eye movement and visual fixation, it is reasonable to expect that it will also play a role in perceptual experience. And this is indeed what we find. Consistent with RLDM - but not predictive processing - when the stimulus or percepts are rewarded, this produces perceptual dominance. That is, participants are more likely to perceptually experience rewarded stimuli and rewarded percepts (Balcetis, Dunning, & Granot, 2012; Wilbertz, Van Slooten, & Sterzer, 2014; Marx & Einhauser, 2015). Moreover, a complementary phenomenon occurs for punished percepts: participants experience perceptual dominance for the *non*-punished percept in the pair, suggesting that the reward or punishment is *not* simply additional information taken into consideration by Bayes-like predictive processing, as a predictive processing view might suggest (Wilbertz, Van Slooten, & Sterzer, 2014)).

According to traditional conceptions, perceptual systems are exclusively thought to produce roughly veridical representations of distal stimuli, i.e., where the function of perception is to "get the world right" (Hohwy 2013, p. 2) and capture "what's going on out there" (Friston 2018, p. 1019) ([xxxx], 2021). By contrast, acknowledging the role of reward in binocular rivalry and, by extension, in perception more broadly, challenges the commitment to exclusive veridicality. There is thus more to perception than just information; perception produces representations *conditional on certain goals*. In this way, taking up an RLDM lens opens the way to characterizing perceptual systems as generating representations of distal stimuli for the purposes of action.

4.2 Desire

In addition to discussions in perception, principles from reinforcement learning are also making an impact on philosophical debates regarding the nature of desire. In the context of philosophical folk psychology, desire is broadly understood as a mental state associated with wanting, motivation, and action.[12] But exactly how desire is defined depends on the theory of desire one subscribes to. For example, on an action-based theory of desire, for

[12] Here, philosophical folk psychology specifically refers to philosophical theories describing human behaviors in terms of mental states such as intentions, beliefs, and desires. For a detailed discussion of philosophical folk psychology, specifically in relation to naturalism in philosophy, see [xxxx] (2020).

an organism to desire *p* is for the organism to be *disposed to act* so as to bring about *p* (Smith, 1994). These theories have the advantage of being relatively straightforward, and of capturing many of the features commonly associated with everyday descriptions of desire. But they face a number of explanatory challenges, including capturing the difference between goodness and desire, explaining why people act out of duty, desires regarding the past, and so on (Schroeder, 2004).

One notable solution for dealing with these challenges has been the *Reward-based Theory of Desire* (Schroeder, 2004; Schroeder and Arpaly, 2014). Drawing on evidence from computational neuroscience, the reward-based theory of desire contends that, "just as H2O is the unfamiliar essence of water, so… states of the reward system are the unfamiliar essence of desire" (Schroeder and Arpaly 2014, p. 187). Specifically, the view holds that "to have an intrinsic (positive) desire that *P* is to use the capacity to perceptually or cognitively represent that *P* to constitute *P* as a reward," and that "to be averse to it being the case that *P* is to use the capacity to perceptually or cognitively represent that *P* to constitute *P* as a punishment" (Schroeder, 2004, p. 131). Here, the perceptual or cognitive representations are cashed out in cellular and systems neuroscientific terms, with a special emphasis placed on the functioning of the dopaminergic system in the brain. Hence, the theory preserves the traditional, philosophical folk psychological notion of desire, but specifies it in contemporary computational and empirical terms.

The Reward Theory of Desire (RTD) is able to account for many of the standard explanatory desiderata, including the distinction between positive and negative desires (valence), the strength of desire, the acquisition and extinction of desires, and the nature of feeling and desires regarding the past - though notable not instrumental desire. For example, the traditional approach to explaining desire strength is to cash it out in terms of the strength of a disposition to act. By contrast, Schroeder argues, "strong desires cause powerful behavioral tendencies in human beings because, all else being equal, they generate powerful reward signals, and powerful reward signals have a powerful impact upon the motor striatum, and the motor striatum, in turn, has a powerful impact on movement" (2004, 139). By extension, strong desires have a stronger influence on movement than weak desires. Expanding on this view, Schroeder and Arpaly argue that "the reward system causes what desires cause," namely, actions, feelings, and cognitions (2014, 137-142).

Notably, however, in virtue of its appeal only to the construct of reward, RTD characterizes only intrinsic desire. As Schroeder (2004, p. 132) observes,

> This is a theory of intrinsic desire. That is, it is a theory of what it is to desire, say, that my father be happy, that my favorite team win the tournament, that I not smell the odors from the cat litter, and so on: to desire things for their own sakes. It is not a theory of what it is to desire things merely as a means to some end, not a theory of desiring to take the bus today in order to benefit the environment, say, or of desiring to smell the cat litter in order to determine how urgently it needs cleaning […] It is not the aim of RTD to say what instrumental desire is. Intrinsic desire will be desire enough.

As a close theoretical alternative, I have defended the *Valuational Theory of Desire* (VTD) ([xxxx], in prep). Whereas reward theory characterizes desire in terms of a single-place relation, i.e., by identifying desires with *reward*, the valuation theory holds that desire is best expressed as a two-place relation, namely, in terms of both reward *and* expected value, as characterized above (Section 2). Hence, VTD holds that

> For an organism to desire that *P* is for it to subpersonally attribute a subjective *reward* or *expected value* to *P*.

For example, for an agent to desire to drink a cup of coffee is for that agent to subpersonally attribute subjective reward or expected value to drinking a cup of coffee.

Like its reward-based alternative, the valuation-based view can account for the core features of desire, including the valence and strength of desire, the acquisition and extinction of desires, and fleeting as well past desires. However, because it accommodates both reward and expected value - that is, by appealing to the comprehensive computational foundations of reinforcement learning, rather than to the narrower neuroscientific characterization of the so-called 'reward' system - the valuation-based view can account for both intrinsic *and* instrumental desires, whereas its reward counterpart can only explain the former.

Building on these kinds of views, it may be the case that we should eventually *replace* the philosophical folk psychological notion of desire with the technical notions of reward and expected value. The advantage of doing so will be that, rather than nesting the technical notions of reward and punishment within the more familiar philosophical notion of desire, these notions instead play an explicit role in resolving puzzles and debates in the philosophy of action, such as those regarding weakness of the will, synchronic self-control, and attributions of blameworthiness in the context of addiction ([xxxx], 2018, 2020). That is, by substituting the notion of desire with the technical notions of reward and expected value, we will be able to draw on these notions' explanatory power directly in order to address our philosophical concerns.

Conclusion

This opinionated review has sought to provide one roadmap for philosophers hoping to engage with the principles and findings associated with RLDM, particularly the central notions of reward and values, as highlighted in the foregoing philosophical discussion of perception and desire.

References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53-63.

Anderson, B. A., & Kim, H. (2018). Mechanisms of value-learning in the guidance of spatial attention. *Cognition*, *178*, 26-36.

Arpaly, N., & Schroeder, T. (2014). *In praise of desire*. Oxford University Press.

Balcetis, E., Dunning, D., & Granot, Y. (2012). Subjective value determines initial dominance in binocular rivalry. *Journal of Experimental Social Psychology*, *48*(1), 122-129.

Berns, G. S., Bell, E., Capra, C. M., Prietula, M. J., Moore, S., Anderson, B., ... & Atran, S. (2012). The price of your soul: neural evidence for the non-utilitarian representation of sacred values. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 754-762.

Bostrom, N. (2017). *Superintelligence*. Dunod.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology*, *66*.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625.

Chelazzi, L., Eštočinová, J., Calletti, R., Gerfo, E. L., Sani, I., Della Libera, C., & Santandrea, E. (2014). Altering spatial priority maps via reward-based learning. *Journal of Neuroscience*, *34*(25), 8594-8604.

Chiew, K. S., & Braver, T. S. (2014). Dissociable influences of reward motivation and positive emotion on cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 509-529.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Colombo, M. (2014). Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, *45*, 57-67.

Colombo, M., & Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition*, *112*, 3-12.

Copeland, B. J., & Proudfoot, D. (2006). Part One The History and Development of Artificial Intelligence. In *Philosophy of Psychology and Cognitive Science*, Ed. P. Thagard.

Cubillo, A., Makwana, A. B., & Hare, T. A. (2019). Differential modulation of cognitive control networks by monetary reward and punishment. *Social cognitive and affective neuroscience*, *14*(3), 305-317.

Cummins, D. D. (2000). *Minds, Brains, and Computers: An Historical Introduction to the Foundations of Cognitive Science.* Wiley.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12), 1704-1711.

Dayan, P., & Abbott, L. F. (2001). Theoretical neuroscience: computational and mathematical modeling of neural systems. *Computational Neuroscience Series.*

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185-196.

Dayan, P. (2011). Interactions Between Model-Free and Model-Based Reinforcement Learning,' Seminar Series from the Machine Learning Research Group. University of Sheffield, Sheffield. Lecture recording. <http://ml.dcs.shef.ac.uk/>.

Della Libera, C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological science*, *20*(6), 778-784.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312-325.

Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory. *Trends in cognitive sciences*, *24*(5), 388-397.

Friston, K. (2018). Does predictive coding have a future?. *Nature neuroscience*, *21*(8), 1019-1021.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in behavioral neuroscience*, *9*, 135.

Glimcher, P. W., & Fehr, E. (Eds.). (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). Deep Learning, *Vol. 1* (MIT Press).

Haugeland, J. (Ed.). (1997). *Mind design II: philosophy, psychology, artificial intelligence*. MIT press.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4), 188-194.

Hickey, C., Chelazzi, L., & Theeuwes, J. (2010). Reward guides vision when it's your thing: Trait reward-seeking in reward-mediated visual priming. *PloS one*, *5*(11), e14087.

Hikosaka, O., Nakamura, K., & Nakahara, H. (2006). Basal ganglia orient eyes to reward. *Journal of neurophysiology*, *95*(2), 567-584.

Hohwy, J. (2013). *The predictive mind.* Oxford: Oxford University Press.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, 108(3), 687-701.

Huebner, B. (2012). Surprisal and valuation in the predictive brain. *Frontiers in psychology*, *3*, 415.

Huebner, B. (2015). Do emotions play a constitutive role in moral cognition?. *Topoi*, 34(2), 427-440.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol*, *7*(4), e1002028.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol*, *8*(3), e1002410.

Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual review of neuroscience*, *38*, 1-23.

Hyman, S. E. (2005). Addiction: a disease of learning and memory. *American Journal of Psychiatry*, *162*(8), 1414-1422.

Juechems, K., & Summerfield, C. (2019). Where does value come from?. *Trends in cognitive sciences*, *23*(10), 836-850.

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: consensus and controversy. *Neuron*, 63(6), 733-745.

Katai, S., Maruyama, T., Hashimoto, T., & Ikeda, S. (2003). Event based and time based prospective memory in Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*(6), 704-709.

Kitcher, P. (1982). *Abusing science: The case against creationism*. MIT Press.

Kliegel, M., Phillips, L. H., Lemke, U., & Kopp, U. A. (2005). Planning and realisation of complex intentions in patients with Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(11), 1501-1505.

Krishnan, H. S., & Shapiro, S. (1999). Prospective and retrospective memory for intentions: A two-component approach. *Journal of Consumer Psychology*, *8*(2), 141-166.

Mackintosh, N. J. (1983). *Conditioning and associative learning* (p. 316). Oxford: Clarendon Press.

Mahmut, M. K., Homewood, J., & Stevenson, R. J. (2008). The characteristics of non-criminals with high psychopathy traits: Are they similar to criminal psychopaths?. *Journal of Research in Personality*, *42*(3), 679-692.

Marx, S., & Einhäuser, W. (2015). Reward modulates perception in binocular rivalry. *Journal of vision*, *15*(1), 11-11.

T. Metzinger & W. Wiese (Eds.). (2017). *Philosophy and Predictive Processing*. Frankfurt: MIND Group.

Montague PR (2007). *Why Choose This Book?* New York: Penguin.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual review of neuroscience*, *34*, 333-359.

Padoa-Schioppa, C., & Schoenbaum, G. (2015). Dialogue on economic choice, learning theory, and neuronal representations. *Current opinion in behavioral sciences*, *5*, 16-23.

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & memory*, *24*(1), 65-69.

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *The Behavioral and brain sciences*, *31*(4), 415.

Redish, A. D. (2013). *The mind within the brain: How we make decisions and how those decisions go wrong*. Oxford University Press.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist*, 43(3), 151.

Savine, A. C., & Braver, T. S. (2010). Motivated cognitive control: reward incentives modulate preparatory neural activity during task-switching.

Silver, D. (2015). *Reinforcement learning* [Video]. YouTube. Available at: https://www.youtube.com/watch?v=2pWv7GOvuf0

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

Schroeder, T., & Arpaly, N. (2013). Addiction and blameworthiness. *Addiction and self-control. Oxford University Press, New York*, 214-238.

Schultz, W. (2015). Neuronal reward and decision signals: from theories to data. *Physiological reviews*, *95*(3), 853-951.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, *529*(7587), 484-489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. Nature *550*, 354–359.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanc- tot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforce- ment learning algorithm that masters chess, shogi, and Go through self-play. Science *362*, 1140–1144.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, *3*(1), 9-44.

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

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Walter, S., & Meier, B. (2014). How important is importance for prospective memory? A review. *Frontiers in psychology*, *5*, 657.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one*, *8*(7), e68210.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS one*, *8*(7), e68210.

Westbrook, A., Lamichhane, B., & Braver, T. (2019). The subjective value of cognitive effort is encoded by a domain-general valuation network.*Journal of Neuroscience*, *39*(20), 3934-3947.

Wilbertz, G., van Slooten, J., & Sterzer, P. (2014). Reinforcement of perceptual inference: Reward and punishment alter conscious visual perception during binocular rivalry. *Frontiers in psychology*, *5*, 1377.