Behavioural Processes xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Behavioural Processes

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

The cognitive mechanisms of optimal sampling

Stephen E.G. Lea[a,*], Ian P.L. McLaren[a], Susan M. Dow[b], Donald A. Graft[c]

[a] Psychology (CLES), University of Exeter, Washington Singer Laboratories, Exeter EX4 4QG, United Kingdom
[b] Bristol Zoo Gardens, Bristol BS8 3HA, United Kingdom
[c] STMicroelectronics, Schaumberg, IL 60173, USA

ARTICLE INFO

Article history:
Received 20 April 2011
Received in revised form 5 October 2011
Accepted 8 October 2011

Keywords:
Foraging
matching law
Neural networks
Optimal sampling
Reinforcement learning
Rescorla–Wagner model
Two-armed bandit

ABSTRACT

How can animals learn the prey densities available in an environment that changes unpredictably from day to day, and how much effort should they devote to doing so, rather than exploiting what they already know? Using a two-armed bandit situation, we simulated several processes that might explain the trade-off between exploring and exploiting. They included an optimising model, dynamic backward sampling; a dynamic version of the matching law; the Rescorla–Wagner model; a neural network model; and ε-greedy and rule of thumb models derived from the study of reinforcement learning in artificial intelligence. Under conditions like those used in published studies of birds' performance under two-armed bandit conditions, all models usually identified the more profitable source of reward, and did so more quickly when the reward probability differential was greater. Only the dynamic programming model switched from exploring to exploiting more quickly when available time in the situation was less. With sessions of equal length presented in blocks, a session-length effect was induced in some of the models by allowing motivational, but not memory, carry-over from one session to the next. The rule of thumb model was the most successful overall, though the neural network model also performed better than the remaining models.

© 2011 Published by Elsevier B.V.

1. Introduction

1.1. The relation between foraging, operant psychology, and animal cognition

Optimal foraging theory (for review, see Stephens and Krebs, 1986; Stephens et al., 2007) has been one of the most fruitful ideas in recent behavioural ecology. Quite soon after its formulation by authors such as MacArthur and Pianka (1966) and Charnov and Orians (1973), its ideas were introduced into operant psychology (e.g. Collier and Rovee-Collier, 1981; Lea, 1979; Abarca and Fantino, 1982). This was an early example of the cross-fertilisation between behavioural ecology and comparative psychology that characterised the last decades of the twentieth century, and it provided a two-way benefit. For comparative psychology, optimal foraging theory provided a plausible evolutionary explanation of the laws of classical and operant conditioning: since almost all animals need to find food, if the known principles of associative learning under positive reinforcement enable them to do so in a near-optimal manner, then that is a sufficient reason for those phenomena following the laws that they do follow, and for them being so widespread in the animal kingdom. Conversely, for behavioural ecology, conditioning principles provided a mechanism by which near-optimal foraging could plausibly be produced, and also a non-arbitrary explanation for situations where foragers' behaviour departs from the predictions of optimality (Lea, 1985).

The bulk of the applications of conditioning principles to questions in foraging, however, have been concerned with steady state behaviour rather than with the acquisition of new information: they have been used to address issues in decision-making rather than learning. Some questions about the role of learning and memory in foraging certainly have been explored, but more in relation to unusual or advanced cognitive capacities, such as spatial memory (e.g. Kamil and Balda, 1985), episodic memory (e.g. Clayton and Dickinson, 1998) and social learning (e.g. Fisher and Hinde, 1949), than in situations where simple associative principles can be applied directly.

Yet there are undoubtedly many situations where near-optimal foraging demands that animals acquire information in ways to which simple associative learning principles can be applied directly. Optimality models of such common foraging problems as diet selection, patch use, and commuting to and from a central place include as parameters quantities such as prey value, handling times and prey densities. It has been widely recognized (e.g. Houston et al., 1982; McNamara, 1996) that these are very unlikely to be either stable across time or directly perceptible, and must therefore be learned and relearned repeatedly within the forager's lifetime.

* Corresponding author. Tel.: +44 1392 724612; fax: +44 1392 724623.
E-mail address: s.e.g.lea@exeter.ac.uk (S.E.G. Lea).

0376-6357/$ – see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.beproc.2011.10.004

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

Accounting for optimal foraging is therefore not just a matter of taking a static (though possibly stochastic, cf. Oaten, 1977; McNamara, 1982) situation and applying to it performance principles, whether those principles are derived from optimal decision theory or from the descriptive laws of operant choice (cf. Herrnstein, 1970). It is also a problem in animal cognition, and specifically in learning; how is a forager to acquire information about where it should look for prey?

1.2. The two-armed bandit

Acquiring information is not necessarily cost free. It may involve devoting time to sampling the environment instead of feeding from sources whose properties are already known. So finding out the properties of the environment may involve an optimising process in itself, in which the opportunity costs of exploration must be offset against the higher rates of food gain that can be expected given better information. Krebs et al. (1978) introduced into behavioural ecology consideration of one model situation that poses just such an explore/exploit problem, the "two-armed bandit". This is an example of an optimal sampling problem. It was first considered in statistics and applied to questions about the most efficient design of medical experiments (e.g. Thompson, 1933; Bradt and Karlin, 1956; Colton, 1963; Fox, 1974). Suppose we have two experimental treatments for a disorder (the two "arms" of the bandit). Neither of them is completely effective, but both work sometimes; however the probabilities that they are effective are not known. Obviously we need to experiment with both, to get some idea of their probabilities of success. But how long should we go on experimenting with both of them before deciding to offer the more successful one to all subsequent patients? If we continue experimenting for too long, we will give some patients a less effective treatment unnecessarily; but equally, if we decide which treatment is better on the basis of too little experience, we risk choosing the less effective treatment and giving it to all subsequent patients.

Applying this model to foraging, Krebs et al. (1978) considered the two arms of the bandit as representing two patches of prey, each of which delivers food at a fixed probability when the forager responds within it. The forager's problem is then to decide how long to spend exploring both patches, before deciding to spend the rest of its time exploiting the patch that has been found to yield more prey during exploration. A key difference from the original medical application is that the forager is considered to have only a limited time available in the situation, and on next exposure to the situation, the prey probabilities are once again unknown, so learning has to start again.

Krebs et al. (1978) considered a solution to the two-armed bandit problem offered by backwards dynamic programming (Bellman, 1956). In this algorithm, the forager is always in one of two states, exploring or exploiting. While exploring, it responds to both patches. After some number of exploratory responses, the forager decides, on the basis of the cumulative outcomes from the two patches, whether to switch to exploitation, and once it does that, it makes all subsequent responses to the preferred patch until the available time is exhausted. Although the principle of such explore/exploit strategies is simple, there is currently no general solution to identifying the optimal point for switching from exploring to exploiting (Cohen et al., 2007). The particular algorithm Krebs et al. used imposed a constraint on the general explore/exploit strategy, such that during the sampling phase the forager always makes one response to each option, and evaluation of whether to continue sampling occurs after each pair of sampling trials; with this constraint, the optimal strategy can be derived. It involves a backward dynamic programming algorithm, which starts from the point where the forager only has time to make two more responses; clearly these should both be made to the patch that has yielded more prey so far. The algorithm then works backwards, working out the expected benefit from continuing sampling further trials or switching to exploitation, for every combination of numbers of prey obtained so far from the two patches, in order to specify the point at which the forager should stop exploring and exploit what currently seems to be the better food source. Krebs et al. carried out simulations of behaviour in the two-armed bandit situation using the backwards dynamic programming algorithm, and showed that it had two key properties. The switch from exploring to exploitation occurred earlier (a) if the difference between payoff probabilities in the two patches was greater and (b) if the time allowed in the patches was less.

1.3. Animal behaviour in two-armed bandit situations

To test whether their backwards dynamic programming model was relevant to the behaviour of real foragers, Krebs et al. (1978) exposed great tits (*Parus major*) to a laboratory environment that had the properties of a two-armed bandit. An aviary had two perches, one at each end. If the tit hopped onto either perch, food was delivered, at a different probability for each perch. The payoff probabilities for the two perches always added up to 0.5 (so that during random exploration, the tits would obtain food for a quarter of hops), but the payoff differential varied from 0.5 (probabilities of 0.5 and 0) to 0.1 (probabilities of 0.3 and 0.2). The number of hops was effectively limited by the birds' becoming satiated, which occurred after about 150 hops on average. The birds showed an approximate explore/exploit pattern of performance, although they did not strictly alternate between the perches in the explore phase, and they tended to continue to make some hops on the less preferred perch throughout the session. Because of this second tendency, Krebs et al. introduced a criterion of 90 out of 100 successive hops being made to the same perch to determine the point at which they would regard a tit as having switched to exploitation. Using this criterion, they observed that the tits showed the payoff differential effect predicted by the optimal sampling model. They also found that the tits switched to exploitation at about the same point as their simulated optimal samplers did in a 150-hop session, but Plowright and Plowright (1987) subsequently showed that Krebs et al. had overestimated the amount of sampling that the optimal model predicted, so that their tits had in fact over-sampled relative to the optimal strategy. Kacelnik (1979) reported that, at least when the two payoff probabilities were rather similar, the tits took longer to commit themselves to one patch in longer sessions; the failure to find such an effect with larger payoff differentials can be attributed to a floor effect, since exploration was typically quite brief under such conditions.

In the language of operant psychology, the fixed payoff probabilities in a two-armed bandit are concurrent random ratio schedules of reinforcement. Dow and Lea (1987b, Experiment 1) carried out a further test of Krebs et al.'s backwards dynamic programming model, using pigeons working for food in a conventional Skinner box on concurrent random ratio schedules arranged on a single key with pecks on a separate key causing changes between the schedules (Findley, 1958). They fixed the length of the session at either 256 pecks or 1024 pecks, and in accordance with the predictions of optimal sampling theory, found that fewer changeovers between schedules were made in the short sessions than in the first 256 pecks of the long sessions, implying that exploration continued for longer in the latter. However, Shettleworth and Plowright (1989) were unable to find any effect of session length on pigeons' behaviour on concurrent random ratio schedules, even though they later reported such effects in a prey choice problem (Plowright and Shettleworth, 1991), and further demonstrations of a session-length effect have not been forthcoming. The empirical status of the

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

session-length effect is therefore unclear, despite the fact that it is a clear prediction of optimal sampling theory that it should occur.

1.4. Learning mechanisms for near-optimal sampling

Although the major predictions of optimal sampling theory have some empirical support, it is obvious that backward dynamic programming as such is not a plausible mechanism for a bird to use to solve the two-armed bandit problem. However, it would be possible for animals to evolve cognitive processes that would produce the kinds of behaviour that backward dynamic programming would require. Two kinds of process have to underlie performance in a two-armed bandit situation: learning rules (which determine how an animal's estimates of the reward probabilities on the two arms are updated by experience), and performance or decision rules (which determine how those estimates are translated into response probabilities; cf. Tolman, 1955). Houston et al. (1982) and Lea and Dow (1984) considered several examples of performance and learning rules that should do a reasonably good job. Houston et al. (1982) simulated the effect of a wide variety of decision rules, but did not explicitly consider the mechanism for updating parameter estimates. Lea and Dow (1984) considered only a single performance principle, Herrnstein's (1961, 1970) matching law for concurrent schedules of reinforcement. Herrnstein and Loveland (1975) pointed out that, under stable conditions, the only behaviour in concurrent ratio schedules that is consistent with the matching law is absolute preference for one schedule or the other – that is, exploitation. However, at the beginning of a session in a two-armed bandit, the reinforcement rates for the two arms are unknown, so some mechanism must be provided for them to be learned. Lea and Dow (1984) accordingly added an explicit learning model to the matching law, used the linear operator model introduced by Bush and Mosteller (1951); this generates a changing estimate (an exponentially weighted moving average, or EWMA) of relative reinforcement rates which relative response rates should, according to the matching law, equal. Most authors who have considered transitions between schedules of reinforcement have used some variant of the Bush–Mosteller model; examples of dynamic matching models include those of Myerson and Miezin (1980), Mazur (1992) and Aparicio and Baum (2009). An example of the use of a linear operator model in a foraging context can be found in McNamara and Houston (1985).

Dow and Lea (1987a) describe the results of computer simulations of a dynamic matching model, applied to a number of different foraging situations. In the case of the two-armed bandit situation, their simulations reproduced some but not all of the behaviour both of Krebs et al.'s (1978) simulations of the backward dynamic programming algorithm, and of the birds studied by Krebs et al. and Dow and Lea (1987b). However, dynamic matching is not the only model that should be considered. The study of associative learning has developed substantially within comparative psychology since the 1980s. Dow and Lea (1987a) did not consider models of learning that were already highly influential within the field of Pavlovian conditioning, such as the model of Rescorla and Wagner (1972). Keasar et al. (2002) have subsequently applied the Rescorla–Wagner model to simulate the behaviour of bumble bees foraging in a two-armed bandit situation, so models of this kind clearly need to be included.

Furthermore, during the 1980s, the study of unsupervised learning – that is, learning where the agent has to find out what to do from experience within a situation, rather than being instructed – became an important field within artificial intelligence, where it is known as "reinforcement learning". The two-armed bandit, and in particular the dilemma of exploration versus exploitation, is a central issue in research within this field (Sutton and Barto, 1998, chapter 1). A large number of models and algorithms have been proposed that might produce near-optimal sampling without requiring the agent to carry out backward dynamic programming, or to have any internal model of the situation.

1.5. The present investigations

The present paper therefore extends the work of Dow and Lea (1987a) by considering additional learning models and investigating whether they can reproduce the key features of behaviour of the optimal sampler or of real birds in the two-armed bandit situation. The models considered are:

(1) The backwards dynamic programming algorithm as specified by Krebs et al. (1978). This was included as a standard of comparison. It is however at a different conceptual level than the remaining models, in that it does not propose a realistic mechanism by which near-optimal behaviour can be achieved, but rather seeks to specify what the behaviour should be like.
(2) The dynamic matching model used by Dow and Lea (1987a). A distinctive feature of matching models is that they work in terms of reinforcement rates (reinforcements per unit time), not reinforcement probabilities (reinforcements per response). All reinforcement learning models will specify that, when the agent responds to an arm in a two-armed bandit situation, the value estimate for that arm will go up if reinforcement follows and down if it does not. In a matching model, however, the value estimate for the non-chosen arm will also be reduced, regardless of whether reward is given on the chosen arm. In the model tested, the adjustments to the values of the two arms were made using the same linear operator model with the same learning rate parameter.
(3) A simple version of the Rescorla–Wagner conditioning model. For application in a choice situation, this model requires the addition of a performance rule; we use the softmax rule with a Boltzmann distribution (cf. Sutton and Barto, 1998, p. 30) to convert the associative strengths of the two outcomes into response probabilities. This transformation requires a parameter t usually referred to as "temperature" which affects how discriminating the agent is between the two outcomes; the natural way to implement this in a Rescorla–Wagner context is to set t equal to the parameter λ which represents the total associative strength that the reinforcer can support, and this is what we did.
(4) A connectionist model, derived from Wills and McLaren (1997). This also uses Rescorla–Wagner model for updating the value estimates for the two arms of the model, but it uses a process of iterated mutual inhibition between actions, with a random element, followed by a winner-takes-all rule, to generate actions from the values. In addition to the parameter for the Rescorla–Wagner learning process, this model requires a decay rate parameter, d.
(5) An ε-greedy algorithm with linear-operator value adjustment and rule of thumb absorption onto one or the other arm. A "greedy" algorithm is one that always chooses the action that is associated with the higher current value estimate, though value need not correspond directly to a current estimate of reinforcement probability; it may be affected by an exploratory component, the possibility that making an action would yield information that would increase future reinforcement rate. An ε-greedy algorithm does the same, except that on a proportion ε of trials it chooses at random (Sutton and Barto, 1998, pp. 26–28).
(6) A rule of thumb model implementing a sample-then-exploit strategy, switching from sampling to exploiting when the difference between current estimates of reinforcement probability on the two arms exceeds a threshold, L. This is a kind

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

of satisficing model (cf. Simon, 1956; Ward, 1992), but not of the usual kind: rather than seeking a satisfactory (but not necessary optimal) amount of food, the model seeks a satisfactory (but not necessary optimal) amount of information as to which patch is better. A rule for updating the estimates of patch density is needed in any such model; in the present case, the linear operator rule was used.

Obviously these are only a small selection of the models that could be considered. They are chosen because they represent a range of different approaches to the problem, all of which have been influential in one or another area of the study of learning. The versions of these models used are also all at roughly the same level of complexity: all require the fitting of either one or two parameters. Thus if one or two of the models currently being considered do consistently better than the others, we can reasonably conclude that those classes of models deserve fuller exploration.

It is scarcely necessary to demonstrate that these models will typically find the arm with the better reward probability in a two-armed bandit; any learning model can be expected to do that. The key questions are whether the models will show the payoff differential and session-length effects found by Krebs et al. (1978), Kacelnik (1979) and Dow and Lea (1987b), and how well the different models do in terms of the overall number of rewards they manage to extract from the situation, since this is the currency that natural selection will operate on. It is especially interesting to compare different learning models on this metric, and to compare them all with the outcome achieved by the backwards dynamic programming model.

A priori reasoning suggests that all the reinforcement learning models described above are likely to show the payoff differential effect, but it is not obvious how any of them would lead to the session-length effect. There are three ways in which they might do so:

First, all models include parameters that directly or indirectly affect the rate of learning. Adjusting such parameters as a function of session length will clearly alter the balance between exploration and exploitation. However, how could such adjustment occur? Killeen (1984) has suggested that the learning rate parameter in dynamic matching models might itself be subject to learning, by a second linear-operator process. If we do not adopt some such proposal, adjustment of the parameter seems to be an ad hoc solution.

The remaining possibilities make use of the fact that a session-length effect can only be expected when the animal has some means of knowing the likely session length in advance. Both in natural situations and in the laboratory, the most likely way for this to happen is for session length to be roughly constant over a series of sessions. Shettleworth and Plowright (1989) pointed out that if this is done, either or both of two carry-over effects may occur, either of which might cause a session-length effect. Following a longer session, the forager's memory of which arm was the better in the previous session may be more firmly established, and this might lead to more protracted sampling. Alternatively, the longer previous session might cause the forager's motivation to seek food to be less. This might have an impact on the learning rate parameter in some models; in others it might affect the relation between reward on an arm and the value attached to it. Shettleworth and Plowright showed empirically that the session-length effects reported by Kacelnik (1979) and Dow and Lea (1987b) could well have been due to such carry-over effects.

2. Procedures

Each of the models was first tested under conditions typical of both Krebs et al. (1978) and Dow and Lea (1987b). We then focused on the conditions of Krebs et al., and varied payoff differential, session length, memory carry-over and motivational carry-over. We did not attempt a fully factorial variation of these conditions. Instead we defined a baseline procedure using intermediate values for payoff differential and session length, and zero values for memory and motivational carry-over. We then varied each condition from this baseline.

Krebs et al. (1978) used reinforcement probabilities that always added up to 0.5 between the two arms, varying in payoff differential from 0.1 (reward probabilities of 0.2 and 0.3) to 0.5 (reward probabilities of 0 and 0.5). We did not want to use the extinction condition (reinforcement probability zero), so we used 0.1 as the low payoff differential, and 0.4 (reward probabilities of 0.05 and 0.45) as the high differential. The baseline condition was therefore set at a payoff differential of 0.25 (reward probabilities of 0.125 and 0.375). In Dow and Lea's (1987b) experiment, the lower payoff probability ranged from 0.02 to 0.05, and the higher from 0.029 to 0.1, with the higher always at least twice the lower. We therefore set the low payoff differential at 0.035 (reward probabilities of 0.035 and 0.070) and the high differential at 0.07 (reward probabilities of 0.02 and 0.09). The baseline condition was set as a payoff differential of 0.05, with reward probabilities of 0.03 and 0.08.

Krebs et al. (1978) report results from a session length of 150 hops. Because their criterion of absorption to a perch was 90 out of 100 hops being made to it, we took 150 hops as the shorter session length, and 600 as the longer, when varying session length; the baseline condition was set at 300 hops.

In baseline conditions, there was no carry-over in either memory or motivation from the previous simulation run. The models started each session with both reward probabilities set at an arbitrary estimate of 0.5, and motivation set to a value of 1.0. In the memory carry-over condition, these estimates were set at the mean of 0.5 and the final reward probability estimate for the same arm from the previous run. In motivation carry-over sessions, motivation at the beginning of the session was set at the mean of the baseline value of 1.0, and a value derived from the total rewards earned within the previous session. The value carried over was calculated so that it would be 1.0 if the forager was perfectly efficient under baseline conditions. Note that we did not include within-session motivational changes in the modelling, so what we are modelling here is the situation where the forager is loading quickly with food but only digesting it later. This is realistic for many bird species.

The effects of variations in motivation were implemented in different ways for different models. In the matching model, the increment of estimated reinforcement rate when reward was earned was multiplied by the motivation parameter (which was thus treated as equivalent of the parameter k in Herrnstein's (1970), systematic development of the law of effect). In the Rescorla–Wagner model (whether implemented analytically or through a neural network), the total associative strength that the reinforcer could support (represented by the parameter λ) was reduced as a function of the number of rewards obtained in the previous run; note that in our simple Rescorla–Wagner model we use λ as the temperature parameter in the transformation of associative strengths into response probabilities, so motivation might have a dual effect in this model. In the ε-greedy model, the increment of estimated reinforcement probability when reward was earned was multiplied by the motivation parameter. Finally, in the rule of thumb model, the value increment resulting from each reward was reduced as a function of the total rewards gained in the previous run.

For each model, parameter variables were first varied systematically to find the values that worked best under each of the baseline conditions derived from Krebs et al. (1978) and Dow and Lea (1987b), using different values for the two conditions if necessary, and with no carry-over of memory or motivation from

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

Table 1
Outcomes of model simulations under two sets of baseline conditions. Entries are mean results from 200 simulation runs. Payoff probabilities on the two arms were 0.125 and 0.375. p(abs) is the probability of absorption; $p(C)$ the probability of initial absorption on the objectively preferable arm; R(abs) the median responses to initial absorption; $\rho(C)$ the mean Herrnstein et al. ρ towards the objectively preferable arm; and $p(Rf)$ the overall reward probability.

Condition	Model	Parameters		p(abs)	$p(C)$	R(abs)	$\rho(C)$	$p(Rf)$
Krebs et al.	Backwards dynamic programming			1.00	1.00	2	0.97	0.365
baseline	Dynamic matching	α 0.06		1.00	0.99	27	0.93	0.354
	Rescorla–Wagner	β 0.55		1.00	0.99	26.5	0.89	0.353
	Connectionist	β 0.08	d 0.20	0.99	1.00	2	0.92	0.362
	ε-greedy	α 0.07	ε 0.02	1.00	1.00	18	0.89	0.358
	Rule of thumb	α 0.035	L 0.08	1.00	1.00	2	0.97	0.366
Dow and Lea	Backwards dynamic programming			1.00	0.97	57	0.91	0.075
baseline	Dynamic matching	α 0.075		1.00	0.85	32	0.82	0.071
	Rescorla–Wagner	β 0.55		0.83	0.95	185.5	0.80	0.070
	Connectionist	β 0.065	d 0.35	0.01	0.60	513	0.52	0.059
	ε-greedy	α 0.08	ε 0.05	0.99	0.93	129.5	0.82	0.070
	Rule of thumb	α 0.03	L 0.05	1.00	0.93	29	0.88	0.073

the previous session. In almost all models, there was a learning rate parameter that could be varied. Some models had additional parameters, for example the decay rate in the connectionist model, and the threshold value in the rule of thumb model. The parameter values chosen were those that gave the best outcome, defined in terms of total reward rate. The models varied in how sensitive their performance was to parameter values, but in all cases the values chosen came from within a range where there was little variation in performance to either side of the chosen values. Once parameter values had been chosen, new simulations were run with those values, and the outcomes of these new simulations are reported.

To test the effects of varying payoff probability differential and session length, the best-performing parameter values for those conditions were found under baseline conditions, and simulations run with the high and low payoff differentials and sessions lengths. To test the effects of memory and motivational carryover, each carryover was implemented separately with the baseline conditions (payoff probabilities of 0.125 and 0.375, session length of 300) and a new sweep of parameters undertaken; the best parameter values found were then used with sessions lengths of 150 and 600, to see whether the models would predict a session-length effect under these conditions.

3. Results

Table 1 gives key results, averaged from 200 simulations of each model under the two baseline conditions. The results reported are:

(a) The probability that the model reached the absorption criterion on either of the arms, p(abs).
(b) The probability that the model absorbed first onto the better arm, $p(C)$. Absorbing onto an arm was defined using the criterion introduced by Krebs et al. (1978), of 90 responses out of 100 being made to a single arm. It should be noted that it was possible in some conditions that a model would meet this criterion on one arm but subsequently make more responses to the other arm.
(c) The median number of responses made before the first run of responses that met the absorption criterion. Medians rather than means are reported to deal with the case where p(abs) was less than one. If a model did not absorb in a particular run, the number of responses to absorption was set at the session length plus one; thus if p(abs) was less than 0.5, so that the model failed to absorb in a majority of cases, the median number of responses made would be recorded as session length plus one.
(d) $\rho(C)$, the mean value of the normalised Mann–Whitney U statistic introduced by Herrnstein et al. (1976), here used to express the tendency for responses to the objectively better arm to occur later in the session than responses to the objectively less good arm. ρ is a measure of the speed of learning to go to the better arm; if a bird first made an uninterrupted series of responses to the worse arm, and then an uninterrupted series of responses to the better arm, ρ would equal 1.0, whereas if its responses to the two arms were randomly distributed within the session, ρ would equal 0.5.
(e) The mean probability of reinforcement obtained in the session, $p(Rf)$.

Table 1 shows that with reward probabilities in the range used by Krebs et al., all models can be adjusted to give performance that is close to that of the backward dynamic programming algorithm, and one of them (the rule of thumb model) does marginally better than the backwards dynamic programming algorithm, though the difference is within likely sampling error. With reward probabilities in the range used by Dow and Lea, however, performance of the models is much more variable, especially with regard to convergence rate. The connectionist model largely failed to converge to the correct solution, and the Rescorla–Wagner model did not do so in every case. Only the rule of thumb model performed as well as the backwards dynamic programming algorithm in terms of overall reinforcement rate obtained.

Table 2 shows the effects of varying payoff probability differential in an environment like that of Krebs et al. (1978), in the absence of either memory or motivational carry-over, and with session length constant at 300. The parameters used for each model were held constant at the values found to give the best performance in the preceding set of tests. It can be seen that all the models predict appropriate behaviour, with absorption on a preferred outcome occurring earlier when the payoff differential was higher. Under both conditions, all models continued to perform relatively well against the standard of the backward dynamic programming algorithm, even though, when the payoff differential was small, only the dynamic programming and rule of thumb models produced convergence in every test run. For all models (including the backwards dynamic programming algorithm), performance was notably better with the bigger payoff differential, both in terms of the speed and accuracy with which the models found the objectively better arm, and in terms of the proportion of the maximum possible payoff achieved. With reward probabilities of 0.2 and 0.3, a purely random forager would achieve a reward rate of 0.25, while an omniscient one would achieve a rate of 0.30; the models (other than the dynamic programming model) achieved reward rates between 0.275 and 0.278, i.e. between 50% and 56% of the possible gains from non-randomness. With reward probabilities of 0.1 and 0.4, between 87% and 96% of the possible gains were achieved.

Table 3 shows the effects of varying session length from 150 to 600 responses, in conditions otherwise similar to those used

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

Table 2
Outcomes of model simulations under two levels of payoff differential, under conditions like those of Krebs et al. (1978). Entries are mean results from 200 simulation runs. p(abs) is the probability of absorption; $p(C)$ the probability of initial absorption on the objectively preferable arm; R(abs) the median responses to initial absorption; $\rho(C)$ the mean Herrnstein et al. ρ towards the objectively preferable arm; and $p(Rf)$ the overall reward probability.

Condition	Model	Parameters		p(abs)	$p(C)$	R(abs)	ρ(C)	$p(Rf)$
Payoff probabilities 0.2 and 0.3	Backwards dynamic programming			1.00	0.88	16	0.83	0.284
	Dynamic matching	α 0.06		0.98	0.88	47.5	0.81	0.275
	Rescorla–Wagner	β 0.55		0.77	0.89	99.5	0.74	0.276
	Connectionist	β 0.08	d 0.20	0.61	0.86	104.5	0.73	0.276
	ε-greedy	α 0.07	ε 0.02	0.93	0.90	80.5	0.77	0.277
	Rule of thumb	α 0.035	L 0.08	1.00	0.83	12.5	0.80	0.278
Payoff probabilities 0.1 and 0.4	Backwards dynamic programming			1.00	1.00	2	0.98	0.390
	Dynamic matching	α 0.06		1.00	1.00	17	0.94	0.380
	Rescorla–Wagner	β 0.55		1.00	1.00	16.5	0.91	0.382
	Connectionist	β 0.08	d 0.20	1.00	1.00	2	0.95	0.389
	ε-greedy	α 0.07	ε 0.02	1.00	1.00	11	0.90	0.382
	Rule of thumb	α 0.035	L 0.08	1.00	1.00	2	0.98	0.394

by Krebs et al. (1978), using the baseline reward probabilities of 0.125 and 0.375. Again model parameters were set at the values that had been found to give the best performance under baseline conditions, with a session length of 300. In terms of overall reinforcement rate, all the models did better in the longer sessions, as would be expected; the difference was substantially less for the backwards dynamic programming algorithm and the rule of thumb model, and somewhat less for the connectionist model, than for the others, mainly because these models did markedly better than the others in the short sessions. The enhanced performance of these models can be attributed to the fact that they led to earlier absorption than the others, without much loss in accuracy of finding the objectively preferable arm. Several models failed to produce 100% absorption in the shorter sessions. As it should, the dynamic programming algorithm absorbed after fewer trials in the shorter sessions than in the longer ones. None of the other models showed such an effect.

Table 4 shows the effect of introducing a 50% memory carry-over from the previous session. Reward probabilities were set at 0.125 and 0.375. The values of model parameters that gave the highest overall reward probability with a session length of 300 were found by a sweep of parameter values, and the models were then tested with these parameter values at session lengths of 150 and 600. Once again the connectionist and rule of thumb models produce markedly earlier absorption than the others in the short sessions, and without a compensating loss of accuracy, so that they do better on the ultimate objective of securing the maximum overall reinforcement rate. The backwards-programming model again shows a session-length effect on the number of responses before initial absorption, though median absorption was very rapid even with the longer session length so the session length effect was small and not statistically significant. The other models show no such effect, any differences in responses to absorption between the two session lengths being similar to those seen in Table 3.

Table 5 shows the effect of introducing a 50% motivational carry-over. Again the connectionist and rule of thumb models show earlier absorption and a correspondingly better outcome in the short sessions than the other models. In the longer sessions, the rule of thumb model does better than the others, securing an outcome that is actually better than that of the dynamic programming algorithm under the same conditions. Several models show an apparent session-length effect. Fig. 1 shows distributions of absorption points from a separate set of simulations for the Wills–McLaren connectionist model, the ε-greedy and the rule of thumb models in long and short sessions, confirming that the difference of median absorption points is not due to curtailment of the distributions.

4. Discussion

All the models simulated were successful in the sense that, with the environmental parameters used by Krebs et al. (1978), they consistently found the better arm of the two-armed bandit in almost every simulated session. The environmental parameters used by Dow and Lea (1987b) were more challenging, and in particular the connectionist model of Wills and McLaren (1997) failed to converge with these lower reinforcement probabilities.

With the Krebs et al. (1978) environment, all models showed a clear payoff-differential effect, converging more quickly and finding the better arm more often with the larger differential. This is consistent with experimental results on many species, including Krebs et al.'s great tits, sticklebacks (Thomas et al., 1985) and bumble bees (Keasar et al., 2002). When session length was varied, all models coped almost equally well with the long (600-response)

Table 3
Outcomes of model simulations at two session lengths, under conditions like those of Krebs et al. (2008). Payoff probabilities on the two arms were 0.125 and 0.375. Entries are mean results from 200 simulation runs. p(abs) is the probability of absorption; $p(C)$ the probability of initial absorption on the objectively preferable arm; R(abs) the median responses to initial absorption; $\rho(C)$ the mean Herrnstein et al. ρ towards the objectively preferable arm; and $p(Rf)$ the overall reward probability.

Condition	Model	Parameters		p(abs)	$p(C)$	R(abs)	ρ(C)	$p(Rf)$
Session length 150 responses	Backwards dynamic programming			1.00	0.94	2	0.91	0.349
	Dynamic matching	α 0.06		0.86	1.00	24	0.86	0.336
	Rescorla–Wagner	β 0.55		0.81	0.99	19	0.79	0.337
	Connectionist	β 0.08	d 0.20	0.87	0.98	2	0.86	0.350
	ε-greedy	α 0.07	ε 0.02	0.71	0.99	28	0.76	0.339
	Rule of thumb	α 0.035	L 0.08	0.99	1.00	2	0.95	0.355
Session length 600 responses	Backwards dynamic programming			1.00	1.00	6	0.98	0.367
	Dynamic matching	α 0.06		1.00	1.00	25	0.96	0.362
	Rescorla–Wagner	β 0.55		1.00	1.00	21	0.95	0.368
	Connectionist	β 0.08	d 0.20	1.00	1.00	3	0.97	0.368
	ε-greedy	α 0.07	ε 0.02	1.00	1.00	20.5	0.95	0.368
	Rule of thumb	α 0.035	L 0.08	1.00	0.99	2	0.98	0.369

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

Table 4
Outcomes of model simulations at two session lengths, under conditions like those of Krebs et al. (2008), assuming 50% memory of reinforcement rates from the previous session. Payoff probabilities on the two arms were 0.125 and 0.375. Entries are mean results from 200 simulation runs. Model parameters were chosen to give the best performance for a session length of 300, with memory present. p(abs) is the probability of absorption; $p(C)$ the probability of initial absorption on the objectively preferable arm; R(abs) the median responses to initial absorption; $\rho(C)$ the mean Herrnstein et al. ρ towards the objectively preferable arm; and $p(Rf)$ the overall reward probability.

Condition	Model	Parameters		p(abs)	$p(C)$	R(abs)	ρ(C)	$p(Rf)$
Session	Backwards dynamic programming			0.99	0.96	1	0.93	0.345
length 150	Dynamic matching	α 0.065		0.93	0.98	14	0.87	0.341
responses	Rescorla–Wagner	β 0.50		0.65	0.94	21.5	0.77	0.325
	Connectionist	β 0.08	d 0.15	0.90	1.00	3	0.85	0.348
	ε-greedy	α 0.075	ε 0.025	0.88	0.99	6.5	0.79	0.348
	Rule of thumb	α 0.05	L 0.075	1.00	0.93	2	0.90	0.348
Session	Backwards dynamic programming			1.00	0.99	2.5	0.98	0.366
length 600	Dynamic matching	α 0.065		1.00	0.99	19	0.96	0.361
responses	Rescorla–Wagner	β 0.50		1.00	1.00	24.5	0.95	0.358
	Connectionist	β 0.08	d 0.15	1.00	1.00	2	0.97	0.370
	ε-greedy	α 0.075	ε 0.025	1.00	1.00	9	0.95	0.370
	Rule of thumb	α 0.05	L 0.075	1.00	0.99	1	0.98	0.371

session, and all approached or even exceeded the performance of the backwards dynamic programming algorithm. However, in the short (150-response) sessions, differences were much more apparent, with the connectionist and rule of thumb models outperforming the others (Table 3). There were few differences between the models with a 300-response session (Tables 1 and 2), so it appears that a session of that length is asymptotically long with the reward probabilities we used.

None of the models were able to respond to the variation in session length by producing quicker absorption in a shorter session (Table 3), except of course for the dynamic programming model. Simulating runs of sessions of constant length, and allowing the system to remember the reinforcement rates at the end of the previous session, did not enable any of the other models to respond to session length (Table 4). Allowing motivational carry-over in blocks of sessions of constant length, however, enabled some models to produce markedly slower absorption in longer sessions, when motivation would be lower (Table 5, Fig. 1). This effect was clearly present for the connectionist, ε-greedy and rule of thumb models – and the first and last of these were the models that were most successful overall. Shettleworth and Plowright (1989) suggested that memory and motivation carry-over effects, between them, might account for the session-length effects reported by Dow and Lea (1987b) and Kacelnik (1979); the present results suggest that, if learning in birds proceeds in anything like any of the ways we have modelled, it is motivational carry-over which is the more important.

Which of these models provides the most plausible account of the behaviour of actual foraging animals? Backward dynamic programming can be ruled out as a mechanism that could be used by a great tit or pigeon; the question is whether other models will produce equivalent behaviours. None of the other models we considered place unrealistic demands on the learning capacities of a bird. The rule of thumb model, arguably the most successful across the board, is open to the objection that it assumes that the animal knows that it is facing a two-armed bandit; it might not be a good strategy to adopt in a different kind of environment, whereas all the other models are quite general. Only an animal whose environment regularly allowed it highly limited access to several different, mutually exclusive sources of food could be expected to evolve such a learning mechanism – but in a world filled with competitors and predators, such environments are not that unusual. Among the more general models, the most successful was the connectionist model based on Wills and McLaren (1997), but it clearly requires modification to cope with low reward probabilities.

It is not clear how important it is that a model should be able to predict a session-length effect in a two-armed bandit situation, since the evidence that animals show such an effect is limited (Dow and Lea, 1987b; Kacelnik, 1979) and open to alternative interpretations (Shettleworth and Plowright, 1989). At a minimum, however, the present results suggest that foragers need to be able to respond differently to patches where their access is sharply limited, and those where they have more or less unlimited time – or at any rate, as much time as they can use. It is quite common for animals with a rich food supply to have to pause because their oesophageal or gut capacity has been reached, as has been demonstrated for waders (Kersten and Visser, 1996; Zwarts and Dirksen, 1990). Furthermore, it is in situations where time of access is limited that it seems to matter most how an animal learns. In extended sessions, any of the models we tested would secure almost as high an overall reward rate as the backwards dynamic programming algorithm. In shorter

Table 5
Outcomes of model simulations at two session lengths, under conditions like those of Krebs et al. (2008), assuming 50% carryover of motivational state from the end of the previous session. Payoff probabilities on the two arms were 0.125 and 0.375. Entries are mean results from 200 simulation runs. p(abs) is the probability of absorption; $p(C)$ the probability of initial absorption on the objectively preferable arm; R(abs) the median responses to initial absorption; $\rho(C)$ the mean Herrnstein et al. ρ towards the objectively preferable arm; and $p(Rf)$ the overall reward probability. Model parameters were chosen to give the best performance for a session length of 300, with motivational carry-over assumed.

Condition	Model	Parameters		p(abs)	$p(C)$	R(abs)	ρ(C)	$p(Rf)$
Session	Backwards dynamic programming			1.00	0.96	2	0.93	0.351
length 150	Dynamic matching	α 0.09		0.98	0.95	7	0.88	0.340
responses	Rescorla–Wagner	β 0.55		0.78	0.98	21	0.79	0.344
	Connectionist	β 0.10	d 0.45	0.94	1.00	1	0.90	0.358
	ε-greedy	α 0.08	ε 0.03	0.82	0.99	14.5	0.79	0.349
	Rule of thumb	α 0.04	L 0.08	1.00	0.96	1	0.93	0.351
Session	Backwards dynamic programming			1.00	0.99	2.5	0.98	0.365
length 600	Dynamic matching	α 0.09		1.00	0.99	12	0.97	0.365
responses	Rescorla–Wagner	β 0.55		1.00	1.00	20	0.95	0.365
	Connectionist	β 0.10	d 0.45	0.99	1.00	10	0.92	0.365
	ε-greedy	α 0.08	ε 0.03	1.00	1.00	39	0.95	0.364
	Rule of thumb	α 0.04	L 0.08	1.00	1.00	7	0.99	0.372

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

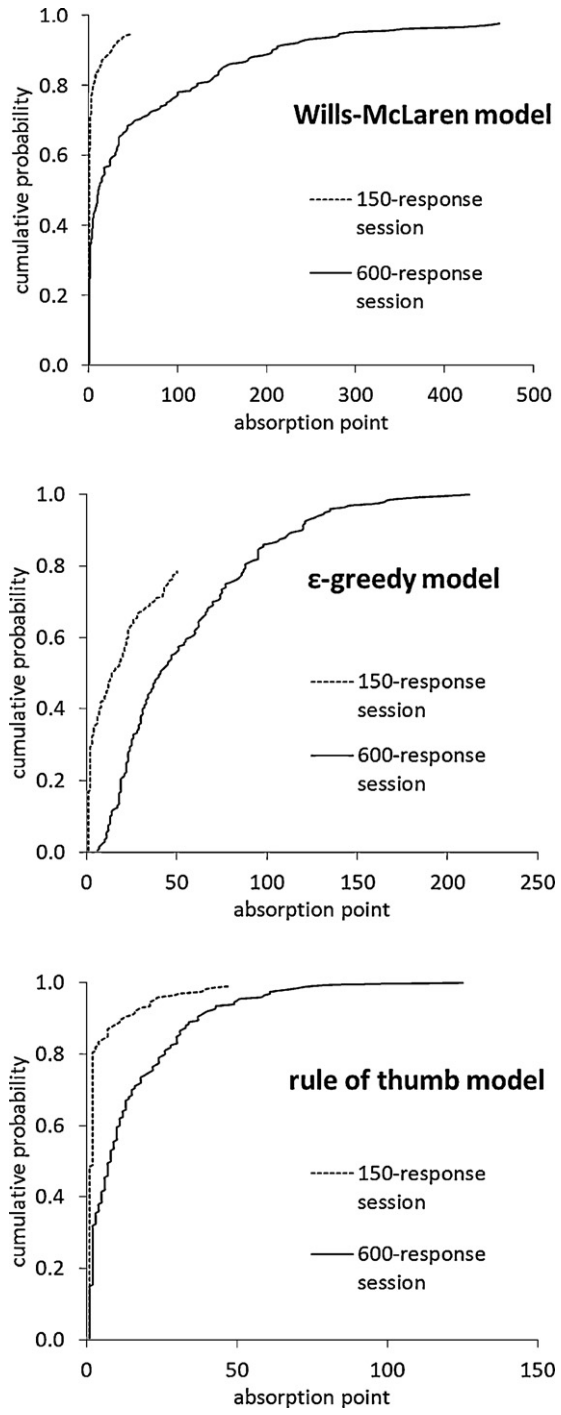

Fig. 1. Distribution of number of trials to initial absorption in short and long sessions under conditions similar to those of Krebs et al. (1978), as simulated by various models, with motivational state carried over from the end of the previous session.

sessions, the differences between the learning models, and the difference between even the best of them and the performance of the backwards dynamic programming algorithm, was greater.

In any of the models, the variation in learning rate required to adapt to different session lengths could be achieved by varying one or more parameters, but this poses the question of how that could happen in a natural situation. It is possible that the parameter variation arises from a learning process, as Killeen (1984) suggests; however the present simulations have shown that it can arise naturally if longer sessions are, as it seems they often would be in natural conditions, associated with a lower level of food motivation. This means that if experiments are carried out to test for capacities to respond to varying session length cognitively, it will always be important to try to neutralise the impacts of session length on food deprivation, as indeed Krebs et al. (1978) sought to do. It would be preferable to separate sessions by relatively long intervals, to allow motivation to return to a baseline state, rather than merely giving compensatory food after longer sessions, since in principle animals might learn that such food was coming. However one should not think of motivational shifts as merely an artefact; under natural conditions, variations in deprivation might provide a simple way of making adaptive adjustments to the balance between exploring and exploiting.

It does appear that relatively simple principles of associative learning are sufficient to produce even quite sophisticated behavioural adjustments, ensuring that foraging in a range of different environments is, if not optimal, at least so close to the optimal in outcome that the selective pressure to evolve more complex cognitive processes must be quite slight.

We have not, of course, explored all possible models of the way in which animals adjust their foraging behaviour to changing environments. The elaboration of new learning models continues (e.g. Dawson et al., 2009; Gross et al., 2008); it would be an impossible task to test them all. Some of the most interesting recent developments in the study of human decision making have been concerned with situations where gaining information rapidly is of the essence of the problem (e.g. Gigerenzer and Goldstein, 1996). Some of this work has been specifically concerned with explore/exploit conflicts where people must determine key parameters from experience (Hills and Hertwig, 2010), and this line of investigation is offering yet more new models for testing (e.g. Erev and Barron, 2005). The two-armed bandit is only one special case of the more general problem of how animals can keep track of varying and changing prey densities. In some respects the two-armed bandit may be an unusual situation, in that the better option does not change within a foraging session; in many natural situations, this may not be true, and therefore maintaining some level of exploration while exploiting may be preferable to complete absorption. Furthermore, the learning of prey densities is only one of many ways in which information gain is required for successful foraging (Stephens, 2007), and foraging is only one of many situations where information is crucial in ecology (Dall et al., 2005).

Although the present comparison has thus been concerned with a restricted field, it has been sufficient to suggest the kinds of model that might do best, and that might most repay further exploration. The two most interesting possibilities are Wills-McLaren model and the rule of thumb model. The Wills-McLaren model has the advantage that it could be applied without modification to other situations where animals need to acquire both food and information simultaneously. This model differs from the others more in the performance principle than in the learning principle it employs. The rule of thumb model, on the other hand, requires the animal in some sense to understand what kind of situation it is in. That might have seemed an inappropriate demand when Krebs et al. (1978) first drew the ideas of optimal sampling to the attention of animal behaviour researchers. After a further three decades of research into animal cognition, however, it does not seem so far-fetched. At least some species of animal seem to be capable of learning the rules of a situation to which they are repeatedly exposed. A recent example is in the work of Rayburn-Reeves et al. (2011, in press) on within-session reversals; they find marked differences in the apparent ability of pigeons and rats to respond appropriately to this situation, which is consistent with the long-held view that the capacity to learn the rules of a situation is one of the ways in which animal intelligence varies across taxa (cf. Bitterman et al., 1958; Mackintosh et al., 1985). We might therefore expect that this more specific kind of model is required to get a good fit to the behaviour

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004

of those animals that have shown such a capacity for understanding the structure of a choice situation.

Acknowledgements

An earlier version of this paper was delivered as part of a symposium in honour of the contributions of Alex Kacelnik at the Conference on Comparative Cognition, Melbourne, FL, March 2011. We are grateful to Alex Kacelnik, Elliott Ludvig, and to others present at the conference for their comments on the paper.

References

Abarca, N., Fantino, E., 1982. Choice and foraging. J. Exper. Anal. Behav. 38, 117–123.
Aparicio, C.F., Baum, W.M., 2009. Dynamics of choice: relative rate and amount affect local preference at three different time scales. J. Exp. Anal. Behav. 91, 293–317.
Bellman, R., 1956. A problem in the sequential design of experiments. Sankhya 16, 221–229.
Bitterman, M.E., Wodinsky, J., Candland, D.K., 1958. Some comparative psychology. Am. J. Psychol. 71, 94–110.
Bradt, R.N., Karlin, S., 1956. On the design and comparison of certain dichotomous experiments. Ann. Math. Stat. 27, 390–409.
Bush, R.R., Mosteller, F., 1951. A mathematical model for simple learning. Psychol. Rev. 58, 313–323.
Charnov, E.L., Orians, G.H., 1973. Optimal Foraging: Some Theoretical Explorations. Author, Vancouver, BC.
Clayton, N.S., Dickinson, A., 1998. Episodic-like memory during cache recovery by scrub jays. Nature 395, 272–274.
Cohen, J.D., McClure, S.M., Yu, A.J., 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Proc. Royal Soc. B 362, 933–942.
Collier, G.H., Rovee-Collier, C.K., 1981. A comparative analysis of optimal foraging behavior: laboratory simulations. In: Kamil, A.C., Sargent, T.D. (Eds.), Foraging Behavior. Garland, New York, pp. 39–76.
Colton, T., 1963. A model for selecting one of two medical treatments. J. Am. Stat. Assoc. 58, 388–400.
Dall, S.R.X., Giraldeau, L.A., Olsson, O., McNamara, J.M., Stephens, D.W., 2005. Information and its use by animals in evolutionary ecology. Trends Ecol. Evol. 20, 187–193.
Dawson, M.R.W., Dupuis, B., Spetch, M.L., Kelly, D.M., 2009. Simple artificial neural networks that match probability and exploit and explore when confronting a multiarmed bandit. IEEE Trans. Neural Networks 20, 1368–1371.
Dow, S.M., Lea, S.E.G., 1987a. Foraging in a changing environment: simulations in the operant laboratory. In: Commons, M.L., Kacelnik, A., Shettleworth, S.J. (Eds.), Quantitative Analyses of Behavior, vol. 6: Foraging. Erlbaum, Hillsdale, NJ, pp. 89–113.
Dow, S.M., Lea, S.E.G., 1987b. Sampling of schedule parameters by pigeons: tests of optimizing theory. Anim. Behav. 35, 102–114.
Erev, I., Barron, G., 2005. On adaptation, maximization, and reinforcement learning among cognitive strategies. Psychol. Rev. 112, 912–931.
Findley, J.D., 1958. Preference and switching under concurrent scheduling. J. Exp. Anal. Behav. 1, 123–144.
Fisher, J., Hinde, R.A., 1949. The opening of milk bottles by birds. Br. Birds 42, 347–357.
Fox, B.L., 1974. Finite horizon behavior of policies for two-arm bandits. J. Am. Stat. Assoc. 69, 963–965.
Gigerenzer, G., Goldstein, D.G., 1996. Reasoning the fast and frugal way: models of bounded rationality. Psychol. Rev. 103, 650–669.
Gross, R., Houston, A.I., Collins, E.J., McNamara, J.M., Dechaume-Moncharmont, F.X., Franks, N.R., 2008. Simple learning rules to cope with changing environments. J. Royal Soc. Interface 5, 1193–1202.
Herrnstein, R.J., 1961. Relative and absolute strengths of response as a function of frequency of reinforcement. J. Exp. Anal. Behav. 4, 267–272.
Herrnstein, R.J., 1970. On the law of effect. J. Exp. Anal. Behav. 13, 243–266.
Herrnstein, R.J., Loveland, D.H., 1975. Maximizing and matching on concurrent ratio schedules. J. Exp. Anal. Behav. 24, 107–116.
Herrnstein, R.J., Loveland, D.H., Cable, C., 1976. Natural concepts in pigeons. J. Exp. Psychol. Anim. Behav. Proc. 2, 285–302.
Hills, T.T., Hertwig, R., 2010. Information search in decisions from experience: do our patterns of sampling foreshadow our decisions? Psychol. Sci. 21, 1787–1792.
Houston, A.I., Kacelnik, A., McNamara, J., 1982. Some learning rules for acquiring information. In: McFarland, D.J. (Ed.), Functional Ontogeny. Pitman, Boston, pp. 140–191.
Kacelnik, A., 1979. Studies of foraging behaviour and time budgeting in great tits (*Parus major*). Unpublished DPhil thesis, University of Oxford.
Kamil, A.C., Balda, R.P., 1985. Cache recovery and spatial memory in Clark's nutcrackers (*Nucifraga columbiana*). J. Exp. Psychol. Anim. Behav. Proc. 11, 95–111.
Keasar, T., Rashkovich, E., Cohen, D., Shmida, A., 2002. Bees in two-armed bandit situations: foraging choices and possible decision mechanisms. Behav. Ecol. 13, 757–765.
Kersten, M., Visser, W., 1996. The rate of food processing in the oystercatcher: food intake and energy expenditure constrained by a digestive bottleneck. Funct. Ecol. 10, 440–448.
Killeen, P.R., 1984. Incentive theory: III. Adaptive clocks. Ann. N. Y. Acad. Sci. 423, 515–527.
Krebs, J.R., Kacelnik, A., Taylor, P., 1978. Test of optimal sampling by foraging great tits. Nature 275, 27–31.
Lea, S.E.G., 1979. Foraging and reinforcement schedules in the pigeon: optimal and non-optimal aspects of choice. Anim. Behav. 27, 875–886.
Lea, S.E.G., 1985. Operant psychology and ethology: failures and successes in interdisciplinary interaction. In: Lowe, C.F., Richelle, M., Blackman, D.E., Bradshaw, C.M. (Eds.), Behaviour Analysis and Contemporary Psychology. Erlbaum, London, pp. 43–51.
Lea, S.E.G., Dow, S.M., 1984. The integration of reinforcements over time. Ann. N. Y. Acad. Sci. 423, 269–277.
MacArthur, R.H., Pianka, E.R., 1966. On the optimal use of a patchy environment. Am. Nat. 100, 603–609.
Mackintosh, N.J., Wilson, B., Boakes, R.A., 1985. Differences in mechanisms of intelligence among vertebrates. Phil. Trans. Royal Soc. B: Biol. Sci. 308, 53–65.
Mazur, J.E., 1992. Choice behavior in transition: development of preference with ratio and interval schedules. J. Exp. Psychol. Anim. Behav. Proc. 18, 364–378.
McNamara, J., 1982. Optimal patch use in a stochastic environment. Theor. Popul. Biol. 21, 269–288.
McNamara, J.M., 1996. Risk-prone behavior under rules which have evolved in a changing environment. Am. Zool. 36, 484–495.
McNamara, J.M., Houston, A.I., 1985. Optimal foraging and learning. J. Theor. Biol. 117, 231–249.
Myerson, J., Miezin, F.M., 1980. The kinetics of choice: an operant systems analysis. Psychol. Rev. 87, 160–174.
Oaten, A., 1977. Optimal foraging in patches: case for stochasticity. Theor. Popul. Biol. 12, 263–285.
Plowright, C.M.S., Plowright, R.C., 1987. Oversampling by great tits: a critique of Krebs, Kacelnik, and Taylor's (1978) test of optimal sampling by great tits. Can. J. Zool. 65, 1282–1283.
Plowright, C.M.S., Shettleworth, S.J., 1991. Time horizon and choice by pigeons in a prey-selection task. Anim. Learn. Behav. 19, 103–112.
Rayburn-Reeves, R.M., Molet, M., Zentall, T.R., 2011. Simultaneous discrimination reversal learning in pigeons and humans: anticipatory and perseverative errors. Learn. Behav. 39, 125–137.
Rayburn-Reeves, R.M., Stagner, J.P., Kirk, C.R., Zentall, T.R., in press. Mid-session simultaneous discrimination reversal differences between rats and pigeons. J. Comp. Psychol.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York, pp. 64–99.
Shettleworth, S.J., Plowright, C.M.S., 1989. Time horizons of pigeons on a two-armed bandit. Anim. Behav. 37, 610–623.
Simon, H.A., 1956. Rational choice and the structure of the environment. Psychol. Rev. 63, 129–138.
Stephens, D.W., 2007. Models of information use. In: Stephens, D.W., Brown, J.S., Ydenberg, R.C. (Eds.), Foraging. Chicago University Press, Chicago, pp. 31–58.
Stephens, D.W., Krebs, J.R., 1986. Foraging Theory. Princeton University Press, Princeton, NJ.
Stephens, D.W., Brown, J.S., Ydenberg, R.C. (Eds.), 2007. Foraging. Chicago University Press, Chicago.
Sutton, R.S., Barto, A.B., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Thomas, G., Kacelnik, A., Van Der Meulen, J., 1985. The three-spined stickleback and the two-armed bandit. Behaviour 93, 227–240.
Thompson, W.R., 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25, 285–294.
Tolman, E.C., 1955. Principles of performance. Psychol. Rev. 62, 315–326.
Ward, D., 1992. The role of satisficing in foraging theory. Oikos 63, 312–317.
Wills, A.J., McLaren, I.P.L., 1997. A connectionist account of differences in gradient after discriminative and non-discriminative training. Q. J. Exp. Psychol. 50, 607–630.
Zwarts, L., Dirksen, S., 1990. Digestive bottleneck limits the increase in food intake of whimbrels preparing to migrate from the Banc d'Arguin, Mauritania. Ardea 78, 257–278.

Please cite this article in press as: Lea, S.E.G., et al., The cognitive mechanisms of optimal sampling. Behav. Process. (2011), doi:10.1016/j.beproc.2011.10.004