Evolution of behavioural flexibility and the forming and breaking of habits

Olof Leimar[1, *], Sasha R. X. Dall[2], Peter Hammerstein[3], Alasdair I. Houston[4], Bram Kuijper[5], and John M. McNamara[6]

[1]Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden; olof.leimar@zoologi.su.se

[2]Centre for Ecology and Conservation, University of Exeter, Penryn TR10 9FE, UK; S.R.X.Dall@exeter.ac.uk

[3]Institute for Theoretical Biology, Humboldt University Berlin, 10115 Berlin, Germany; p.hammerstein@biologie.hu-berlin.de

[4]School of Biological Sciences, University of Bristol, Bristol, BS8 1TQ, UK; A.I.Houston@bristol.ac.uk

[5]Centre for Ecology and Conservation, University of Exeter, Penryn TR10 9FE, UK; A.L.W.Kuijper@exeter.ac.uk

[6]School of Mathematics, University of Bristol, Bristol, BS8 1UG, UK; john.mcnamara@bristol.ac.uk

[*]Corresponding author; e-mail: olof.leimar@zoologi.su.se

Abstract

The formation of habits, whereby learnt actions come to be performed automatically with repetition and practice, is a well-established focus for studies in psychology. This contrasts with evolutionarily motivated studies of learning, which typically view behaviour as either learnt or fixed to elucidate the ecological conditions where each predominates. Here, we envisage habit formation functioning to free up limited mental resources (e.g., attention), potentially improving an individual's ability to multitask. As an ecologically relevant case, we investigate exploration-exploitation in foraging under predation risk. In our model, a forager does not know the quality of feeding options, but can learn from the rewards they give. When the environment occasionally changes, individuals can attend to exploration of feeding options in the new conditions. The options can then become habitually exploited, freeing attention for antipredatory vigilance. Via evolutionary simulations, we show that evolutionarily stable forming and breaking of foraging habits can substantially reduce mortality from predation, without drastically reducing foraging success, when environmental conditions remain stable enough between changes. We therefore argue that exploration-exploitation strategies involving the forming and breaking of habits is a type of behavioural flexibility that is likely to be selected for under a range of ecological conditions.

Keywords: Exploration-exploitation trade-off, behavioural flexibility, multi-tasking, reinforcement learning, multi-armed bandits.

Introduction

Habits, as studied in experimental psychology, are modes of behaviour where individuals come to perform actions as more-or-less automatic responses, triggered by stimuli or circumstances they repeatedly encounter. The formation of habits is typically guided by goal-directed learning processes [1, 2, 3, 4, 5], making it a common feature of human task learning. Indeed, habit learning mechanisms seem to be conserved across the mammalian species that have been subject to laboratory reinforcement learning research [6] and are suggested to be shared by all vertebrates [7, 8]. The continuum between learnt, habitual and fixed behaviour suggests a more nuanced view of learnt behaviour than is typical for evolutionarily motivated studies of learning. Despite early ideas about the inheritance of habits for the evolution of species-specific behaviours [9], behavioural ecologists typically consider the evolutionary stability of learning strategies against innate, fixed behavioural responses [10, 11]. Here we suggest that the forming and breaking of habits represent important types of behavioural flexibility and deserve to be taken into account in evolutionary investigations of animal behaviour.

An important property of habits is that, in terms of mental resources, they are efficient ways of performing tasks, by letting actions be guided reflexively by stimuli and contexts [5]. In contrast to non-habitual behaviour, which involves continuous learning and updating, habitual task performance enables individuals to free cognitive resources like attention to devote to other concerns [3, 5]. It is generally recognised that limited attention can constrain multitasking [12, 13]. Here we use evolutionary simulations to show that habit-enhanced multitasking can be selectively favoured, by letting animals efficiently combine foraging and vigilance against predators. Our work is a further development of the study of foraging-vigilance trade-offs, which has long been a central topic in behavioural ecology [14, 15, 16].

To analyse the forming and breaking of habits, we use the idea of a trade-off between exploration and exploitation. This trade-off figures prominently in the fields of psychology, behavioural ecology, and machine learning [17, 18, 19, 20, 21, 22], but there is no consensus about the cognitive mechanisms that control such behaviour. Our approach is to assume that transitions between exploration and exploitation correspond to the forming of new habits and the breaking of old habits [1, 2, 5, 23, 6].

We examine situations where the reward distributions of different feeding options are unknown but remain constant over a period of time, after which they can change. In nature, a change could for instance be due to a new season or moving to a new location, and there could be several such environmental changes over an individual's lifetime. We assume that information about an option, and whether it has changed,

is only provided by the rewards it delivers. This is just one possibility, and in reality individuals might also use other cues, such as just having arrived to a new location, as indicators of a new environment. If each successive environment is constant for long enough, it would be advantageous to first learn about the feeding options and then to become habitual in foraging, thus balancing the attentional demands of foraging and vigilance. Such situations could favour the evolution of psychological mechanisms for the forming and breaking of habits, and it is our aim to establish the conditions under which these mechanisms are advantageous.

In psychological studies of learning, changing environments are often discussed in terms of volatility of rewards [24], with reversal of rewards perhaps being the most studied [25, 26, 2], together with the case of continual shifting of the reward values or the probabilities of reward [17, 27]. While these forms of volatility correspond to aspects of environments encountered in nature, they might be less important as selection pressures on cognitive mechanisms for the forming and breaking of habits than wholesale environmental changes, which we study here. There are also psychological studies of this latter kind [28].

The elements of our models include psychological mechanisms of exploration and learning, as well as rules for the forming and breaking of habits. For exploration, we use traditional action-value learning, but modified such that the initial estimated values of options are high [20]. We refer to the approach as optimistic exploration (OE). This approach is well established in theoretical work on reinforcement learning [20, 29, 30, 31, 32, 33, 34] and has the advantage of a close correspondence to traditional assumptions about learning in experimental psychology. For comparison we also show results with so-called Thompson sampling [35, 36] as the exploration strategy. Thompson sampling (TS) is a Bayesian statistical approach with near optimal or optimal exploration-exploitation performance in certain situations [36], but it is not known if it is approximated by cognitive mechanisms in real animals. Both the OE and TS procedures originated in the study of so-called multi-armed bandits [20, 36].

To implement habit formation, we use the well-established idea in cognitive modelling [2, 37] that habits are characterised by consistency of choice, i.e. the tendency to choose a given action in a given situation. In our simulations, we expressed the consistency of choice as a running estimate of the probability to repeat the most recent action. We assume that as an individual's estimate of its consistency of choice passes a threshold, the individual switches to habitual foraging, at the same time increasing its vigilance against predators. For the reverse transition from habitual foraging to exploration, we take inspiration from work in neuroscience on detecting and reacting to unexpected changes in the world [18, 38, 39]. An habitual forager can use suitable

cues from the environment to break its habit and start exploring. In our models, we assume that a lowering of the rate of rewards below a threshold causes a transition to exploration.

In the following, we illustrate the kinds of environmental volatility we investigate and show results from individual-based evolutionary simulations of the performance of exploration-exploitation strategies. We find that the forming and breaking of foraging habits has the potential to significantly reduce mortality from predation, without drastically reducing foraging success. We therefore argue that there is likely to be substantial natural selection for this form of behavioural flexibility.

Results

Figure 1 shows a situation that could promote the forming and breaking of habits. A forager is faced with a number of feeding options (in our simulations we use 8 feeding options). In each of a number of foraging trials, there is a choice of feeding option to use in that trial. Each feeding option delivers positive, log-normally distributed rewards, with a certain mean and (log-scale) random variation around this mean (Fig. 1a). In our analyses we assess performance over different numbers of successive environments (including a single environment). Figure 1b shows a case of four environments where, in each environment, most options deliver low rewards (for simplicity assumed equal), and a few options yield higher rewards. Ideally, in such situations a forager should explore to find the highest yielding option in an environment. The intuition behind optimistic exploration (OE) is sketched in Fig. 1c. By starting with high initial estimates of reward rates, which then are reduced through the sampling of options, a forager is incentivised to sample all options and, eventually, to focus on the highest yielding one. If a sufficiently thorough exploration can be achieved early in the period of an environment, and foraging then becomes habitual, the cost of increased predation risk from reduced vigilance during exploration and learning can be fairly small (Fig. 1d; survival for full exploration and full vigilance in the panel apply to all examples).

It can be helpful to first examine exploration and the evolution of habit formation for the case of a single (unknown) environment, which is shown in Fig. 2. The environment is randomly generated, in the sense that the mean reward for each of 8 options is independently drawn from a log-normal distribution. To aid understanding, the figure shows both a case of full-attention exploration (OE; Fig. 2a,c), and a case where individuals become habitual in foraging when their consistency of choice exceeds a threshold (Fig. 2b,d).

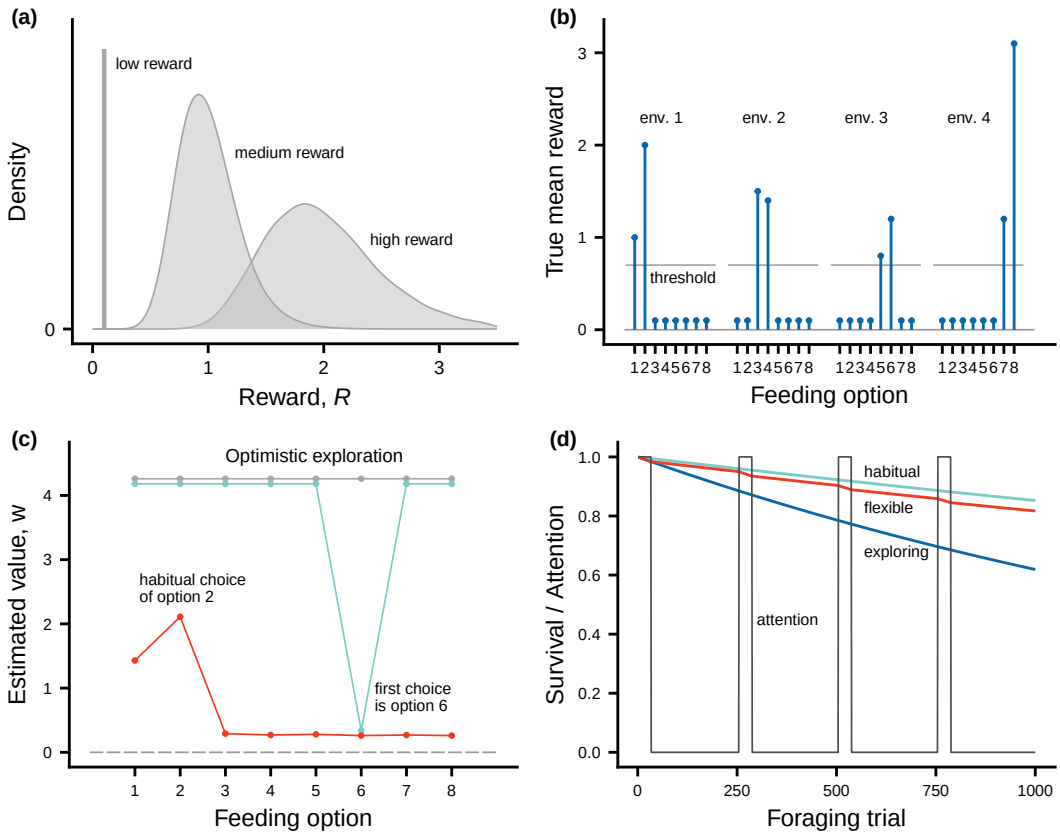

Figure 1: Sketch of model elements. Individuals make and break foraging habits, thereby trading off vigilance and exploration of food sources. As an example, there are four successive environments, each with 250 foraging trials, resulting in a maximal foraging lifetime of 1000 trials. In each successive environment, there are eight feeding options, or stimulus types. The options vary in reward distributions between environments. By devoting attention to the exploration of feeding options individuals can increase their feeding rates, at a cost of decreased vigilance. **(a)** Distribution of rewards for different options (log-normal with different means and $\sigma_R = 0.25$), corresponding to the situation in environment 1. **(b)** Mean reward values for the feeding options in each successive environment. The thresholds (grey horizontal lines; $\hat{R}$) illustrate a strategy of giving full attention to exploration if the estimated rate of reward drops below the threshold. **(c)** Illustration of a learning strategy of optimistic exploration. The curves show an individual's estimates of the the value of options in environment 1 at different points in time. Estimates start high (grey curve). The first choice is option 6, after which estimates are given by the greenish curve. The estimates after all options are explored are given by the red curve (curves are shifted slightly to avoid overlap). The individual's behaviour may then become habitual. **(d)** Survival for full habitual foraging (full vigilance; greenish curve; 0.85 at trial 1000), full attention to exploration (blue curve; 0.62 at trial 1000), and an intermediate case of initially exploring each environment and then becoming habitual (red curve). The grey curve shows a corresponding switching on and off of attention for a flexible strategy.

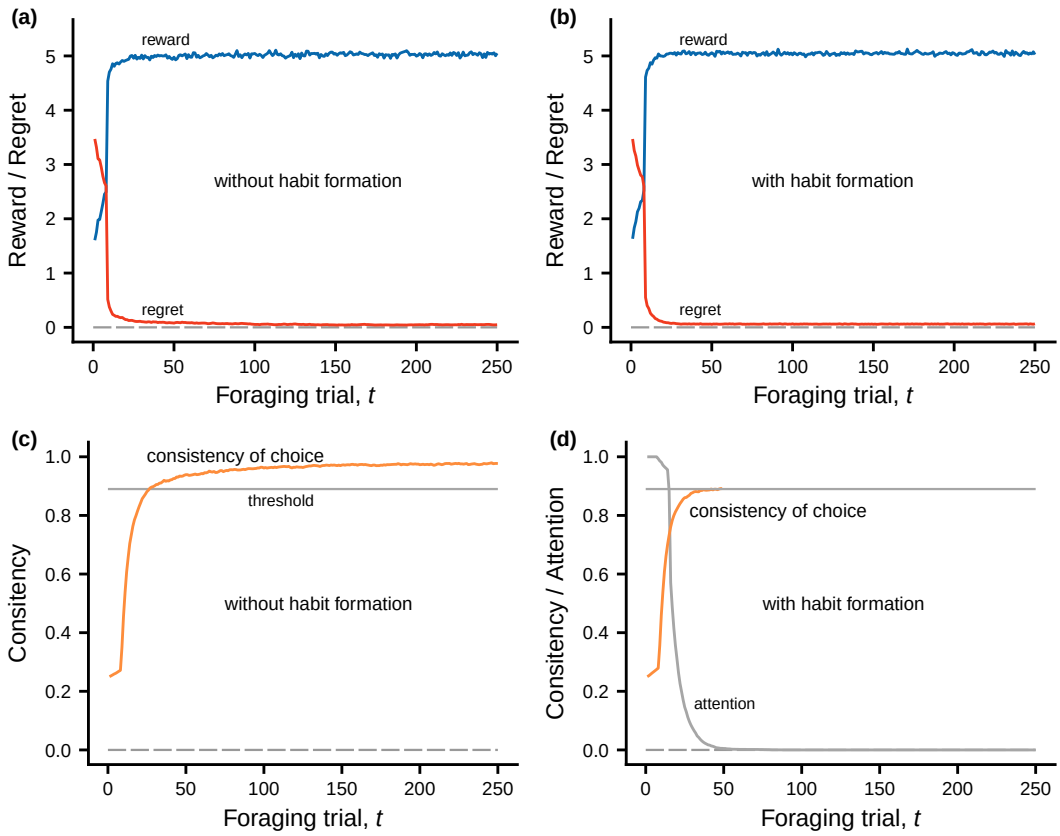

Figure 2: Illustration of habit formation when individuals face a single (unknown) environment during their lifetime. The panels show averages over the first 250 foraging trials (of a total of 1000 trials) in an environment with 8 feeding options, with groups of 10 individuals replicated 400 times (making up a total of 4000 individuals; see Fig. S1 in SI for examples of individual explorative choices). In each replicate the mean rewards of the options are independently drawn from a log-normal distribution (with parameters $\mu_0 = 0$, $\sigma_0 = 1$). Panels **(a)** and **(c)** show OE without habit formation, and **(b)** and **(d)** show corresponding results when individuals become habitual (turning off attention to exploration), when consistency of choice goes above an evolved threshold $\hat{h} = 0.89$ (after which the update of consistency stops). The averages over all individuals of the reward per trial (blue curves) and of the so-called regret (red curves) are shown in **(a)** and **(b)**. An individual's regret in a trial measures how much the individual lost by not choosing the option with the maximum expected reward (zero regret indicates an optimal choice). The average over all individuals of the consistency of the chosen option (orange curves) is shown in **(c)** and **(d)**. The consistency of an option measures the degree to which the current choice is the same as the most recent choice. Panel **(d)** also shows the average over individuals of the attention to exploration. The OE parameters are $w_0 = 8.5$, $\alpha = 0.95$, $\beta = 14.0$ (see also Fig. S2 in SI for corresponding TS results). The average survival over 1000 trials for the habit forming individuals is around 0.85, close to the maximal value for full vigilance, and better than the expected survival of 0.62 for full attention to exploration.

This habit formation threshold and the traits of the OE strategy are at evolutionary equilibrium for the situation in Fig. 2b,d (as determined by individual-based simulations). The example illustrates that, in a constant environment, habit formation can give approximately the same level of rewards as for maintained exploration, while having an advantage in terms of vigilance. The expected loss of rewards from not choosing the best option is referred to as 'regret' in the study of multi-armed bandits [36, 40]. The red curves in the figure show that this regret becomes nearly as small with habit formation as without (Fig. 2a,b).

Next, to investigate the evolution of both forming and breaking of habits, we combine (*i*) an exploration strategy (OE or, for comparison, TS), (*ii*) a habit formation strategy, implemented as a consistency-of-choice threshold trait $\hat{h}$, for turning off attention to foraging (as illustrated in Fig. 2), and (*iii*) a habit-breaking strategy with traits that allow the exploration of new environments. We have implemented habit breaking as a threshold trait $\hat{R}$, such that an individual may turn on attention to exploration if an estimated rate of reward R_{e} goes below the threshold. The individual updates the estimate with an evolving trait λ. Further, to control the maximum frequency of switching on and off of attention, we introduced an evolving trait $\hat{t}$, such that there needs to be at least $\hat{t}$ trials of habitual foraging before the habit-breaking threshold is applied.

The sequence of environments shown in Fig. 1a,b have clear differences between the environments and the options within them, which should make the forming and breaking of habits particularly advantageous. This contention is supported by the results in Fig. 3. For the sequence of the two first environments in Fig. 1b and for all four, the rewards collected are near optimal (the values of regret become close to zero) and the survival over 1000 foraging trials is near the maximal value (for full vigilance).

Changing the situation such that the mean rewards from the foraging options in successive environments are randomly drawn, as in Fig. 2, increases the challenge of foraging-vigilance multitasking somewhat. This is shown by the results in Fig. 4, with greater mean regrets (Fig. 4a,b) and more shifting on and off of attention (Fig. 4c,d; see also Figs. S5, S6 in the SI for examples of individual foragers). As a consequence of the greater challenge, the survival at trial 1000 is around 0.83 for two and 0.82 for four successive environments, slightly lower than in Fig. 3. Further increasing the challenge by shortening the duration of each successive environment results in increased regrets and reduced survival (see Fig. SX in SI).

Comparing our results with OE as the exploration strategy (Figs. 2, 3, 4) with the corresponding results for TS (Figs. S2, S3, S4 in SI), we find the same qualitative tendencies in terms of rewards, regrets, and survival. Overall, this supports a general

conclusion about selection for the forming and breaking of habits: a clear separation among the reward distributions of available feeding options, allowing rapid habit formation, together with readily available habit-breaking indicators of new environments, and with environments staying relatively constant for an appreciable period, act to promote the evolution of habit-mediated exploration-exploitation strategies.

Finally, an example of many successive fairly short-duration environments (Fig. S8 in SI) illustrates that if individuals have a way to identify environment changes, the forming and breaking of habits becomes more efficient (in the example, environments last 50 trials, and individuals have evolved a value of $\hat{t}$ to correspond to this).

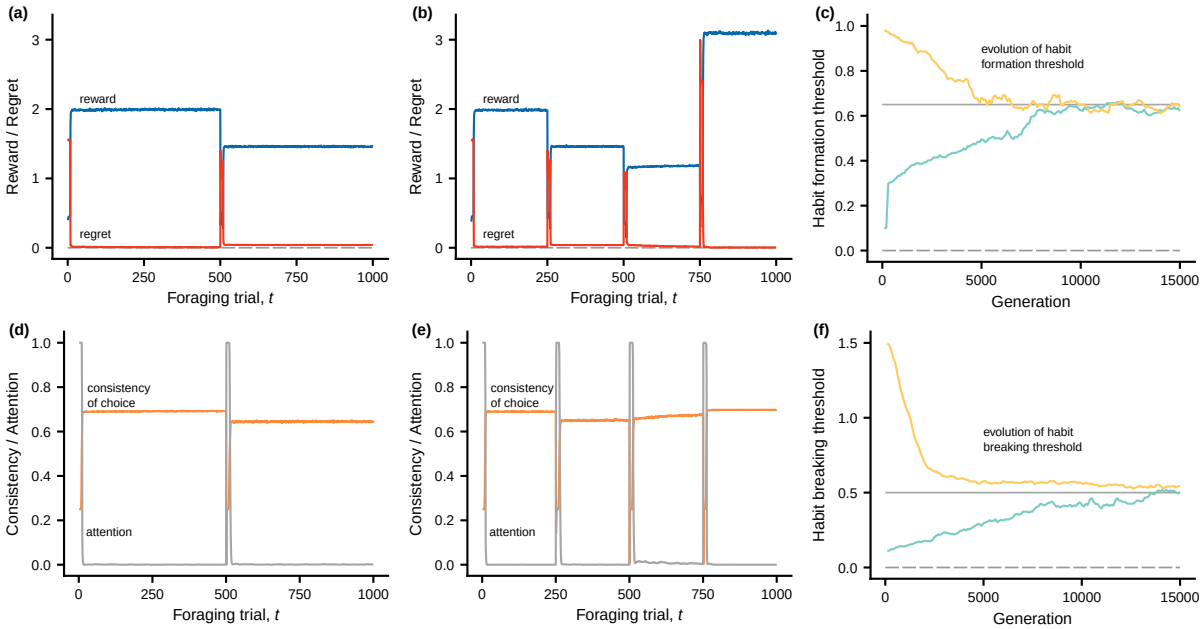

Figure 3: The forming and breaking of habits for the environments in Fig. 1b. Panels **(a)** and **(d)** show the dynamics of reward/regret and attention to exploration for OE, given as means over all simulated individuals, for the sequence of the first two environments in Fig. 1b, with each environment lasting 500 trials. Panels **(b)** and **(e)** show the corresponding results for the sequence of all four environments in Fig. 1b, each lasting 250 trials. For this latter case, panel **(c)** shows two simulated evolutionary trajectories for the habit-forming trait $\hat{h}$, and panel **(f)** shows two trajectories for the habit-breaking trait $\hat{R}$. At evolutionary equilibrium, the trait values are $w_0 = 4.4$, $\alpha = 0.98$, $\beta = 11.5$, $\hat{h} = 0.64$, $\hat{R} = 0.62$, $\lambda = 0.88$, $\hat{t} = 17.5$ for two successive environments, and $w_0 = 4.7$, $\alpha = 0.98$, $\beta = 12.0$, $\hat{h} = 0.65$, $\hat{R} = 0.50$, $\lambda = 0.82$, $\hat{t} = 13.8$ for four successive environments (see also Fig. S3 in SI for corresponding TS results). The survival over 1000 trials is around 0.84 for both these cases.

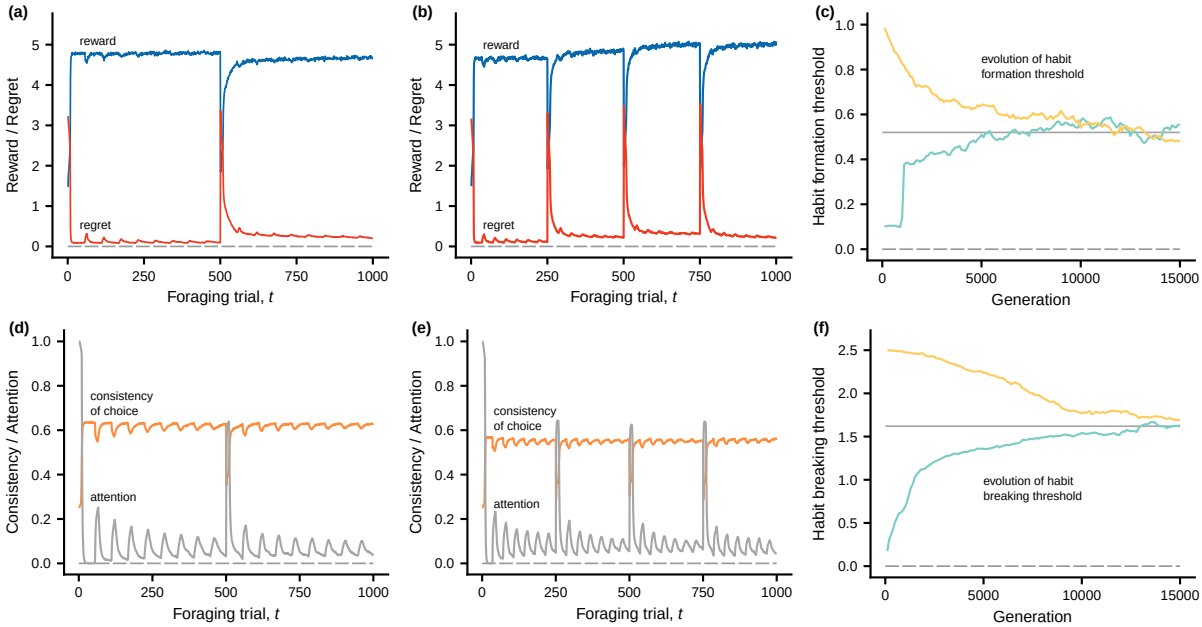

Figure 4: The forming and breaking of habits for sequences of randomly generated environments. For each environment in a sequence, the expected rewards for each of 8 foraging options is drawn from a log-normal distribution, in a similar way as for Fig. 2. Panels **(a)** and **(d)** show the dynamics of reward/regret and attention to exploration for OE, given as means over all simulated individuals, for a sequence of two randomly drawn environments, with each environment lasting 500 trials. Panels **(b)** and **(e)** show the corresponding results for a sequence of four environments, each lasting 250 trials. For this latter case, panel **(c)** shows two simulated evolutionary trajectories for the habit-forming trait $\hat{h}$, and panel **(f)** shows two trajectories for the habit-breaking trait $\hat{R}$. At evolutionary equilibrium, the trait values are $w_0 = 8.7$, $\alpha = 0.98$, $\beta = 11.9$, $\hat{h} = 0.60$, $\hat{R} = 1.64$, $\lambda = 0.90$, $\hat{t} = 55.8$ for two successive environments, and $w_0 = 8.1$, $\alpha = 0.98$, $\beta = 9.7$, $\hat{h} = 0.52$, $\hat{R} = 1.62$, $\lambda = 0.94$, $\hat{t} = 35.4$ for four successive environments (see also Fig. S4 in SI for corresponding TS results). The survival over 1000 trials is, respectively, around 0.83 and 0.82 for two and four environments.

Discussion

Based on our evolutionary analyses, we propose that the forming and breaking of habits is an important type of behavioural flexibility. The flexibility is advantageous over a range of situations, including when there are wholesale environmental changes. Our evolutionary perspective corroborates current views on habits in psychology [3, 4, 5]. An example of a point of agreement is that habit formation is known to be promoted by periods of environmental stability, and also by the repetition of actions that yield predictable rewards [2, 5].

Different meanings have been given to the term behavioural flexibility [41, 42, 43], with different experimental measures or tests of such flexibility. Reversal learning is perhaps the most commonly used test, and it is in this general sense we are suggesting that the forming and breaking of habits is a form of behavioural flexibility. The specific habit-forming and -breaking mechanisms assumed in our model are likely to be simplifications of real mechanisms. Thus, habits may form gradually [2, 5] rather than through a threshold mechanism and the breaking of habits may rely on other cues than assumed in our model. Inspired by theories of exploration-exploitation behaviour [18, 38, 39], we assumed that a lowering of the rate of rewards below a threshold can break a habit and trigger renewed exploration. In practice, other cues of a change of context likely contribute to the breaking of habits [3, 4, 5]. Such additional cues have been found to influence exploration-exploitation behaviour [28].

There is a wide field of investigation into mechanisms of learning. The switching on and off of learning in our models can be seen as a version of the much-studied idea that learning rates should respond to environmental characteristics, for instance with higher learning rates in more volatile environments [44, 45, 46, 47]. Further, the concept of attention has long been used to develop ideas about learning [48, 49, 50, 51, 52, 53]. Our model assumption of a simple switching on and off of attention to exploration of feeding options should be seen as a first-step simplification of possibly complex dynamics, involving gradation and shifts between different reward-predicting stimulus dimensions. Even so, there is an empirically well-supported tradition in psychology and neuroscience of so-called dual-process models, with essentially discrete modes of behaviour [1, 2, 5, 54, 55, 56]. The exploratory and habitual behaviours in our model are meant to represent such modes.

In addition to these models, there could to be other ways of being efficient in combining foraging and vigilance. One example is a model where individuals can commit to foraging over a period of several trials, while at the same time being sensitive to indirect, predator-related cues that can interrupt the foraging [57]. Individuals can then forage while also maintaining a certain vigilance, by being sensitive to cues, such as rustling sounds, indicating the presence of a predator. This need not be related to habit formation, but it is also possible that habitual foraging allows an individual to direct more attention to a range of cues that can give information about predation risk.

In general, exploration, foraging, and vigilance are tasks with several demands on mental capacities. For instance, exploration in nature could involve paying attention to and remembering locations and surroundings of different feeding options, in this way showing complexities beyond operating a multi-armed bandit.

The essential role played by habitual foraging in our model is to enhance multi-

tasking, through improved vigilance with largely maintained foraging success. The combination of foraging and vigilance is sometimes studied in psychology [57, 58] and, as mentioned, has long been of interest in behavioural ecology [14, 15]. This interest includes the role of attention in detecting predators [12, 13, 59] and the importance of vigilance for survival [60]. There are experimental approaches to measuring variation in vigilance, for instance comparing the time until detection of a (model) predator between different kinds of aggressive behaviour in prey animals [61], or a similar comparison for fish prey surrounded by either familiar or unfamiliar conspecifics [62]. Such approaches, involving measuring the reaction time to the presentation of a model predator, could serve to diagnose whether or not a foraging habit has been formed. This would then be an addition to current experimental methods of diagnosing habitual behaviour in psychological research [3, 4, 5].

A characteristic property of a habits is of course the tendency to use a particular action in a given situation. This could be used to diagnose and thus establish a time scale for the shift from exploration to exploitation of a foraging individual. For habit formation to be advantageous in nature, this time scale should be shorter than the time scale of significant environmental change. In addition to seasonal changes and new locations, one needs to also take into account other changes, such as those deriving from foraging itself. The latter will be particularly relevant for group-living animals with frequency-dependence in the rewards of foraging.

In addition to foraging, there are other ecologically significant activities that sometimes need to be combined with antipredatory vigilance, for instance social behaviour (like aggression), caring for offspring, and attracting mates. The formation of habits for these activities might then be advantageous. This points to the general conclusion that further study of habits should be worthwhile for evolutionary and behavioural ecology, including through modelling, field observation, and ecologically-motivated experiments.

Methods

Here we give an overview of the models we analyse, including the type of environments we study and the strategies individuals use in these environments. The strategies have parameters, and some of these parameters are assumed to be genetically determined traits. The evolution of these traits is studied in individual-based simulations, where the reproductive success is taken to be proportional to the total amount of reward an individual accumulates over its lifetime, up to death by predation or, otherwise, over

the total number of trials assumed for the model.

Distribution of rewards from a feeding option

As is illustrated in Fig. 1a, there is random variation in the rewards from a particular feeding option. We assume that rewards are log-normally distributed, which prevents negative values. Thus,

$$R_k(t) = \exp\big(\rho_k + z(t)\big), \tag{1}$$

where $z(t)$ are independent and normally distributed with mean zero and standard deviation σ_R and with ρ_k such that $\exp(\rho_k + \sigma_R^2/2) = \bar{R}_k$, where $\bar{R}_k$ is the mean reward from the option. This gives $\rho_k = \log(\bar{R}_k) - \sigma_R^2/2$. As illustrated in Fig. 1a, we use $\sigma_R = 0.25$ in all our examples. The mean values $\bar{R}_k$ of options vary between environments (see for instance Fig. 1b). We sometimes use the notation $R(t)$ to mean $R_k(t)$ for the chosen option k in trial t.

Predation risk and attention dynamics

In principle, individuals could pay attention to and learn about certain stimulus dimensions or features while ignoring others, but as a simplification we consider all learning about feeding options as regulated in an on/off fashion. Attention is indicated by a 0/1 variable a, with $a(t) = 1$ in trial t if the individual is exploring and learning in that trial (using either of the two approaches described below). The rate of predation in trial t, as a function of attention, is then

$$\mu_{\text{pred}}(t) = \mu_0 + \mu_1 a(t). \tag{2}$$

For our analyses we have used $\mu_0 = 0.00016$ and $\mu_1 = 0.00032$. Over a total of 1000 trials, this gives a survival of 0.62 with full attention to exploration and 0.85 for pure habitual foraging. This means that habit formation allows a forager to decrease predation mortality by at most around 60% of its value for full attention to exploration.

For deciding when to turn off attention and become habitual, we assume that while learning an individual keeps track of its consistency of choice h_k for each feeding option, as follows:

$$h_k(t+1) = h_k(t) + \alpha_h\big(1 - h_k(t)\big) \tag{3}$$

for the chosen option k in trial t and

$$h_l(t+1) = h_l(t) + \alpha_h\big(0 - h_l(t)\big) \tag{4}$$

for each non-chosen option $l \neq k$ (h_k can be thought of as a quantitative measure of 'habit strength'). We assume $\alpha_h = 0.25$ for the consistency updating rate, and we sometimes write consistency of choice to mean the consistency h_k of the chosen option on a given trial. This form of consistency of choice has been used previously [2] and has also been referred to as a 'choice kernel' [37].

An individual turns off attention to exploration in trial t if

$$h_k(t) > \hat{h} \tag{5}$$

for the chosen option k. If attention is off in trial t, the individual may turn it on if an estimated rate of reward R_e goes below a threshold $\hat{R}$:

$$R_\mathrm{e}(t) < \hat{R}. \tag{6}$$

The update for the estimated rate of reward is $R_\mathrm{e}(t+1) = R_\mathrm{e}(t) + \lambda\big(R(t) - R_\mathrm{e}(t)\big)$. This turning on of attention is a full reset of the individual's state, including of any learned estimates of the value of feeding options. In our evolutionary simulations, the update parameter λ evolved to be close to 1, so $R_\mathrm{e}(t) \approx R(t)$ for our results. To avoid the possibility of a rapid switching on and off of attention, for simulations with OE we introduce the requirement that there has been at least $\hat{t}$ trials of habitual foraging before applying condition (6). The parameters $\hat{h}$, $\hat{R}$, $\hat{t}$, and λ are implemented as genetically determined traits in our evolutionary simulations (for simulations with TS, the trait $\hat{t}$ is under very weak selection and was, for simplicity, kept at zero).

Optimistic exploration (OE)

We use action value learning [20] for the exploration part of the strategy, with the added element that the starting estimated values of the feeding options is relatively high ('optimistic initial values'; [20]). This is essentially a version of classical Rescorla-Wagner learning [63], adapted to operant conditioning.

So if $w_k(t)$ is an individual's estimate of the reward value of choosing option k in trial t, the starting value is assumed to be $w_k(0) = w_0$. Ideally, for optimistic exploration, w_0 should be higher than any of the actual mean reward values of options in the current environment (cf. Fig. 1c). If option k is chosen in trial t, the individual

updates its estimate to $w_k(t+1)$, where

$$w_k(t+1) = w_k(t) + \alpha \big(R_k(t) - w_k(t)\big). \tag{7}$$

Here α is the learning rate, for simplicity assumed to be constant in time and the same for all options, and $R_k(t)$ is the reward perceived by the individual from interacting with a stimulus of type k at time t. The quantity

$$\delta_k(t) = R_k(t) - w_k(t) \tag{8}$$

is referred to as the prediction error and is the difference between the reward $R_k(t)$ currently experienced by the individual and its prior estimate $w_k(t)$ of the reward. The change in the estimated feature value w_k in equation (7) is thus the learning rate α times the prediction error, and tends to move the estimate towards a 'true value'. In our simulations, the learning rate α evolved to be close to 1.

Soft-max choice

In each trial t, there is a choice between $n_\mathrm{s} = 8$ feeding options, or stimulus types. An individual chooses option k with probability

$$p_k(t) = \frac{\exp\big(\beta w_k(t)\big)}{\sum_m \exp\big(\beta w_m(t)\big)}, \tag{9}$$

where β is referred to as an 'inverse temperature' parameter. This procedure is typically referred to as a soft-max rule, going from estimated values to a probability of choice, and it is commonly used in reinforcement learning models [20].

Genetically determined traits

For evolutionary analysis, we assume that certain of the learning parameters are evolving traits. For optimistic exploration, the initial value w_0, the learning rate α, and the inverse temperature parameter β are such traits. In addition to these, there are also traits for the forming and breaking of habits, as described above.

Thompson sampling (TS)

Thompson sampling [35, 36] is a much-studied approach, with near optimal or optimal exploration-exploitation performance in certain situations, including for our example

in Fig. S2. We use this form of exploration for some of the results in the Supplementary Information, to give a comparison to the OE approach in the main text.

As mentioned, we assume that the different feeding options have log-normal reward distributions. So if option k, $k = 1, \ldots, K$, is chosen in trial t, we assume that the reward $R_k(t)$ is drawn from a log-normal distribution with parameters ρ_k and σ_R, i.e., $y_k(t) = \log(R_k(t))$ is normal with mean ρ_k and standard deviation σ_R. For convenience we sometimes use the precision $\tau_R = 1/\sigma_R^2$ as parameter instead of σ_R. In order to have analytic expressions for the posterior distributions, in some of our examples we further assume as prior that the parameters ρ_k of the options are (independently) drawn from a normal distribution with parameters ρ_0 and $\tau_0 = 1/\sigma_0^2$. Having observed y_{ki}, $i = 1, \ldots, n_k$ or, as a sufficient summary of observations, $\bar{y}_k = 1/n_k \times \sum_{i=1}^{n_k} y_{ki}$, the posterior distribution for ρ_k,

$$\mathrm{p}\big(\rho_k \,|\, y_{k1}, \ldots, y_{kn_k}\big),$$

is then normal with mean

$$\hat{\rho}_{kn_k} = \frac{\tau_0 \rho_0 + n_k \tau_R \bar{y}_k}{\tau_0 + n_k \tau_R} \tag{10}$$

and precision

$$\hat{\tau}_{kn_k} = \tau_0 + n_k \tau_R, \tag{11}$$

with $\hat{\sigma}^2_{kn_k} = 1/\hat{\tau}_{kn_k}$. This posterior is then used for the sampling of parameters.

The sampling and action choice procedure

The Thompson sampling procedure is, for each k, to 'draw' parameters $\hat{y}_k$ from the normal distributions with parameters from equations (10, 11), and then to choose the option with the largest posterior mean reward, i.e., with the largest value of

$$\hat{R}_{kn_k} = \exp\big(\hat{y}_{kn_k} + \sigma^2/2\big), \tag{12}$$

which is the same as choosing the option with the largest value of $\hat{y}_{kn_k}$.

Habitual foraging

For an OE forager, a transition to habitual foraging means that choices continue to be made according to equation (9), but there is no learning, i.e. there is no updating of

the estimates w_k of the kind given in equation (7). For TS, we assume that habitual foraging means that the last choice before the transition is repeated. In either case, the attention variable $a(t)$ in equation (2) is equal to zero during habitual foraging, leading to increased vigilance.

Order of events in a trial

In a trial, the following occurs: (*i*) choice of feeding option; (*ii*) reward; (*iii*) updating of consistency of choice and reward rate; (*iv*) if attention to exploration is on, check if it should be turned off; (*v*) if attention to exploration is off, check if it should be turned on; (*vi*) for optimistic exploration with attention on, perform learning updates; (*vii*) risk of predation, as in equation (2).

Evolutionary simulations

A simulated population has non-overlapping generations. For computational efficiency (multi-threading), the population is split into groups of size 10, with 400 groups, making up a total of 4000 individuals. Individuals are sexually reproducing, with one diploid locus for each trait. For simplicity, we assume that individuals are hermaphrodites, mating is random across the entire population, and the different loci are unlinked. Also for simplicity, we assume that reproductive success accumulates proportional to collected rewards over all trials (so that an individual dying from predation still has reproductive success from its rewards collected until death; other assumptions about the life history, such as discrete reproductive seasons, are of course possible).

Simulations were implemented as C++ programs, with separate programs for OE and TS. The genetically determined traits were w_0, α, β, $\hat{h}$, $\hat{R}$, λ, $\hat{t}$ for OE, and $\hat{h}$, $\hat{R}$, λ for TS. We implemented mutation as potentially occurring for each allele of a newly formed individual, with probability 0.002 and normally distributed mutational increments. The standard deviation of the increments was adjusted to the range of variation of a given trait.

Code availability

C++ source code for the individual-based simulations is available at GitHub, together with instructions for compilation on a Linux operating system (https://github.com/oleimar/habits1)

References

[1] Ray J. Dolan and Peter Dayan. Goals and habits in the brain. *Neuron*, 80(2): 312–325, 2013. doi: 10.1016/j.neuron.2013.09.007.

[2] Kevin J. Miller, Amitai Shenhav, and Elliot A. Ludvig. Habits without values. *Psychological Review*, 126(2):292–311, 2019. doi: 10.1037/rev0000120.

[3] Mark E. Bouton. Context, attention, and the switch between habit and goal-direction in behavior. *Learning & Behavior*, 49(4):349–362, 2021. doi: 10.3758/s13420-021-00488-z.

[4] Mark E. Bouton. Habit and persistence. *Journal of the Experimental Analysis of Behavior*, 121(1):88–96, 2024. doi: 10.1002/jeab.894.

[5] Eike K. Buabang, Kelly R. Donegan, Parnian Rafei, and Claire M. Gillan. Leveraging cognitive neuroscience for making and breaking real-world habits. *Trends in Cognitive Sciences*, 29(1):41–59, 2025. doi: 10.1016/j.tics.2024.10.006.

[6] Wendy Wood and Dennis Rünger. Psychology of habit. *Annual Review of Psychology*, 67:289–314, 2016. doi: 10.1146/annurev-psych-122414-033417.

[7] Sten Grillner and Brita Robertson. The basal ganglia over 500 million years. *Current Biology*, 26(20):R1088–R1100, 2016. doi: 10.1016/j.cub.2016.06.041.

[8] Sten Grillner. How circuits for habits are formed within the basal ganglia. *Proceedings of the National Academy of Sciences*, 122(13):e2423068122, 2025. doi: 10.1073/pnas.2423068122.

[9] Mariagrazia Portera and Mauro Mandrioli. Who's afraid of epigenetics? Habits, instincts, and Charles Darwin's evolutionary theory. *History and Philosophy of the Life Sciences*, 43(1):20, 2021. doi: 10.1007/s40656-021-00376-9.

[10] Slimane Dridi and Laurent Lehmann. Environmental complexity favors the evolution of learning. *Behavioral Ecology*, 27(3):842–850, 2016. doi: 10.1093/beheco/arv184.

[11] Magdalena Kozielska and Franz J. Weissing. A neural network model for the evolution of learning in changing environments. *PLOS Computational Biology*, 20 (1):e1011840, 2024. doi: 10.1371/journal.pcbi.1011840.

[12] Reuven Dukas. Behavioural and ecological consequences of limited attention. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1427):1539–1547, 2002. doi: 10.1098/rstb.2002.1063.

[13] Reuven Dukas. Causes and consequences of limited attention. *Brain Behavior and Evolution*, 63(4):197–210, 2004. doi: 10.1159/000076781.

[14] Manfred Milinski and Rolf Heller. Influence of a predator on the optimal foraging behaviour of sticklebacks (*Gasterosteus aculeatus* L.). *Nature*, 275(5681):642–644, 1978. doi: 10.1038/275642a0.

[15] Alasdair Ian Houston, John M. McNamara, and John M. C. Hutchinson. General results concerning the trade-off between gaining energy and avoiding predation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 341(1298):375–397, 1993. doi: 10.1098/rstb.1993.0123.

[16] Guy Beauchamp. *Animal vigilance: monitoring predators and competitors*. Academic Press, 2015. doi: 10.1016/C2014-0-01423-2.

[17] Nathaniel D. Daw, John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006. doi: 10.1038/nature04766.

[18] Jonathan D Cohen, Samuel M McClure, and Angela J Yu. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):933–942, 2007. doi: 10.1098/rstb.2007.2098.

[19] Katja Mehlhorn, Ben R. Newell, Peter M. Todd, Michael D. Lee, Kate Morgan, Victoria A. Braithwaite, Daniel Hausmann, Klaus Fiedler, and Cleotilde Gonzalez. Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, 2(3):191–215, 2015. doi: 10.1037/dec0000033.

[20] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction second edition*. MIT Press, Cambridge, MA, 2018.

[21] Debojyoti Biswas, Andrew Lamperski, Yu Yang, Kathleen Hoffman, John Guckenheimer, Eric S. Fortune, and Noah J. Cowan. Mode switching in organisms for solving explore-versus-exploit problems. *Nature Machine Intelligence*, 5(11): 1285–1296, 2023. doi: 10.1038/s42256-023-00745-y.

[22] Victor Ajuwon, Tiago Monteiro, Alexandra K. Schnell, and Nicola S. Clayton. To know or not to know? Curiosity and the value of prospective information in animals. *Learning & Behavior*, 53(1):114–127, 2025. doi: 10.3758/s13420-024-00647-y.

[23] Wendy Wood and David T. Neal. A new look at habits and the habit-goal interface. *Psychological Review*, 114(4):843–863, 2007. doi: 10.1037/0033-295X.114.4.843.

[24] Matthew R. Nassar, Robert C. Wilson, Benjamin Heasly, and Joshua I. Gold. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37):12366–12378, 2010. doi: 10.1523/JNEUROSCI.0822-10.2010.

[25] A. Izquierdo, J. L. Brigman, A. K. Radke, P. H. Rudebeck, and A. Holmes. The neural basis of reversal learning: An updated perspective. *Neuroscience*, 345: 12–26, 2017. doi: 10.1016/j.neuroscience.2016.03.021.

[26] Alireza Soltani and Alicia Izquierdo. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 20(10):635–644, 2019. doi: 10.1038/s41583-019-0180-y.

[27] Timothy E. J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F. S. Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, 2007. doi: 10.1038/nn1954.

[28] Deborah Racey, Michael E. Young, Dennis Garlick, Jennifer Ngoc-Minh Pham, and Aaron P. Blaisdell. Pigeon and human performance in a multi-armed bandit task in response to changes in variable interval schedules. *Learning & Behavior*, 39(3):245–258, 2011. doi: 10.3758/s13420-011-0025-7.

[29] Mark Steyvers, Michael D. Lee, and Eric-Jan Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, 2009. doi: 10.1016/j.jmp.2008.11.002.

[30] Ronen I. Brafman and Moshe Tennenholtz. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[31] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 881–888, New

York, NY, USA, June 2006. Association for Computing Machinery. doi: 10.1145/1143844.1143955.

[32] Oded Berger-Tal and Tal Avgar. The glass is half-full: overestimating the quality of a novel environment is advantageous. *PLoS One*, 7(4):e34578, 2012. doi: 10.1371/journal.pone.0034578.

[33] Sam Lobel, Omer Gottesman, Cameron Allen, Akhil Bagaria, and George Konidaris. Optimistic initialization for exploration in continuous control. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7612–7619, 2022. doi: 10.1609/aaai.v36i7.20727.

[34] Sheelabhadra Dey, James Ault, and Guni Sharon. Continual optimistic initialization for value-based reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, pages 453–462, Richland, SC, May 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 979-8-4007-0486-4.

[35] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. doi: 10.2307/2332286.

[36] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson Sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018. doi: 10.1561/2200000070.

[37] Robert C Wilson and Anne GE Collins. Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547, 2019. doi: 10.7554/eLife.49547.

[38] Angela J. Yu and Peter Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692, 2005. doi: 10.1016/j.neuron.2005.04.026.

[39] Peter Dayan and Angela J. Yu. Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4):335–350, 2006. doi: 10.1080/09548980601004024.

[40] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002. doi: 10.1023/A:1013689704352.

[41] Jean-Nicolas Audet and Louis Lefebvre. What's flexible in behavioral flexibility? *Behavioral Ecology*, 28(4):943–947, 2017. doi: 10.1093/beheco/arx007.

[42] Stephen E. G. Lea, Pizza K. Y. Chow, Lisa A. Leaver, and Ian P. L. McLaren. Behavioral flexibility: A review, a model, and some exploratory tests. *Learning & Behavior*, 48(1):173–187, 2020. doi: 10.3758/s13420-020-00421-w.

[43] Lucina Q. Uddin. Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, 22(3):167–179, 2021. doi: 10.1038/s41583-021-00428-w.

[44] Peter Dayan, Sham Kakade, and P. Read Montague. Learning and selective attention. *Nature Neuroscience*, 3(11):1218–1223, 2000. doi: 10.1038/81504.

[45] Samuel J. Gershman. A unifying probabilistic view of associative learning. *PLOS Computational Biology*, 11(11):e1004567, 2015. doi: 10.1371/journal.pcbi.1004567.

[46] Payam Piray and Nathaniel D. Daw. A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12(1):6587, 2021. doi: 10.1038/s41467-021-26731-9.

[47] Olof Leimar, Andrés E Quiñones, and Redouan Bshary. Flexible learning in complex worlds. *Behavioral Ecology*, 35(1):arad109, 2024. doi: 10.1093/beheco/arad109.

[48] N. J. Mackintosh. The effects of overtraining on a reversal and a nonreversal shift. *Journal of Comparative and Physiological Psychology*, 55(4):555–559, 1962. doi: 10.1037/h0046275.

[49] N. J. Mackintosh. Further analysis of the overtraining reversal effect. *Journal of Comparative and Physiological Psychology*, 67(2, Pt.2):1–18, 1969. doi: 10.1037/h0026784.

[50] N. J. Mackintosh. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4):276–298, 1975. doi: 10.1037/h0076778.

[51] Elijah Lovejoy. An attention theory of discrimination learning. *Journal of Mathematical Psychology*, 2(2):342–362, 1965. doi: 10.1016/0022-2496(65)90009-X.

[52] Elijah Lovejoy. Analysis of the overlearning reversal effect. *Psychological Review*, 73(1):87–103, 1966. doi: 10.1037/h0022687.

[53] John M. Pearce and Geoffrey Hall. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6):532–552, 1980. doi: 10.1037/0033-295X.87.6.532.

[54] Nathaniel D. Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005. doi: 10.1038/nn1560.

[55] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Value representations in the rodent orbitofrontal cortex drive learning, not choice. *eLife*, 11:e64575, 2022. doi: 10.7554/eLife.64575.

[56] Ted Moskovitz, Kevin J. Miller, Maneesh Sahani, and Matthew M. Botvinick. Understanding dual process cognition via the minimum description length principle. *PLOS Computational Biology*, 20(10):e1012383, 2024. doi: 10.1371/journal.pcbi.1012383.

[57] Kevin Lloyd and Peter Dayan. Interrupting behaviour: Minimizing decision costs via temporal commitment and low-level interrupts. *PLOS Computational Biology*, 14(1):e1005916, 2018. doi: 10.1371/journal.pcbi.1005916.

[58] Sahiti Chebolu, Peter Dayan, and Kevin Lloyd. Vigilance, arousal, and acetylcholine: Optimal control of attention in a simple detection task. *PLOS Computational Biology*, 18(10):e1010642, 2022. doi: 10.1371/journal.pcbi.1010642.

[59] Sasha R. X. Dall, Burt P. Kotler, and Amos Bouskila. Attention, 'apprehension' and gerbils searching in patches. *Annales Zoologici Fennici*, 38(1):15–23, 2001.

[60] William H. Ellsworth, Scott D. Peacor, Richard B. Chandler, L. Mike Conner, Elina P. Garrison, Karl V. Miller, and Michael J. Cherry. Measuring the benefit of a defensive trait: Vigilance and survival probability. *Ecology*, 105(11):e4429, 2024. doi: 10.1002/ecy.4429.

[61] Sven Jakobsson, Olle Brick, and Cecilia Kullberg. Escalated fighting behaviour incurs increased predation risk. *Animal Behaviour*, 49(1):235–239, 1995. doi: 10.1016/0003-3472(95)80172-3.

[62] S. W. Griffiths, S. Brockmark, J. Höjesjö, and J. I. Johnsson. Coping with divided attention: the advantage of familiarity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1540):695–699, 2004. doi: 10.1098/rspb.2003.2648.

[63] Robert A. Rescorla and Alan R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical conditioning II: current research and theory*, pages 64–99. Appleton-Century-Crofts, New York, 1972.

Acknowledgements

We thank Redouan Bshary, Peter Dayan, and Alex Thornton for helpful comments on a previous version of the manuscript.

Author contributions

O.L., S.R.X.D., A.I.H., and J.M.M. designed research; O.L. performed research; O.L. wrote the original draft; O.L., S.R.X.D., P.H., A.I.H., B.K,, and J.M.M. discussed results and contributed to the editing of the final manuscript.

Competing interests

The authors declare that they have no competing interest.

Supplementary information

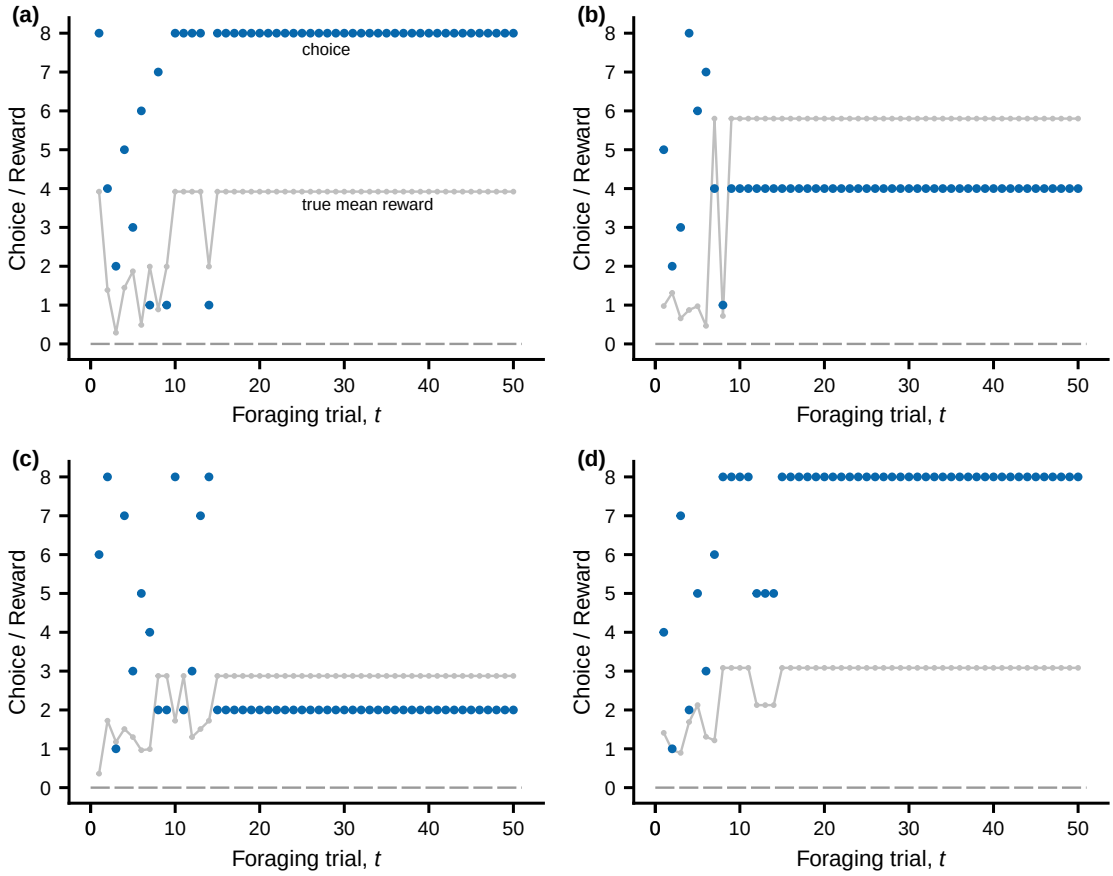

Figure S1: Examples of the sequence of choices of four different individuals that use optimistic exploration (OE), taken from the data in Fig. 2a. The individuals appear in panels **(a)** to **(d)**, which show the individual choices of feeding options in the first 50 trials (blue points). The true mean rewards of the options are also shown (in grey). For instance, in panel **(b)**, the individual chooses option 5 in trial 1, which gives a low reward, then chooses options 2, 3, 8, 6, and 7, which also give low rewards, and then option 4 which has the maximum reward. After also sampling option 1, which gives a low reward, the individual settles on choosing option 4. This can be compared with the case in panel **(d)**, where option 8 gives the maximum reward and option 5 a somewhat lower reward, resulting in a total of three choices of option 5 before settling on option 8.

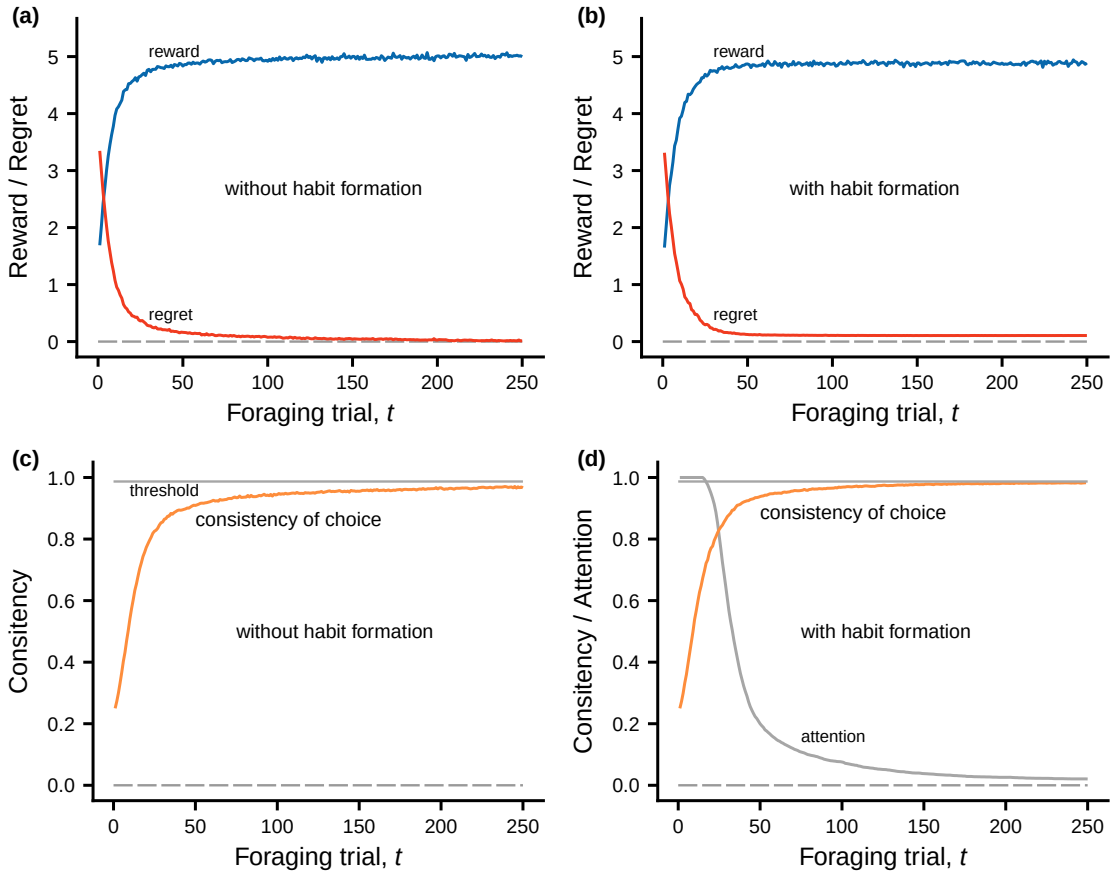

Figure S2: Habit formation when individuals that use Thompson sampling (TS) face a single environment during their lifetime (compare with Fig. 2 in the main text, where OE is used). The panels show averages over the first 250 foraging trials (of a total of 1000 trials) in an environment with 8 feeding options, with groups of 10 individuals replicated 400 times (making up a total of 4000 individuals; see Fig. S5 for examples of individual explorative choices). In each replicate the mean rewards of the options are independently drawn from a log-normal distribution (with parameters $\mu_0 = 0$, $\sigma_0 = 1$). Panels **(a)** and **(c)** show TS without habit formation, and **(b)** and **(d)** show corresponding results when individuals become habitual (turning off attention to exploration), when consistency of choice goes above an evolved threshold $\hat{h} = 0.987$ (after which the update of consistency stops). The averages over all individuals of the reward per trial (blue curves) and of the so-called regret (red curves) are shown in **(a)** and **(b)**. An individual's regret in a trial measures how much the individual lost by not choosing the option with the maximum expected reward (zero regret indicates an optimal choice). The average over all individuals of the consistency of the chosen option (orange curves) is shown in **(c)** and **(d)**. The consistency of an option measures the degree to which the current choice is the same as the most recent choice. Panel **(d)** also shows the average over individuals of the attention to exploration. The average survival over 1000 trials for the habit forming individuals is around 0.84, fairly close to the maximal value for full vigilance, and better than the expected survival of 0.62 for full attention to exploration.

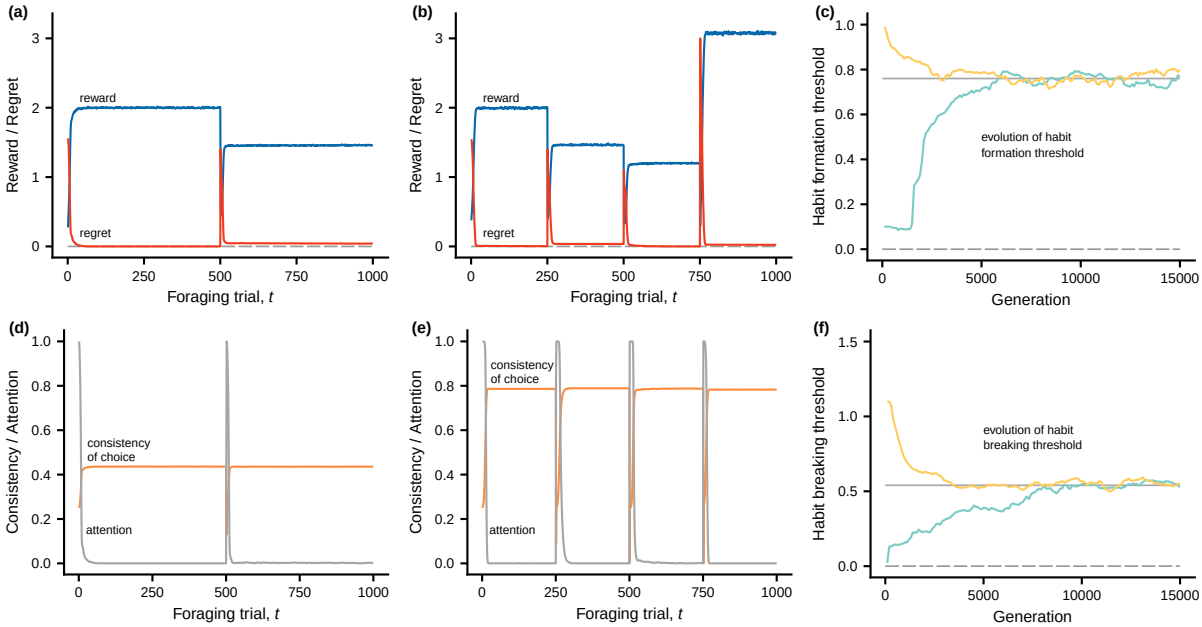

Figure S3: The forming and breaking of habits for the environments in Fig. 1b when individuals use TS for exploration. Panels **(a)** and **(d)** show the dynamics of reward/regret and attention to exploration for TS, given as means over all simulated individuals, for the sequence of the first two environments in Fig. 1b, with each environment lasting 500 trials. Panels **(b)** and **(e)** show the corresponding results for the sequence of all four environments in Fig. 1b, each lasting 250 trials. For this latter case, panel **(c)** shows two simulated evolutionary trajectories for the habit-forming trait $\hat{h}$, and panel **(f)** shows two trajectories for the habit-breaking trait $\hat{R}$. At evolutionary equilibrium, the trait values are $\hat{h} = 0.36$, $\hat{R} = 0.82$, $\lambda = 0.62$ for two successive environments, and $\hat{h} = 0.76$, $\hat{R} = 0.54$, $\lambda = 0.82$ for four successive environments. The survival over 1000 trials is near the maximum of 0.85 for two successive environment and around 0.83 for four environments.

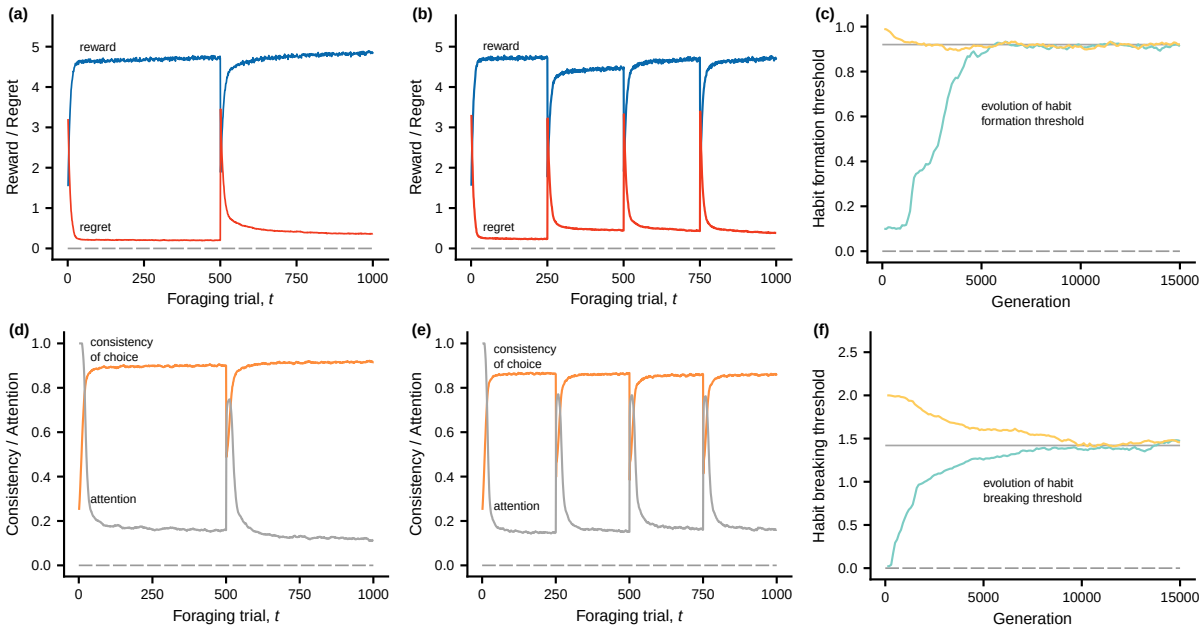

Figure S4: The forming and breaking of habits for sequences of randomly generated environments when individuals use TS for exploration. For each environment in a sequence, the expected rewards for each of 8 foraging options is drawn from a log-normal distribution, in a similar way as for Fig. S2. Panels **(a)** and **(d)** show the dynamics of reward/regret and attention to exploration for TS, given as means over all simulated individuals, for a sequence of two randomly drawn environments, with each environment lasting 500 trials. Panels **(b)** and **(e)** show the corresponding results for a sequence of four environments, each lasting 250 trials. For this latter case, panel **(c)** shows two simulated evolutionary trajectories for the habit-forming trait $\hat{h}$, and panel **(f)** shows two trajectories for the habit-breaking trait $\hat{R}$. At evolutionary equilibrium, the trait values are $\hat{h} = 0.94$, $\hat{R} = 1.30$, $\lambda = 0.94$ for two successive environments, and $\hat{h} = 0.96$, $\hat{R} = 1.42$, $\lambda = 0.96$ for four successive environments. The survival over 1000 trials is, respectively, around 0.80 and 0.79 for two and four environments.

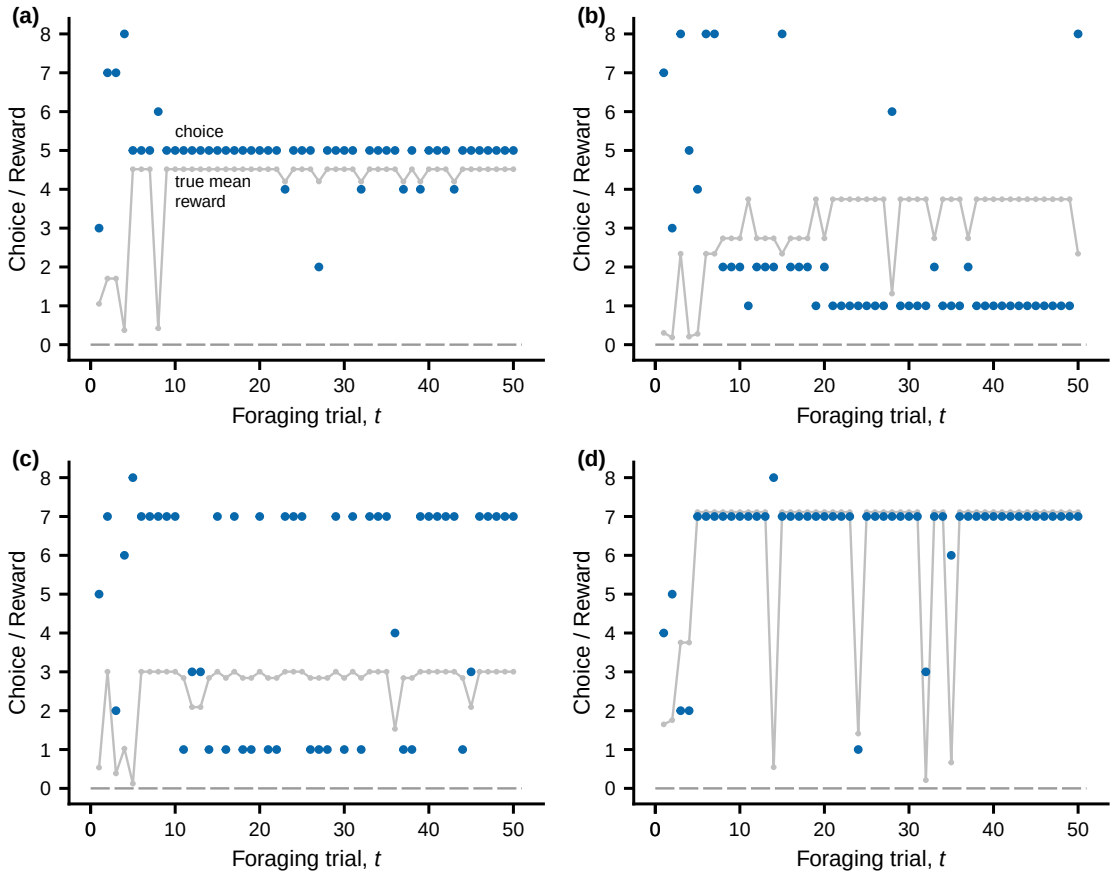

Figure S5: Examples of the sequence of choices of four different individuals that use Thompson sampling (TS), taken from the data in Fig. S2a. The individuals appear in panels **(a)** to **(d)**, which show the individual choices of feeding options in the first 50 trials (blue points). The true mean rewards of the options are also shown (in grey). The Bayesian inspired choice procedure for TS is described in the Methods section and, compared to OE, results in somewhat slower but ultimately more thorough exploration of the feeding options.

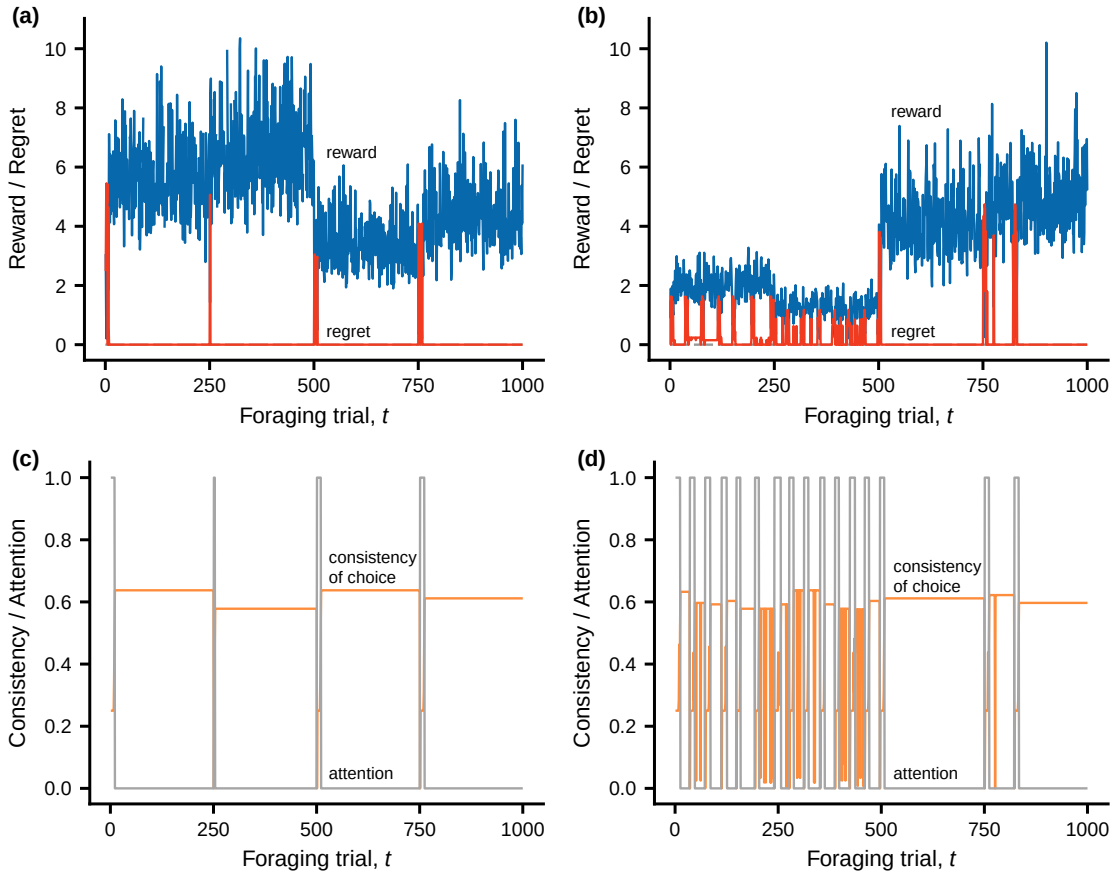

Figure S6: Examples of rewards (blue), regrets (red), consistency of choice (orange), and the attention to exploration (grey) of two different individuals that use OE over four successive randomly generated environments, taken from the data in Fig. 4b,e. One individual appears in panels **(a)** and **(c)** and the other in **(b)** and **(d)**.

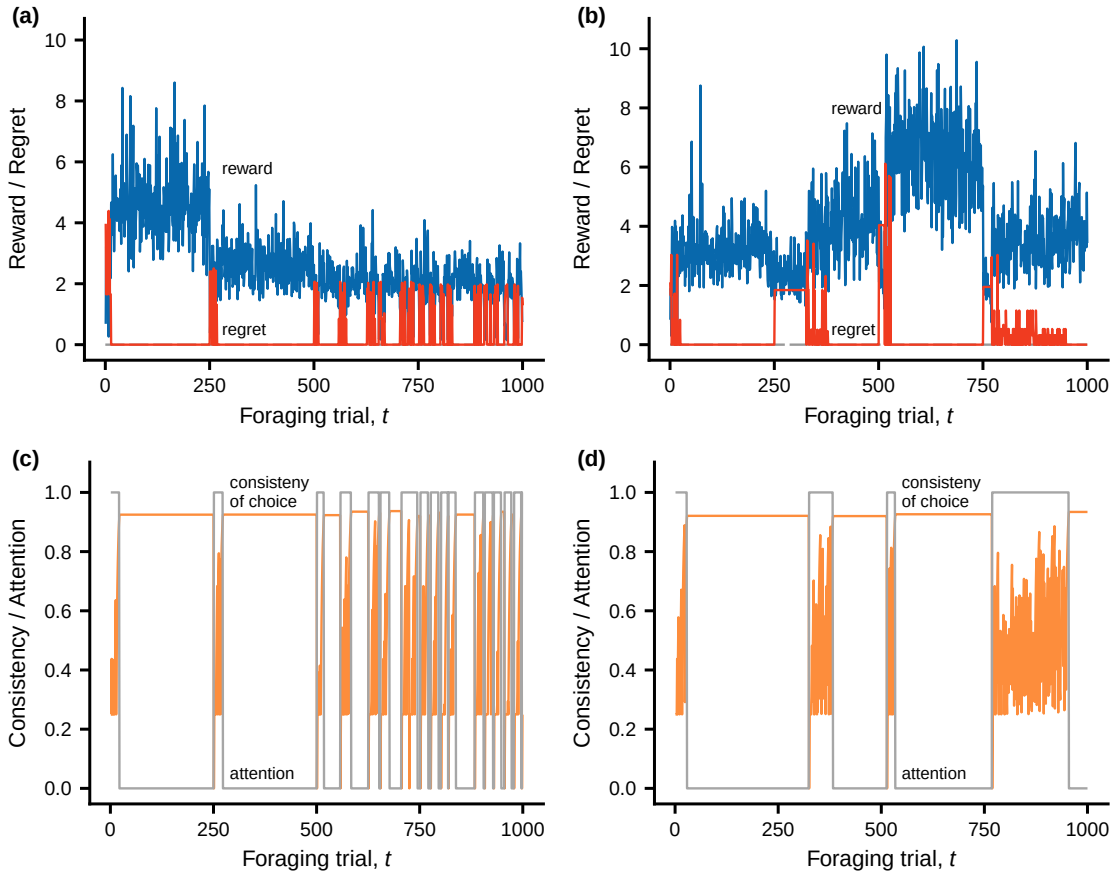

Figure S7: Examples of rewards (blue), regrets (red), consistency of choice (orange), and the attention to exploration (grey) of two different individuals that use TS over four successive randomly generated environments, taken from the data in Fig. S4b,e. One individual appears in panels **(a)** and **(c)** and the other in **(b)** and **(d)**.

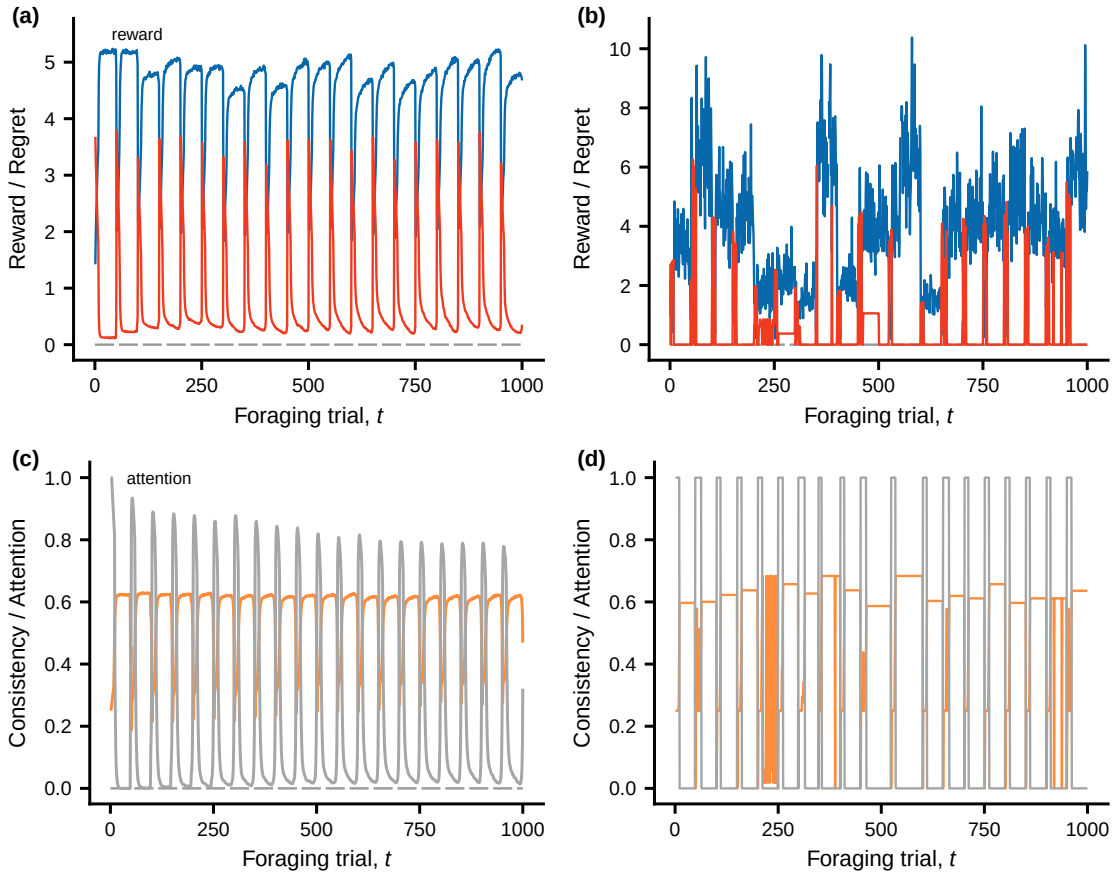

Figure S8: Estimated rewards (blue), regrets (red), consistency of choice (orange), and the attention to exploration (grey) for individuals that use OE. Individuals face a sequence of 20 randomly generated environments. Panels **(a)** and **(c)** show mean values over all individuals (4000 individuals split into 400 groups), and **(b)** and **(d)** show an example of a single individual. At evolutionary equilibrium, the trait values are $w_0 = 5.5$, $\alpha = 0.93$, $\beta = 16.4$, $\hat{h} = 0.59$, $\hat{R} = 2.95$, $\lambda = 0.87$, $\hat{t} = 48.5$. The average survival over 1000 trials is around 0.80. One reason for this relatively successful strategy is the rather large value of the trait $\hat{t}$, which is close to the duration of an environment (i.e., 50 trials).

Comparing optimistic exploration with Thompson sampling

Figures S2, S3, and S4 show results that correspond to Figs. 2, 3, and 4 in the main text, with optimistic exploration (OE) replaced by Thompson sampling (TS) as the exploration strategy. From these results, the main differences between OE and TS appear to be as follows.

Comparing Fig. 2a with Fig. S2a, we see that the regret decreases more rapidly for OE than for TS, but that TS is more thorough in approaching zero regret, so that in the long run in a given environment, TS will outperform OE. If we then consider the evolution of habit formation (comparing Fig. 2b,d with Fig. S2b,d), it seems that the faster initial exploration by OE allows for more rapid habit formation: the mean attention to exploration decreases more rapidly for OE than for TS (Fig. 2d vs. Fig. S2d). Overall, there might then be a slight advantage of OE over TS if one considers the exploration-exploitation trade-off.

For the combined evolution of habit formation and habit breaking, in the simpler-to-handle environments illustrated in Fig. 1b, OE and TS appear to perform almost equally well (compare Fig. 3 with Fig. S3). For the slightly more challenging randomly generated environments, it seems that OE does slightly better than TS. More rapid (but less thorough) exploration by OE than TS could be the explanation for this difference.