Reinforcement Learning Theory Reveals the Cognitive Requirements for Solving the Cleaner Fish Market Task

Andrés E. Quiñones,1,* Olof Leimar,2 Arnon Lotem,3 and Redouan Bshary4

1. Laboratorio de Biología Evolutiva de Vertebrados, Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia; 2. Department of Zoology, Stockholm University, Stockholm, Sweden; 3. School of Zoology, Tel Aviv University, Tel Aviv, Israel; 4. Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland

Submitted September 28, 2018; Accepted September 17, 2019; Electronically published March 2, 2020

Online enhancements: supplemental PDF. Dryad data: https://doi.org/10.5061/dryad.pnx0k6h5.

Abstract: Learning is an adaptation that allows individuals to respond to environmental stimuli in ways that improve their reproductive outcomes. The degree of sophistication in learning mechanisms potentially explains variation in behavioral responses. Here, we present a model of learning that is inspired by documented mechanisms potentially explains variation in behavioral responses. Here, we present a model of learning that is inspired by documented intra- and interspecific variation in the performance of a simultaneous two-choice task, the biological market task. The task presents a problem that cleaner fish often face in nature: choosing between two client types, one that is willing to wait for inspection and one that may leave if ignored. The cleaner’s choice hence influences the future availability of clients (i.e., it influences food availability). We show that learning the preference that maximizes food intake requires subjects to represent in their memory different combinations of pairs of client types rather than just individual client types. In addition, subjects need to account for future consequences of actions, either by estimating expected long-term reward or by experiencing a client leaving as a penalty (negative reward). Finally, learning is influenced by the absolute and relative abundance of client types. Thus, cognitive mechanisms and ecological conditions jointly explain intra- and interspecific variation in the ability to learn the adaptive response.

Keywords: biological markets, mutualism, cognition, reinforcement learning, decision-making.

Introduction

Animals must face and appropriately respond to the ever-changing nature of environmental conditions. Cognitive processes are broadly defined as the mechanisms by which animals acquire, process, and store environmental information in order to properly respond to changing environments (Shettleworth 2010). From an evolutionary perspective, we often assume that interspecific variation in cognition is brought about by the distinct action of natural selection in each environment (Chittka et al. 2012). However, there is often also intraspecific variation in cognitive abilities (Thornton and Lukas 2012). Furthermore, the same species often faces different environments within its distribution. This raises the question of whether differences in behavior and in cognitive abilities more generally are due to local cognitive adaptation or to similar cognitive machineries responding to different environmental situations. One step toward answering this question is to have formalized expectations of how different cognitive systems behave in different environmental situations.

Associative learning is a key cognitive mechanism that allows individuals to associate rewards with environmental stimuli and appropriate behavior (Staddon 2016). Using these associations to respond to environmental changes is particularly adaptive in complex environments (Dridi and Lehmann 2016). Furthermore, we expect natural selection to shape the cognitive machinery in charge of learning processes to use the information available in a particular environment (Shettleworth 2009). The study of associative learning has traditionally been pursued by academic traditions related to the field of psychology (Enquist et al. 2016; Staddon 2016), where researchers use a combination of experiments and mathematical models to infer the structural and functional characteristics of the learning machinery of animals (Rescorla and Wagner 1972; Staddon 2016). However, the aim of these studies is not typically to understand intra- and interspecific variation, and rarely do they include the ecology of the species as an explanatory variable (Kamil 1983; McAuliffe and Thornton 2015). To make such links, we need a strong interaction between empirical data on intra- and interspecific behavioral variation
and modeling of learning processes. Here, we present a foraging model in the tradition of associative learning. Unlike typical evolutionary models in behavioral ecology that capture the action of natural selection in changing behavioral preferences, we model how the preferences of individuals develop over their lifetimes through learning. Furthermore, we compare how different learning mechanisms shape the developed preference under different ecological conditions. We see this as a necessary stepping-stone to develop models of the evolution of cognitive machineries, which is in line with recent interest in combining behavioral mechanisms and evolutionary considerations (Fawcett et al. 2013; McNamara 2013).

In our model, individuals face the decision of choosing between an ephemeral food source and a more permanent one in a simultaneous two-choice task. The task was originally designed to test the predictions derived from biological market theory on the cleaner fish Labroides dimidiatus (Bshary and Grutter 2002). Biological market theory emphasizes the importance of partner choice options for explaining payoff distributions among cooperating partners (Noë and Hammerstein 1995; Hammerstein and Noë 2016). Variation in partner choice options is a feature of the cleaning mutualism involving L. dimidiatus. The cleaners occupy small territories, so-called cleaning stations, in which they remove ectoparasites from “client” fishes (Côté 2000). Based on their home range sizes, most client species can be categorized as either residents (small home range with access to only one cleaning station) or visitors (large home range with access to two or more cleaning stations; Bshary 2001). Only visitors can choose among cleaners based on the quality of the service provided. One aspect of service quality is receiving priority: because cleaners have about 2,000 interactions per day (Grutter 1994), it regularly happens that two clients simultaneously seek the service of the same cleaner, so one has to wait for inspection (Bshary 2001). In such situations, residents have to wait; they lack alternative options, while visitors can use the partner choice options and swim off to visit another station if initially ignored (Bshary and Schäffer 2002). Consequently, cleaners can increase food intake by giving priority to visitors. This decision rule has been observed in nature (Bshary 2001).

In laboratory experiments that mimic the behavioral rules of residents and visitors, cleaners readily learn to give preference to an ephemeral food-offering plate over a more permanent one in a simultaneous two-choice task (Bshary and Grutter 2002). While this result has been reproduced three times (Salwiczek et al. 2012; Wismer et al. 2014; Triki et al. 2018), it has become clear that juvenile cleaners and cleaners from some habitats fail to readily learn the solution to this biological market task (Salwiczek et al. 2012; Wismer et al. 2014; Triki et al. 2018). As cleaners are open-water spawners with pelagic eggs and larvae (i.e., they live in well-mixed populations), this documented intraspecific variation in learning performance is difficult to explain by local genetic adaptation. Instead, it may better be accounted for by different ontogenetic trajectories developed under different conditions. In addition, it has become clear that various relatively large-brained species perform poorly at this task (chimpanzees, orangutans, capuchins, rats, and pigeons), with African gray parrots being an exception (Pepperberg and Hartsfield 2014). Therefore, it is of great interest to model the biological market task and to generate testable predictions regarding the learning mechanisms and/or the ecological conditions that may explain intra- and interspecific variation in the ability to solve an apparently complex foraging problem.

The Model

Overview

Given that the biological market task was designed to capture the interactions between cleaner fish and their client reef fish, our model is set up with this ecological problem in mind. We model a cleaner’s learning as an associative process, where cleaners associate behavioral actions in specific environmental states (hereafter, simply “states”) with the level of reward obtained from those actions. The states are the clients available to the cleaner; the actions are choosing which of those clients the cleaner serves; and the reward is the amount of food the cleaner obtains by cleaning. The choice of client might determine which clients are available in the next step (fig. 1A). In the model, we refer to the cleaner as an “agent.” During learning, an agent experiences a series of states (sets of clients), chooses actions (which client to clean), and obtains food, which triggers rewards. These rewards allow the agent to learn the long-term reward that can be expected from those states. Hereafter, we refer to that long-term expectation as a “value.” Consequently, to maximize rewards, agents use values to choose the actions, where the choice of action may be viewed as the decision-making process.

The learning dynamics are captured using the reinforcement learning (RL) formalism developed in machine learning (Sutton and Barto 2018). RL provides the theoretical basis for many important advances in artificial intelligence. More importantly, core concepts in RL have found parallels in the physiological mechanisms of the mammalian brain and fit many psychological phenomena of learning (Glimcher 2011; Sutton and Barto 2018). The update procedure that improves estimates and preferences in RL is driven by the prediction error, referred to as a “temporal-difference (TD) error.” The prediction error
is the difference between the expected and the observed value. Notably, the idea that certain populations of dopaminergic neurons communicate a quantity like the prediction error has support in the neurobiological literature (Montague et al. 1996; Niv 2009; Glimcher 2011). Dopamine activity is known to be related to the reward system in the brain. Key experiments show that changes in the phasic activity of dopaminergic neurons correlate not with the amount of reward but with the mismatch between the obtained and the expected reward. Therefore, in principle, animals could use the signal of dopaminergic neurons to improve their estimates of reward, just like RL algorithms use the prediction error. Furthermore, one particular family of RL algorithms uses the prediction error in two different modules: one in charge of improving value estimation (the critic) and one in charge of decision-making (the actor). These are the so-called actor-critic approaches. In the brain, dopaminergic neurons are also known to influence two separate subdivisions of the striatum: the ventral striatum and the dorsal striatum. Dopamine seems to alter patterns of neural plasticity differently in both areas. The dorsal striatum is involved in neural processes related to action selection, while the ventral striatum is thought to be critical for reward processing (Sutton and Barto 2018). Overall, TD algorithms, and the actor-critic approach in particular, capture important details of animal learning.

Figure 1: General description of the model setup. A. When facing two options in the cleaning station, clients must choose to prioritize one of the options. This decision can partly influence the presence of clients in the following time step. B. Learning agents use rewards obtained from their interaction in a particular state (client combination) to estimate values representing the expected long-term reward. Hence, a value is the sum of current and future rewards, discounting the later ones according to how far into the future they are obtained (depicted in the definition of “true” value, \( \hat{V}(S_t) \)). To estimate \( V(S_t) \), the value of a state at time \( t \), agents associate the state with the reward obtained and the expected value of the state that follows. The value estimate of the following time step, \( V(S_{t+1}) \), includes the one after that, \( V(S_{t+2}) \), and so on, so by adding the value of two states following each other, agents estimate the long-term reward. In C and D, we show the actor-critic approach, in which a prediction error (\( \delta \), see text for details) determines the updates of the value estimates for each state (the critic) as well as the preferences (the actor). We model two types of learning agents. Partially aware agents estimate the value of states defined by the client type chosen, irrespective of other options available (C); in contrast, fully aware agents estimate the value of the states defined by the pair of clients present in the cleaning station (D).
(Takahashi et al. 2008). In the following subsections, we describe in more detail how we adapt the actor-critic approach to the cleaner system, with particular emphasis on four key components of the model, namely, the states, the agent’s representation of those states, the decision rule, and the update rule. There are, however, a vast variety of RL algorithms that differ in specific details of the learning sequence. Thus, in the supplemental PDF, available online, we show an alternative implementation of the model using the SARSA algorithm (Sutton and Barto 2018), with a softmax function as the policy component. SARSA stands for state-action-reward-state-action, which gives a short description of the sequence of events that leads to learning in this algorithm. In contrast to the actor-critic approach presented here, the SARSA algorithm is part of a set of models, such as the Rescorla-Wagner (1972) or the relative payoff sum (Harley 1981; Maynard Smith 1982) learning rules, that use values directly in the decision process. In other words, these models do not explicitly implement an actor module; instead, agents choose actions according to their relative value. A clear limitation of the SARSA algorithm is that with a reasonably flexible decision rule, such as the softmax function that is commonly applied, the preference developed by the agent for one option is determined by the difference in value between that option and the others. Hence, in cases where one of the options is better but the value difference is small (such as in the market task), the SARSA algorithm will not develop a strong preference for the better option. Because of this limitation, we consider the actor-critic approach a more suitable implementation, and we present it in the main text. However, SARSA and related models have had a long tradition in the psychological literature (see previous references) and an appeal in behavioral ecology due to their simplicity, which facilitates analytical tractability (Dridi and Lehmann 2014; Enquist et al. 2016; Dridi and Akçay 2017; Wubs et al. 2018). Thus, we present the SARSA implementation in the supplemental PDF. Overall, our main results hold, irrespective of the specific algorithm.

States

The ecological foraging environment of cleaners is the population of clients that swim into their cleaning stations. We define a state as the set of clients available to the cleaner at a given time point. We assume that up to two clients can be together at the cleaning station. For the sake of simplicity and fitting the laboratory experiments, we assume that all clients offer identical amounts of food but belong to two distinct species, which can also be classified into two types: residents and visitors. Furthermore, we assume that the two types differ in a morphological character that is easily perceived by the cleaners, for example, color. Thus, cleaners distinguish between the two types.

The potential states that an agent faces are the different combinations of client types, including the absence of a client, in the spots available; these combinations sum to six states (fig. 2).

We use two different processes for assigning clients to the cleaning station: one that mimics the cleaners’ natural environment and one that mimics the experimental setup in which their cognitive abilities are assessed (Salwiczek et al. 2012; Wismer et al. 2014; Triki et al. 2018). In the experimental setup, agents face the resident-visitor choice in the first step. If they choose the visitor, they have access to the resident alone in the following time step. If they choose the resident, the visitor will not be available for cleaning; thus, the cleaner faces an empty cleaning station. Right after either of those situations, the agent again faces the resident-visitor option, and once again, the following state is determined by their choice. Note that in the experimental setup, cleaners face only three of all the possible states. In the natural setup, residents and visitors visit the cleaning station according to their relative frequency in the environment. A resident fills up one of the two empty spots in the station with probability \( p_r \), a visitor does so with probability \( p_v \), and the spot remains empty with probability \( p_e = 1 - p_r - p_v \). Hence, the states, corresponding to pairs of clients (and not to individual spots), reach the cleaning station with probabilities analogous to the Hardy-Weinberg equilibrium for one locus with three alleles. Residents and visitors leave the station if they are not given priority with probability \( l_r \) and \( l_v \), respectively. Visitors are clients with large range.

| States          | Actions  |
|-----------------|----------|
| Client 1       | Client 2 |
| resident       | visitor  |
| resident       | -        |
| visitor        | -        |
| resident       | resident |
| visitor        | visitor  |
| -              | -        |

Figure 2: Possible combination of clients visiting an agent’s cleaning station. The combination of both clients is the environmental state, while which of those clients the agent serves is the action that the agent takes. Reinforcement learning literature refers to the combination of the environmental state and the action taken as the state-action pair. We present a model with six states and seven state-action pairs.
sizes and access to more than one cleaning station, while residents have access to only one station; thus, visitors are generally more likely to leave if unattended (l_v > l_r; fig. 1A). Clients that leave the station unattended head to a different cleaning station to seek cleaning service.

All of these parameters (p_r, p_v, l_v, l_r), together with the cleaner’s decision of which client to attend to when two clients are present simultaneously, determine the transition probabilities between the different states. Technically, the stochastic process that determines the transition between the different states is Markovian, or memoryless, because states faced before the current one do not influence which state will come in the following time step. Formally, such a process is called a “Markov decision process” (Puterman 2014).

Agents’ State Representations

RL comprises a set of techniques developed to find optimal behaviors in Markov decision processes. A useful way to find such optimal solutions is to estimate the value of acting in a particular way when exposed to a specific state and to prefer the action among those available that yields the higher value gain. In figure 2, we show the list of states and their available actions for our model of cleaning agents. An agent that has a full representation of the states would save separate values for each state in memory. This full representation of the states implies that the agent has the capacity to recognize combinations of clients as different from each other. This capacity, even though a reasonable choice from an engineering point of view, cannot be assumed when dealing with our particular setup if the aim is to understand biological agents. In fact, when considered more generally, the number of possible state combinations of stimuli in the environment can be extremely high, which means that representing all of them in memory is not feasible and selecting the relevant state combinations that should be learned is a known cognitive challenge (Goldstein et al. 2010). A cognitively simpler and more biologically sound approach is to assume that by default agents estimate a value for the action of choosing a particular client type (resident, visitor, or absence). Such a learning agent implicitly pulls together the values of all states that include a particular type of client. Hence, an agent is unaware of the potentially different values that the choice to serve a client can result in depending on the other client it is matched with. Thus, we refer to them hereafter as “partially aware agents” (PAAs; fig. 1C). In contrast, we refer to agents with a representation in memory of all the client combinations (states) as “fully aware agents” (FAAs; fig. 1D). We use our model to investigate how learning influences the preferences of these two types of agent.

Decision Rule

Agents choose between alternative actions according to the set of preferences they develop over the learning process. In cases where one of the two types of client is alone, we assume that the cleaner decides to attend to the available client. In cases where both options are of the same type, we assume that these options are indistinguishable for the cleaner; thus, it chooses randomly with a 0.5 probability. Thus, there is only one state (the heterotypic case) that offers different actions to choose from: the state in which there is a resident and a visitor. For this state, the agent weighs its preference for the visitor (θ_v) against the preference for the resident (θ_r) to make a decision. The agent translates its preferences into probability using a logistic function. Concretely, the probability of choosing the action i over the action j is given by

\[ \pi_i = \frac{1}{1 + e^{-(\theta_i - \theta_j)}}, \]

and the complementary probability (\(\pi_j = 1 - \pi_i\)) is the probability of choosing the action j.

Update Rule

As we mentioned earlier, among RL methods, we use TD methods because of the similarities found in neuroscience between the function of the prediction error and the role of dopaminergic neurons in natural learning processes. Among the TD methods, we develop our model based on the actor-critic approach. The actor-critic approach assumes that learning processes are separated into two modules (fig. 1C, 1D): one in charge of estimating values (the critic) and one defining preferences and actions (the actor). These two modules, however, are not independent. The update rule used by both modules is based on the prediction error, a quantity measuring the mismatch between the obtained and the expected reward.

Learning agents update their estimate of value for a state every time they face it, and they update their preference for an action every time they take it. These updates are proportional to the prediction error. The prediction error corresponds to the difference between expected and observed value of a state. The expected value is simply the current estimate backed up in the agent’s memory. As for the observed value, agents obviously cannot directly observe the long-term value obtained from an action because the value includes both current and future rewards. Instead, the observed value is the sum of the immediate reward obtained from the action taken, which the agent directly experiences, and the discounted estimate of the value obtained from the state that the agent will face in
the next time step (fig. 1B). Formally, the prediction error is given by

$$\delta_t = R_t + \gamma V(S_{t+1}) - V(S_t), \quad (2)$$

where $\gamma$ is the time-discounting parameter, $V(S_t)$ is the estimated value at time $t$ of the state faced at time $t$ by FAAs (or the client type chosen at time $t$ by PAAs), and $V(S_{t+1})$ is the estimated value at time $t$ of the state faced at time $t + 1$. Finally, $R_t$ is the total reward obtained by the agent from taking an action at time $t$.

In the overview of the model, we mentioned that reward for the agents in our model represents the amount of food for the cleaner. In other words, we assume that cleaners are hardwired to use food as a primary reinforcer, an environmental input that innately motivates the behavioral choices of animals. Note, however, that reward is not directly the amount of food but may rather be viewed as an evolved neuronal signal in the cognitive system triggered by food. The underlying assumption is that the cognitive machinery has been shaped by natural selection to use certain environmental inputs as drivers of behavior in ways that improve reproductive outcomes (Robson and Samuelson 2011; Schultz 2015). Thus, which environmental stimuli an animal uses as a reinforcer will depend on the ecology that has shaped the cognitive machinery over evolutionary time (Shettleworth 2009; Dridi and Akçay 2017). Moreover, animals may be selected to use more than one reinforcer when assigning value. In the cleaner example, estimating long-term reward ($\gamma > 0$) might be a cognitively expensive process, while using behavioral cues from clients to estimate the future consequences of actions might be cheaper. Cleaners, for instance, could use the event of a client leaving their station without being served as a penalty (negative reward), a numerical signal, just like a reward, that when obtained reduces the estimated value associated with an action (fig. 1A). This would allow them to have the information about the consequence of their choice without going through a more complicated, and perhaps costlier, estimation. To allow for this, we consider two sources of reward: one positive and one negative. We let the total reward be given by

$$R_t = P_t + N_t, \quad (3)$$

where $P_t$ and $N_t$ are the magnitudes of the positive reward and the negative reward (penalty), respectively, obtained from taking an action at state $S$ at time $t$. Because each client offers the same amount of food ($P_t$ has the same numerical value for residents and visitors), cleaners maximize food intake by maximizing the time spent by clients in their station. As we mentioned earlier, cleaners maximize the time spent by clients in the station by choosing visitors (the ephemeral choice) in the heterotypic state. As for the penalty, it is triggered whenever a client leaves the cleaning station without being served; hence, it depends on the probabilities with which each of the client types leaves the station if unattended ($l_i, l_j$). We vary the numerical value of the penalty to assess its effect on the learning process.

To avoid confusion, we would like to note that in the psychological literature, negative reward may frequently be referred to as “negative reinforcer” or “punishment.” We have decided to follow Sutton and Barto’s (2018) terminology and their distinction between reinforcers and rewards (see the discussion in chap. 14 of their book). Accordingly, “reinforcers” are environmental stimuli that cause a subsequent increase (decrease if negative) in the frequency of a behavior. In contrast, “reward” is used to refer to the numerical value of the signal triggered in the cognitive system when encountering the reinforcer. In that sense, “reward” is a property of the cognitive system rather than of the environment itself. This conceptualization of “reward” is in line with its use in neuroscience and economics (Robson and Samuelson 2011; Schultz 2015). We also refrain from using the term “punishment” to describe negative rewards. In behavioral ecology in general and in the cleaner fish system in particular, “punishment” often refers to a strategy that individuals taking part in social dilemmas might use to alter the payoff of their social partners (Clutton-Brock and Parker 1995; Bshary and Grutter 2005; Raihani et al. 2012). Thus, we decided to use “penalty” to refer to negative reward because of its intuitive appeal across fields.

The prediction error, defined in equation (2), is the basis for the update of the estimated value by the critic and for the update of the behavioral preference by the actor. The critic module uses the prediction error to improve the value estimates according to

$$\Delta V(S_t) = \alpha \delta_t, \quad (4)$$

where $\alpha$ is the speed of learning value estimates and $\Delta V(S_t)$ refers to the change in the value estimated during one time step for the state faced at time $t$ ($\Delta V(S_t) = V_{t+1}(S_t) - V_t(S_t)$). Through the actor module, the agent updates the preferences for the actions available according to

$$\Delta(\theta_i - \theta_j) = \beta \delta_t 2(1 - \pi_t), \quad (5)$$

where $\beta$ is the speed of learning preferences; $\pi_t$ is the probability of taking the action $i$, given in equation (1); and $\Delta(\theta_i - \theta_j)$ refers to the overall change in preference for the two mutually exclusive options during one time step ($\Delta(\theta_i - \theta_j) = (\theta_i - \theta_j)_{t+1} - (\theta_i - \theta_j)_t$). In other words, a positive value in equation (5) means an increment in the preference for $i$ and a decrement in the preference for $j$. This last expression (eq. [5]) derives from calculating the first derivative of the probability of taking an action
The American Naturalist

(eq. [1]) with respect to preference and dividing it by the original probability function (Sutton and Barto 2018). This captures the effect that changes in preference have on the probability of an action.

There is an important difference to note in the implementation of the actor module for PAAs and FAAs. PAAs, because their state representation includes individual client types and not client type combinations, do not strictly perceive actions as belonging to one particular state. We assume that they nevertheless update a preference for the client types they choose and use this preference to make decisions. However, the update for one type occurs under all the combinations of client types that include that type. In contrast, FAAs in our model update the preference for residents and visitors only when these are faced in the context of the heterotypic state. When residents and visitors are faced in another context (e.g., with another client of the same type), the preference update would not influence the choice in the heterotypic case; thus, we do not account for those changes in preference.

Simulations

To assess the influence of the different parameters of the model on the types of agents, we ran a set of 100 stochastic simulations for each parameter combination. We assume that the agents start with “educated” guesses for the values of states. In other words, cleaners know that clients provide food and more or less how much they provide based on a possible combination of innate predisposition and earlier experience. We are interested in how the agents develop a preference for one of the two options, not the overall estimation process; hence, the exact magnitude of the initial value is less important than the fact that all of the states that involve at least one client are initialized with the same value. Specifically, they are initialized depending on the environmental and cognitive parameters according to

\[
V_0 = P + \sum_{n=1}^{\infty} (1 - P_u)P \gamma^n = P \left( 1 + \frac{(1 - P_u)\gamma}{1 - \gamma} \right).
\]

Equation (6) captures the long-term expected reward of an action that provides \( P \) units of primary reward. The second term captures future rewards, which are exponentially discounted according to parameter \( \gamma \). Furthermore, the second term is weighted by the probability that there is at least one client in the station. In the simulation, agents experienced states (pairs of clients) and chose one of the two options for cleaning. Each decision made by the agent triggered an update in the estimated value for the option taken.

The code used for the simulations can be found in Quiñones (2019), and the data generated from the code are provided in the Dryad Digital Repository (https://doi.org/10.5061/dryad.pnvx0k6h5; Quiñones et al. 2019).

Results

Figure 3 shows the proportion of visitors chosen over residents in periods of 1,000 time steps in learning runs for both FAAs (fig. 3A, 3C) and PAAs (fig. 3B, 3D). The states that the agents face mimic either the ecological environment of cleaners (fig. 3A, 3B) or the experimental setup in which their learning abilities are assessed (fig. 3C, 3D). For each type of agent and environment, we show four different cognitive conditions. Specifically, we show simulations where agents include only future positive reward in their estimates (\( \gamma = 0.8, N = 0 \), yellow), agents use a client leaving as a penalty but do not include long-term reward in their estimates (\( \gamma = 0, N = -0.5 \), blue), agents include both long-term reward and penalty in their estimates (\( \gamma = 0.8, N = -0.5 \), red), and agents include neither of them (\( \gamma = 0, N = 0 \), black). When agents use only positive reward to estimate value, FAAs (fig. 3A, 3C) develop a preference for visitors over residents only if they account for future rewards (yellow). However, when FAAs use both reward and penalty in their estimation, they develop a preference for the visitor even if the estimation does not take future actions into account (fig. 3, blue). In contrast to FAAs, PAAs (fig. 3B, 3D) do not develop a preference for either option in the learning process; this is irrespective of the level of future discounting and the source of reward. In figure S2 (figs. S1–S4 are available online), we show that the results presented in figure 3 hold qualitatively for agents that use the SARSA algorithm with a softmax policy.

In figure 4, we show that the preference developed by the agents for visitors in the resident-visitor context is indeed caused by the visitor’s tendency to leave if unattended (for FAAs under natural conditions). We vary the probability that visitors leave the station after a resident has been given priority and find that the higher this probability is, the higher the preference agents develop for the visitor. If visitors do not leave the station (\( l_v = 0 \)), agents learning with different cognitive mechanisms do not differ from each other, and they do not develop a preference for either option. In figure S3, we show that agents that learn with the SARSA implementation qualitatively fit the patterns described for the actor-critic case.

To assess the effect of client abundance on the capacity to develop a preference for the visitor, we systematically varied the probability of having residents and visitors. We present the outcome of this analysis in figure 5 only for FAAs because PAAs do not develop a preference for the visitor irrespective of client abundance (data not shown). In figure 5A, we show how the choice of the learning agent
is affected by the overall client abundance, that is, the sum of the abundances of visitors and residents (we assume residents and visitors are equally abundant: $p_r = p_v$). This analysis confirms that either estimating long-term reward or using penalty is necessary to develop a preference for the visitor. However, agents using exclusively one of these two options develop different levels of preference depending on the overall client abundance. Agents that estimate long-term reward develop the highest preference under intermediate client abundances, while agents using only penalty develop higher preferences for the visitor with high overall abundance. We also extend the analysis to situations where visitors and residents are not equally abundant. We present that analysis in two triangles in figure 5B and 5C. Each side of a triangle represents increasing probability for one of the three options: resident, visitor, and absence of clients. Color depicts how the probability of choosing the visitor varies with client abundance. Figure 5B shows the results for agents that estimate long-term reward and use only positive reward ($\gamma = 0.8$). In this scenario, preference for the visitor is mainly influenced by the overall client abundance (inverse absence axis). Intermediate to low overall client abundance (center of triangle) yields the highest level of preference. The lowest preference for the visitor is found where the overall client abundance is very high irrespective of whether visitors or residents contribute to the overall abundance (lower part of triangle). The triangle in figure 5C corresponds to simulations where agents do not estimate long-term reward but include penalty in their estimations. In such conditions, agents develop a preference for the visitor over all of the parameter space. In the supplemental PDF (fig. S4), we show the same exploration for...
agents that use the SARSA algorithm. In this case, preferences for the visitor are overall much lower. However, the effect of client abundance goes in the same direction as in the actor-critic approach. Generally, high client abundance diminishes the preference for the visitor when agents use only positive reward.

**Discussion**

We have evaluated the cognitive toolbox that is needed for a learning agent to develop a preference for an ephemeral food source that is of equal value to a permanent food source, both in an experimental simultaneous two-choice task (the biological market task) and under conditions that capture the natural interactions between cleaner fish and their client reef fish. While the natural conditions are more complex, the same set of cognitive tools is needed in both situations; this is irrespective of the use of the actor-critic approach or the SARSA approach. Most importantly, our analyses suggest that simple associative learning is not enough to solve the task, despite the fact that prioritizing the ephemeral food source over the permanent one yields double the amount of food in each trial. Instead, agents need two additional tools. First, they need to be fully aware agents (i.e., represent each different situation separately and hence calculate values for each situation separately). This form of representation has been termed “configurational representation” (Sutherland and Rudy 1989) or “chunking” (Kolodny et al. 2015b). Second, agents need to use one mechanism that helps them to incorporate the future effects of a choice. Below, we first discuss these two additional tools, and then we ask how far the results may explain the observed intra- and interspecific variation in performance in the biological market task.

**Configurational Representation or Chunking as a Relevant Cognitive Tool**

Our choice of agents, FAAs and PAAs, highlights an important question in associative learning: how do animals choose the stimuli they associate with reward? In our model, PAAs associate only the stimulus from residents with the reward obtained from residents, irrespective of which other clients are present. In contrast, FAAs associate the stimuli of both clients present with the reward obtained. In practice, the key distinction FAAs make is to discriminate, during
the associative process, the context in which a choice is made. This allows the agent to associate different values with different contexts. This distinction is analogous to the one made in experimental psychology between elemental, or component, representation and configurational, or holistic, representation. In an elemental representation, while an animal learns, it develops associative strength with each individual stimulus such that the linear combination of stimuli

---

**Figure 5**: Mean probability of choosing a visitor over a resident under different levels of client abundances using the actor-critic approach and fully aware agents. In A, we show the effect of varying the overall client abundance ($p_r + p_v$), assuming residents and visitors are equally abundant. We show the proportion of visitors chosen over residents for the different cognitive adaptations evaluated before. Each point represents the average proportion over 100 replicates, and error bars show the interquartile range. The proportion in each replicate corresponds to only the last tenth of the time steps. In B and C, we extend that analysis for client combinations where residents and visitors are not equally abundant. To do that, we systematically varied the probability of resident and visitor (as well as the complementary probability of absence of clients). The axis of each triangle shows the relative abundance of the three possible options (visitor, resident, and absence). Color indicates the probability of choosing a visitor over a resident. B corresponds to simulations where agents use only positive reward and estimate long-term reward ($N = 0, \gamma = 0.8$; yellow points in A); here, agents have the highest probability when the abundances of clients are intermediate. C corresponds to simulations where agents use reward and penalty, but they estimate only short-term reward ($N = -0.5, \gamma = 0$; blue points in A); here, agents obtain higher discrimination between the visitor and the resident, and the abundance of clients has a much lower impact on the result. Black points inside the triangles show the location corresponding to the data plotted in A. Values were obtained by interpolating a series of learning simulations where we systematically varied the probability of having a resident ($p_r$) and a visitor ($p_v$). We varied these probabilities from 0.1 to 0.8 in increments of 0.1. For each combination of $p_r$ and $p_v$, we ran 30 replicates. To calculate the probability, we used only the last quarter of the choices, when most replicates had reached a steady state. Other parameter values are $l_r = 0$, $l_v = 1$, $\alpha = 0.01$, and $\beta = 0.01$. 
provides the total association with reward (Rescorla and Wagner 1972; Mackintosh 1975). In contrast, in a configurational representation an animal assigns all the associative strength to the combination of stimuli (Sutherland and Rudy 1989). This configurational representation allows animals to discriminate in nonlinear ways the different stimuli present at a given point. An example of a learning task that requires configurational learning is negative pattern discrimination (Woodbury 1943; Sutherland and Rudy 1989).

In this task, animals are presented with various stimuli, which when presented in isolation trigger reward but when presented all together do not. The idea here is that the subjects must distinguish between the compound stimuli in order to respond correctly. Using such an experimental paradigm, Sutherland and Rudy (1989) showed that the hippocampal formation is involved in configurational representation in rats. Given the evolutionary links among vertebrate brains (O’Connell and Hofmann 2011), it would be interesting to know whether cleaners perform better in general in tasks related to configurational representation and whether the region of their brain that is analogous to the hippocampal formation is involved in such an ability.

Another concept related to our choice of FAAs (and to configurational learning) is chunking (Miller 1956). Chunking is in a broad sense the capacity to collect individual units into sets and make these sets units of cognitive processes themselves. Early studies argued that chunking allows humans to use memory more efficiently and thus improve retention and processing of information (Miller 1956; Simon 1974). More recently, chunking is perceived as a central process in language acquisition, problem-solving, and experts’ memory (Conway and Christiansen 2001; Gobet et al. 2001; Kolodny et al. 2015a). For instance, experts’ above-average performance in certain tasks is explained by the way they chunk information related to their particular field of expertise (Gobet et al. 2001), and animal problem-solving ability is expected to be sensitive to the size of the chunks in their memory representation (Kolodny et al. 2015b). This perspective highlights the dynamical aspect of chunking. Nevertheless, we note that a mechanism that is proposed to be key for the human ability to acquire language appears to be a fundamental requirement for the ability to learn to solve the biological market task. As cleaner fish from some habitats excel at this task, it thus appears that a basic unit for language acquisition may either have a long evolutionary history among vertebrates or readily evolve in response to ecological need (i.e., rather independently of brain size).

**Future Consequences and Learning**

Being an FAA with the ability to chunk is not sufficient to solve the market task. FAAs must also connect the choice made in the current time step with its consequences in the following time step in order to solve the market task. It has been suggested that the capacity to make decisions while taking into account the future consequences of actions is an exclusive human capacity (Suddendorf and Corballis 2007). However, experiments on animals have challenged that view (Raby et al. 2007; Osvath and Osvath 2008; Osvath 2009). Part of the issue is that future planning is seen as an advanced cognitive capacity, different from associative learning (Osvath and Osvath 2008). From a theoretical point of view, models of associative learning that estimate long-term expected rewards either through chaining (Enquist et al. 2016) or through continuous learning of statistical regularities (Singh et al. 2010; Kolodny et al. 2014) can readily allow agents to solve tasks that require planning sequences of actions over time. Here, we have implemented a model in the philosophy of Enquist et al. (2016) by allowing agents to consider future consequences by estimating expected long-term reward (γ > 0). The analyses show that taking the future explicitly into account is indeed one possible tool to solve the market task (to develop a preference for an ephemeral food source). However, we also found an alternative, potentially much simpler mechanism that is at least equally efficient: negative RL. Interestingly, if an agent perceives a client swimming off or the removal of a plate as a penalty, it will learn to avoid the action (choosing the permanent food source) that leads to the removal of the ephemeral food source. In other words, the context of a cleaner’s ecology and of the biological market experiment is such that the penalty acts as a correlate of knowing the consequence in the following time step. Thus, by using client behavior as a source of primary reward, our virtual agents can bypass the problem of accounting for the future through future positive rewards.

In some sense, the penalty may be viewed as a useful heuristic that aids the estimation process that individuals perform through associative learning. Note, however, that such a primary reward is expected to evolve only when there is a benefit to considering future consequences. For instance, one should not expect it to evolve in an ecology where client-to-cleaner relative density is so high that there are always clients in a cleaning station.

**On Intra- and Interspecific Variation in Learning the Biological Market Task**

When searching for explanations for the observed performances in the biological market task, it is important to keep in mind that we might look at different reasons for variation. Cleaners are open-water spawners with pelagic egg and larval stages, making ontogenetic explanations for intraspecific variation highly likely. For interspecific comparisons, evolved differences in cognitive machineries may be more important.
According to current evidence, a cleaner’s performance in the biological market task can be predicted by the intraspecific variation observed in cleaners that some individuals have learned to solve the general problem under natural conditions while other individuals have not. The former are able to apply the learned rule to the market experiment and hence perform well. In contrast, those that have not would need to learn the solution from scratch and hence fail within the limited number of trials run in the experiment. The sophisticated cleaners may need to generalize from real client fish to heretofore unknown Plexiglas plates. Generalized rule learning has been documented for cleaners under ecologically relevant conditions (Wismer et al. 2016), as well as for other fish species (Ferrari et al. 2007), making its use by cleaners in the market context a realistic possibility. Alternatively, the ability to chunk, as well as the development of an aversion to visitor clients leaving, is tightly linked to the frequency of visitor clients leaving. Our model assumes that FAAs have the ability to chunk, but we did not model the learning process leading to chunking. Low frequencies of visitors leaving because of low client availability and/or low probability of visitors leaving may prevent the formation of chunks in nature and hence prevent solving the biological market task. Similarly, low frequencies of visitors leaving may prevent cleaners from developing a strong aversion to the stimulus and hence provide another route to failure in the task.

For interspecific comparisons, some ontogenetic explanations developed above may also hold. However, we also have to consider the potential role of ecological relevance. While all species tested so far have, beyond doubt, the capacity to learn through negative reinforcement, the challenge might be to identify the relevant stimulus that acts as a negative reinforcer. This possible explanation is supported by follow-up experiments that showed that capuchins can readily learn to solve the biological market task if the information about permanent versus ephemeral is coded in the color of the (otherwise identical) food items while the plates are the same color (Prêtot et al. 2016). These positive results provide strong evidence that the ability to chunk information is not per se a limiting factor, at least not in capuchins. Indeed, configurational representation and chunking are cognitive phenomena that have been demonstrated in nonhuman animals (Woodbury 1943; Sutherland and Rudy 1989; Terrace 1991; Conway and Christiansen 2001). Perhaps the key point is not whether an animal is able to form the full representation of options like FAAs do but in what context it will. Clearly, representing all possible state combinations in the environment as different chunks is not feasible, and even representing too many of them may not be adaptive because of memory limitations and computational challenges (Lotem et al. 2017). Moreover, chunking in general may be adaptive only under specific conditions (Kolodny et al. 2015). It seems logical, therefore, that the default representation is that of PAAs unless a condition of high ecological relevance warrants the ability to chunk.

In conclusion, it appears that the evolved machinery to filter information and to attach salience to the provided cues may explain most interspecific variation in the performance of this task. Although the biological market experiment presents an optimal foraging problem, we hypothesize that the interspecific social aspect (i.e., clients as active agents that react to cleaner decisions) is key to understanding why cleaners can readily solve the task. Indeed, we see various reasons why chunking could readily evolve in a social context. Most importantly, the value of specific group members may often be modified by the presence of third parties. For example, a young male monkey may value a female as a grooming partner if she is isolated but should avoid her in the presence of a dominant male. Under such conditions, calculating an average value for each possible action toward the female would yield a rather suboptimal behavioral rule for the young male (always grooming or never grooming). The cleaner’s ecology, as well as this example, highlights the general point that social animals must include the social context when making a valuation of their strategic decisions. Thus, the cognitive tools described here might be generally important for social animals within ecologically relevant contexts.

Acknowledgments

We thank Zegni Triki and Noa Truskanov for input on earlier versions of the models.

Literature Cited

Bshary, R. 2001. The cleaner fish market. Pages 146–172 in J. A. R. A. M. Van Hooff, P. Hammerstein, and R. Noë, eds. Economics in nature: social dilemmas, mate choice and biological markets. Cambridge University Press, Cambridge.

Bshary, R., and A. S. Grutter. 2002. Experimental evidence that partner choice is a driving force in the payoff distribution among cooperators or mutualists: the cleaner fish case. Ecology Letters 5:130–136.

———. 2005. Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. Biology Letters 1:396–399.
Bhary, R., and D. Schäffer. 2002. Choozy reef fish select cleaner fish that provide high-quality service. Animal Behaviour 63:557–564.

Chittka, L., S. J. Rossiter, P. Skorupski, and C. Fernando. 2012. What is comparable in comparative cognition? Philosophical Transactions of the Royal Society B 367:2677–2685.

Clutton-Brock, T. H., and G. A. Parker. 1995. Punishment in animal societies. Nature 373:209–216.

Conway, C. M., and M. H. Christiansen. 2001. Sequential learning in non-human primates. Trends in Cognitive Sciences 5:539–546.

Côté, I. M. 2000. Evolution and ecology of cleaning symbioses in the sea. Oceanography and Marine Biology Annual Review 38:311–355.

Dridi, S., and E. Akçay. 2017. Learning to cooperate: the evolution of social rewards in repeated interactions. American Naturalist 191:58–73.

Dridi, S., and L. Lehmann. 2014. On learning dynamics underlying the evolution of learning rules. Theoretical Population Biology 91:20–36.

Enquist, M., J. Lind, and S. Ghirlanda. 2016. The power of associative learning and the ontogeny of optimal behaviour. Royal Society Open Science 3:160734.

Enquist, M., J. Lind, S. Ghirlanda, and O. Kolodny. 2017. The evolution of cognitive mechanisms in response to cultural innovation and creativity. Current Zoology 61:350–368. Lotem, A., J. Y. Halpern, S. Edelman, and O. Kolodny. 2017. The evolution of cognitive mechanisms in response to cultural innovations. Proceedings of the National Academy of Sciences of the USA 114:7915–7922.

Mackintosh, N. J. 1975. A theory of attention: variations in the associability of stimuli with reinforcement. Psychological Review 82:276–298.

Maynard Smith, J. 1982. Evolution and the theory of games. Cambridge University Press, Cambridge.

McAuliffe, K., and A. Thornton. 2015. The psychology of cooperation in animals: an ecological approach. Journal of Zoology 295:23–35.

McNamara, J. M. 2013. Towards a richer evolutionary game theory. Journal of the Royal Society Interface 10:20130544.

Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63:81–97.

Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of Neuroscience 16:1936–1947.

Niv, Y. 2009. Reinforcement learning in the brain. Journal of Mathematical Psychology 53:139–154.

Noé, R., and P. Hammerstein. 1995. Biological markets. Trends in Ecology and Evolution 10:336–339.

O’Connell, L. A., and H. A. Hofmann. 2011. The vertebrate mesolimbic reward system and social behavior network: a comparative synthesis. Journal of Comparative Neurology 519:3599–3639.

Osvald, M. 2009. Spontaneous planning for future stone throwing by a male chimpanzee. Current Biology 19:R190–R191.

Osvald, M., and H. Osvald. 2008. Chimpanzee (Pan troglodytes) and orangutan (Pongo abelii) foreshadow: self-control and pre-experience in the face of future tool use. Animal Cognition 11:661–674.

Pepperberg, I. M., and L. A. Hartsfield. 2014. Can grey parrots (Psittacus erithacus) succeed on a “complex” foraging task failed by nonhuman primates (Pan troglodytes, Pongo abelii, Sapajus apella) but solved by wrasse fish (Labroides dimidiatus)? Journal of Comparative Psychology 128:298–306.

Préfôt, L., R. Bhary, and S. F. Bronsan. 2016. Comparing species decisions in a dichotomous choice task: adjusting task parameters improves performance in monkeys. Animal Cognition 19:819–834.

Puterman, M. L. 2014. Markov decision processes: discrete stochastic dynamic programming. Wiley Series in Probability and Statistics. Wiley, New York.

Quiñones, A. E. 2019. Cognitive mechanisms—cleaner version 1.0. Zenodo, https://doi.org/10.5281/zenodo.3381646.

Quiñones, A. E., O. Leimar, A. Lotem, and R. Bhary. 2019. Data from: Reinforcement learning theory reveals the cognitive requirements for solving the cleaner fish market task. American Naturalist, Dryad Digital Repository, https://doi.org/10.5061/dryad.pnvx0k6h3.

Raby, C. R., D. M. Alexis, A. Dickinson, and N. S. Clayton. 2007. Planning for the future by western scrub-jays. Nature 445:919–921.

Rahmani, N. J., A. Thornton, and R. Bhary. 2012. Punishment and cooperation in nature. Trends in Ecology and Evolution 27:288–295.

Rescorla, R. A., and A. R. A. Wagner. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. Pages 64–99 in A. H. Black and W. F. Prokasy, eds. Classical conditioning II: current research and theory. Appleton-Century-Crofts, New York.

Robson, A. J., and L. Samuelson. 2011. The evolutionary foundations of preferences. Pages 221–310 in J. Benhabib, A. Bisin,
Salwiczek, L. H., L. Prétôt, L. Demarta, D. Proctor, J. Essler, A. I. Pinto, S. Wismer, et al. 2012. Adult cleaner wrasse outperform capuchin monkeys, chimpanzees and orang-utans in a complex foraging task derived from cleaner–client reef fish cooperation. PLoS ONE 7:e49068.

Schultz, W. 2015. Neuronal reward and decision signals: from theories to data. Physiological Reviews 95:853–951.

Shettleworth, S. J. 2010. Cognition, evolution, and behavior. 2nd ed. Oxford University Press, Oxford.

Simon, H. A. 1974. How big is a chunk? by combining data from several experiments, a basic human memory unit can be identified and measured. Science 183:482–488.

Singh, S., R. L. Lewis, A. G. Barto, and J. Sorg. 2010. Intrinsically motivated reinforcement learning: an evolutionary perspective. IEEE Transactions on Autonomous Mental Development 2:70–82.

Staddon, J. E. R. 2016. Adaptive behavior and learning. 2nd ed. Cambridge University Press, Cambridge.

Suddendorf, T., and M. C. Corballis. 2007. The evolution of foresight: what is mental time travel, and is it unique to humans? Behavioral and Brain Sciences 30:299–313 (“Open Peer Commentary,” pp. 313–351).

Sutherland, R. J., and J. W. Rudy. 1989. Configural association theory: the role of the hippocampal formation in learning, memory, and amnesia. Psychobiology 17:129–144.

Sutton, R. S., and A. G. Barto. 2018. Reinforcement learning: an introduction. 2nd ed. MIT Press, Cambridge, MA.

Takahashi, Y., G. Schoenbaum, and Y. Niv. 2008. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Frontiers in Neuroscience 2:86–99.

Terrace, H. S. 1991. Chinking during serial learning by a pigeon. I. Basic evidence. Journal of Experimental Psychology: Animal Behavior Processes 17:81–93.

Thornton, A., and D. Lukas. 2012. Individual variation in cognitive performance: developmental and evolutionary perspectives. Philosophical Transactions of the Royal Society B 367:2773–2783.

Triki, Z., S. Wismer, E. Levorato, and R. Bshary. 2018. A decrease in the abundance and strategic sophistication of cleaner fish after environmental perturbations. Global Change Biology 24:481–489.

Wismer, S., A. Grutter, and R. Bshary. 2016. Generalized rule application in bluestreak cleaner wrasse (Labroides dimidiatus): using predator species as social tools to reduce punishment. Animal Cognition 19:769–778.

Wismer, S., A. I. Pinto, A. L. Vail, A. S. Grutter, and R. Bshary. 2014. Variation in cleaner wrasse cooperation and cognition: influence of the developmental environment? Ethology 120:519–531.

Woodbury, C. B. 1943. The learning of stimulus patterns by dogs. Journal of Comparative Psychology 35:29–40.

Wubs, M., R. Bshary, and L. Lehmann. 2018. A reinforcement learning model for grooming up the hierarchy in primates. Animal Behaviour 138:165–185.

Associate Editor: Maria E. Orive
Editor: Daniel I. Bolnick

Cleaner fish approaching a pair of clients that are looking for cleaning service at the same time. Photo credit: Simon Gingins.