Title: Computational models reveal how cleaner fish adjust decisions in a biological market

Abbreviated title: Cleaner fish cognitive mechanisms in a biological market

Abstract

While it is generally straightforward to quantify individual performance in cognitive experiments, identifying the underlying cognitive processes remains a major challenge. Often, different mechanistic underpinnings yield similar performances, and Lloyd Morgan's cannon warrants acceptance of the simpler explanation. Alternatively, when the different mechanisms interact with environmental conditions, variation in performance across environments might allow to statistically infer the mechanism responsible. We illustrate this point by fitting computational models to experimental data on performance by wild-caught cleaner fish *Labroides dimidiatus* in an ephemeral reward task,

as well as cleaner and client fish densities from the locations of capture. Using Bayesian statistics to fit the model parameters to performance data revealed that cleaner fish most likely estimate future consequences of an action, while it appears unlikely that the removal of the ephemeral reward acts as psychological punishment (negative reinforcement). Incorporating future consequences also yields performances that can be considered the result of locally optimal decision-rules, in contrast to the negative reinforcement mechanism. We argue that the combination of computational models with data is a powerful tool to infer the mechanistic underpinnings of cognitive performance.

Lay summary

Cleaner fish eat ectoparasites off other fishes, so-called clients. It regularly happens that two clients seek a cleaner's service simulatenously. Cleaners benefit from prioritising clients unwilling to wait, so they can feed on those willing to wait. To make the right choice, cleaners must somehow "look" into the future to anticipate consequences of current choices. By combining a learning model with data, we show that cleaners estimate the long-term value of their actions rather than using simpler heuristics. Estimating long-term value is a mechanism involved in human foresight.

Keywords— learning, behavior, cleaners, bayesian statisitics, behavioral mechanisms

Introduction

Often alternative cognitive mechanisms yield similar behavior and/or cognitive performances. This poses a problem for disentangling the mechanistic underpinnings of behavior. This is particularly clear in research aimed at discovering between species variation in *higher* cognitive abilities; or in other words, research on whether non-human animals show cognitive abilities believed to be uniquely human. For instance, when researchers try to find *mental time travel-like* behavior, they usually come up with experiments to show the behavior displayed requires inferences made through past events (Dally, Emery, and Clayton 2006). However, they often face the challenge of alternative scenarios where simpler explanations, like classic associative learning, can bring about the observed behavioral outcome (Suddendorf and Corballis 2007). Similarly, attempts to demonstrate the presence of *theory of mind* in non-human animals face objections justified by alternative mechanisms underpinning similar behavioral results (C. M. Heyes 1998). Such controversies are usually settled by using the principle of parsimony and its cognitive version, Lloyd Morgan's cannon, which states that the simpler explanation (mechanism) should be accepted. Ideally, alternative hypotheses should be evaluated in light of their explanatory power.

Learning is a key overarching cognitive mechanism that allows individuals to associate rewards with environmental stimuli and thus behave adaptively (Staddon 2016; Shettleworth 2009). Associative learning, in particular, exists

in all major vertebrate taxa, and in many invertebrates as well (C. Heyes 2012; Macphail 1982; Staddon 2016; Behrens et al. 2008). Associative learning is not homogeneous throughout its taxonomic distribution, rather there are differences across and within species (Shettleworth 2009; Sih and Giudice 2012). So presumably, the mechanistic underpinnings of learning have been modified by natural selection (Marler and Peters 1989).

One way to formalize the alternative mechanistic underpinnings of associative learning is to develop quantitative models of learning processes. This approach, which started within experimental psychology (Staddon 2016; Brandon, Vogel, and Wagner 2002), has been very fruitful in disentangling the mechanistic structure of cognitive systems. More recently, the development of reinforcement learning theory (Sutton and Barto 2018), has allowed to evaluate the empirical support for alternative mechanistic hypotheses by providing quantitative predictions which are amenable to statistical tests (Farashahi et al. 2020, 2017). Interestingly, these learning models not only have received support from behavioral data, but also are consistent with the current view on reward processing in the brain (Schultz 2015).

From an evolutionary perspective, mechanisms are likely selected because of how they allow individuals to respond to environmental variation. For example, biological market theory predicts that the exchange rate of goods and/or services traded between cooperative partners adjusts to the law of supply and demand, when individuals have some degree of partner choice (Noë

and Hammerstein 1995). Supply and demand conditions, which typically depend on the abundance of the species involved, certainly vary in time and space. Therefore, natural selection should favor the ability to flexibly adjust decisions and behavioral output to current market conditions. Indeed, such adjustments have been documented (Axén, Leimar, and Hoffman 1996). One example of strategic adjustment in a biological market is the marine cleaning mutualism involving the cleaner fish *Labroides dimidiatus* and 'client' fishes. Client fishes seek cleaner fish services at their territory (so-called "cleaning station") and offer themselves as food patches to get their ectoparasites removed, which provides cleaners with food and clients with improved health (Waldie et al. 2011; Ros et al. 2011; Triki et al. 2016; Demairé et al. 2020). Given the capacity of some client fish to swim larger distances and access multiple cleaning stations while others access the only cleaning station in their territory, it is crucial to categorize clients as either "visitors" or "residents", respectively. During cleaning interactions, a cleaner fish often faces a choice between a visitor and a resident client seeking its cleaning services simultaneously. Visitors have the option to switch to another cleaner fish if being made to wait, while residents must wait for inspection. Indeed, visitors have been observed to use their partner choice option in that way (Bshary and Schäffer 2002), which may explain why cleaners give visitors service priority in a field study in the Red Sea (Bshary 2001). Furthermore, in a lab based paradigm, design to mimic the resident-visitor choice (ephemeral reward task), cleaners learned to prefer the cue associated with the epheral food source (visitor)

and hence accessed both food sources, obtaining double the amount of food (Bshary and Grutter 2002). However, further exploration revealed that not all cleaner fish manage to develop a preference for the ephemeral option in the lab (Triki et al. 2018). Over the last decade, over a hundred wild-caught cleaner fish have been tested in the exact same paradigm of the ephemeral reward task (Salwiczek et al. 2012; Wismer et al. 2014; Triki et al. 2018, 2019, 2020). These fish often come from different reef locations. Investigation of the local eco-sociological conditions revealed that cleaner and client fish population densities have a substantial impact on cleaner fish performance in the task. Cleaner fish from reef sites with relatively low densities were more likely to fail at solving the task (Triki et al. 2019, 2018; Wismer et al. 2014). This intra-specific variation is unlikely due to local genetic adaptation, because cleaner fish are open water spawners and the environmental conditions can vary within the lifespan of a fish.

Mechanistic models explicitly designed to mimic the ephemeral reward task have shown that the simplest form of associative learning (operant conditioning) cannot account for a solution to the ephemeral reward task (Prat, Bshary, and Lotem 2022; Quiñones et al. 2019). Operant conditioning is a form of associative learning where individuals use short term reward to associate and choose actions. Such models allow varying the cognitive tool kit and evaluating which minimal kit is necessary to solve the task at hand (e.g. Dubois et al. (2021)). To be able to give visitors priority over residents, cleaners need to be able to assess a client's value separately for the

three possible scenarios (alone, paired with a fish with the same strategic option, paired with a fish with the alternative strategic option) (Quiñones et al. 2019). The ability to distinguish and value one stimulus differently alone from compound versions of it has been termed configurational learning, chunking, or segmentation (see references in Prat, Bshary, and Lotem (2022)). In addition to configurational learning, cleaners also need to account for the future consequences of current decisions. In the model by Quiñones et al. (2019), this could be achieved in two non-mutually exclusive ways: through low temporal discounting of future effects, also termed 'chaining' (Enquist, Lind, and Ghirlanda 2016); and/or through perceiving a visitor client leaving as psychological punishment (i.e. as a negative reinforcer). Chaining is when individuals include in their valuation of an action the reward effects that this will have in the future. This is done by combining in a single valuation the reward obtained in the current time with all the reward that comes after, discounting for how far in the future reward is accrued. 'Chaining' the reward of these different time steps allows individuals to take actions that increase the long-term reward at the sacrifice of short term considerations (Enquist, Lind, and Ghirlanda 2016). Even though, 'chaining' can be readily implemented computationally in learning models (Enquist, Lind, and Ghirlanda 2016; Sutton and Barto 2018), cognitively it seems to be a complex adaptation (Suddendorf and Corballis 2007). On the other hand, using client behavior as a negative reinforcer is, in principle, easier to implement. Thus, the standard logic of Lloyd Morgan's cannon demands that operant

conditioning as the simpler explanation is to be accepted by default. Ideally, however, the two mechanisms should be evaluated in light of how well they explain the available data. Note that different fields interested in cognition and decision making use different words to refer to negative reinforcers (Quiñones et al. 2019; Sutton and Barto 2018). Here, for the sake of simplicity and clarity, we will use the word 'penalty' to refer to this mechanism which includes a negative reinforcer.

In here we used the field and experimental data to fit the parameters of a reinforcement learning model to infer the cognitive mechanism that cleaners use in their interaction with clients. Specifically, our approach of fitting the computational model to the empirical data aimed at: (i), determining which mechanism cleaner fish use to incorporate future consequences of current decisions by testing whether chaining, penalty, or a combination of both best explains their performance; (ii) determining whether the two mechanisms differed with respect to the ecological conditions that are likely to cause high versus low performance in the ephemeral reward task. Additionally, we assessed which mechanism yields optimal performance patterns. Relying on the logic of biological market theory, we predicted that appropriate performance is to show a high preference for visitors only under high local cleaner-to-client ratio.

Methods

The model

The model consists of a set of individual-based simulations where individuals, representing cleaner fish, face a series of choices between two options, which simulate the natural conditions of the cleaning market. Individuals experience a series of discrete time points in which they face different 'states', defined by the number and category of client fish (visitor or residents) inviting for cleaning services. There are six possible states: zero clients, one resident, one visitor, resident-resident, visitor-visitor, and resident-visitor. The probability of each state is largely determined by the relative abundance of cleaner fish, residents and visitors, but to some degree by cleaner fish choices when it faces the resident-visitor combination. This is because residents are willing to queue for cleaning service; while visitors leave the queue (with a certain probability) when made to wait. Individuals obtain a fixed reward from cleaning a client fish regardless of the category. Every time individuals face and make a choice they update the probability of making that same choice. The update is based on the difference between the expected value and the obtained reward - the prediction error (δ_t) - (Sutton and Barto 2018; Rescorla and Wagner 1972). Formally, the prediction error is given by

$$\delta_t = R_t - V_t(S_t) + \gamma V_t(S_{t+1}), \tag{1}$$

where R_t is the sum different reward sources at time t; $V_t(S_t)$ is the estimated value at time t of the the state faced by the agent at time t; similarly $V_t(S_{t+1})$ is the estimated value of the state to come in the following time-step, γ is the discount factor for future rewards. When the estimated value of the current state ($V_t(S_t)$) is equal to the sum of short-term (R_t) and future discounted reward ($\gamma V_t(S_{t+1})$) learning stops for that state. If $\gamma = 0$ the estimates made by the agent only capture short-term reward. We assume short-term reward to two components: positive reward determined by the amount of food obtained from cleaning a client; and negative reward triggered when by a client leaving the station without being cleaned. Formally, we let total reward be given by $R_t = P_t - \eta_t$. Where η is a parameter of the model that determines the the size of the negative reward triggered by unattended clients leaving the station.

The prediction error (Eq. **??**) is used to update the value of each one of the states the agent faces, as well as the preference for the resident and visitor options. The value update is simple the product of the prediction error and the parameter for the speed of learning ($\Delta V(S_t) = \alpha\delta_t$). The change in preference between the resident and visitor is given by $\Delta(\theta_v - \theta_r)_t = \alpha\delta_t 2(1-\pi_v)$, where θ_i represent the preference for one of two options and the difference captures the total change relative to one another; π_v corresponds to the current probability of choosing the visitor. p_i is determined by applying the logistic function to the difference in preferences between the two mutually exclusive options ($\pi_v = \frac{1}{1+e^{-(\theta_v-\theta_r)}}$). This amounts to a preference update

that is carried in the direction that leads to more reward being obtained, given the new information. In the long run, the probability of choosing a visitor over a resident converges in the model. To which probability the model will converge depends on the relative abundance of cleaners, visitors and residents; as well as on the probability of visitors leaving the cleaning station when made to wait. Further details of the model implementation can be found in Quiñones et al. (2019).

The model shows that agents need to find a way to incorporate future consequences of current choices. In the model, this could be achieved with either of two parameters that could also work together. First, γ measures how much individuals include future rewards in their decision updates. If $\gamma = 0$, individuals only use the immediate reward obtained from a cleaning interaction. As γ increases, individuals include more of the reward obtained from the subsequent choices. That amounts to estimating and using for decision making the future expected rewards of an action (chaining). Second, η measures how much individuals include in their reward the fleeing behavior of visitors as a negative component (penalty). Both of these parameters allow individuals to use in their estimates the future effects of their choices.

Empirical data

The empirical data were collected between 2010 and 2019 always during the austral winter months June to August from a total of five study reef sites

(Corner Beach-CB, Horseshoe-HS, Mermaid Cove-MC, Northern Horseshoe-NHS, and The Crest-TC) at Lizard Island ($14.6682^{\circ}S, 145.4604^{\circ}E$), Great Barrier Reef, Australia. The data consist of three sets: fish censuses, field observations of cleaner-client interactions to quantify the probability of visitors leaving if made to wait, and the performance of wild-caught cleaner fish in the ephemeral reward test. In total, we have twelve site/year data sets for fish censuses and corresponding performance in the lab test. Thus, some sites were sampled more than once. To estimate the population density of cleaner fish and their clients at a given site in a given year, Triki et al. (2019) used a series of ten transects of $30m$ each. Observers swam along the transect lines placed on the reef and first counted the visible large-bodied adult fish (species with total length $TL \geq 10cm$) including cleaner fish on a width of $5m$, and then on the return individuals of small-bodied fish species ($TL < 10cm$) on a width of $1m$ (see Triki et al. (2018) for further details on fish censuses data collection). Total length estimates were done by the observer. We then scaled the counts of cleaner fish, small-bodied, and large-bodied clients fish densities per $100m^2$.

The field observation data consisted of video recordings/encodings of the cleaner-client cleaning interactions. There were videos from eight cleaners per site/year of a duration of $30min$ each. Triki et al. extracted information from every event wherein a visitor client was made to wait in favor of another client (visitor or resident), and noted whether or not the visitor left or queued for the cleaning service (Triki et al. 2019, 2020).

The cognitive performance data was from a total of 120 cleaners (10 individuals per 12 site/year) tested in the ephemeral reward task (Triki et al. 2019, 2020). Authors housed all captured cleaners individually in glass aquaria ($62cm \times 27cm \times 37cm$) and provided them with PVC pipes ($10cm \times 1cm$) as shelters. The task consisted of exposing the cleaner fish to substitute models of client fish in the form of two *Plexiglas* plates offering the same amount of food (one item of mashed prawn). The two plates differed in colour and pattern (horizontal green stripes or vertical pink stripes) but had equal size ($10cm \times 7cm$). Importantly, the two plates played different roles as either a visitor (ephemeral food source) or resident (permanent food source). That is, if a cleaner fish inspected the resident plate first, the experimenter withdrew the visitor plate out of the aquarium as a consequence. Choosing first the visitor plate, however, granted access to both plates. The equal size of the plates forced cleaner fish to decide based solely on the association between the behaviour and the collor/pattern cue (Wismer et al. 2019). Triki et al. (2019, 2020) tested the fish for a maximum of 200 trials with 20 trials a day, 10 trials in the morning and 10 trials in the afternoon. They randomized and counterbalanced the plates' spatial location (i.e. left or right) between trials. Similarly, they counterbalanced the plates' decoration (colour and pattern) and the plates' role (visitor or resident) between the tested fish. In the original studies, once a fish reached a learning criterion, that is, performing significantly above chance level in a binomial test ($p - value \leq 0.05$,), they passed to a reversal version of the task where the roles of the visitor/resident

Plexiglas plates were swapped. The reversal phased stopped when the fish performed significatly above chance, or the fish completed the 200 trial together with the initial; see Triki et al.(2019, 2020). Here, we are interested in explaining the total frequency of visitor choices using the model, rather than just the achievement of the criterion. Total frequency of visitor choices naturally comes out of the model, and allows us to use all the variation among cleaners, instead of reducing that to a binomial variable. Thus, we used instead a subset of these data to estimate the final cleaner fish preferences for the visitor plate, even if they do not reach the learning criterion within 200 trials. To do so, we first extracted the trial-by-trial outcomes from the last two sessions (20 trials) of those who never reached the learning criterion for visitor plate ($N = 45$ cleaner fish). For those who reached the learning criterion at some point during the test and passed to a reversal phase, we extracted the trial-by-trial outcomes from the last session (10 trials) before passing to reversal and the last 10 trials they were exposed to in the test ($N = 75$ cleaner fish).

We chose a combination of initial and reversal to quantify preference for the visitor client. However, it could be argued that using only the initial phase gives a better estimation of the cleaner fish preference for the visitor. In the supplementary material (Fig. **??**) we show how using initial and reversal (a), and only initial (b) maps to the previously used criteria. The initial and reversal match better the criteria chosen in previous analysis of the ephemeral reward task experimental set-up (Triki et al. 2019, 2020).

Statistical analysis

The aim of the analyses is to fit the key model parameters γ and η, to the empirical data from Triki et al. (2019, 2020) to test whether each or a combination of these effects is a better explanation for the pattern seen in the data. We used the ecological variables: cleaners, visitor clients, resident clients abundances and visitor clients leaving probability,
as input to the models. As the response variable, we used the frequency with which cleaners chose the visitor option in the ephemeral reward task. Finally, we used the probability with which agents in the model simulations choose the visitor in resident-visitor options as the prediction for the response variable. We kept all other parameter values used for the model simulations constant, see Table **??**.

To capture with the model the relationship between the ecological variables and cleaner fish preferences for visitors, we needed to scale the absolute population densities of cleaner fish from the empirical data to a measure of relative abundances that captures client visitation patterns. This is because, in the model, relative abundances of clients define not only the probability of residents and visitors but also how often the cleaning station is empty (e.g. there are no clients to be cleaned). The frequency with which clients visit the station is another variable influencing the station occupancy, which in nature may vary among different client species depending on their ectoparasite loads. We do not have field estimates for species-specific parasite loads, especially