Behavioural Processes xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Behavioural Processes

journal homepage: www.elsevier.com/locate/behavproc

Enriching behavioral ecology with reinforcement learning methods

Willem E. Frankenhuis[a,*,1], Karthik Panchanathan[b,1], Andrew G. Barto[c,1]

[a] Behavioural Science Institute, Radboud University, Montessorilaan 3, PO Box 9104, 6500, HE, Nijmegen, The Netherlands
[b] Department of Anthropology, University of Missouri, United States
[c] College of Information and Computer Sciences, University of Massachusetts Amherst, United States

ARTICLE INFO

Keywords:
Adaptation
Evolution
Development
Learning
Dynamic programming
Reinforcement learning

ABSTRACT

This article focuses on the division of labor between evolution and development in solving sequential, state-dependent decision problems. Currently, behavioral ecologists tend to use dynamic programming methods to study such problems. These methods are successful at predicting animal behavior in a variety of contexts. However, they depend on a distinct set of assumptions. Here, we argue that behavioral ecology will benefit from drawing more than it currently does on a complementary collection of tools, called reinforcement learning methods. These methods allow for the study of behavior in highly complex environments, which conventional dynamic programming methods do not feasibly address. In addition, reinforcement learning methods are well-suited to studying how biological mechanisms solve developmental and learning problems. For instance, we can use them to study simple rules that perform well in complex environments. Or to investigate under what conditions natural selection favors fixed, non-plastic traits (which do not vary across individuals), cue-driven-switch plasticity (innate instructions for adaptive behavioral development based on experience), or developmental selection (the incremental acquisition of adaptive behavior based on experience). If natural selection favors developmental selection, which includes learning from environmental feedback, we can also make predictions about the design of reward systems. Our paper is written in an accessible manner and for a broad audience, though we believe some novel insights can be drawn from our discussion. We hope our paper will help advance the emerging bridge connecting the fields of behavioral ecology and reinforcement learning.

1. Introduction

Each organism faces a host of *adaptive problems*. These range from short-term problems, like foraging under the risk of predation, to long-term problems, like allocating resources among growth, maintenance, and reproduction. Solving an adaptive problem entails generating an *adaptive phenotype*, ranging from behavioral repertoires to developmental patterns. In behavioral ecology, an adaptive phenotype refers to one that improves *fitness* relative to other phenotypes. The best measure of fitness depends on the species and the conditions in which it evolves. Fitness is often defined as the long-term growth rate of a lineage (Donaldson-Matasci et al., 2008; Lewontin and Cohen, 1969; McNamara et al., 2016; Starrfelt and Kokko, 2012). In practice, biologists often measure proxies for fitness, including general measures like survival and reproduction, or domain-specific measures like energetic returns from foraging.

Here, we focus on processes that solve adaptive problems over two timescales: evolution and development. Evolution refers to changes across generations in the frequencies of types in a population, whether those changes be in genes, developmental systems, or phenotypic traits. Natural selection, the differential reproductive success of inherited variations, is the only known evolutionary process that results in adaptation. Development refers to changes within an organism from conception to death. Learning, the acquisition of new information, abilities, or responses as a result of experience, is one developmental process that can produce adaptive behavior. Across generations natural selection shapes learning mechanisms, which enable organisms to acquire adaptive behavior within their lifetimes. This behavior, in turn, provides the phenotypic variation upon which subsequent natural selection acts. In general, developmental processes, such as learning, both result from and contribute to evolution (Baldwin, 1896; Hinton and Nowlan, 1987; Laland et al., 2001; Maynard Smith et al., 1985; Nolfi et al., 1994; Oudeyer and Smith, 2016; Oyama et al., 2001; Staddon, 2016; Watson and Szathmáry, 2016; West-Eberhard, 2003).

The best way to solve a particular adaptive problem (i.e., to maximize fitness) depends on the properties of the environment. If the environment is constant across time and space, natural selection may favor *fixed, non-plastic traits*, which some call innate, canalized, or

* Corresponding author.
E-mail address: w.frankenhuis@psych.ru.nl (W.E. Frankenhuis).
[1] All authors contributed equally.

https://doi.org/10.1016/j.beproc.2018.01.008
Received 13 August 2017; Received in revised form 5 January 2018; Accepted 10 January 2018

reliably developing (Mameli and Bateson, 2011; Samuels, 2004). For instance, all zebras have stripes. This trait is not developmentally plastic. By contrast, if the environment varies over time and space, natural selection might favor *phenotypic plasticity*: the ability of a genotype to produce a range of different phenotypes depending on local environmental conditions and an individual's state (Schlichting and Pigliucci, 1998). In some cases, phenotypic plasticity can be *cue-driven and switch-like*, in which natural selection equips the organism with innate, rather than learned, instructions for adaptive behavior. Even if an organism learns parameter values of the environment (e.g., the environment is predator rich), it does not learn which behavior is adaptive in this state (e.g., in a predator-rich environment, develop defensive armor). This is precisely what happens in some Daphnia species, in which cues to predators trigger the development of defensive armor (Agrawal et al., 1999). In other cases, phenotypic plasticity is guided by a process of *developmental selection*, in which current phenotypes are shaped by the consequences of past phenotypes, including but not limited to learning from past behaviors (Snell-Rood, 2012). With developmental selection, and learning in particular, a division of labor arises: natural selection shapes the learning mechanisms and the developing organism learns how to behave adaptively. Learning by trying out different behaviors is ubiquitous in nature. For example, stickleback fish have evolved learning mechanisms that allow them to process food more efficiently by generating different behaviors and pursuing those that work best (Dill, 1983).

We make two arguments in this paper. First, behavioral ecology should draw on reinforcement learning (RL) methods (Sutton and Barto, 1998) more than it currently does. Although inspired by learning theories of psychology (Skinner, 1953; Thorndike, 1911), as we treat it here, RL methods refer to a collection of machine learning algorithms, rather than to a theory of how actual animal learning works. Although biologists are increasingly using RL methods to study adaptive behavior (e.g., Dridi and Lehmann, 2014, 2015, 2016; Enquist et al., 2016; Frank, 1996, 1997; Whalen et al., 2015), we are not aware of any papers in behavioral ecology that explain RL methods and compare them with stochastic dynamic programing (SDP) methods, which *are* widely used by behavioral ecologists. We add a note on terminology: because the most widely used RL algorithms can be seen as *approximate* SDP methods (Bertsekas and Tsitsiklis, 1996; Powell, 2007), we use the term 'SDP methods' to refer to '*non-RL* SDP methods'. Second, RL methods are well suited to studying the conditions in which natural selection favors fixed traits, cue-driven-switch plasticity, or developmental selection. By using RL methods in theoretical models, we can study when RL as a mechanism of behavioral adaptation might evolve, as opposed to fixed traits or cue-driven-switch strategies.

Our paper is written in an accessible manner and for a broad audience. Sections 1–3 are designed for students and scholars working either in behavioral ecology or RL, without any background in the other field. Sections 4–5 may include new insights for some experts. In Section 2, we describe sequential, state-dependent decision problems. In Section 3, we discuss SDP methods. In Section 4, we introduce RL methods, and discuss when these can be more suitable than SDP methods. And, in Section 5, we revisit the division of labor between evolutionary and developmental adaptation, and briefly discuss how RL methods can help us understand the evolution of behavioral mechanisms.

We focus on two methods for computing optimal behavior in models: SDP and RL. We do not discuss psychological research related to RL and its integration into behavioral ecology. Other scholars have called for better integration of psychological learning theories and data into optimality models of behavior, thus bridging function and mechanism (Kacelnik, 2012; Kacelnik and Bateson, 1997; McNamara and Houston, 2009; Trimmer et al., 2012). Such integration has great value, because it leads to new discoveries and unifies knowledge from experimental psychology with behavioral ecology; for instance, by illuminating whether well-established animal learning processes are able to perform close to the optima predicted by various models. However, our focus is on computational methods, not biological mechanisms. Our central point is that RL methods allow biologists to make predictions about behavior in situations that are difficult or infeasible to study using SDP methods. Although we greatly appreciate simple models for the insights they provide (many of our own models fit this category), problems of high complexity sometimes require increased model complexity, which may trade-off insight for prediction (Levins, 1966).

2. Sequential, state-dependent decision problems

In the 1970s, behavioral ecologists started using optimization methods to study animal behavior. These methods assume that natural selection has designed organisms so that their behavior maximizes fitness. For example, in the typical foraging model, organisms are assumed to maximize the rate of energy gain (Stephens and Krebs, 1986). Initial models assumed that organisms made only a single decision or multiple decisions that are independent of each other. These models are successful at predicting behavior in a variety of contexts, but they are not well suited for studying sequential, state-dependent decision problems, in which payoffs and costs depend on a sequence of decisions made over extended periods of time. Therefore, biologists turned to *optimal control theory* (Bellman, 1957) to model sequential, state-dependent decision problems (Clark and Mangel, 2000; Houston et al., 1988; Houston and McNamara, 1999; Krebs et al., 1978; Mangel, 1990, 2015; Mangel and Clark, 1988; McNamara and Houston, 1986; for explanation tailored to psychologists, see Frankenhuis et al., 2013).

Consider a bird that needs to find enough food each day to survive the non-breeding season (alternatively, the bird could maximize a balance between survival and reproductive success; the example is taken from Houston et al., 1988). Birds vary day-to-day in their energy reserves, based in part on their previous foraging decisions. Reserves increase through food consumption and decrease through metabolic expenditure. If energy levels drop below a threshold, the bird dies of starvation. On the flip side, energy reserves do not continue to increase indefinitely with consumption; at some point, the fitness benefit of additional reserves is outweighed by the cost. The bird can choose to forage in one of a fixed number of locations on each day. Each location has a characteristic probability of death from predation and a distribution of energetic returns from foraging. Further, the most productive habitats are also the riskiest; because predators will be drawn to the locations where most birds forage, there is a higher probability of being killed in more productive patches. A researcher may wonder: In which location(s) should a bird forage? And, how should this choice vary as a function of current energy reserves and time to the end of the season?

Markov decision processes (MDPs) are discrete-time versions of optimal control problems and provide a mathematical framework for studying sequential, state-dependent decision problems. In these decision problems, time is broken up into discrete steps and outcomes are partly random and partly under the control of the decision maker. An MDP has four components:

1. *States*: A set of possible, relevant configurations of a system (such as an organism and its environment).
2. *Actions*: A set of (perhaps state-specific) actions available to the agent
3. *Transition function*: A function that probabilistically assigns the next state based on the current state and chosen action.
4. *Reward function*: A function that assigns an immediate reward based on the current state and chosen action (note: 'reward' can mean different things depending on the application, and it can include costs too).

Different fields use different terminology in describing MDPs. Behavioral ecologists use the term 'agent' to refer to organisms. They

distinguish between the 'state' of the organism and the 'state' of the environment. In RL, the term 'agent' refers to an entity that uses an algorithm to interact with and learn about its environment. These agents are situated in environments that can be in different states. In RL, the 'environment' includes both the organism itself (i.e., conditions internal to the body envelope) and its external environment (Sutton and Barto, 1998). In Sections 2 and 3, we use the terms 'agent' and 'environment' as used in behavioral ecology: an 'agent' refers to an organism and the 'environment' refers to the organism's external environment. In Section 4, we use the terms 'agent' and 'environment' as they are used in RL.

The possible energy levels may represent the bird's state. The bird has the same action set each day, choosing a location in which to forage. The transition function probabilistically assigns the bird a new energy level based on its current level (current state) and the choice of foraging location (chosen action). The reward function probabilistically determines whether the bird survives a day based on its current energy and the choice of foraging location. By formulating a state-dependent decision problem as an MDP, the task becomes finding an *optimal policy* that, for each possible state, specifies the action choice that maximizes the expected cumulative future return, a measure of total reward over time. In behavioral ecology, this means finding a policy that maximizes expected fitness. In the bird example, an optimal policy instructs the bird where to forage based on its current energy reserve and the current time period, in order to maximize the probability that it survives the non-breeding season.

3. Stochastic dynamic programming

In order to predict animal behavior in sequential, state-dependent decision problems, such as foraging, behavioral ecologists often use Stochastic Dynamic Programming (SDP) methods (Clark and Mangel, 2000; Houston and McNamara, 1999; Mangel, 1990, 2015; Mangel and Clark, 1988; for an explanation tailored to psychologists, see Frankenhuis et al., 2013). In SDP, the decision problem is broken down into a collection of simpler sub-problems, and then each of those sub-problems is solved once and the solutions stored. Behavioral ecologists typically use the method of *backward induction,* starting at the terminal time period and then working backwards to the first time period. The algorithm begins by recording the rewards associated with each of the possible final states. It then takes one step back in time and notes all of the penultimate states it could be in. For each of these states, the algorithm computes and records the action that will maximize the expected reward from here on out. It repeatedly performs this procedure until it reaches the first time period, taking one step back at a time and then computing and recording the best action for each state. Once it reaches the initial time period, the algorithm will have visited all possible states and computed the optimal policy, which instructs the agent on how to behave for each state in each time step. SDP is clearly not intended to mimic psychological processes used by animals. Instead, "it identifies the optimal strategy from the perspective of an observer, without discussing how the decision maker may achieve it" (Kacelnik, 2012, p. 25).

We can use SDP to find an optimal policy for making decisions in a single environmental condition or across a range of different conditions. Environmental variation can take many different forms, and this variation shapes optimal decisions. Here, we consider a stylized environment that cycles through one of two possible states during an organism's lifetime. Suppose a bird resides either in bountiful conditions or in meager conditions (e.g., due to temporal differences in rainfall). And, suppose the bird has no control over which condition it is in. In this MDP, the external environment is characterized by two conditions, each with distinct transition and reward probabilities. In the bountiful condition, the probability of finding food, and/or the energetic value of food, might be higher. As a consequence, a bird foraging in this condition is more likely to move from hunger to satiation than a bird foraging in a meager condition. The set of state variables that describe a bird will include its estimate of the current condition and its energy budget.

An individual bird may never be absolutely certain about the environmental condition it is currently in. However, the bird may have access to cues that provide information about the state of the world, bountiful or meager, and thereby reduce its uncertainty (Dall et al., 2015; Dunlap and Stephens, 2016; McNamara and Houston, 1980; Mangel, 1990; Moran, 1992; Nettle et al., 2013; Stamps and Frankenhuis, 2016; Sultan and Spencer, 2002; Trimmer et al., 2011; Uller et al., 2015). Cues are observations that are more likely to occur in certain conditions than others (e.g., smoke is more likely when there is a fire than when there is not a fire). If cues offer little information for discriminating between conditions, or sampling cues is too costly (trading off with investment in other fitness-relevant activities, such as skill development, avoiding predators, or finding mates), an optimal policy might invest little or nothing at all in information search and instead be a compromise solution. For instance, the organism may develop a generalist phenotype, which performs fairly well across conditions, but not great in any particular condition. If cues are reliable, the policy can include separate instructions (e.g., foraging strategies) for dealing with each condition, which are triggered once the organism is sufficiently confident about the current condition it is in.

In order to use SDP methods, we must make two assumptions. First, the modeler needs to know the transition and reward functions (items 3 and 4 from the list in Section 2). Second, the sets of states, state variables, actions, and transition and reward functions, must not be so large that we cannot feasibly compute an optimal policy. Over the past several decades, behavioral ecologists have studied a great variety of significant problems using models that adhere to these assumptions, producing many profound insights (Clark and Mangel, 2000; Houston and McNamara, 1999; Mangel and Clark, 1988). However, we may not always want to make these assumptions, for at least two reasons.

The first reason is related to the well-known *curse of modeling* (Bellman, 1957): it can be difficult for a modeler to know all the transition and reward probabilities that determine expected values needed for SDP computations. The second reason is related to the *curse of dimensionality* (Bellman, 1957): as the number of states, state variables, actions, and transition and reward functions increases, SDP takes up more and more computational time and resources to find an optimal policy. Many sequential decision problems, vividly illustrated by games such as Backgammon, Chess, and Go, cover such a large space of states, that SDP cannot find an optimal policy within a feasible amount of time. In behavioral ecology, the behavior of real animals might depend on interactions between a sizable number of state variables, or animals might forage in an ecology characterized by a fine-grained and rapidly-changing structure. Such challenges can vastly increase the state space, requiring methods other than SDP, or models of such reduced complexity that they do not represent important details. In particular, there seems to be an emerging interest among behavioral ecologists in RL methods (Dridi and Lehmann, 2014, 2015, 2016; Enquist et al., 2016; Fawcett et al., 2014; Whalen et al., 2015). In the next section, we discuss how RL methods deal with the above challenges.

4. Reinforcement learning

Reinforcement Learning (RL) is a widely used collection of methods that mitigate both the curse of modeling and the curse of dimensionality in ways that allow us to approximate optimal policies for sequential, state-dependent decision problems when the transition and reward probabilities are not explicitly known to the modeler and/or when the decision problem is highly complex, i.e., spanning a massive state space. In such situations, using SDP will not be feasible. As we announced in Section 2, we now use the terms 'agent' and 'environment' as they are used in RL. That is, the term 'agent' refers to an entity that implements an algorithm to interact with and learn about an

environment, and the term 'environment' may include an organism's internal condition as well as its external environment.

An RL agent's goal is to learn how to maximize a measure of the cumulative amount of reward received over the long term while interacting with its environment, updating its policy based on the consequences of its actions. The agent typically knows the actions it can take and the current state of the environment (though, there are RL methods in which the agent does not know the current environmental state; for a tutorial, see Littman, 2009). The agent does not, however, initially know how to behave optimally (i.e., know which action maximizes future reward for any possible state). By interacting with its environment and observing the consequences of its actions, the agent can learn to approximate an optimal policy.

In our bird foraging model, we discussed how SDP based on backward induction finds an optimal policy by starting at the last time period and working back to the first time period. By contrast, an RL agent starts at the beginning by selecting actions over multiple episodes. This process results in gradually improving behavior. A key difference between SDP and RL is that SDP is an 'offline' computation. An offline computation separates finding an optimal policy from using it. How states are visited in an offline computation in order to update their values is unrelated to how an agent would visit states as it would while behaving in the environment. In contrast, an RL algorithm (at least a 'model-free' RL algorithm, something of a misnomer as we discuss below) updates the values of states as the RL agent visits them while behaving in its (real or simulated) environment. The RL agent selects actions that exploit the latest form of its policy, while also introducing exploratory actions in order to improve its policy. RL can mitigate the curse of dimensionality because in many cases the agent avoids visiting vast regions of an immense state space in which expending computational effort does little to improve the approximation of an optimal policy. This is a major reason that RL methods have had such striking success in problems with extremely large state spaces, such as the games Backgammon (Tesauro, 1994) and Go (Silver et al., 2016, 2017). Conventional SDP algorithms are infeasible for problems of this size and complexity. Of course, using RL methods implies abandoning the goal of finding exactly optimal policies, but this is a small price to pay when high-quality approximations are possible.

Model-free RL methods mitigate the curse of modeling by approximating optimal policies on the basis of many *sample trajectories*, which can be generated either by simulations of the agent interacting with its environment, or by actual interactions of the agent with its real environment. The term 'model-free' for this type of RL algorithm is something of a misnomer because unless the learning agent interacts with its real environment, it learns while interacting with a simulation of its environment, which of course requires an environment model. But a key advantage of RL over SDP is that it is much easier, for many types of problems, to simulate an agent interacting with its environment than it is to acquire explicit knowledge of all the probability distributions governing the interaction that would be required for a conventional SDP algorithm. Simulations only require these probability distributions to be *sampled* for specific cases, thus avoiding the need to produce at the start an exhaustive tabulation of all probabilities. Using (pseudo-) random number generators, computer simulations can produce trajectories of agents behaving in their environments according to an MDP's probabilities without ever having to explicitly produce those probabilities. Clearly, in order to accurately simulate trajectories of an MDP the simulation program needs to be able to simulate state transitions and rewards conforming to the MDP's transition and reward probabilities, but these probabilities can be generated by complex processes within the simulation program without ever directly accessing the probabilities themselves. As an example, suppose the next state of an MDP (e.g., predation risk at some future time) is a number whose probability depends on the output of some random process (e.g., infection rates in predator populations). In special cases, it can be easy to compute the probabilities of all the possible next states, but often this is difficult. On the other hand, knowing only how the random process affects a next state's probability, it is easy to generate sample next states. Moreover, samples are needed only for state transitions that occur in the simulations, which, as pointed out above, can be a small fraction of all the possible state transitions.

This simulation-based property of RL methods is important for behavioral ecology because it can be very difficult and error-prone to infer realistic transition and reward probabilities from observing behavior in messy real-world environments (e.g., inferring the probability of being eaten, if choosing to forage, given a particular level of predation risk). However, biologists can more easily develop computer programs to simulate the actions of animals in artificial environments that are based on empirical data. In other cases, a biologist might learn about optimal behavior without building an environment model at all, but rather by collecting sample behavioral trajectories produced by artificial embodied RL agents (e.g., robot insects implementing RL algorithms) interacting with real, complex environments.

RL algorithms that learn from simulated (or real) behavioral trajectories as just described are called 'model-free' because they only use models to produce the simulated behavioral trajectories from which they learn. The simulation model is not really a part of the RL algorithm. A 'model-based' RL algorithm, on the other hand, has internal access to an environment model so that it can combine its trajectory-focused computations with offline computations like SDP. Agents implementing model-based RL algorithms are capable of 'offline planning' by evaluating the consequences of actions that are never actually executed in their (possibly simulated) environments. As we will discuss below, an agent may combine offline planning with model-free learning, starting out with certain evolved parameter settings that are subsequently updated through learning (Enquist et al., 2016; Whalen et al., 2015).

RL approaches allow us to approximate optimal policies for ecologies that are more complex than those for which SDP is feasible. In our bird foraging example, location-specific parameters like predation probability and expected energy gain may vary from season to season. Incorporating this variation into the model means specifying more parameters. We may also want to increase the number of actions available to the agent. Or, we might want to model an ecology with fine-grained spatial and temporal variation in resources. These kinds of assumptions result in vastly larger sets of states and actions. In each of these cases, we can use RL methods to approximate an optimal policy, but not SDP methods. There are, in fact, many different ways in which RL methods can be employed to study the kinds of MDPs discussed above. We do not discuss these methods in detail here (for a thorough introduction, see Sutton and Barto 1998).

An agent using model-based RL can learn or improve its model on the basis of its experiences interacting with its environment. An agent might start out with a partial model, use it for offline planning to decide how to act, all the while improving the model based on its experienced state transitions and rewards (Enquist et al., 2016; Whalen et al., 2015). A model can be learned either passively or actively. Passively learning a model means that the agent's need to improve its model does not figure into how it selects actions. Actively learning a model, on the other hand, means that the agent's actions (at least some of them) are selected for the purpose of obtaining good 'training data' for the purpose of improving the agent's model. For instance, an active agent might select actions that are likely to result in observations that maximally reduce its uncertainty about transition and reward functions (Kruschke, 2008; Oudeyer and Smith, 2016).

Despite the advantages RL methods offer over SDP methods that we discussed above, RL methods are not without their own challenges. They can fail for a variety of reasons. Most obviously, if learning is based on simulated (rather than real) experience, the simulation model may not be an adequate model of the agent's interaction with its actual environment. Simulation models can be easier to formulate than models that rely on explicit enumeration of the relevant probabilities, but they

still must be faithful to the actual circumstances in which the modeled animal behaves. RL algorithms can mitigate the curse of modeling, but they cannot eliminate it. Irrespective of computational method (SDP or RL), one should not confuse models with reality. Models are useful tools for explanation and prediction, which need to be revised if empirical observations deviate from model predictions (Box, 1976).

Another challenge in applying RL methods is that of selecting a way to represent the agent's policy and/or a function, called a value function, that guides the agent's decisions. The usual way to do this is to select a set of numerical features characterizing the problem's states, together with a parameterized functional form, e.g., a multi-layer neural network, that transforms feature descriptions of states into actions and/or values. The RL algorithm incrementally adjusts the parameters as learning proceeds. Although not altogether different from the challenge of selecting and discretizing a state space in order to use SDP, selecting features and parameterized forms able to produce good approximations to optimal behavior is a challenge for any application of RL. An application can fail to produce useful predictions as a result of inappropriate selection of these algorithm details.

In short: RL methods can approximate optimal policies for ecologies that are more complex than those for which SDP methods are feasible. RL methods allow agents to learn from sample trajectories, thus focusing computational effort on regions of the state space that are relevant to the problem, while avoiding regions where computational effort is unlikely to significantly improve the approximation of an optimal policy. In addition, RL methods rely on simulation models of environmental interaction instead of explicit knowledge of all transition and reward probabilities as required by SDP. Moreover, as we will argue in the next section, reinforcement learning methods have an additional benefit: they are well-suited to studying how biological mechanisms solve developmental and learning problems.

5. Integrative models of evolutionary and developmental adaptation

Behavioral ecologists commonly assume that, given enough time, natural selection shapes phenotypic traits, including behavioral strategies, to maximize fitness (which is not to say that every phenotype is optimally adapted; see below). This *phenotypic gambit* (Grafen, 1984) allows researchers to ignore matters of instantiation, such as the underlying genetic architecture or developmental mechanisms (Maynard Smith et al., 1985). The gambit is a methodological stance that provides a starting point for empirical research. If observations deviate from predictions, we need to refine our model; for instance, we may need to better characterize the problem an animal is solving or the resources it has available to do so (Epstein, 2008; Levins, 1966; Smaldino, 2017). The gambit has proven to be useful in behavioral ecology (Gardner, 2009; Mangel, 2015; Maynard Smith, 1978; Mayr, 1983) and is mathematically consistent with tenets of evolutionary theory, such as population genetics (Grafen, 2014; Hammerstein, 1996), yet it remains controversial (Fawcett et al., 2013).

By employing the gambit, behavioral ecologists can avoid the question of whether natural selection equips the organism with instructions for adaptive behavior or with the ability to learn adaptive behavior (but see Kacelnik, 2012; Kacelnik and Bateson, 1997; McNamara and Houston, 2009). There is, of course, an extensive literature on the evolution of phenotypic plasticity (Barrett, 2015; Dunlap and Stephens, 2016; Gomulkiewicz and Kirkpatrick, 1992; McNamara and Houston, 1980; Mangel, 1990; Moran, 1992; Nettle et al., 2013; Schlichting and Pigliucci, 1998; Stamps and Frankenhuis, 2016; Sultan and Spencer, 2002; Trimmer et al., 2011; Uller et al., 2015). Here, biologists use mathematical models to study in which conditions natural selection favors plasticity over fixed phenotypes, and how organisms integrate cues coming from different sources, such as genes, epigenetic inheritance, prenatal effects, and postnatal experiences (Botero et al., 2015; Dall et al., 2015; Jablonka et al., 1995; Lachmann and Jablonka, 1996; Leimar et al., 2006; McNamara et al., 2016; Rivoire and Leibler, 2014; Stamps and Frankenhuis, 2016). Nevertheless, in these models the organism is born with a cue-driven-switch policy, which it executes. Other types of plasticity, such as developmental selection, have received far less theoretical attention (but see Arnold, 1978; Dridi and Lehmann, 2014, 2015; Enquist et al., 2016; Frank, 1996, 1997; Snell-Rood, 2012; Trimmer and Houston, 2014; Trimmer et al., 2012; Whalen et al., 2015).

To this point, we have focused on different methods a researcher can use to find or approximate an optimal policy. We have not discussed how a real animal acquires its behavioral strategy. Whether innate or learned policies are favored will depend on the distribution of environments faced by ancestral generations. If an environment changes very slowly relative to an organism's lifespan, remaining nearly stable across generations, natural selection can build innate adaptations, like the stripes on a zebra. However, if an environment changes at a noticeable rate between generations, but slowly enough within generations for learning to be useful, there needs to be some division of labor between evolutionary and developmental processes. What might this division look like?

The answer depends on how slowly an environment changes relative to an organism's lifetime. If an organism effectively spends its whole life in the same conditions and the number of different conditions is very limited, natural selection might produce a policy that 'switches' an organism into one of several discrete phenotypes. Such *polyphenisms* abound in nature, and include certain predator-induced defences in amphibians and crustaceans (Gilbert and Baressi, 2016). In this case, trait development may occur during a critical period and be irreversible (Botero et al., 2015; English et al., 2016; Fawcett and Frankenhuis, 2015; Frankenhuis and Fraley, 2017; Frankenhuis and Panchanathan, 2011; Panchanathan and Frankenhuis, 2016; Fischer et al., 2014; Pfab et al., 2016). Alternatively, if an organism is likely to experience different conditions during its lifetime (e.g., due to migration or seasonality) and the number of different conditions is limited, natural selection may favor a policy that contains instructions for each condition and the ability to continuously 'switch' instructions based on the current conditions. In this case, trait development may be reversible without a critical period.

With irreversible and reversible switches, the organism does not 'learn' a policy. Natural selection 'discovers' and equips the organism with a policy, which the organism then executes. By contrast, natural selection may result in *organisms that learn policies* when the environment is highly complex, spanning a massive state space. The environment may include a great variety of different patches, each of which is characterized by a different set of transition and reward probabilities, which themselves may change during an organism's lifetime (but not too fast for learning to be useful). In such a complex ecology, natural selection might favor developmental selection rather than a vast collection of pre-specified strategies and a policy for choosing among them, or it might favor a combination of innate instructions and developmental selection. For instance, the organism might use cues to infer the current conditions, but when this fails, switch to a developmental selection strategy. In any case, for developmental selection to be adaptive, there needs to be enough repetition in the organism's experiences to provide a sufficient number of 'learning trials'. When learning involves prediction, future events need to be correlated with past events. Further, the costs of actions should not be too high relative to their benefits. In a foraging context, if locations are far apart or traveling is dangerous, the cost of exploration may always outweigh the marginal gain of finding a better location.

With developmental selection, the agent needs a means for evaluating the consequences of its actions that works within its lifetime. Reward systems, implemented by animal nervous systems, accomplish this by 'rewarding' behavior that tends to lead to adaptive outcomes, reinforcing the production of those behaviors (Barto, 2013; Dridi and Akçay, 2018; Singh et al., 2009, 2010; Sorg, 2011; see also Cosmides

and Tooby, 2013; Niv et al., 2002; Samuelson and Swinkels, 2006). When reproductive success is a distal consequence of specific behaviors, however, the relationship between reward systems and reproductive success may not be easy to discern. Computational experiments using RL methods designed to find reward systems that are optimally suited to furthering reproductive success (possibly assessed by proxies, such as foraging success) suggest that optimal reward systems are likely exquisitely sensitive both to an agent's own capabilities and limitations, as well as to the distribution of environments in which the agent is likely to find itself (Singh et al., 2009, 2010; Sorg, 2011). The experiments just described explored a state space that would have been manageable with SDP methods as well. However, if these experiments were extended to include more state variables (e.g., by making the internal environment of the organism or its external environment more complex, for instance, by adding further physical or social dimensions), RL methods are likely to be up to the task, while SDP methods might not be.

When modeling the evolution of reward systems, we must choose a definition of fitness. As noted, empirical studies often measure fitness as the survival and reproductive success of individuals. However, fitness should be assigned to developmental systems (or strategies or genotypes), not to individuals. Often, the long-term growth rate of a lineage provides the appropriate measure of fitness (Donaldson-Matasci et al., 2008; Lewontin and Cohen, 1969; McNamara et al., 2016; Starrfelt and Kokko, 2012). This measure implies a particular fitness calculation (i.e., the geometric mean fitness of developmental systems across generations), which may favor different outcomes than a fitness calculation based on individual survival and reproduction (i.e., the arithmetic mean fitness across individuals within a generation). Incorporating this insight into RL theorists' studies of the evolution of reward systems, which have relied on arithmetic mean fitness (e.g., Singh et al., 2009, 2010; Sorg, 2011), might be an interesting direction for future research.

The art of modeling is to decide which aspects of an organism and its environment need to be modeled, and which ones can be left out. These decisions depend on the aims and scope of a model (Levins, 1966; Parker and Maynard Smith, 1990). "General models promote understanding of qualitative features. The parameters of such models may not be easy to measure. Specific models are based on a particular system and have parameters that can be measured so that predictions can be made" (Houston and McNamara, 2005, p. 934; see also Frankenhuis and Tiokhin, in press). For both types of models, it is often appropriate to include a small number of state variables, resulting in a compliant state space, which SDP methods can handle. However, in some cases we want to study qualitative patterns that emerge with a larger number of state variables, or quantitative predictions in messy, real-world ecologies. In such cases, behavioral ecology will benefit from drawing on RL methods. Moreover, as RL methods are also well-suited to modeling mechanistic instantiation, these methods hold great promise for studying simple rules that perform well in complex environments, thus illuminating the psychological mechanisms that real animals use in their daily lives.

In summary, the main advantage of RL methods over SDP methods is that they can find good policies for MDPs with massive state spaces and with transition and reward functions that are unknown to the modeler, so that simulation is easier than getting all the probabilities needed for SDP. It is good for behavioral ecologists to know that RL methods allow for more complexity than SDP methods. It remains an open question, however, whether they will gain new insights, and if so which ones, from being able to predict behavior in more complex scenarios. In any case, for those who want to explore large state spaces, RL methods offer a rich and appropriate set of tools.

Funding

This work was supported by the Netherlands Organization for Scientific Research (grant number 016.155.195); the Robert Wood Johnson Foundation (grant number 73657); the James S. McDonnell Foundation (grant number 220020502); and the Jacobs Foundation (grant number 2017 1261 02) to WEF.

Acknowledgments

We thank Alex Kacelnik, John McNamara, Alisdair Houston, Jesse Fenneman, Nicole Walasek, Sarah de Vries, Slimane Dridi, two anynomous reviewers, and the editor Billy Baum, for their valuable feedback on previous versions of this manuscript.

References

Agrawal, A.A., Laforsch, C., Tollrian, R., 1999. Transgenerational induction of defences in animals and plants. Nature 401, 60–63. http://dx.doi.org/10.1038/43425.
Arnold, S.J., 1978. The evolution of a special class of modifiable behaviors in relation to environmental pattern. Am. Nat. 112, 415–427. http://dx.doi.org/10.1086/283283.
Baldwin, J.M., 1896. A new factor in evolution. Am. Nat. 30, 441–451. http://dx.doi.org/10.1086/276408.
Barrett, H.C., 2015. The Shape of Thought: How Mental Adaptations Evolve. Oxford University Press, New York.
Barto, A., 2013. Intrinsic motivation and reinforcement learning. In: Baldassarre, G., Mirolli, M. (Eds.), Intrinsically Motivated Learning in Natural and Artificial Systems. Springer, Berlin, pp. 17–47. http://dx.doi.org/10.1007/978-3-642-32375-1_2.
Bellman, R.E., 1957. Dynamic Programming. Princeton University Press, Princeton, N .J.
Bertsekas, D.P., Tsitsiklis, J.N., 1996. Neuro-Dynamic Programming. Athena Scientific, Belmont, MA.
Botero, C.A., Weissing, F.J., Wright, J., Rubenstein, D.R., 2015. Evolutionary tipping points in the capacity to adapt to environmental change. Proc. Natl. Acad. Sci. 112, 184–189. http://dx.doi.org/10.1073/pnas.1408589111.
Box, G.E., 1976. Science and statistics. J. Am. Stat. Assoc. 71, 791–799. http://dx.doi.org/10.1080/01621459.1976.10480949.
Clark, C.W., Mangel, M., 2000. Dynamic State Variable Models in Ecology. Oxford University Press, Oxford, England.
Cosmides, L., Tooby, J., 2013. Evolutionary psychology: new perspectives on cognition and motivation. Annu. Rev. Psychol. 64, 201–229. http://dx.doi.org/10.1146/annurev.psych.121208.131628.
Dall, S.R., McNamara, J.M., Leimar, O., 2015. Genes as cues: phenotypic integration of genetic and epigenetic information from a Darwinian perspective. Trends Ecol. Evol. 30, 327–333. http://dx.doi.org/10.1016/j.tree.2015.04.002.
Dill, L.M., 1983. Adaptive flexibility in the foraging behavior of fishes. Can. J. Fish. Aquat. Sci. 40, 398–408. http://dx.doi.org/10.1139/f83-058.
Donaldson-Matasci, M.C., Lachmann, M., Bergstrom, C.T., 2008. Phenotypic diversity as an adaptation to environmental uncertainty. Evol. Ecol. Res. 10, 493–515.
Dridi, S., Akçay, E., 2018. Learning to cooperate: the evolution of social rewards in repeated interactions. Am. Nat. 191, 58–73. http://dx.doi.org/10.1086/694822.
Dridi, S., Lehmann, L., 2014. On learning dynamics underlying the evolution of learning rules. Theor. Popul. Biol. 91, 20–36. http://dx.doi.org/10.1016/j.tpb.2013.09.003.
Dridi, S., Lehmann, L., 2015. A model for the evolution of reinforcement learning in fluctuating games. Anim. Behav. 104, 87–114. http://dx.doi.org/10.1016/j.anbehav.2015.01.037.
Dridi, S., Lehmann, L., 2016. Environmental complexity favors the evolution of learning. Behav. Ecol. 27, 842–850. http://dx.doi.org/10.1093/beheco/arv184.
Dunlap, A.S., Stephens, D.W., 2016. Reliability, uncertainty, and costs in the evolution of animal learning. Curr. Opin. Behav. Sci. 12, 73–79. http://dx.doi.org/10.1016/j.cobeha.2016.09.010.
English, S., Fawcett, T.W., Higginson, A.D., Trimmer, P.C., Uller, T., 2016. Adaptive use of information during growth can explain long-term effects of early life experiences. Am. Nat. 187, 620–632. http://dx.doi.org/10.1086/685644.
Enquist, M., Lind, J., Ghirlanda, S., 2016. The power of associative learning and the ontogeny of optimal behavior. Royal Soc. Open Sci. 3, 160734. http://dx.doi.org/10.1098/rsos.160734.
Epstein, J.M., 2008. Why model? J. Artif. Soc. Soc. Simul. 11, 12.
Fawcett, T.W., Fallenstein, B., Higginson, A.D., Houston, A.I., Mallpress, D.E., Trimmer, P.C., McNamara, J.M., 2014. The evolution of decision rules in complex environments. Trends Cogn. Sci. 18, 153–161. http://dx.doi.org/10.1016/j.tics.2013.12.012.
Fawcett, T.W., Frankenhuis, W.E., 2015. Adaptive explanations for sensitive windows in development. Front. Zool. 12 (Suppl. 1), S3. http://dx.doi.org/10.1186/1742-9994-12-S1-S3.
Fawcett, T.W., Hamblin, S., Giraldeau, L.-A., 2013. Exposing the behavioral gambit: the evolution of learning and decision rules. Behav. Ecol. 24, 2–11. http://dx.doi.org/10.1093/beheco/ars085.
Fischer, B., Van Doorn, G.S., Dieckmann, U., Taborsky, B., 2014. The evolution of age-dependent plasticity. Am. Nat. 183, 108–125. http://dx.doi.org/10.1086/674008.
Frank, S.A., 1996. The design of natural and artificial adaptive systems. In: Rose, M.R., Lauder, G.V. (Eds.), Adaptation. Academic Press, New York, pp. 451–505.
Frank, S.A., 1997. The design of adaptive systems: optimal parameters for variation and selection in learning and development. J. Theor. Biol. 184, 31–39. http://dx.doi.org/10.1006/jtbi.1996.0241.
Frankenhuis, W.E., Fraley, R.C., 2017. What do evolutionary models teach us about sensitive periods in psychological development? Eur. Psychol. 22, 141–150.
Frankenhuis, W.E., Panchanathan, K., 2011. Balancing sampling and specialization: an adaptationist model of incremental development. Proc. Royal Soc. B 278, 3558–3565. http://dx.doi.org/10.1006/jtbi.1996.0241.
Frankenhuis, W.E., Panchanathan, K., Barrett, H.C., 2013. Bridging developmental

systems theory and evolutionary psychology using dynamic optimization. Dev. Sci. 16, 584–598. http://dx.doi.org/10.1111/desc.12053.
Frankenhuis, W.E., Tiokhin, L., 2018. Bridging evolutionary biology and developmental psychology: toward an enduring theoretical infrastructure. Child Dev. http://dx.doi.org/10.1111/cdev.13021. Advance online publication.
Gardner, A., 2009. Adaptation as organism design. Biol. Lett. 5, 861–864. http://dx.doi.org/10.1098/rsbl.2009.0674.
Gilbert, S.F., Baressi, M., 2016. Developmental Biology, 11th. Sinauer Associates, Sunderland, MA.
Gomulkiewicz, R., Kirkpatrick, M., 1992. Quantitative genetics and the evolution of reaction norms. Evolution 46, 390–411. http://dx.doi.org/10.1111/j.1558-5646.1992.tb02047.x.
Grafen, A., 1984. Natural selection, kin selection and group selection. In: Krebs, J.R., Davies, N.B. (Eds.), Behavioural Ecology: An Evolutionary Approach, 2nd ed. Blackwell Scientific Publications, Oxford, pp. 62–84.
Grafen, A., 2014. The formal darwinism project in outline. Biol. Philos. 29, 155–174. http://dx.doi.org/10.1007/s10539-013-9414-y.
Hammerstein, P., 1996. Darwinian adaptation, population genetics and the streetcar theory of evolution. J. Math. Biol. 34, 511–532. http://dx.doi.org/10.1007/BF02409748.
Hinton, G.E., Nowlan, S.J., 1987. How learning can guide evolution. Complex Syst. 1, 495–502.
Houston, A., Clark, C., McNamara, J., Mangel, M., 1988. Dynamic models in behavioural and evolutionary ecology. Nature 332, 29–34. http://dx.doi.org/10.1038/332029a0.
Houston, A.I., McNamara, J.M., 1999. Models of Adaptive Behavior: An Approach Based on State. Cambridge University Press, Cambridge.
Houston, A.I., McNamara, J.M., 2005. John Maynard Smith and the importance of consistency in evolutionary game theory. Biol. Philos. 20, 933–950. http://dx.doi.org/10.1007/s10539-005-9016-4.
Jablonka, E., Oborny, B., Molnar, I., Kisdi, E., Hofbauer, J., et al., 1995. The adaptive advantage of phenotypic memory in changing environments. Philos. Trans. R. Soc. B 350, 133–141. http://dx.doi.org/10.1098/rstb.1995.0147.
Kacelnik, A., 2012. Putting mechanisms into behavioral ecology. In: Hammerstein, P., Stevens, J.R. (Eds.), Evolution and the Mechanisms of Decision Making. MIT Press, Cambridge, MA, pp. 21–38.
Kacelnik, A., Bateson, M., 1997. Risk-sensitivity: crossroads for theories of decision-making. Trends Cogn. Sci. 1, 304–309. http://dx.doi.org/10.1016/S1364-6613(97)01093-0.
Krebs, J.R., Kacelnik, A., Taylor, P., 1978. Test of optimal sampling by foraging great tits. Nature 275, 27–31. http://dx.doi.org/10.1038/275027a0.
Kruschke, J.K., 2008. Bayesian approaches to associative learning: from passive to active learning. Learn. Behav. 36, 210–226. http://dx.doi.org/10.3758/LB.36.3.210.
Lachmann, M., Jablonka, E., 1996. The inheritance of phenotypes: an adaptation to fluctuating environments. J. Theor. Biol. 181, 1–9. http://dx.doi.org/10.1006/jtbi.1996.0109.
Laland, K.N., Odling-Smee, J., Feldman, M.W., 2001. Cultural niche construction and human evolution. J. Evol. Biol. 14, 22–33. http://dx.doi.org/10.1046/j.1420-9101.2001.00262.x.
Leimar, O., Hammerstein, P., Van Dooren, T.J.M., 2006. A new perspective on developmental plasticity and the principles of adaptive morph determination. Am. Nat. 167, 367–376. http://dx.doi.org/10.1086/499566.
Levins, R., 1966. The strategy of model building in population biology. Am. Sci. 54, 421–431.
Lewontin, R.C., Cohen, D., 1969. On population growth in a randomly varying environment. Proc. Natl. Acad. Sci. U. S. A. 62, 1056–1060. http://dx.doi.org/10.1073/pnas.62.4.1056.
Littman, M.L., 2009. A tutorial on partially observable Markov decision processes. J. Math. Psychol. 53, 119–125. http://dx.doi.org/10.1016/j.jmp.2009.01.005.
Mameli, M., Bateson, P., 2011. An evaluation of the concept of innateness. Philos. Trans. R. Soc. B 366, 436–443. http://dx.doi.org/10.1098/rstb.2010.0174.
Mangel, M., 1990. Dynamic information in uncertain and changing worlds. J. Theor. Biol. 146, 317–332. http://dx.doi.org/10.1016/S0022-5193(05)80742-8.
Mangel, M., 2015. Stochastic dynamic programming illuminates the link between environment, physiology, and evolution. Bull. Math. Biol. 77, 857–877. http://dx.doi.org/10.1007/s11538-014-9973-3.
Mangel, M., Clark, C.W., 1988. Dynamic Modeling in Behavioral Ecology. Princeton University Press, Princeton, NJ.
Maynard Smith, J.T., 1978. Optimization theory in evolution. Annu. Rev. Ecol. Syst. 9, 31–56. http://dx.doi.org/10.1146/annurev.es.09.110178.000335.
Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D., Wolpert, L., 1985. Developmental constraints and evolution: a perspective from the mountain lake conference on development and evolution. Q. Rev. Biol. 60, 265–287. http://dx.doi.org/10.1086/414425.
Mayr, E., 1983. How to carry out the adaptationist program? Am. Nat. 121, 324–334.
McNamara, J.M., Dall, S.R., Hammerstein, P., Leimar, O., 2016. Detection vs. selection: integration of genetic, epigenetic and environmental cues in fluctuating environments. Ecol. Lett. 19, 1267–1276. http://dx.doi.org/10.1111/ele.12663.
McNamara, J., Houston, A., 1980. The application of statistical decision theory to animal behaviour. J. Theor. Biol. 85, 673–690. http://dx.doi.org/10.1016/0022-5193(80)90265-9.
McNamara, J.M., Houston, A.I., 1986. The common currency for behavioral decisions. Am. Nat. 127, 358–378. http://dx.doi.org/10.1086/284489.
McNamara, J.M., Houston, A.J., 2009. Integrating function and mechanism. Trends Ecol. Evol. 24, 670–675. http://dx.doi.org/10.1016/j.tree.2009.05.011.
Moran, N.A., 1992. The evolutionary maintenance of alternative phenotypes. Am. Nat. 139, 971–989. http://dx.doi.org/10.1086/285369.
Nettle, D., Frankenhuis, W.E., Rickard, I.J., 2013. The evolution of predictive adaptive responses in human life history. Proc. Royal Soc. B 280, 20131343. http://dx.doi.org/10.1098/rspb.2013.1343.
Niv, Y., Joel, D., Meilijson, I., Ruppin, E., 2002. Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. Adaptive Behav. 10, 5–24. http://dx.doi.org/10.1177/10597123020101001.
Nolfi, S., Parisi, D., Elman, J.L., 1994. Learning and evolution in neural networks. Adapt. Behavior 3, 5–28. http://dx.doi.org/10.1177/105971239400300102.
Oudeyer, P.Y., Smith, L.B., 2016. How evolution may work through curiosity-driven developmental process. Topics Cogn. Sci. 8, 492–502. http://dx.doi.org/10.1111/tops.12196.
Oyama, S., Griffiths, P.E., Gray, R.D. (Eds.), 2001. Cycles of Contingency: Developmental Systems and Evolution. MIT Press, Cambridge, MA.
Panchanathan, K., Frankenhuis, W.E., 2016. The evolution of sensitive periods in a model of incremental development. Proc. Royal Soc. B 283, 20152439. http://dx.doi.org/10.1098/rspb.2015.2439.
Parker, G.A., Maynard Smith, J., 1990. Optimality theory in evolutionary biology. Nature 348, 27–33. http://dx.doi.org/10.1038/348027a0.
Pfab, F., Gabriel, W., Utz, M., 2016. Reversible phenotypic plasticity with continuous adaptation. J. Math. Biol. 72, 435–466. http://dx.doi.org/10.1007/s00285-015-0890-3.
Powell, W.B., 2007. Approximate Dynamic Programming: Solving the Curses of Dimensionality. John Wiley & Sons.
Rivoire, O., Leibler, S., 2014. A model for the generation and transmission of variations in evolution. Proc. Natl. Acad. Sci. 111, E1940–E1949. http://dx.doi.org/10.1073/pnas.1323901111.
Samuels, R., 2004. Innateness in cognitive science. Trends Cogn. Sci. 8, 136–141. http://dx.doi.org/10.1016/j.tics.2004.01.010.
Samuelson, L., Swinkels, J.M., 2006. Information: evolution and utility. Theor. Econ. 1, 119–142.
Schlichting, C.D., Pigliucci, M., 1998. Phenotypic Evolution: A Reaction Norm Perspective. Sinauer Associates, Sunderland.
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, 484–489. http://dx.doi.org/10.1038/nature16961.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Chen, Y., 2017. Mastering the game of Go without human knowledge. Nature 550, 354–359. http://dx.doi.org/10.1038/nature24270.
Singh, S., Lewis, R.L., Barto, A.G., 2009. Where do rewards come from. In: Taatgen, N.A., van Rijn, H. (Eds.), Proceedings of the 31 St Annual Conference of the Cognitive Science Society. Austin TX. pp. 2601–2606.
Singh, S., Lewis, R.L., et al., 2010. Intrinsically motivated reinforcement learning: an evolutionary perspective. IEEE Trans. Auton. Ment. Dev. 2, 70–82. http://dx.doi.org/10.1109/TAMD.2010.2051031.
Skinner, B.F., 1953. Science and Human Behavior. NY: Macmillan, New York.
Smaldino, P.E., 2017. Models are stupid, and we need more of them. In: Vallacher, R.R., Read, S.J., Nowak, A. (Eds.), Computational Social Psychology, pp. 311–331 Routledge New York.
Snell-Rood, E.C., 2012. Selective processes in development: implications for the costs and benefits of phenotypic plasticity. Integr. Comp. Biol. 52, 31–42. http://dx.doi.org/10.1093/icb/ics067.
Sorg, J.D., 2011. The Optimal Reward Problem: Designing Effective Reward for Bounded Agents, Doctoral Dissertation. University of Michigan.
Staddon, J.E., 2016. Adaptive Behavior and Learning. Cambridge University Press.
Stamps, J., Frankenhuis, W.E., 2016. Bayesian models of development. Trends Ecol. Evol. 31, 260–268. http://dx.doi.org/10.1016/j.tree.2016.01.012.
Starrfelt, J., Kokko, H., 2012. Bet-hedging –a triple tradeoff between means, variances and correlations. Biol. Rev. 87, 742–755. http://dx.doi.org/10.1111/j.1469-185X.2012.00225.x.
Stephens, D.W., Krebs, J.R., 1986. Foraging Theory. Princeton University Press.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Sultan, S.E., Spencer, H.G., 2002. Metapopulation structure favors plasticity over local adaptation. Am. Nat. 160, 271–283. http://dx.doi.org/10.1086/341015.
Tesauro, G.J., 1994. TD-Gammon, a self-teaching backgammon program, achieves master- level play. Neural Comput. 6, 215–219. http://dx.doi.org/10.1162/neco.1994.6.2.215.
Thorndike, E.L., 1911. Animal Intelligence: Experimental Studies. Macmillan, New York.
Trimmer, P.C., Houston, A.I., 2014. An evolutionary perspective on information processing. Topics Cogn. Sci. 6, 312–330. http://dx.doi.org/10.1111/tops.12085.
Trimmer, P.C., Houston, A.I., Marshall, J.A., Mendl, M.T., Paul, E.S., McNamara, J.M., 2011. Decision-making under uncertainty: biases and Bayesians. Anim. Cogn. 14, 465–476. http://dx.doi.org/10.1007/s10071-011-0387-4.
Trimmer, P.C., McNamara, J.M., Houston, A.I., Marshall, J.A.R., 2012. Does natural selection favour the Rescorla-Wagner rule? J. Theor. Biol. 302, 39–52. http://dx.doi.org/10.1016/j.jtbi.2012.02.014.
Uller, T., English, S., Pen, I., 2015. When is incomplete epigenetic resetting in germ cells favoured by natural selection? Proc. R. Soc. B 282, 20150682. http://dx.doi.org/10.1098/rspb.2015.0682.
Watson, R.A., Szathmáry, E., 2016. How can evolution learn? Trends Ecol. Evol. 31, 147–157. http://dx.doi.org/10.1016/j.tree.2015.11.009.
West-Eberhard, M.J., 2003. Developmental Plasticity and Evolution. Oxford University Press, New York.
Whalen, A., Cownden, D., Laland, K., 2015. The learning of action sequences through social transmission. Anim. Cogn. 18, 1093–1103. http://dx.doi.org/10.1007/s10071-015-0877-x.