INVITED PAPER

The Misbehavior of Reinforcement Learning

In this paper, the authors compare two separate approaches to operant learning in terms of computational power and flexibility, putative neural correlates, and the ability to account for human behavior as observed in repeated-choice experiments.

By Gianluigi Mongillo, Hanan Shteingart, and Yonatan Loewenstein

ABSTRACT | Organisms modify their behavior in response to its consequences, a phenomenon referred to as operant learning. The computational principles and neural mechanisms underlying operant learning are a subject of extensive experimental and theoretical investigations. Theoretical approaches largely rely on concepts and algorithms from reinforcement learning. The dominant view is that organisms maintain a value function, that is, a set of estimates of the cumulative future rewards associated with the different behavioral options. These values are then used to select actions. Learning in this framework results from the update of these values depending on experience of the consequences of past actions. An alternative view questions the applicability of such a computational scheme to many real-life situations. Instead, it posits that organisms exploit the intrinsic variability in their action-selection mechanism(s) to modify their behavior, e.g., via stochastic gradient ascent, without the need of an explicit representation of values. In this review, we compare these two approaches in terms of their computational power and flexibility, their putative neural correlates, and, finally, in terms of their ability to account for behavior as observed in repeated-choice experiments. We discuss the successes and failures of these alternative approaches in explaining the observed patterns of choice behavior. We conclude by identifying some of the important challenges to a comprehensive theory of operant learning.

KEYWORDS | Computational intelligence; decision making; gradient methods; learning (artificial intelligence); learning systems; machine learning; Markov decision process; neural networks; reinforcement learning

I. INTRODUCTION

> "He who spares the rod hates his son, but he who loves him is careful to discipline him"—Proverbs 13:24.

Operant learning refers to a process of behavior modification in which the likelihood of a specific behavior is increased or decreased through positive or negative reinforcement, each time the behavior is exhibited. Operant learning has been practiced for millennia. Animals have been trained to assist humans in work and war for many centuries, indicating that the use of *carrots and sticks* to shape behavior is certainly not new [1]. These insights were not restricted to animal training. For example, Assyrian parents, more than 2500 years ago, were encouraged to use canning as means of educating children, i.e., of inducing long-term changes in their behavior: "Spare the rod, spoil the child" (Ahiqar 6:81) [2], a practice which is today both illegal in many countries and is strongly discouraged [3].

By contrast, *quantitative* studies on how rewards and punishments shape behavior have awaited the seminal work of the American psychologist Edward Thorndike at the end of the 19th century. Thorndike placed cats in small cages and measured the time it took the animals to open the cage, a feat requiring a particular action from the animal such as a lever press. The study of how the escape time decreased with practice in different conditions, as well as other experiments, led Thorndike to formulate the

Manuscript received August 31, 2013; revised February 12, 2014; accepted February 12, 2014. Date of publication March 14, 2014; date of current version March 25, 2014. This work was supported by the Israel Science Foundation under Grant 868/08, a grant from the Ministry of Science and Technology, Israel, and the Ministry of Foreign and European Affairs, the Ministry of Higher Education and Research, France, and the Gatsby Charitable Foundation.

G. Mongillo is with the Laboratory of Neurophysics and Physiology, Paris Descartes University, Paris, France and also with CNRS, UMR 8119, 75006 Paris, France.
H. Shteingart is with the Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel.
Y. Loewenstein is with the Department of Neurobiology, the Alexander Silberman Institute of Life Sciences, the Edmond and Lily Safra Center for Brain Sciences, the Department of Cognitive Science, and the Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem 91904, Israel (e-mail: yonatan.loewenstein@mail.huji.ac.il).

Digital Object Identifier: 10.1109/JPROC.2014.2307022

0018-9219 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Law of Effect: "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur" [4].

The theoretical foundations to the understanding of operant learning were laid in the middle of the 20th century in three lines of research. First, simple quantitative operant learning experiments in humans and animals have motivated mathematical psychologists to construct quantitative phenomenological models of operant learning in these experiments [5]. At the same time, the pioneers of artificial intelligence began to explore trial-and-error learning as an engineering principle. Finally, developments in the field of optimal control, most notably the development of dynamic programming by Richard Bellman, enabled the later development of what is today standard reinforcement learning (RL) techniques [6].

In this review, we examine some of the different models for operant learning in view of the observed behavior of animals and humans, and briefly discuss the neural correlates of this learning behavior. The review is organized in the following way. In Section II, we discuss alternative models for operant learning that are motivated by normative considerations. We also discuss their putative or plausible neural basis. In Section III-A, we discuss the results of discrete-trial operant learning experiments and relate them to the existing models. In Section III-B, we discuss free-operant learning experiments and the difficulty of relating them to existing models. In Section IV, we discuss phenomenological models. In Section V, we conclude by identifying some of the important challenges to a comprehensive theory of operant learning.

II. REINFORCEMENT LEARNING

In this section, we give a short, and incomplete, overview on RL. Its aim is to introduce some basic concepts and learning algorithms which will provide a framework for the discussion of behavioral experiments in Section III. The interested reader is referred to [6] and [7] for a more general and comprehensive treatment of these topics. We also shortly survey the literature about putative neuronal substrates of these RL algorithms.

The RL problem can be formulated as follows. Consider an agent interacting with an environment through consecutive *perception–action* cycles. In each cycle, the agent gathers information about the environment (referred to as an *observation*) and performs an action. The action can have both immediate and long-term consequences. The immediate consequence is that following the action, the agent receives a reward (a scalar signal). However, the action can also affect the environment, and thus affect future observations, actions, and rewards. The goal of the agent is to choose actions so as to maximize some measure of the overall collected rewards, for instance, the average future reward per action.

RL is a collection of methods devised to find the optimal policy, a (possibly stochastic) mapping from observations to actions, that realizes the goal of the agent. The applicability and/or the effectiveness of the different methods depend on the complexity of the agent-environment interaction. Roughly speaking, this complexity depends on how well the agent can predict (in a statistical sense) the effects of its actions on the environment given the observations.

In some RL problems, the next state of the environment and the reward obtained is a (possibly stochastic) function of *only* the current state and action. An RL problem which satisfies the above *Markov property* is called a Markov decision process (MDP) [Fig. 1(a)]. If the agent has access to this state (e.g., the current observation reveals the current state of the environment) then the agent can select actions optimally by considering only the current state.

By contrast, there are situations where the observations fail to disclose the true (hidden) state of the environment. For instance, this could happen because the agent receives only partial information about the current state of the environment. The information is partial in the sense that it does not allow the agent to unambiguously distinguish among different states. This, in general, severely complicates the learning and control problems, as described below. A model for this kind of situations is obtained by assuming that the probability of a given observation depends on the current (hidden) state of the environment. Such an RL problem is called partially observable MDP (POMDP).

A. Value-Based Learning

In the case of finite MDPs (i.e., with a finite number of states and actions), one can univocally associate to each state–action pair a *value*. The value of a state–action pair, under a given policy, is a prediction about the future cumulative reward the agent will obtain by taking that action in that state. To improve the policy, the agent can then search for actions that increase the value function. Hereafter, we refer to RL methods that search the policy space via the value function as value-based methods.

Learning algorithms in this class determine the optimal policy by computing the values of the different actions for each possible state. Let $Q(a,s)$ be the future cumulative reward the agent will obtain by taking action a in state s, and then following the current policy. Because of the Markov property, i.e., the probability of reaching a given state depends only on the current state and on the action taken, the Q's satisfy the following consistency conditions:

$$Q(s,a) = r(a,s) + \gamma \sum_{s',a'} T(s',a,s)\pi(a',s')Q(s',a') \quad (1)$$

where $r(a,s)$ is the average reward from action a in state s, $T(s',a,s)$ is the probability that the next state will be s'

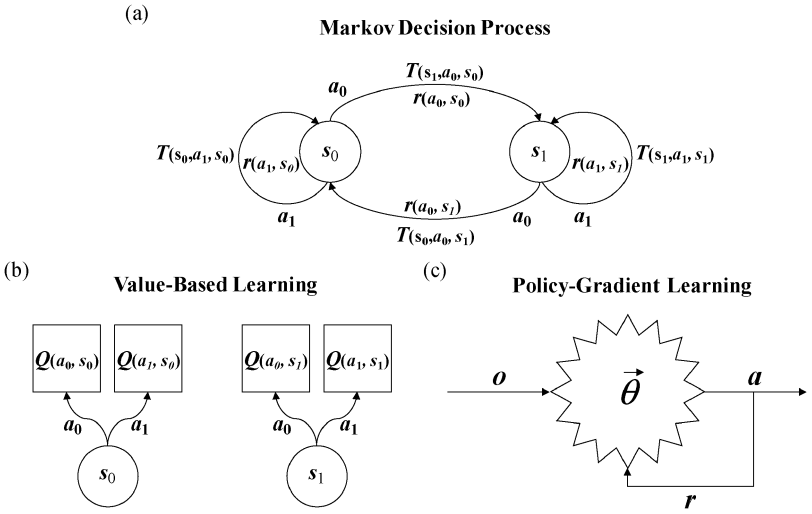

Fig. 1. *(a) Markov decision process example: For clarity, only some of the possible transitions and reward contingencies are represented. (b) Value-based learning: The Q's are an estimate of the average future cumulative reward obtained by choosing a given action in a given state. (c) Policy-gradient learning: Policy parameters are adjusted by gradient ascent so as to maximize the average reward. See main text for details.*

after taking action a in state s, and $\pi(a', s')$, which describes the policy, is the probability of taking action a' in state s'. The parameter γ denotes the discounting of future rewards. For simplicity, we assume that there is no temporal discounting and $\gamma = 1$.[1] To improve the policy, the agent can choose in each state the action which maximizes the right-hand side of (1). This operation is called *policy improvement*. The policy improvement theorem guarantees that the new policy is not worse than the old one, and it is better unless the old policy is the optimal one [6]. This changing the policy, however, will change the values of the states which have to be recomputed under the new policy. This operation is called *policy evaluation*. By interleaving improvement and evaluation, which is called *policy iteration*, one is guaranteed to find the optimal policy in a finite number of iterations.

Policy evaluation using (1) requires the knowledge of the transition probabilities $T(s', a, s)$ and of the rewards contingencies $r(a, s)$. Thus, one possibility for the agent is to learn the transition probabilities and the reward contingencies (i.e., learn the model of the environment), and then to use the model to *compute* the optimal policy, as described above. Methods of solution that explicitly learn the contingencies of the environment, that is, the parameters of the MDP model, are referred to as model-based methods [6], [7]. Alternatively, the agent can *directly* estimate the state–action values [the Q's in (1); Fig. 1(b)], and use these values for action–selection. Methods of solution that do not require learning a model of the environment (e.g., in terms of state transition probabilities) are referred to as model-free methods. Here, we focus on model-free methods as these dominate the field of operant learning.

[1] $\gamma < 1$ is needed to guarantee the existence of the value function in the case of continuing tasks.

The basic idea behind model-free RL is to solve (1) by stochastic approximation, while using estimates of the values to choose actions. In this section, we describe a popular algorithm, called SARSA, that achieves this objective [6]. We consider the sequence of states, actions, and rewards of an agent interacting with an environment. In each cycle, the agent, being in state s and taking action a, updates its estimate of the corresponding state–action value function $Q(a, s)$ according to

$$Q(a, s) \leftarrow Q(a, s) + \eta\delta \tag{2}$$

where $\eta > 0$ is the learning rate, and $\delta \equiv r + Q(a', s') - Q(a, s)$ is the reward prediction error (RPE), with a' and s' being the next action and the next state, respectively, and r is the obtained reward in the cycle. The RPE is a basic quantity that plays a central role in all value-based methods. Roughly speaking, it is a measure of how good the agent is at predicting the consequences of its behavior. A positive RPE is a *good surprise* for it indicates that, as a consequence of taking action a in state s, the agent received a larger-than-expected reward and/or reached a state with larger-than-expected value. That is, the agent was *underestimating* the value of taking action a in state s. Similarly, a negative RPE is a *bad surprise* because lower-than-expected reward has been received and/or a state with lower-than-expected value has been reached, indicating that the agent was *overestimating* the value of taking action a in state s.

Note that the update in (2) requires information about the state in the beginning of the cycle s, the action taken a, the immediate reward r, the next state s', and the next action a' $(s \rightarrow a \rightarrow r \rightarrow s' \rightarrow a')$, giving this algorithm its

name, SARSA. If the policy is kept fixed while η is properly decayed to 0 (see [8] for a more precise statement) and each state–action pair is sampled infinitely often, then (2) is bound to converge to the solution of (1) with probability 1.

The policy can be improved concurrently with the estimation of the Q function by making action–selection dependent on the current estimate of the Q's. Typically, this is achieved by utilizing a stochastic policy that balances *exploitation* and *exploration*. Exploitation corresponds to choosing the action associated with the highest value with the rationale that if the values are accurate, then a *greedy* policy (always choosing actions associated with the highest value) will maximize the average cumulative reward. By contrast, when the values are not accurate, exploration, which corresponds to choosing actions that are currently suboptimal, is useful for improving the current estimates. The most widely used action–selection functions, both in applications and when explaining operant learning, are the *ϵ-greedy* and the *soft-max* action–selection functions. In ϵ-greedy, the agent chooses the action a associated with the higher value [i.e., a for which $Q(a,s)$ is maximal] with a probability $1-\epsilon$ and chooses randomly and uniformly between all actions with a probability ϵ. In soft-max, the probability of selecting action a in state s is given by

$$\pi(a,s) = \frac{\exp[\beta Q(a,s)]}{\sum_{a'} \exp[\beta Q(a',s)]} \tag{3}$$

where the sum is over all action a' available in state s, and $\beta > 0$ is a parameter controlling the stochasticity of action–selection. The parameters ϵ and $1/\beta$ control the tradeoff between exploration and exploitation. The smaller these parameters are, the more dominant is exploitation.

Compared to MDPs, POMDPs are computationally more difficult to solve. Even if the agent has a complete knowledge of the dynamics of the POMDP, finding the optimal policy in the general case is actually impossible—a completely general exact algorithm could be used to solve the halting problem [9]. One solution is to use SARSA or SARSA-like algorithms, treating the observations as if they were states (i.e., as if they satisfied the Markov property). However, this algorithm may converge to a solution that is far from optimality or may fail to converge altogether [10], [11]. For a survey of POMDPs solution techniques, see [12] and references therein.

1) Putative Neural Correlates: The largely dominant hypothesis within the fields of neuroscience and neuroeconomics is that the implementation of value-based algorithms by the brain underlies much of operant learning. It has been suggested that different brain regions support the different computations required for value-based learning [13]. The cerebral cortex, and more specifically the prefrontal regions, learn and represent the "states of the world," which are task relevant [14], [15]. The values of the states, or of the state–action, pairs, are learned and represented in the basal ganglia, which is a subcortical structure known to be involved in action–selection [16]. Different proposals further partition the basal ganglia into different functional components. According to one proposal, the values are coded in the striatum (the input structure of the basal ganglia) and are directly used to select the actions [17], [18]. Another proposal suggests, instead, that the values coded in the striatum are used to update the policy which would be maintained in the nucleus accumbens (which is a subdivision of the striatum) and in the pallidum (which is the output structure of the basal ganglia) [19], [20]. The evidence supporting these hypotheses is largely based on experiments demonstrating that the neural activity in these brain regions is correlated with some of the variables of the value-based RL model, as computed by fitting the RL model to the experimentally measured behavior. For a detailed discussion of this methodology, see [21].

However, the most influential support to the hypothesis that the brain implements value-based RL stems from a series of beautiful experiments suggesting that a particular group of neurons, the dopamine neurons located in the midbrain, encodes the RPE (for review, see [22]–[24]). In these experiments, monkeys were trained in a classical conditioning paradigm, in which, repeatedly, visual or auditory cues were followed by a reward, e.g., a drop of juice. Note that in classical conditioning, unlike operant conditioning, the reward is contingent upon the stimulus rather than upon the response [25]. In the naive animals, dopamine neurons respond with a transient increase of their firing rates to reward delivery but not to cue presentation. By contrast, in the trained animals, they respond to the cue but not to the reward. Remarkably, in the trained animal, dopamine neurons respond with a transient decrease in their firing rate to reward omission after the cue. Theoretical modeling reveals that this pattern of activity is to be expected if the activity of these neurons represents the RPE [22]. A more recent study has demonstrated a causal link between the dopamine signal and operant learning by replacing reward delivery with optogenetic activation of dopamine neurons in mice. The results of this study provide a strong support for the role of dopamine neurons in driving operant learning [26].

The pattern of dopamine neurons activity, in the specific experimental conditions described above, does undeniably mimic the behavior of a signal putatively related to RPE [22]–[24]. However, in other respects, that same pattern of activity seems inconsistent with a *bona fide* RPE signal. One evident problem with the hypothesis of a one-to-one relation between the (phasic) activity of dopamine neurons and the RPE is the *asymmetry* with which positive and negative errors can be signaled. The ability to signal negative errors (through a transient decrease of the firing rates) is significantly lower than the ability to signal positive errors (through a transient

increase of the firing rates), due to the low baseline firing rates of dopamine neurons (∼3–5 Hz) [27], [28]. Two workarounds to this problem have been suggested. One is that the relationship between dopamine firing rate and RPE is strongly nonlinear (however, see [29]), and the other is that a different system is in charge of signaling negative RPE, with the dopamine system primarily signaling *only* positive RPE [30].

Another experimental result that appears inconsistent with the dopamine-RPE hypothesis is the finding that dopamine neurons also respond with a transient activity increase to salient stimuli not associated with reward as well as to stimuli predicting negative reward [27], [31]. These, and the short latency in dopamine response, suggest that dopamine neurons activity might be instrumental in discovering stimuli and/or motor responses that could be task relevant, rather than encoding the RPE [27], [32]. Finally, given the substantial heterogeneity in the responses of dopamine neurons, it is unclear how such a diverse neural population could broadcast a global signal as the RPE [28].

B. Policy-Gradient Learning

One alternative approach for finding the optimal policy for an MDP is by searching the policy space without the *intermediate* step of computing the value function. A widely used and effective way of performing such a search is to consider a suitable parametric family of policies, and then find the optimal parameters by gradient ascent (see [33], [34], and references therein). Hereafter, we refer to RL methods that *directly* search the policy space by gradient ascent as policy-gradient methods. An important advantage of policy-gradient methods over value-based methods is that they retain their convergence guarantees under very general conditions when applied to POMPDs [12].

For purpose of illustration, we consider below the application of policy-gradient methods to a simplified problem, in which the reward depends only on the most recent observation and action, and the observations are temporally independent and generated according to some time- and action-independent distribution [35]. There are generalizations to the policy-gradient approach that relax these assumptions. These are, however, beyond the scope of this paper. The interested reader is again referred to [33], [34], and references therein.

Let $\pi_{\boldsymbol{\theta}}(a,o)$, the probability of taking action a upon observation o, be a policy parametrization, and $\boldsymbol{\theta}$ be the vector of free parameters to be optimized. The average reward per action obtained by following policy $\pi_{\boldsymbol{\theta}}$ is given by

$$R(\boldsymbol{\theta}) = \sum_{o,a} p(o)\pi_{\boldsymbol{\theta}}(a,o)r(a,o) \tag{4}$$

where $p(o)$ is the probability of observing o, and $r(a,o)$ is the average reward obtained by taking action a upon observation o. Performance can be improved by iterating

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) \tag{5}$$

where $\eta > 0$ is the learning rate. If the learning rate converges to zero slowly enough [8], then such learning is guaranteed to converge to a local maximum of the average reward. There are different approaches to estimate the gradient of $R(\boldsymbol{\theta})$. One approach consists in making small perturbations to the parameters and estimating the average reward obtained by following the corresponding policy over a suitably long time interval. From the average reward estimates so obtained, one can then compute an estimate of the gradient $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$ by finite-difference methods. This is equivalent to *batch* learning, where parameters are changed only after a large amount of *training data* (i.e., consecutive observation–action–reward triplets) has been experienced. An alternative approach, which is also more biologically plausible, consists in changing the parameters as soon as new experience is acquired. This is equivalent to *online* learning, where parameters of the policy are changed after each observation–action–reward triplets.

A class of online algorithms can be derived by noting that

$$\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \sum_{o,a} p(o)r(a,o)\pi_{\boldsymbol{\theta}}(a,o)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a,o). \tag{6}$$

Therefore, if one changes $\boldsymbol{\theta}$ according to

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta r \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a,o) \tag{7}$$

every time that action a is taken upon observation o, then the average change in the parameters is proportional to $\nabla_{\boldsymbol{\theta}} R(\boldsymbol{\theta})$ [see (6)]. Equation (7) is a special form of a general class of online learning rules called REINFORCE algorithms [35].

1) Putative Neural Correlates: The use of gradient-based techniques for learning in artificial systems, and particularly in artificial neuronal networks, has a long and successful history [36], [37]. It is unclear, however, whether biological systems implement this kind of learning and, if that is the case, through which mechanisms. In order to gain insight into potential mechanisms, we focus first on the process of action–selection.

It is generally believed that action–selection emerges from competition between different neural populations, each coding for a different motor response. As a result of this competition, the neurons of the *winning* population become active, firing at high rates, while the neurons corresponding to the *losing* populations are quiescent. Consequently, the action associated with the winning population is executed [38]. The substantial neural variability plays an important role in this competition process [39]. This variability manifests itself as variability

in the outcome of the competition, naturally implementing a stochastic policy. The outcome of the competition is influenced by the inputs to the different populations. These inputs may carry sensory and memory information about the current state of the world and thus enable the system to respond differently in different circumstances. Finally, the properties of the competing networks, e.g., the relative strengths of the intrapopulation (positive) and interpopulation (negative) feedbacks can bias the network in favor and against different populations.

In this framework, any physiological mechanism that modulates the winner-take-all dynamics underlying the action–selection process would result in a change in the policy of the agent, thus affecting the likelihood that a given action is selected in a specific situation. One plausible candidate mechanism is synaptic plasticity. It is well established that the efficacies of synapses change as a function of the activities of the corresponding presynaptic and/or postsynaptic neurons [40]–[42]. There is also evidence that activity-dependent synaptic plasticity is modulated by the reward-dependent dopamine signal [43]–[46]. This raises the possibility that gradient-based algorithms operates already at the level of synapses, in other words, that the *tunable parameters* in (7) are the synaptic efficacies. Note, however, that it is possible that gradient learning is a good description of operant learning but that its implementation in the nervous system is not through synaptic modifications but at a more macroscopic level [Fig. 1(c)].

Policy-gradient learning may be achieved by exploiting the network internal stochasticity or variability. Roughly speaking, the idea is that neural variability results in policy variability, which in turn results in variability in the rate of delivered reward. Policy-gradient learning is achieved if, on average, synaptic efficacies are changed so as to increase the likelihood of those patterns of network activity that were correlated with an increased rate of rewards. There have been several proposals that implement variants of the REINFORCE family of algorithms that are based on this idea, where the necessary neural variability results from the stochastic release of neurotransmitter [47], from the irregularity of the spiking processes [48], [49], or even from the purposeful injection of synaptic noise from other brain areas [50].

REINFORCE algorithms, when implemented at the microscopic synaptic level, are a special case of a more general class of synaptic plasticity rules, where changes in the synaptic efficacies are driven by the covariance of reward and neural activity [51], [52]. Covariance-driven synaptic plasticity is relatively easy to implement in the biological "hardware," and in many cases converges to the gradient solution [51], [52]. Remarkably, it has been proven that operant matching, which describes the behavior in many operant learning tasks is a generic and robust outcome of covariance-driven synaptic plasticity (see Section III-B) [51], [53], [54].

To test directly whether policy-gradient learning is indeed implemented by synaptic modifications requires quantitative measurements of the *synaptic plasticity rules* in the living brain. However, such experiments are currently technically too demanding and therefore await future research.

C. Further Considerations

For finite MDPs, value-based methods are guaranteed to find the optimal policy while policy-gradient methods are only guaranteed to find a local maximum which may correspond to a suboptimal policy. However, it is important to consider how the qualities of the solutions degrade in face of the "practicalities" required for real-world applications, or when the assumptions about the dynamics of the agent–environment interactions are relaxed.

It is often the case for real-world problems that the observation–action space is very large. This poses several problems to value-based methods. One is the amount of memory one needs in order to store the value function. Another is that, albeit policy iteration allows one to effectively search the policy space for the optimal policy, it is computationally expensive and thus becomes rapidly unpractical as the dimension of the state–action space increases. Finally, during learning, there is the problem of the amount of experience (i.e., time) one needs in order to achieve a reliable estimate of the value function. A standard solution consists in resorting to some form of approximation to represent the value function (usually in the form of some parametric family denoted "function approximation"). In these cases, however, value-based methods are not guaranteed to converge and may even dramatically misbehave (e.g., value function updates diverge with function approximation) [6], [55] (but see [56] and [57]). Policy-gradient methods, instead, are well behaved with function approximation, and they can usually be shown to converge under mild regularity conditions.

Another important real-world situation is the case of POMDPs. As we have already noted, in POMDPs, value-based methods have no convergence guarantees and can even return strongly suboptimal solutions. Partial information, on the other hand, has less severe effects on policy-gradient methods.

III. BEHAVIORAL EXPERIMENTS

While RL methods provide a framework to study operant learning in arbitrarily complex settings, most research on humans' and animals' operant learning has focused on relatively simple repeated-choice experiments. In this paradigm, subjects repeatedly choose among different alternatives (typically two) and are rewarded according to some schedule unknown to them. There are two basic settings for repeated-choice experiments: discrete trials and free operant. In the discrete-trial setting, the experiment is divided into temporally separated trials

and the subject makes a single choice every trial, e.g., by pressing one of several buttons. By contrast, in the free-operant setting, the subject can respond repeatedly without any constraints.

The discrete-trial setting is widely used in conjunction with electrophysiological recordings or functional imaging because this setting allows the experimentalist to finely control the choice time as well as all the preceding behavior in the trial, which is important in order to correlate brain activity with behavior. The free-operant setting, on the other hand, represents a more ecologically relevant condition. In fact, it is often considered that choice behavior in free-operant settings can be likened to foraging behavior of animals in the wild, where they must make choices about foraging locations as well as about the amount of time to spend in them.

We consider below the behavioral results in repeated-choice learning experiments in both discrete-trial and free-operant settings. We discuss the ability of the RL models described above to account for the observed patterns of behavior in these experiments, and whether the observed behavior can differentiate between alternative models.

A. Discrete-Trial Operant Learning

A popular discrete-trial task utilizes the *two-armed bandit* reward schedule. The participant, human or animal, repeatedly chooses between two actions which are rewarded, typically in a binary way (i.e., reward or no reward), with constant probabilities that depend on the actions. The name two-armed bandit reflects the resemblance of these tasks to the problem of choosing between two slot machines in a casino. As predicted by the Law of Effect, with practice, the participants shift their preference in favor of the alternative associated with the higher reward probability (see [5] and references therein). In one of the earlier studies, Grant *et al.* [58] instructed human participants to repeatedly predict whether a lamp would turn on. There were five groups of participants, and the probability that the light would turn on for the different groups was $q_1 = 0$, 0.25, 0.5, 0.75, or 1. In other words, participants chose between two actions, $a_1 =$ "predict ON" and $a_2 =$ "predict OFF," and the probability of reward associated with the two actions was q_1 and $q_2 = 1 - q_1$, respectively. Grant *et al.* recorded the fraction of trials, in which the participants predicted that the lamp would turn on. The fraction of trials (computed over the last five trials of an experiment composed of 60 trials) in which the participants predicted that the lamp would turn on was roughly equal to the probabilities that the lamp would indeed turn on: 0, 0.25, 0.53, 0.77, and 1, respectively. Formally, denoting N_i the number of times alternative i was chosen, $N_1/N_2 \simeq q_1/q_2$. This pattern of choice behavior is commonly referred to as *probability matching* (not to be confused with operant matching mentioned above) because the participant matches her probability of choosing an alternative to the probability that this alternative would be rewarded.

Note that probability matching deviates from the *optimal* policy, the one that maximizes the total number of correct answers. For example, if $q_1 = 0.75$, then responding a_1 on every trial would result in 75% success, on average. By contrast, probability matching would yield the correct answer on average in $0.75 \times 0.75 + 0.25 \times 0.25 = 0.625$ of the trials. The observation that people probability match rather than maximize has attracted the attention of many theorists interested in rational choice theory: "We have here an experimental situation which is essentially of an economic nature in the sense of seeking to achieve a maximum of expected reward, and yet the individual does not in fact, at any point, even in a limit, reach the optimal behavior" [59]. Therefore, this pattern of choice behavior has been studied extensively, yielding contradictory results. The unpleasant truth is that, after more than half of a century of experiments, whether probability matching is an asymptotically stable behavior (i.e., maintained after extensive practice) is still a matter of debate. Gallistel has argued that whether participants match depends on the feedback available [25]. Full feedback about both the reward associated with the chosen action (*obtained* reward) and that associated with the nonchosen action (*foregone* reward) leads to probability matching, whereas participants maximize if the feedback they receive is restricted to the obtained reward [25], [60], [61]. There are claims that, at least in humans, probability matching is not a robust phenomenon and that participants choose the more rewarding alternative more frequently than expected by probability matching [62]. Moreover, it has been argued that the larger the rewards, the stronger is the tendency to maximize [63]. By contrast, there are numerous accounts of probability matching, not only in laboratory settings but also when humans gamble substantial amounts of money on the outcome of real-life situations [64].

It is instructive to consider the asymptotic behavior predicted by the different RL algorithms in the two-armed bandit task. Let us first consider the SARSA algorithm (Section II-A) for an MDP with a single state and two actions a_1 and a_2, randomly rewarded (either $r = 0$ or $r = 1$) with constant probabilities q_1 and q_2, respectively. This is a *minimal* MDP description for the two-armed bandit task (see more on this below). For a sufficiently small learning rate, the values of the two actions will converge to the probabilities of reward associated with the two actions $Q(a_1) \to q_1$ and $Q(a_2) \to q_2$, where we have dropped the dependence on s [see (2)].[2] If the actions are selected according to the soft-max function [see (3)], then the (asymptotic) probability of choosing a_1, $p \equiv \pi(a_1)$, is given by

$$p = \frac{1}{1 + e^{\beta(q_2 - q_1)}}. \tag{8}$$

[2]More precisely, the dynamics will converge to those values plus a constant which represents the amount of future rewards. However, in the model, this will not effect behavior because only the difference between the Q values is used in the action–selection rule.

Thus, the higher the probability of reward associated to an action, the more often that action will be chosen (i.e., $p > 0.5$ for $q_1 > q_2$, and $p < 0.5$ for $q_1 < q_2$). Nevertheless, as can be seen from (8), the less-rewarding alternative will be chosen with nonzero probability, even after a large number of trials (as long as $\beta > 0$ and finite). This behavior is qualitatively consistent with probability matching, but not quantitatively. In fact, according to (8), it will be

$$\frac{N_1}{N_2} = \frac{e^{\beta q_1}}{e^{\beta q_2}} \neq \frac{q_1}{q_2}. \tag{9}$$

If the actions are selected according to the ϵ-greedy function, the asymptotic behavior remains qualitatively the same. The less-rewarding alternative will be chosen with $\epsilon/2 > 0$ probability, but the pattern of choices will still be quantitatively inconsistent with probability matching. In fact, it will be $N_1/N_2 = 2/\epsilon - 1$, assuming that a_1 is the most-rewarding action. It should be noted that there are other value-based algorithms not discussed here, such as actor–critic, whose asymptotic behavior in the two-armed bandit will converge to exclusive choosing of the most rewarding alternative [65].

For comparison, we consider now the asymptotic behavior that would result from policy-gradient learning (Section II-B). Let the policy be parametrized by the probability p of choosing a_1. The average reward as a function of p is given by $R(p) = p \times q_1 + (1 - p) \times q_2$, and the *gradient* with respect to p is given by

$$\frac{dR}{dp} = q_1 - q_2. \tag{10}$$

Thus, performing gradient ascent with respect to p on the average reward [see (5)] will converge to choosing the more rewarding alternative exclusively (i.e., to the optimal policy). The maximum of $R(p)$ is achieved for $p^* = 1$ when $q_1 > q_2$, and for $p^* = 0$ when $q_1 < q_2$. Similarly, performing gradient ascent on any tunable parameter θ, with $p \equiv p(\theta)$, that allows saturation of the probability of choice will converge to the same behavior. Similarly to the pattern of asymptotic choice behavior predicted by SARSA, this behavior also is in contrast with probability matching behavior.

As mentioned above, the *simple* two-armed bandit task elicits quite different patterns of choice behavior depending on the details of the experimental settings. According to theory, these details are predicted to be computationally benign, in the sense that they should have little or no effect in the resulting asymptotic behavior. One possibility is that the details of the experimental setting strongly affect the subjects' *internal model* of the task, thereby producing significant differences in the resulting behavior [66], [67]. Several studies lend support to this hypothesis. For example, one study has demonstrated that human participants probability match when instructed to repeatedly predict whether a lamp will light on. By contrast, they tend toward maximizing when presented with the same sequence of random binary events if they fully understand the stochastic mechanism that maps actions to rewards (but not its parameters) [68]. These results indicate that participants probability match because they suspect that the sequence of events may be nonrandom. Along the same lines, another interesting observation is that 3–4 year old children choose the most rewarding alternative more often than college students, who instead tend to probability match [69]. Again, young children may do better than students because they may be less suspicious. More recently, Laquitaine *et al.* [70] studied choice behavior of monkeys in a two-armed bandit schedule. On average, animals approximately probability match. However, probability matching is not the *typical* behavior. In some sessions, animals maximized, whereas in others they chose the two alternatives with an equal probability. Thus, in these experiments, approximate probability matching is the outcome of averaging over the (substantial) session-to-session variability in the ability to learn to identify and choose the more rewarding alternative. Interestingly, that study also identified a neural correlate of this session-to-session variability. The higher the neural activity in a brain region known as the dorsal putamen (which is part of the basal ganglia) at the beginning of the session, the higher is the performance at the end of the session [70].

In the framework of value-based MDPs, such a sensitivity of the behavior to computationally immaterial variations in the experimental setting could be understood as an *erroneous* modeling of the task by the subject. While from the experimentalist point of view, the two-armed bandit task is a single-state MDP with two actions, from the participants' point of view it could be a POMDP of arbitrarily complex structure [71]. For example, the state could depend on the history of actions [72]. Even the number of possible actions, from the participants' point of view, could be different. In fact, humans and animals are known to develop idiosyncratic and stereotyped superstitious behaviors even in simple laboratory settings, highlighting the difficulty in utilizing the correct model of states and actions in operant learning [73], [74].

B. Free-Operant Learning

In Section III-A, we described operant learning in a discrete-trial design, in which the decision time is dictated by the experimentalist. However, there is a long tradition of free-operant experiments that are devoid of discrete trials. In these experiments, the animal freely moves back and forth between two targets, harvesting rewards that are delivered according to a predefined stochastic schedule. Often, the concurrent variable-interval (VI) schedule is used, and we will focus on these experiments. The concurrent VI schedule is more complicated than the two-armed bandit schedule described above because the choices that the subject makes change the probability of reward. Specifically, a target in this schedule can be either baited or empty. When the subject chooses a baited target, the subject is rewarded immediately

and the target becomes empty. An empty target is re-baited probabilistically such that the time to re-bait is drawn from an exponential distribution. Once baited, a target remains baited until it is chosen. The experimentalist controls the means of the two exponential distributions, thus determining whether a target will be "rich" or "poor." The more time the subject spends in one target, the higher is the probability of obtaining a reward in the *other* target. As a result, animals have an incentive to switch between the two targets, which indeed they do. However, while the policy that maximizes the average reward predicts *regular* alternations between the targets [75], the actual pattern of choice behavior is *irregular*, even asymptotically. This irregularity manifests in the distributions of stay durations in the targets, which are approximately exponential [54], [76]–[78].

This result also highlights an important difference between decisions that are made in discrete time, as in Section III-A, and decisions in free-operant experiments. An exponential-like distribution of stay times implies that, while the subject is at one of the targets, the probability of switching must be infinitesimally small (more precisely, in each small time interval Δt, it must be of the order of Δt). Thus, the subject is continuously choosing between action "*stay*," which has a finite probability [i.e., $1 - O(\Delta t)$] to be selected, and action "*leave*," which has instead an $O(\Delta t)$ probability of being selected. In the framework of value-based learning, this could be achieved if the parameter that controls exploration in the action–selection function (see Section II-A) is of the order of Δt. An alternative possibility is to assume that the actions correspond to choosing the time at which to leave. In this case, the set of possible actions becomes infinite.

Another interesting observation reported in these experiments is that the fraction of the total time subjects spend in a target matches the fraction of rewards harvested from that target, a behavior known as Herrnstein's matching law or operant matching [54], [79]–[81]. Despite the similarity in the name (which led to a lot of confusion over the years), probability matching and Herrnstein's matching law are not the same phenomena. In fact, they are inconsistent. To see that, we reconsider the case of the two-armed bandit schedule in which targets 1 and 2 yield a binary reward with (fixed) probabilities q_1 and q_2, respectively. The average number of rewards harvested from target i, I_i, also known as the income, is given by $I_i = q_i \times N_i$. According to probability matching, $N_1/N_2 = q_1/q_2 = (I_1/I_2)^{1/2}$. By contrast, according to Herrnstein's matching law, $N_1/N_2 = I_1/I_2$. It should also be noted that in the case of a two-armed bandit schedule with fixed probabilities of reward, because $I_i = q_i \times N_i$, Herrnstein's matching law equation can only be satisfied if $N_1 \times N_2 = 0$. In other words, in the case of fixed probabilities of reward, the only behavior that is consistent with Herrnstein's matching law is choosing one of the alternatives exclusively. Therefore, maximizing behavior, but not probability matching, observed in some of the discrete-time two-armed bandit experiments, is consistent with Herrnstein's matching law [72].

Operant matching has been repeatedly demonstrated not only in free-operant tasks [54], [78], [81] but also in discrete-trial tasks [81], [82], not only in the laboratory but also in free ranging animals [83]. Nevertheless, deviations from this rule have also been observed [80]. Baum [84] has proposed a generalized form of the matching law. In its symmetrical form, the generalized matching law predicts that $N_1/N_2 \simeq (I_1/I_2)^{\alpha}$ where α is a parameter. Typically, when estimated from behavioral data, $\alpha < 1$, which corresponds to a pattern of choice behavior called under-matching [80]. Note that both operant matching and probability matching adhere to the generalized matching law, with $\alpha = 1$ in the former and $\alpha = 1/2$ for the latter.

Compared with discrete-time operant learning, continuous-time operant learning has received little theoretical attention, in particular in view of standard RL algorithms. This may be due to the difficulty in accounting for behavior in free-operant experiments using value-based RL, as mentioned above. However, it turns out that the pattern of choice behavior observed in free-operant experiments is readily explainable in the framework of covariance-based synaptic plasticity, discussed in Section II-B. It turns out that operant matching naturally emerges from covariance-based synaptic plasticity [51], which is closely related to policy-gradient RL. The same framework can also naturally explain under-matching, as resulting from mistuning of the parameters of the covariance-based plasticity [53].

A recent study has investigated the pattern of choice behavior resulting from a network model composed of two competing neuronal subpopulations (corresponding to the choices of the two targets; see above) in the presence covariance-based synaptic plasticity. The pattern of choice behavior of the model reproduces in detail many of the characteristics of the experimentally observed behavior, such as the exponential distribution of stay durations and the operant matching behavior. It also quantitatively accounts for the dynamics of learning in response to changes in the parameters of the reward schedule [54].

IV. PHENOMENOLOGICAL MODELS OF OPERANT LEARNING

The RL framework discussed in Section II makes explicit assumption about the algorithms used by the brain in order to relate the history of choices and rewards to the probability of choosing an action. However, these approaches may be too restrictive. An alternative approach is to determine the mapping of that history to actions directly from the data, without assuming any particular model. In two noteworthy examples of this approach [85], [86], monkeys were trained in the concurrent VI schedule to repeatedly choose between two targets for a liquid reward. Studying many tens of thousands of choices made by each monkey over many days, the two groups of

researchers constructed linear–nonlinear probabilistic models [37] of the monkeys' behavior. In both studies, it was found that the probability of choosing an alternative action is well approximated by a function of the difference between the rates of reward associated to the two alternatives [85], [86]. It is not clear how to interpret these results in view of value-based RL. If the participant's internal model of the task is a single-state two-action MDP, and the actions' values are learned by SARSA, then one would predict that the probability of choice would depend on the difference in the average *returns* (rate of rewards divided by rate of choices) associated with the two targets and not on the difference in the rate of rewards. However, concurrent VI schedule is a rather complicated POMDP, and, therefore, it is difficult to draw concrete conclusions to the applicability of value-based RL to this problem. One important issue worthwhile considering when attempting to construct such phenomenological models of behavior is that these models require a considerable number of trials, which are collected over many days. However, if the behavior is not stationary over the period in which the data were collected, then the resultant model will necessarily be inaccurate. Indeed, a recent study has reanalyzed the behavior of the monkeys in one of the studies described above [86] and demonstrated substantial nonstationarity over multiple time scales [87].

Another approach is to consider simple, biologically and computationally plausible *heuristics* to account for patterns of behavior observed during operant learning. For example, a recent study has found that the predictive power of the simple win-stay/lose-shift, in which a participant tends to repeat after a positive reward and tends to switch after a negative reward in a four-armed bandit schedule is comparable to that of RL models [88]. In a comprehensive study of a very large number of human participants in different operant tasks, Erev and Haruvy have proposed a complex phenomenological model that accounts for many different characteristics of behavior in these experiments [89].

In contrast to the success of phenomenological models over RL models, a study of a very large data set of 200 human participants making 240 000 two-armed bandit choices has tested the power of SARSA-like algorithms to describe and predict human operant learning. As a first step, the action–selection function used by humans was characterized nonparametrically and was shown to be well approximated by a combination of ϵ-greedy and soft-max. That study demonstrated that SARSA-like algorithms describe the behavior better than competing heuristics, if the model assumes that first experience resets the initial conditions in (2) that describes the dynamics of learning the values, and if the experimentally measured action–selection function is used [90]. Nevertheless, this model does not explain all aspects of behavior. For example, it has been demonstrated that surprising positive payoffs reduce the rate of repeating the previous choice and surprising negatively payoffs increase it, both in the stock market and in simple repeated two-alternative force choice tasks (two-armed bandit) [91]. This result does not seem to be explainable by standard RL algorithms, and is even a challenge to the Law of Effect.

V. CONCLUDING REMARKS

In this review, we considered two families of RL models. Value-based learning that is well suited to learn the optimal policy in MDPs and policy-gradient learning (which is a direct policy search method) that is more flexible, being also applicable to POMDPs. The attempt to apply these models of learning to explain patterns of choice behavior in repeated-choice experiments yields mixed results.

Value-based learning accounts for some aspects of behavior and neural activity in discrete-time operant learning but leaves other unexplained. Policy-gradient learning, implemented using covariance-based synaptic plasticity, can be used successfully to explain behavior in free-operant learning in some experiments, but does not fully account for the behavior in others. The phenomenological models we discussed fare no better. Similarly to RL models, they account for some aspects of behavior and fail at explaining others. We are forced to conclude that, after almost a century of intense experimental and theoretical investigations, we are still far from understanding, both computationally and at the level of neurons and synapses, learning behavior in simple (apparently simple, one would be tempted to say) situations as the repeated-choice experiments discussed above.

Additional advancements in RL theory will most certainly further our understanding of the computations underlying operant learning. One interesting example that has received attention recently is hierarchical RL. This method breaks the learning task into a hierarchy of simpler learning problems [92]. There is even some neuroimaging evidence suggesting that, in fact, the brain may utilize a similar approach [93].

An important issue to consider in applying RL models to explain behavior (which is instead often neglected) is that of identifying the states and the actions which are relevant to the task. Models of operant learning typically take as given that the subject *knows* what are the relevant states and actions. However, identifying the states is a difficult task in the laboratory, and to a larger extent in real life [94]. Along these lines, it has been suggested that operant learning is a two-step process. In the first step, a low-dimensional set of states is identified in an unsupervised way, based on the statistical properties of the inputs. The second step utilizes RL algorithms to find the optimal policy given the set of states so extracted [95]. A plausible alternative is that the relevant state–action sets and the policy are learned in parallel [96].

Over the past few years, there has been a shift in emphasis from value-dependent reinforcement learning to

more generic formulations of inference. This is reflected in several attempts to cast RL and optimal control in terms of (Bayes) optimal inference [97]–[100]. Perhaps the best example of this is the notion of active inference, in which rewards are absorbed into the problem of inferring hidden states of the world (cf., POMDP) by associating them with prior beliefs about the states an agent should occupy [98]. By converting the RL problem into an inference problem, one can then call upon a plethora of neuronally plausible schemes for approximate Bayesian inference [99]. This may be an important development from the current perspective, because active perceptual inference, and its neuronal implementation, can be formulated using the same sort of gradient–ascent schemes that we have discussed in the context of RL. As such, they provide a direct link to cognitive dynamics, a link noted by Kalman a half century ago, when he emphasized the formal equivalence between optimal control, Kalman filtering, and, implicitly, Bayesian belief updating. Furthermore, treating RL as an inference problem allows one to cast "heuristics" as prior beliefs, placing them in a formal and normative framework.

We believe that part of the failure of RL models stems from a more fundamental reason. The RL stance is essentially a behaviorist one, in that it depicts the organism as a general-purpose learning system whose behavior can be *arbitrarily* shaped via stimulus–response–reward associations. It seems difficult to account for the strong sensitivity of behavior to computationally immaterial details in the experimental settings using such a general-purpose learning model (as discussed in Section III). Stating the obvious, humans and animals do not approach the learning problem as blank slates. Millions of years of evolution have favored the formation of particular associations and not others [101]. One striking example manifests in a phenomenon commonly known as taste aversion [102]. This phenomenon is studied in experiments in which animals are presented with a novel food and, hours later, are exposed to gastrointestinal distress (e.g., as a result of radiation and other illness-inducing agents, such as lithium chloride). Consequently, the animal will avoid the new food when it is presented to it again. This form of learning is extremely powerful, and learning to avoid the new food is achieved after a single trial. By contrast, it is impossible to teach an animal to avoid a lever press by associating it with a malaise hours later or to teach the animal to avoid ingesting a particular food by associating it with an electric shock delivered hours later.[3] If food and tones are considered simply as different observations, and the electric shock and the gastrointestinal distress as negative reinforcers, then RL models do not predict these peculiarities. Rather, a model that explains this behavior should assume that the two negative reinforcers are *qualitatively* different, allowing only the latter to be associated with food consumption hours earlier. It is easy to envision how evolutionary pressure resulted in such a specific learning mechanism that can help animals learn to avoid poisonous food. There are numerous other such peculiarities of learning. A particularly amusing one is the failure of birds in the following task. Food-deprived cockerel chicks were trained in a straight runway, in which the food cup moved when the chicks moved, in the same direction at double the speed. In this "looking glass" setting, the appropriate behavior to obtain the food is to move *away* from it. The animals were unable to learn the task [103]. Other similar examples have been reported by animal trainers [104].

The results discussed in this review highlight the power, as well as the limitations of relating RL to operant learning. It is our opinion that addressing these limitations will require incorporating into RL ecologically and biologically informed models of the learning organism. ■

Acknowledgment

The authors would like to thank I. Erev, P. Dayan, T. Neiman, and S. Seung for discussions, and E. Tartaglia for comments on a previous version of the manuscript.

[3]Taste aversion is considered by many as an instance of classical conditioning rather than operant conditioning, but the two are not distinguishable in the way that the experiment is performed.

REFERENCES

[1] P. D. McGreevy and R. A. Boakes, *Carrots and Sticks: Principles of Animal Training*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[2] V. H. Matthews and D. C. Benjamin, *Old Testament Parallels: Laws and Stories From the Ancient Near Est*. Mahwah, NJ, USA: Paulist Press, 1991.

[3] I. A. Hyman, *The Case Against Spanking*. San Francisco, CA, USA: Jossey-Bass, 1997.

[4] E. L. Thorndike, *Animal Intelligence*. Darien, CT, USA: Hafner, 1911.

[5] R. R. Bush and F. Mostfeller, *Stochastic Models for Learning*. New York, NY, USA: Wiley, 1955.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[7] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[8] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407, 1951.

[9] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems," in *Proc. 16th Nat. Conf. Artif. Intell.*, Orlando, FL, USA, 1999, pp. 541–548.

[10] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Learning without state-estimation in partially observable Markovian decision processes," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 284–292.

[11] S. W. Hasinoff, "Reinforcement learning for problems with hidden state," Dept. Comput.

Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2003.

[12] K. P. Murphy, "A survey of POMDP solution techniques," Dept. Comput. Sci., Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep., 2000.

[13] K. Doya, "Reinforcement learning: Computational theory and biological mechanisms," *HFSP J.*, vol. 1, pp. 30–40, 2007.

[14] J. Duncan, "An adaptive coding model of neural function in prefrontal cortex," *Nature Rev. Neurosci.*, vol. 2, pp. 820–829, 2001.

[15] E. K. Miller and T. J. Buschman, "Rules through recursion: How interactions between the frontal cortex and basal ganglia may build abstract, complex, rules from concrete, simple, ones," in *The Neuroscience of Rule-Guided Behavior*. Oxford, U.K.: Oxford Univ. Press, 2007.

[16] P. Redgrave, T. J. Prescott, and K. Gurney, "The basal ganglia: A vertebrate solution to the selection problem?" *Neuroscience*, vol. 89, pp. 1009–1023, 1999.

[17] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, "Representation of action-specific reward values in the striatum," *Science*, vol. 310, pp. 1337–1340, 2005.

[18] T. Schönberg, N. D. Daw, D. Joel, and J. P. O'Doherty, "Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making," *J. Neurosci.*, vol. 27, pp. 12 860–12 867, 2007.

[19] G. E. Alexander and M. D. Crutcher, "Functional architecture of basal ganglia circuits: Neural substrates of parallel processing," *Trends Neurosci.*, vol. 13, pp. 266–271, 1990.

[20] S. M. Nicola, "The nucleus accumbens as part of a basal ganglia action selection circuit," *Psychopharmacology*, vol. 191, pp. 521–550, 2007.

[21] G. Corrado and K. Doya, "Understanding neural coding through the model-based analysis of decision making," *J. Neurosci.*, vol. 27, pp. 8178–8180, 2007.

[22] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593–1599, 1997.

[23] W. Schultz, "Dopamine signals for reward value and risk: Basic and recent data," *Behav. Brain Funct.*, vol. 6, 2010, DOI: 10.1186/1744-9081-6-24.

[24] P. W. Glimcher, "Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis," *Proc. Nat. Acad. Sci. USA*, vol. 108, pp. 15 647–15 654, 2011.

[25] C. R. Gallistel, *The Organization of Learning*. Cambridge, MA, USA: MIT Press, 1990.

[26] K. M. Kim, M. V. Baratta, A. Yang, D. Lee, E. S. Boyden, and C. D. Fiorillo, "Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement," *PLoS ONE*, vol. 7, 2012, e33612.

[27] E. S. Bromberg-Martin, M. Matsumoto, and O. Hikosaka, "Dopamine in motivational control: Rewarding, aversive, alerting," *Neuron*, vol. 68, pp. 815–834, 2010.

[28] J. Roeper, "Dissecting the diversity of midbrain dopamine neurons," *Trends Neurosci.*, vol. 36, pp. 336–342, 2013.

[29] H. M. Bayer and P. W. Glimcher, "Midbrain dopamine neurons encode a quantitative reward prediction error signal," *Neuron*, vol. 47, pp. 129–141, 2005.

[30] B. Seymour, N. D. Daw, J. P. Roiser, P. Dayan, and R. Dolan, "Serotonin selectively modulates reward value in human decision-making," *J. Neurosci.*, vol. 32, pp. 5833–5842, 2012.

[31] J. C. Horvitz, "Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events," *Neuroscience*, vol. 96, pp. 651–656, 2000.

[32] P. Redgrave and K. Gurney, "The short-latency dopamine signal: A role in discovering novel actions?" *Nature Rev. Neurosci.*, vol. 7, pp. 967–975, 2006.

[33] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, Cambridge, MA, USA: MIT Press, 1999, pp. 1057–1063.

[34] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *J. Artif. Intell. Res.*, vol. 15, pp. 319–350, 2001.

[35] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, pp. 229–256, 1992.

[36] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon Press, 1995.

[37] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA, USA: MIT Press, 2001.

[38] P. Cisek and J. F. Kalaska, "Neural mechanisms for interacting with a world full of action choices," *Annu. Rev. Neurosci.*, vol. 33, pp. 269–298, 2010.

[39] A. A. Faisal, L. P. J. Selen, and M. D. Wolpert, "Noise in the nervous system," *Nature Rev. Neurosci.*, vol. 9, pp. 292–303, 2008.

[40] R. S. Zucker and W. G. Regehr, "Short-term synaptic plasticity," *Annu. Rev. Physiol.*, vol. 64, pp. 355–405, 2002.

[41] Y. Dan and M.-M Poo, "Spike timing-dependent plasticity of neural circuits," *Neuron*, vol. 44, pp. 23–30, 2004.

[42] R. C. Malenka and M. F. Bear, "LTP and LTD: An embarrassment of riches," *Neuron*, vol. 44, pp. 5–21, 2004.

[43] T. M. Jay, "Dopamine: A potential substrate for synaptic plasticity and memory mechanisms," *Progr. Neurobiol.*, vol. 69, pp. 375–390, 2003.

[44] V. Pawlak and J. N. D. Kerr, "Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity," *J. Neurosci.*, pp. 2435–2446, 2008.

[45] J. R. Wickens, "Synaptic plasticity in the basal ganglia," *Behav. Brain Res.*, vol. 199, pp. 119–128, 2009.

[46] J.-C. Zhang, P.-M. Lau, and G.-Q. Bi, "Gain in sensitivity and loss in temporal contrast of STDP by dopaminergic modulation at hippocampal synapses," *Proc. Nat. Acad. Sci. USA*, vol. 106, pp. 13 028–13 033, 2009.

[47] H. S. Seung, "Learning in spiking neural networks by reinforcement of stochastic synaptic transmission," *Neuron*, vol. 40, pp. 1063–1073, 2003.

[48] X. Xie and H. S. Seung, "Learning in neural networks by reinforcement of irregular spiking," *Phys. Rev. E*, vol. 69, 2004, 041909.

[49] R. Urbanczik and W. Senn, "Reinforcement learning in populations of spiking neurons," *Nature Neurosci.*, vol. 12, pp. 250–252, 2009.

[50] I. R. Fiete and H. S. Seung, "Gradient learning in spiking neural networks by dynamic perturbation of conductances," *Phys. Rev. Lett.*, vol. 97, 2006, 048104.

[51] Y. Loewenstein and H. S. Seung, "Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity," *Proc. Nat. Acad. Sci. USA*, vol. 103, pp. 15 224–15 229, 2006.

[52] Y. Loewenstein, "Synaptic theory of replicator-like melioration," *Front. Comput. Neurosci.*, vol. 4, no. 17, 2010, DOI: 10.3389/fncom.2010.00017.

[53] Y. Loewenstein, "Robustness of learning that is based on covariance-driven synaptic plasticity," *PLoS Comput. Biol.*, vol. 4, 2008, e1000007.

[54] T. Neiman and Y. Loewenstein, "Covariance-based synaptic plasticity in an attractor network model accounts for fast adaptation in free operant learning," *J. Neurosci.*, vol. 33, pp. 1521–1534, 2013.

[55] M. Fairbank and E. Alonso, "The divergence of reinforcement learning algorithms with value-iteration and function approximation," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2012, pp. 3070–3077.

[56] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 993–1000.

[57] H. R. Maei and R. S. Sutton, "Gq(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces," in *Proc. 3rd Conf. Artif. Gen. Intell.*, vol. 1, pp. 91–96.

[58] D. A. Grant, H. W. Hake, and J. P. Hornseth, "Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement," *J. Exp. Psychol.*, vol. 42, 1951, DOI: 10.1037/h0054051.

[59] K. J. Arrow, "Utilities, attitudes, choices: A review note," *Econometrica*, vol. 26, pp. 1–23, 1958.

[60] R. J. Herrnstein and D. H. Loveland, "Maximizing and matching on concurrent ratio schedules," *J. Exp. Anal. Behav.*, pp. 107–116, 1975.

[61] W. L. Woolverton and J. K. Rowlett, "Choice maintained by cocaine or food in monkeys: Effects of varying probability of reinforcement," *Psychopharmacology*, vol. 138, pp. 102–106, 1998.

[62] N. Vulkan, "An economists perspective on probability matching," *J. Econ. Surv.*, vol. 14, pp. 101–118, 2000.

[63] D. R. Shanks, R. J. Tunney, and J. D. McCarthy, "A re-examination of probability matching and rational choice," *J. Behav. Decision Making*, vol. 15, pp. 233–250, 2002.

[64] S. M. McCrea and E. R. Hirt, "Match madness: Probability matching in prediction of the ncaa basketball tournament," *J. Appl. Soc. Psychol.*, vol. 39, pp. 2809–2839, 2009.

[65] Y. Sakai and T. Fukai, "The actor-critic learning is behind the matching law: Matching versus optimal behaviors," *Neural Comput.*, vol. 20, pp. 227–251, 2008.

[66] C. S. Green, C. Benson, D. Kersten, and P. Schrater, "Alterations in choice behavior

by manipulations of world model," *Proc. Nat. Acad. Sci. USA*, vol. 107, pp. 16401–16406, 2010.

[67] D. J. Koehler and G. James, "Probability matching in choice under uncertainty: Intuition versus deliberation," *Cognition*, vol. 113, pp. 123–127, 2009.

[68] E. B. Morse and W. N. Rundquist, "Probability matching with an unscheduled random sequence," *Amer. J. Psychol.*, vol. 73, pp. 603–607, 1960.

[69] P. L. Derks and M. I. Paclisanu, "Simple strategies in binary prediction by children and adults," *J. Exp. Psychol.*, vol. 73, pp. 278–285, 1967.

[70] S. Laquitaine, C. Piron, D. Abellanas, Y. Loewenstein, and T. Boraud, "Complex population response of dorsal putamen neurons predicts the ability to learn," *PloS One*, vol. 8, 2013, e80683.

[71] H. Shteingart and Y. Loewenstein, "Reinforcement learning and human behavior," *Current Opinion Neurobiol.*, vol. 25, pp. 93–98, 2014.

[72] Y. Loewenstein, D. Prelec, and S. H. Seung, "Operant matching as a Nash equilibrium of an intertemporal game," *Neural Comput.*, vol. 21, pp. 2755–2773, 2009.

[73] K. Ono, "Superstitious behavior in humans," *J. Exp. Anal. Behav.*, vol. 47, pp. 261–271, 1987.

[74] B. F. Skinner, "'Superstition' in the pigeon," *J. Exp. Psychol.*, vol. 38, pp. 168–172, 1948.

[75] A. I. Houston and M. J, "How to maximize reward rate on two variable-interval paradigms," *J. Exp. Anal. Behav.*, vol. 35, pp. 367–396, 1981.

[76] G. M. Heyman, "Is time allocation unconditioned behavior?" *Quantitative Analyses of Behavior*, vol. 2, Cambridge, MA, USA: Ballinger Press, 1982.

[77] J. Gibbon, "Dynamics of time matching: Arousal makes better seem worse," *Psychonom. Bull. Rev.*, vol. 2, pp. 208–215, 1995.

[78] C. R. Gallistel, T. A. Mark, A. P. King, and P. E. Latham, "The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect.," *J. Exp. Psychol. Animal Behav. Process.*, vol. 27, pp. 354–372, 2001.

[79] R. J. Herrnstein, "Relative and absolute strength of response as a function of frequency of reinforcement," *J. Exp. Anal. Behav.*, vol. 4, pp. 267–272, 1961.

[80] M. Davison and D. McCarthy, *The Matching Law: A Research Review*. Hillsdale, NJ, USA: Erlbaum, 1988.

[81] R. J. Herrnstein, *The Matching Law*. New York, NY, USA: Russell Sage Foundation, 1997.

[82] L. P. Sugrue, G. S. Corrado, and W. T. Newsome, "Matching behavior and the representation of value in the parietal cortex," *Science*, vol. 304, pp. 1782–1787, 2004.

[83] W. M. Baum, "Choice in free-ranging wild pigeons," *Science*, vol. 185, pp. 78–79, 1974.

[84] W. M. Baum, "On two types of deviation from the matching law: Bias and undermatching," *J. Exp. Anal. Behav.*, vol. 22, pp. 231–242, 1974.

[85] B. Lau and P. W. Glimcher, "Value representations in the primate striatum during matching behavior," *Neuron*, vol. 58, pp. 451–463, 2005.

[86] G. S. Corrado, L. P. Sugrue, H. S. Seung, and W. T. Newsome, "Linear-nonlinear-Possion models of primate choice dynamics," *J. Exp. Anal. Behav.*, vol. 84, pp. 581–617, 2005.

[87] K. Iigaya, L. P. Sugrue, G. S. Corrado, Y. Loewenstein, W. T. Newsome, and S. Fusi, "Deviations from the matching law reflect reward integration over multiple timescales," in *Proc. Cosyne Abstracts 2013*, Salt Lake City, UT, USA, 2013, pp. 93–98.

[88] D. A. Worthy, M. J. Hawthorne, and A. R. Otto, "Heterogeneity of strategy use in the Iowa gambling task: A comparison of win-stay/lose-shift and reinforcement learning models," *Psychonom. Bull. Rev.*, vol. 20, pp. 364–371, 2013.

[89] I. Erev and E. Haruvy, "Learning and the economics of small decisions," in *The Handbook of Experimental Economics*, J. H. Kagel and A. E. Roth, Eds. Princeton, NJ, USA: Princeton Univ. Press.

[90] H. Shteingart, T. Neiman, and Y. Loewenstein, "The role of first impression in operant learning," *J. Exp. Psychol. Gen.*, vol. 142, pp. 476–488, 2013.

[91] I. Nevo and I. Erev, "On surprise, change, the effect of recent outcomes," *Front. Psychol.*, vol. 3, no. 24, 2012, DOI: 10.3389/fpsyg.2012.00024.

[92] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dyn. Syst.*, vol. 13, pp. 341–379, 2003.

[93] C. Diuk, K. Tsai, J. Wallis, M. Botvinick, and Y. Niv, "Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia," *J. Neurosci.*, vol. 33, pp. 5797–5805, 2013.

[94] T. Neiman and Y. Loewenstein, "Reinforcement learning in professional basketball players," *Nature Commun.*, vol. 2, 2011, DOI: 10.1038/ncomms1580.

[95] R. Legenstein, N. Wilbert, and L. Wiskott, "Reinforcement learning on slow features of high-dimensional input streams," *PLoS Comput. Biol.*, vol. 6, 2010, e1000894.

[96] O.-A. Maillard, R. Munos, and D. Ryabko, "Selecting the state-representation in reinforcement learning," in *Proc. Neural Inf. Process. Syst. Conf.*, 2011.

[97] M. Toussaint and A. Storkey, "Probabilistic inference for solving discrete and continuous state Markov decision processes," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 945–952.

[98] K. J. Friston, J. Daunizeau, and S. J. Kiebel, "Reinforcement learning or active inference?" *PloS One*, vol. 4, 2009, e6421.

[99] K. Friston, P. Schwartenbeck, T. FitzGerald, M. Moutoussis, T. Behrens, and R. J. Dolan, "The anatomy of choice: Active inference and agency," *Front. Human Neurosci.*, vol. 7, 2013, DOI: 10.3389/fnhum.2013.00598.

[100] P. A. Ortega and D. A. Braun, "Thermodynamics as a theory of decision-making with information-processing costs," *Proc. Roy. Soc. A*, vol. 469, no. 2153, 2013, DOI: 10.1098/rspa.2012.0683.

[101] P. Dayan, Y. Niv, B. Seymour, and N. D. Daw , "The misbehavior of value and the discipline of the will," *Neural Netw.*, vol. 19, pp. 1153–1160, 2006.

[102] A. Verendeev and A. L. Riley, "Conditioned taste aversion and drugs of abuse: History and interpretation," *Neurosci. Biobehav. Rev.*, vol. 36, pp. 2193–2205, 2012.

[103] W. A. Hershberger, "An approach through the looking-glass," *Animal Learn. Behav.*, vol. 14, pp. 443–451, 1986.

[104] K. Breland and M. Breland, "The misbehavior of organisms," *Amer. Psychol.*, vol. 16, pp. 681–684, 1961.

ABOUT THE AUTHORS

Gianluigi Mongillo received the M.Sc. degree in theoretical physics and the Ph.D. degree in neurophysiology from the University of Rome "La Sapienza," Rome, Italy, in 2000 and 2005, respectively.

He has been Chargé de Recherche with the French National Centre for Scientific Research (CNRS), Paris, France, since 2009. His research interests include analytical approach to the dynamics of large neuronal networks, synaptic plasticity and its functional impact on network dynamics, models of working memory, and, more recently, reinforcement learning and decision making.

Hanan Shteingart received the B.Sc. degree in physics and electrical engineering and the M.Sc. degree in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2001 and 2007, respectively. Since 2008, he has been working toward the Ph.D. degree at The Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel.

Yonatan Loewenstein received the B.Sc. degree in physics and the Ph.D. degree in computational neuroscience from The Hebrew University of Jerusalem, Jerusalem, Israel, in 1996 and 2004, respectively.

Between 2004 and 2006, he was a Postdoctoral Fellow at the Massachusetts Institute of Technology, Cambridge, MA, USA. Since 2007, he has held a joint position as an Assistant Professor in the Department of Neurobiology, the Department of Cognitive Sciences, and the Edmond and Lily Safra Center for Brain Sciences (ELSC), Hebrew University of Jerusalem, where he heads the Ph.D. program. He is also a member of the Center for the Study of Rationality at the Hebrew University of Jerusalem, Jerusalem, Israel.