INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Barendregt NW, Gold JI, Josic K, Kilpatrick ZP. 2026 Information-seeking decision strategies mitigate risk in dynamic, uncertain environments. *J. R. Soc. Interface* **23**: 20251308.
https://doi.org/10.1098/rsif.2025.1308

Received: 15 December 2025
Accepted: 23 February 2026

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
biomathematics

Keywords:
sequential decision-making, explore–exploit trade-offs, information seeking, partially observable Markov decision processes, foraging theory

Author for correspondence:
Zachary P. Kilpatrick
e-mails: zpkilpat@colorado.edu; zacharykilpatrick@gmail.com

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8400720.

THE ROYAL SOCIETY PUBLISHING

Information-seeking decision strategies mitigate risk in dynamic, uncertain environments

Nicholas W. Barendregt[1], Joshua I. Gold[2], Kresimir Josic[3] and Zachary P. Kilpatrick[1]

[1]Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO, USA
[2]Department of Neuroscience, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[3]Department of Mathematics, University of Houston, Houston, TX, USA

NWB, 0000-0002-3268-9426; KJ, 0000-0002-1975-3913; ZPK, 0000-0002-2835-9416

To survive in dynamic and uncertain environments, individuals must develop effective decision strategies that balance information gathering and decision commitment. Models of such strategies often prioritize either optimizing tangible payoffs, like reward rate, or gathering information to support a diversity of (possibly unknown) objectives. However, our understanding of the relative merits of these two approaches remains incomplete, in part because direct comparisons have been limited to idealized, static environments that lack the dynamic complexity of the real world. Here, we compare the performance of normative reward- and information-seeking strategies in a dynamic foraging task. Both strategies show similar transitions between exploratory and exploitative behaviours as environmental uncertainty changes. However, we find disparities in the actions they take, resulting in meaningful performance differences: whereas reward-seeking strategies generate slightly more reward on average, information-seeking strategies provide more consistent and predictable outcomes. Our findings support the adaptive value of information-seeking behaviours that can mitigate risk with minimal reward loss.

1. Introduction

Effective behavioural strategies are adaptive [1,2]. This adaptivity is often cast in terms of normative frameworks, which use objective functions that represent preferences in both actions and outcomes [3–7]. However, the objectives that drive organisms' natural behaviours can be complex and remain poorly understood. One challenge is that strategies with different objectives can produce similar behaviours. For example, it can be difficult to distinguish a strategy that aims to maximize a particular outcome, like reward rate, from one that aims to maximize information that is useful for achieving that outcome. This challenge is particularly acute in naturalistic settings, where there is often a complex relationship between particular outcomes and information that is relevant to those outcomes.

This interplay between rewards and information is evident in strategies for resource foraging. Foraging typically requires sequential decisions that alternate between choosing to gather evidence (e.g. about local resource availability) and using this information to take actions (e.g. keep exploiting local resources or explore more broadly for other resources) [8–10]. Many theoretical accounts of foraging have focused on the optimization of a reward-based objective. A notable example is the pervasive marginal value theorem (MVT), which is based on optimizing the average rate of reward (e.g. food/resource) intake [11,12]. These reward-optimal approaches have provided key insights into foraging behaviours [10,13–16]. However, debates continue about whether animals behave

in a reward-rational manner [17] or even capable of implementing normative reward-maximizing strategies [18,19].

The relative value of information versus reward depends on environmental uncertainty, which is central in dynamic foraging. Empirical studies show that animals sometimes adjust search behaviour to increase information about resource distributions [20], and computational models similarly find that information-seeking can be advantageous when signals are noisy or intermittent [21]. Evidence-accumulation models of foraging can outperform purely reward-driven rules under uncertainty [22], and computational work on olfactory search shows that information-directed sampling improves reliability in turbulent environments [21,23]. These results suggest that information-seeking can offer reliability advantages under unreliable feedback, motivating our comparison with reward-seeking strategies.

Alternatively, animals could approximate reward optimization by appealing to other intermediate objectives that serve this outcome in flexible ways [24]. One class of such strategies is information maximization, which drives foragers to become as knowledgeable as they can about the environment. These strategies do not optimize explicitly for a tangible outcome, such as immediate reward intake, but can indirectly produce reliably favourable outcomes, accommodate environmental changes and uncertainty, and be less error-prone than fine-tuned normative strategies. One example of an information-maximizing strategy is infotaxis, in which individuals move to locations they expect to provide the largest reduction in their uncertainty about the environment [21]. This strategy is particularly effective for source-tracking tasks [25,26], producing behavioural trajectories that qualitatively replicate experimental findings [20,27]. However, direct comparisons between information-seeking strategies like infotaxis and reward-seeking strategies like those based on the MVT are rare [28] and are typically restricted to search in static environments. How these model classes compare in more naturalistic (e.g. dynamic, sequential and uncertain) settings remains unknown.

Although related frameworks, most notably active inference, provide a unified account of exploratory behaviour by combining instrumental (e.g. expected reward) and epistemic (e.g. expected information gain) motives into a single expected free energy objective [29–31], they typically treat sampling and reward acquisition as consequences of the same action, as in multi-armed bandits and related formulations. This structure does not capture the distinction in our task between gathering evidence without committing to an outcome and sampling probabilistic reward only at commitment. As a result, existing work does not separately characterize the optimal reward-maximizing and information-maximizing policies we study here or map how their differences depend on environmental volatility and feedback reliability. By isolating these components, our approach identifies where the consequences of pursuing the two objectives diverge and how those divergences shape explore–exploit behaviour and risk profiles in dynamic environments. This recursive evaluation is conceptually related to the backward induction used in active inference's sophisticated inference formulation [29], although here the two terms are treated as fully separable utilities.

We compared how sequential versions of normative reward-maximizing ('rewardmax') and infotaxis ('infomax') strategies perform in dynamic foraging environments. Simulated agents performed a sequential, two-alternative foraging task that included manipulations of both the probability that the correct alternative changes between trials and the probability that choosing the correct alternative yields a reward [32]. Our results show that both models predict qualitatively similar behaviours as the environmental uncertainty changes, recreating the explore–exploit trade-off observed in many classic foraging studies. As expected, the rewardmax model maximizes average reward. However, the infomax model obtains reward more reliably than the rewardmax model, thereby mitigating risk. This variance reduction is analogous to bet-hedging, in which organisms trade short-term gains for lower variability in long-term outcomes [33]. This reliability is consistent across both environmental and model perturbations. Together, these results illustrate an important advantage of information maximization as a foraging strategy in naturalistic settings and point to relevant task designs where these two strategies can be further disentangled.

2. Methods

The rewardmax and infomax models each use the same (Bayesian) evidence-gathering process to update beliefs, obtained from the posterior probabilities about the environment. The models differ in terms of the rules they use to terminate evidence gathering and commit to a choice. Below, we first describe a common evidence-gathering process evolving within and across trials and then describe the different commitment rules for each model that we obtain via dynamic programming.

2.1. Sequential task development and evidence integration

Our sequential task can be interpreted as a simplified two-patch foraging problem. A forager may gather non-rewarding sensory cues, such as brief visual or olfactory inspection, that inform patch quality without yielding food. A commitment corresponds to exploiting one patch, producing probabilistic reward. Environmental volatility represents changes in patch quality across visits, whereas feedback reliability corresponds to the stochastic success of extracting food once a patch is chosen. This ecological interpretation grounds the task in a common foraging scenario while still allowing exact characterization of optimal policies. Formally, this task defines a finite-horizon partially observable Markov decision process (POMDP) [34,35], in which observation and commitment actions are distinct and update the agent's beliefs through separate channels.

We consider a sequence of foraging decisions in an environment that, for each decision, takes on one of two possible hidden states, $s^i \in \{s_+, s_-\}$, where the superscript, i, denotes the ith state in the sequence. Each state determines which of two options is more likely to provide a reward to a foraging agent. For instance, the state could represent which one of two resource patches is more likely to yield edible plants [36,37]. To infer the state of the environment, the agent makes independent and identically distributed observations, ξ_j^i, each of which follows a state-dependent distribution, $f_\pm\left(\xi_j^i\right) = f\left(\xi_j^i | s^i = s_\pm\right)$. This set-up corresponds to a forager rapidly probing two distant patches whose underlying yield probabilities (one high, one low) change only slowly, so

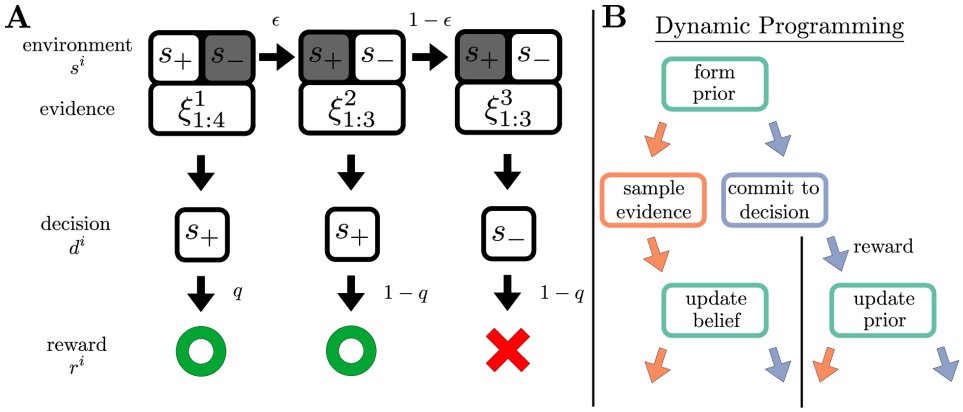

Figure 1. Dynamic foraging task and modelling approach. (A) Schematic of an example realization of the dynamic foraging task. Between decisions, the environment is in a constant, hidden state s^i that the agent infers from a sequence of observations, $\xi^i_{1:n}$. Once sufficiently confident, the agent makes a decision d^i and receives a probabilistic reward r^i with reliability q. The environment then changes state with probability ϵ, and the process begins again. (B) Schematic of the dynamic-programming approach for objective-maximizing behaviour. At each time step, the agent uses prior information to calculate the expected utility of sampling evidence and committing to a decision. The agent then takes the action that maximizes expected utility (either reward or information) and updates their belief based on the information gained from the environment (in the case of sampling) or from probabilistic reward (in the case of commitment).

each quick probe provides a new conditionally independent sample. In the future, models with temporal correlations or depletion effects could be considered to relax this assumption. A Bayesian agent deduces which state is more likely after n observations by computing the log-likelihood ratio (LLR), y^i_n, which we will also refer to as the agent's 'belief'. The best possible estimate of the state from the sequence of noisy observations $\xi^i_{1:n}$ [4,38,39] is determined by computing

$$y^i_n = \ln \frac{\Pr\left(s^i = s_+ | \xi^i_{1:n}\right)}{\Pr\left(s^i = s_- | \xi^i_{1:n}\right)} = \ln \frac{f_+\left(\xi^i_n\right)}{f_-\left(\xi^i_n\right)} + y^i_{n-1}. \tag{2.1}$$

To obtain equation (2.1), we used the conditional independence of the observations $\xi^i_{1:n}$ to obtain an iterative update rule for the LLR. This update rule involves a prior belief, y^i_0, carried forward from the previous trial when $i = 2, 3, ...$, as described below. When $i = 1$, we assume $y^1_0 = 0$; both states are considered equally likely to be drawn as the initial state s^1.

An agent's decision $d^i \in \{s_+, s_-\}$ about the hidden state s^i triggers two events. First, the agent receives a probabilistic reward or punishment $r^i \in \{\circ, \times\}$ based on their choice, determined by a Bernoulli random variable conditioned on the correctness of the decision (correct, $d^i = s^i$, or incorrect $d^i \neq s^i$) [32]

$$\begin{aligned} \Pr\left(r^i = \circ | d^i = s^i\right) &= \alpha, \\ \Pr\left(r^i = \times | d^i \neq s^i\right) &= \beta. \end{aligned} \tag{2.2}$$

For convenience, we will take $\alpha = \beta = q$ throughout this work. Because q is the probability that the feedback matches the correctness of the agent's response, we can interpret this parameter as the reliability of patch resource yields. Second, once the agent makes a decision, the environment changes state with a fixed probability, ϵ [6]

$$\Pr\left(s^{i+1} = s_\pm | s^i = s_\mp\right) = \epsilon. \tag{2.3}$$

This change probability captures the timescale of temporal fluctuations present in foraging environments, such as temperature changes that can make some resource patches more bountiful, or the appearance of predators that reduce the chances of safe and successful foraging. Here, the environmental state reflects slowly varying background conditions (e.g. habitat quality) rather than immediate resource depletion, so state transitions are not directly triggered by the agent's choice. Models of depletion-driven dynamics would instead require asymmetric, choice-dependent transitions. After feedback is delivered and the state updates, the $(i+1)$th trial begins. The observer may continue to make observations, starting with ξ^{i+1}_1, until they make decision d^{i+1}. An example realization of this process is schematized in figure 1A.

Having defined the environmental feedback and transition processes in equations (2.2) and (2.3), we can now give the expression for the updated prior, y^{i+1}_0, after the ith decision. This prior depends on the belief y^i_n that determined decision d^i, as well as the ensuing probabilistic reward, r^i. Using the same representation of the belief as the LLR used in equation (2.1), we define the prior after the ith decision as

$$y^{i+1}_0\left(y^i_n, d^i, r^i\right) = \ln \frac{\Pr\left(s^{i+1} = s_+ | y^i_n, d^i, r^i\right)}{\Pr\left(s^{i+1} = s_- | y^i_n, d^i, r^i\right)}. \tag{2.4}$$

Using the Law of Total Probability and conditional independence, we obtain the following update rule:

$$y_0^{i+1}\left(y_n^i, d^i, r^i\right) = \text{sign}\left(d^i\right)\ln\begin{cases}\dfrac{(1-\epsilon)qe^{|y_n^i|}+\epsilon(1-q)}{\epsilon qe^{|y_n^i|}+(1-\epsilon)(1-q)}, & r^i = \circ\\ \dfrac{(1-\epsilon)(1-q)e^{|y_n^i|}+\epsilon q}{\epsilon(1-q)e^{|y_n^i|}+(1-\epsilon)q}, & r^i = \times\end{cases}, \tag{2.5}$$

where $\text{sign}\left(d^i\right) = \pm 1$ corresponds to $d^i = s_\pm$. Together, equations (2.1) and (2.5) determine a Bayesian agent's belief in our dynamic foraging task.

2.2. Commitment rules: rewardmax and infomax models

Having established the Bayesian agent's belief-update rules, we now define the different commitment rules for the rewardmax (reward-maximizing) and infomax (information-maximizing) strategies under a fixed time-step budget (N time steps). To construct these rules, we adopt a dynamic-programming perspective (figure 1B). Specifically, an agent with current belief y_n^i chooses between two actions: (i) make a new observation, ξ_{n+1}^i, and update the belief to y_{n+1}^i, while the environment stays in the same state; or (ii) commit to a decision, d^i, receive probabilistic feedback (reward) and update the belief to y_0^{i+1} based on the feedback, with the environmental state changing with probability ϵ. Both of these actions (sampling and commitment, respectively) have an associated time-step cost, defined below. The action-selection process repeats until the agent exhausts their budget of N time steps. To determine which action to take, the agent uses a belief-dependent utility function U, which depends on the agent's objective function, detailed below.

We focus on the two-state case for analytical tractability and to allow exact characterization of the optimal policies. Extending the task to more than two patches would introduce a higher-dimensional belief space and probably increase the divergence between rewardmax and infomax strategies.

Rewardmax commitment rule. To define the rewardmax utility function, we assume that the agent acts to optimize the average reward rate, ρ, over their whole time-step budget, which we define as

$$\rho = \frac{\langle R\rangle}{\langle T_t\rangle + \langle T_i\rangle}, \tag{2.6}$$

where $\langle R\rangle$ is the average reward, $\langle T_t\rangle$ is the average total time of a decision and $\langle T_i\rangle$ is the average inter-decision interval [3,5]. All averages are taken across the entire N time-step budget. The N total time steps in the agent's budget restrict their ability to forage (e.g. limited seasonal availability of foraging resources). Our problem is thus a finite-horizon optimization, and, as such, the denominator of equation (2.6) will always equal N. Thus, only the terms in the numerator need to be maximized.

To optimize the reward-based utility function, U_{rm}, we use Bellman's equation and backward induction on the time-step budget [3,40,41]. At the final time step, the agent has nothing to gain from making an observation. Thus, the utility function at time step N is determined by the utility of committing to a decision,

$$\begin{aligned} U_{\text{rm}}\left(p_n^i; N\right) &= \max\left\{U_{\text{rm}}^+\left(p_n^i; N\right), U_{\text{rm}}^-\left(p_n^i; N\right)\right\},\\ &= \max\begin{Bmatrix} p_n^i\left[qR_+ + (1-q)R_-\right] + \left(1-p_n^i\right)\left[(1-q)R_+ + qR_-\right],\\ \left(1-p_n^i\right)\left[qR_+ + (1-q)R_-\right] + p_n^i\left[(1-q)R_+ + qR_-\right]\end{Bmatrix}, \end{aligned} \tag{2.7}$$

where $U_{\text{rm}}^\pm$ is the utility of committing to the decision $d^i = s_\pm$, $p_n^i = \Pr(s^i = s_+|\xi_{1:n}^i) = \frac{1}{1+e^{-y_n^i}}$ is the probability the ith environmental state is s_+ given the sequence of observations $\xi_{1:n}^i$ (which we refer to as the state likelihood), and $R_\pm$ is the reward/punishment delivered if the feedback is $r^i = \pm$. Because the state likelihood, p_n^i, and the belief, y_n^i, are equivalent, we use the state likelihood in our utility function definitions for notational simplicity.

With the final utility function constructed, we can now perform backward induction. Consider time-step number $k < N$, and assume that the agent's belief is y_n^i. In this case, we must modify the commitment utility functions $U_{\text{rm}}^\pm$ to reflect the utility of committing, which triggers the transition of the environmental state from i to $(i+1)$. If the agent chooses to commit, the new belief will be y_0^{i+1}. For U_{rm}^+, the new utility is given by

$$\begin{aligned} U_{\text{rm}}^+\left(p_n^i; k\right) = p_n^i\Big[&q\left\{R_+ + \gamma U_{\text{rm}}\left(p_0^{i+1}\left(p_n^i, s_+, +\right); k+\tau_d\right)\right\}\\ &+(1-q)\left\{R_- + \gamma U_{\text{rm}}\left(p_0^{i+1}\left(p_n^i, s_+, -\right); k+\tau_d\right)\right\}\Big]\\ &+\left(1-p_n^i\right)\Big[(1-q)\left\{R_+ + \gamma U_{\text{rm}}\left(p_0^{i+1}\left(p_n^i, s_+, +\right); k+\tau_d\right)\right\}\\ &+q\left\{R_- + \gamma U_{\text{rm}}\left(p_0^{i+1}\left(p_n^i, s_+, -\right)\right)\right\}\Big], \end{aligned} \tag{2.8}$$

where $\tau_d > 0$ is the time-step cost of *commitment* (decision execution), representing the time needed to travel to a patch and exploit it (e.g. handle and consume a resource), during which no additional evidence can be gathered. The parameter $\gamma \in [0,1]$ is a

discounting term on future utility: smaller values place little value on delayed gains, while $\gamma = 1$ corresponds to valuing future and immediate rewards equally. As $\gamma \to 0$, the model converges to a greedy (i.e. single-action optimal) strategy that does not take possible future utility gains into account. We can obtain an expression for U^{-}_{rm} following the same reasoning.

We must also introduce the utility of sampling an additional piece of environmental evidence U^{s}_{rm}, which is equal to the average utility obtained via sampling [3,5]

$$U^{s}_{\mathrm{rm}}\left(p^{i}_{n};k\right) = \gamma\left\langle U_{\mathrm{rm}}\left(p^{i}_{n+1};k+\tau_{s}\right)|p^{i}_{n}\right\rangle_{p^{i}_{n+1}} = \gamma\int_{0}^{1} U_{\mathrm{rm}}\left(p^{i}_{n+1};k+\tau_{s}\right) f_{p}\left(p^{i}_{n+1}|p^{i}_{n}\right)\, dp^{i}_{n+1}, \tag{2.9}$$

where $\tau_s > 0$ is the time-step cost of *sampling* (evidence gathering), representing the time needed to obtain an additional non-rewarding cue about patch quality (e.g. brief visual/olfactory inspection or probing), during which no reward is collected, and f_p is the conditional distribution of obtaining the future state likelihood p^{i}_{n+1} given the current likelihood p^{i}_{n}. Together, τ_s and τ_d capture opportunity costs associated with gathering information versus acting: τ_s governs the time spent acquiring cues without reward, and τ_d governs the time spent executing a choice to obtain probabilistic reward/feedback. These costs can differ across ecological contexts (e.g. quick inspection before approaching a patch: $\tau_s < \tau_d$; time-consuming scouting/probing: $\tau_s > \tau_d$). The discount factor γ reflects how strongly future utility is devalued, corresponding to impatience, shallow planning horizons and/or uncertainty about whether future decision steps will remain available in volatile environments.

This conditional distribution depends on the evidence-generating distribution $f_{\pm}(\xi)$ (for a Gaussian-distributed example, see [3]). For our task, we assume that the observations of the environmental state, ξ^{i}_{j}, are Bernoulli-distributed with parameter h so that

$$f_{\pm}(\xi) = \begin{cases} h, & \xi = \pm 1 \ \wedge \ s^{i} = s_{\pm} \\ 1-h, & \xi = \mp 1 \ \wedge \ s^{i} = s_{\pm} \end{cases}. \tag{2.10}$$

Given this evidence distribution, the conditional state-likelihood distribution f_p is

$$f_{p}\left(p^{i}_{n+1}|p^{i}_{n}\right) = \begin{cases} p^{i}_{n}h + \left(1-p^{i}_{n}\right)(1-h), & p^{i}_{n+1} = \dfrac{hp^{i}_{n}}{hp^{i}_{n} + (1-h)\left(1-p^{i}_{n}\right)} \\ p^{i}_{n}(1-h) + \left(1-p^{i}_{n}\right)h, & p^{i}_{n+1} = \dfrac{(1-h)p^{i}_{n}}{(1-h)p^{i}_{n} + h\left(1-p^{i}_{n}\right)} \end{cases}, \tag{2.11}$$

which can be obtained via a stochastic change of variables. This distribution describes the conditional probability that a subsequent sample will update the belief to a particular value, given the current belief.

Given equations (2.8) and (2.9), we can define the general rewardmax utility function as

$$U_{\mathrm{rm}}\left(p^{i}_{n};k\right) = \max\left\{U^{+}_{\mathrm{rm}}\left(p^{i}_{n};k\right), U^{-}_{\mathrm{rm}}\left(p^{i}_{n};k\right), U^{s}_{\mathrm{rm}}\left(p^{i}_{n};k\right)\right\}. \tag{2.12}$$

At every time step, an ideal (Bayesian) foraging agent calculates this utility function given their current belief (or, equivalently, the state likelihood) and takes the action that has maximal utility. Here, we assume the forager knows the statistical structure of the task—reward probabilities, Markov transition rates and the delays τ_s and τ_d—which is sufficient for evaluating expected utilities under this model.

Infomax commitment rule. As in the traditional infotaxis model [21], the infomax model assumes that the agent seeks to minimize the uncertainty about the environmental state as described by the state entropy [42]

$$\begin{aligned} H\left(p^{i}_{n}\right) &= -\sum_{s\in s_{\pm}} \Pr\left(s^{i}=s|\xi_{1:n}\right)\log_{2}\Pr\left(s^{i}=s|\xi_{1:n}\right) \\ &= -p^{i}_{n}\log_{2}\left(p^{i}_{n}\right) - \left(1-p^{i}_{n}\right)\log_{2}\left(1-p^{i}_{n}\right). \end{aligned} \tag{2.13}$$

Using equation (2.13), we can again apply Bellman's equation and backward induction, using uncertainty minimization as the objective function, to derive the infomax utility function. Unlike standard infotaxis, which is formulated for static or slowly varying environments to maximize expected information gain at each time step, our infomax utility function optimizes information gain over a future sequence of environmental states under the agent's budget of N time steps. This approach allows us to establish a proper comparison with the rewardmax model. By introducing discounting on future utility and setting $\gamma = 0$, we can recover the traditional infotaxis objective that accounts for only the next time step.

As with the rewardmax utility function, we start by considering the final time step in the agent's budget. At this point, future information gains are not possible (i.e. $U_{\mathrm{im}}\left(p^{i}_{n};N\right) = 0$). For time-step number $k < N$, we can calculate the utility $U^{+}_{\mathrm{im}}(p^{i}_{n};k)$ of a decision $d^{i} = s_{+}$, when the state likelihood is p^{i}_{n} based on the information obtained from the probabilistic feedback r^{i}. However, this information gain is not simply the state entropy of p^{i+1}_{0}, because even if feedback is completely unreliable and decoupled from the agent's performance (i.e. $q = 0.5$), the previous state (s^{i}) still partially predicts the next (s^{i+1}) as long as $\epsilon \neq 0.5$. Therefore, we

must normalize the information gain using the case of non-informative feedback, $q = 0.5$, as a baseline,

$$\begin{aligned} U^{+}_{\mathrm{im}}\left(p^i_n;k\right) = {} & H\left(p^i_n + \epsilon - 2p^i_n\epsilon\right) \\ & - \left[qp^i_n + (1-q)\left(1-p^i_n\right)\right] H\left(p^{i+1}_0\left(p^i_n, s_+, +\right)\right) \\ & - \left[q\left(1-p^i_n\right) + (1-q)p^i_n\right] H\left(p^{i+1}_0\left(p^i_n, s_+, -\right)\right) \\ & + p^i_n\left[q\gamma U_{\mathrm{im}}\left(p^{i+1}_0\left(p^i_n, s_+, +\right); k+\tau_d\right) + (1-q)\gamma U_{\mathrm{im}}\left(p^{i+1}_0\left(p^i_n, s_+, -\right); k+\tau_d\right)\right] \\ & + \left(1-p^i_n\right)\left[(1-q)\gamma U_{\mathrm{im}}\left(p^{i+1}_0\left(p^i_n, s_+, +\right); k+\tau_d\right) + q\gamma U_{\mathrm{im}}\left(p^{i+1}_0\left(p^i_n, s_+, -\right); k+\tau_d\right)\right], \end{aligned} \tag{2.14}$$

where the expression $p^i_n + \epsilon - 2p^i_n\epsilon$ is the simplified form of p^{i+1}_0 from equation (2.5) when $q = 0.5$. As with the reward-maximizing strategy, the utility U^{-}_{im} for a decision $d^i = s_-$ can be obtained in the same manner.

We then calculate the utility of making another observation as the reduction in uncertainty expected from updating the belief about the current state, s^i, based on that observation,

$$\begin{aligned} U^{s}_{\mathrm{im}}\left(p^i_n;k\right) &= H\left(p^i_n\right) - \left\langle H\left(p^i_{n+1}\right)|p^i_n\right\rangle_{p^i_{n+1}} + \gamma\left\langle U\left(p^i_{n+1};k+\tau_s\right)|p^i_n\right\rangle_{p^i_{n+1}} \\ &= H\left(p^i_n\right) - \int_0^1 \left[H\left(p^i_{n+1}\right) - \gamma U\left(p^i_{n+1};k+\tau_s\right)\right] f_p\left(p^i_{n+1}|p^i_n\right)\, dp^i_{n+1}, \end{aligned} \tag{2.15}$$

where the conditional distribution f_p is the same as the distribution given by equation (2.11). Combining equations (2.14) and (2.15) gives the general infomax utility function

$$U_{\mathrm{im}}\left(p^i_n;k\right) = \max\left\{U^{+}_{\mathrm{im}}\left(p^i_n;k\right), U^{-}_{\mathrm{im}}\left(p^i_n;k\right), U^{s}_{\mathrm{im}}\left(p^i_n;k\right)\right\}. \tag{2.16}$$

3. Results

We compared how the rewardmax and infomax strategies perform in our dynamic foraging task. The two strategies exhibit similar, but distinguishable, behavioural patterns under different task conditions governed by changes in environmental stability, ϵ, and feedback reliability, q. These differences reflect the overall benefits of rewardmax for obtaining more reward, on average, but of infomax for obtaining rewards more reliably, and go beyond standard analyses that focus on mean rewards. A consolidated summary of the parameter ranges studied and the corresponding performance patterns for both strategies is provided in electronic supplementary material, section S1.

3.1. Rewardmax and infomax strategies prescribe similar explore–exploit trade-offs

Single stochastic simulations of the rewardmax strategy on our dynamic foraging task (figure 2A) show that the regions of belief space where either commitment or sampling is the optimal action vary non-trivially as the agent uses actions from their budget. From these individual realizations, we can extract two behavioural metrics (schematized in figure 2B): (i) the number of commitment actions within a sequence of consecutive commitments, which we call a 'commit burst'; and (ii) the number of sample actions within a sample sequence, which we call a 'sample burst'. Both of these burst lengths are non-negative integers, because, for instance, even a single commitment action that is both preceded and followed by sampling actions counts as a sample burst of length one (figure 2). The statistics of these two quantities indicate the agent's preference for resource exploitation and environmental exploration, respectively. Previous work in foraging and psychology [43] has demonstrated that agents often navigate a trade-off between these two stereotypical behaviours, as in the classic 'win-stay, lose-switch' strategy [44,45].

By manipulating the environmental change probability, ϵ, and feedback reliability, q, we find that both rewardmax and infomax behaviours undergo a phase transition (figure 2C,D). Specifically, as ϵ decreases and q increases, both strategies prescribe a transition from sampling evidence at least occasionally (exploring) to never sampling evidence and committing to a decision with every action (exploitation; figure 2B,C). As these environmental parameters approach this phase transition, the reliability of feedback and the stability of the environment both increase, allowing the agent to carry a stronger belief into the next trial. In these environments, agents maximizing reward commit to high-value actions because doing so offers both immediate reward and information about future rewards. Agents maximize information by committing to high-value actions because the feedback obtained from probabilistic rewards carries more information, on average, than passive observations of the environment. Accordingly, the phase transition's boundary can be approximated by comparing the average change in belief magnitude from feedback versus sampling, given that the agent believes either state is equally likely (i.e. for LLR belief $y = 0$) as the level set

$$(1-h)\left[(1-\epsilon)q + \epsilon(1-q)\right] = h\left[\epsilon q + (1-\epsilon)(1-q)\right]. \tag{3.1}$$

For details of how to obtain this approximation, see electronic supplementary material, section S2. A separate feature visible in figure 2C is the broad grey plateau where the sampling–commitment ratio equals one. This plateau occurs when ϵ is close to 0.5, so the evidence from each sample is only weakly correlated with the latent state. In this near-memoryless regime, additional samples offer almost no marginal information gain. Consequently, the optimal policy reduces to a one-sample strategy: the agent takes a single observation to break the initial symmetry in its belief and then immediately commits. This strategy produces large

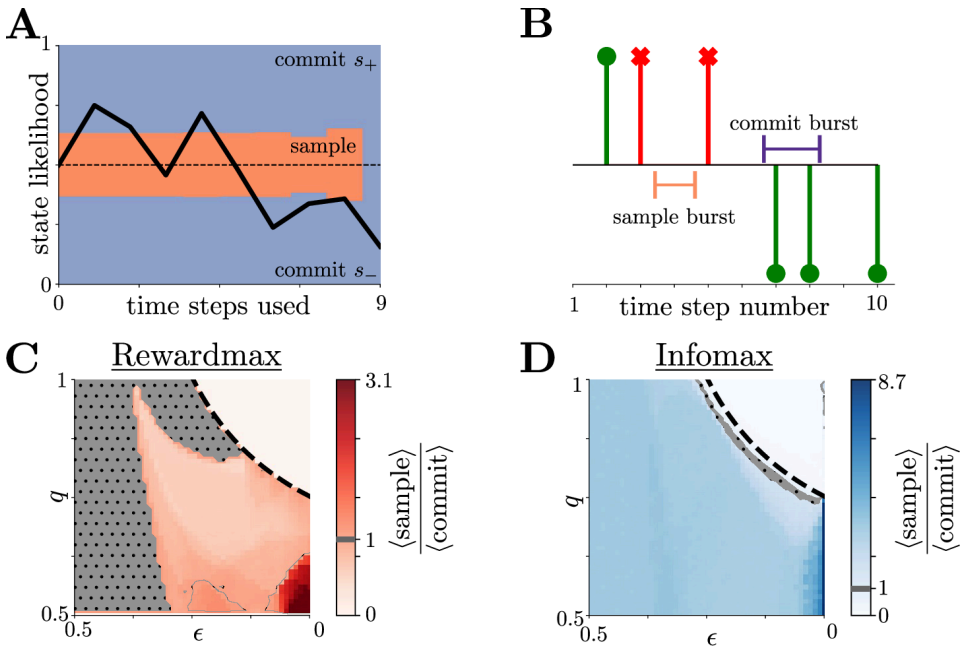

Figure 2. Both strategies predict explore–exploit phase transitions. (A,B) Example realization of the reward-maximizing strategy in state-likelihood space p_n^i (A) and action space (B). In (B), upward (downward) stems denote decisions towards s_+ (s_-), and green stems with circles (red spines with strikes) denote decisions that are rewarded (punished). Action numbers without stems denote actions for which the agent samples the environment for evidence. From each realization, we extract two behavioural metrics: (i) the average number of sequential commitments within a commit burst (e.g. purple), and (ii) the average number of sequential samples within a sample burst (e.g. orange). In this particular realization, there are four commit bursts of lengths 2, 1, 2 and 1, and there are four sample bursts of length 1. (C) Ratio of average sample burst length to average commit burst length for rewardmax behaviour as functions of environmental stability (ϵ) and reward reliability (q). In the realization shown in (B), the average sample burst length is 1, and the average commit length is 1.5, resulting in a ratio metric of 2/3. In the grey, dotted region of parameter space, this ratio is unity. Values larger than unity indicate more exploratory behaviour, whereas values smaller than unity indicate more exploitative behaviour. A black, dashed line shows the approximate location of the phase-transition boundary given by equation (3.1). As the environment becomes more stable (ϵ decreases) and provides more reliable feedback (q increases), the agent opts to never sample, relying on pure exploitation that reliably delivers reward and evidence. Results generated with 10^4 realizations, each with Bernoulli-evidence parameter $h = 0.75$, time-step budget $N = 10$, action time-step costs $(\tau_d, \tau_s) = (1, 1)$, reward structure $(R_c, R_i) = (100, -100)$ and no discounting ($\gamma = 1$). (D) Same as (C), but for infomax behaviour. Infomax behaviour exhibits a similar, but less sharp, phase transition to pure exploitation than rewardmax.

parameter regions in which the average number of samples and the average number of commitments are identical. Once feedback increments belief more than sampling, both strategies generally prefer to always commit.

The phase transition's boundary shifts in predictable ways as other task parameters are changed (see electronic supplementary material, figures S1 and S2). For example, if the commit time cost τ_d is larger than the sample time cost τ_s, the exploitation region shrinks, because commitment must produce a larger benefit to balance out the higher cost. The opposite holds true when $\tau_s > \tau_d$; i.e. the exploitation region grows. Similarly, the exploitation region shrinks as the environmental evidence's Bernoulli parameter h increases, because environmental evidence carries more information than information from reward. The opposite holds true when h decreases; i.e. the exploitation region grows. Furthermore, if the magnitude of reward $|R_c|$ is larger than the magnitude of punishment $|R_i|$, the rewardmax model's exploitation region expands, because commitment produces a higher average reward utility. The opposite holds true when $|R_i| > |R_c|$ (i.e. the exploitation region shrinks). The infomax model's exploitation region does not change with reward or punishment, because information utility operates independently from reward. The phase transition for both models does not appear to have a noticeable dependence on the agent's time-step budget N.

Despite these similarities, task-dependent phase transitions between exploitative and exploratory behaviours, there are noticeable behavioural differences in the way the two strategies perform the tasks. Using the same experimental parameter space, we define 'action alignment' between the strategies as the fraction of times that both prescribe the same action, given the same environmental belief (see figure 3A for a schematic). Action alignment is highest when ϵ is small and q is large, where commitment consistently yields highly informative feedback (figure 3B). Under these conditions, both strategies commit on every time step. However, the strategies differ across the remainder of parameter space, with lowest alignment when ϵ is at an intermediate level and q is large. This difference increases as future utility is more strongly discounted ($\gamma \to 0$), with the two strategies producing different behaviours across all but the most exploitation-dominant task conditions when only immediate rewards are considered.

3.2. Infomax produces more robust rewards than rewardmax

We further explored how the two strategy classes performed in terms of obtaining reward as their levels of temporal discounting were varied. Over an ensemble of belief realizations, both the rewardmax and infomax models produce distributions, $f_{\text{rm}}(\tilde{\rho})$ and $f_{\text{im}}(\tilde{\rho})$, respectively, over empirical reward rates $\tilde{\rho} = \frac{R}{T_t + T_i}$. We examined not simply the average, $\langle \tilde{\rho} \rangle$ (or, equivalently, ρ as defined in equation (2.6)), but the entire distribution of reward outcomes for each environment and strategy.

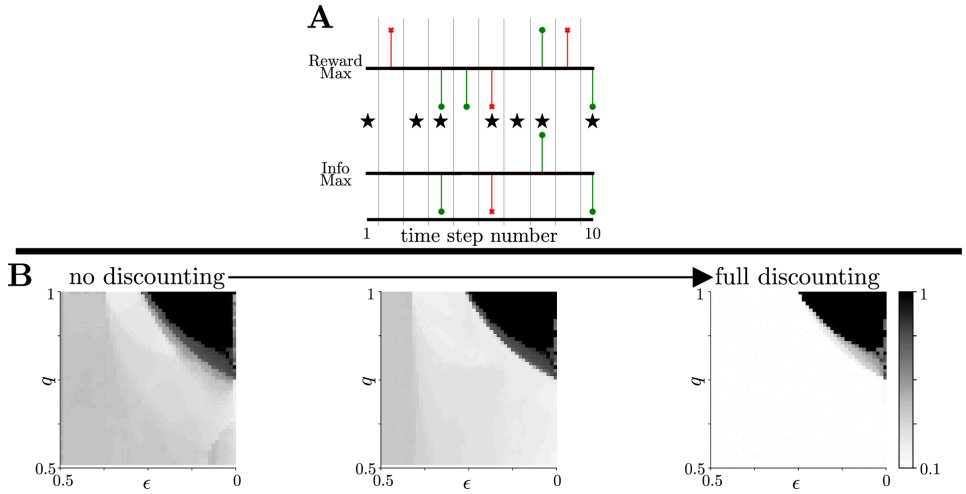

Figure 3. Action alignment of rewardmax and infomax behaviour. (A) Schematic of the alignment metric. For both models, we generated an action-space representation (using the same notation as in figure 2B) from a common environmental belief realization. We quantified the similarity of behaviours generated by the two models as 'action alignment', defined as the proportion of identical actions (marked with stars) when presented with the same belief about the current environmental state. (B) Alignment of rewardmax and infomax behaviours as a function of environmental stability (ϵ) and reward reliability (q) and progressing from no temporal discounting (left) to full temporal discounting of reward (right). Strategies are most distinct (i.e. alignment is lowest) in environments with intermediate stability (moderate ϵ) and high reliability (large q). This distinctness is emphasized as future utility is increasingly discounted ($\gamma \to 0$). Results generated with 10^4 realizations of common belief trajectories with $\gamma \in \{1, 0.5, 0\}$ (left-to-right panels) and all other task parameters as in figure 2.

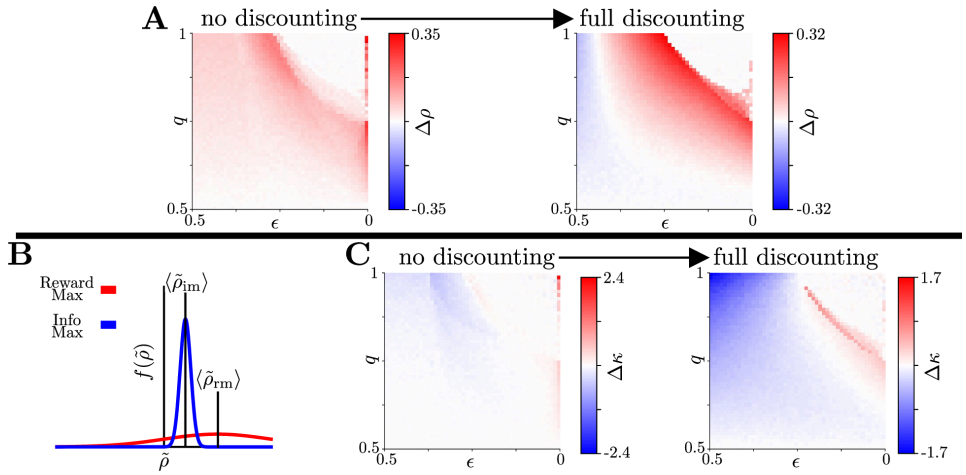

Figure 4. Robustness of the infomax model. (A) Average normalized reward rate differential $\Delta\rho = \frac{\rho_{rm} - \rho_{im}}{R_c}$ as a function of environmental stability, ϵ, and reward reliability, q, progressing from no temporal discounting ($\gamma = 1$, left) to full temporal discounting ($\gamma = 0$, right) of reward. (B) Schematic of reward rate distributions for both models. The rewardmax strategy tends to maximize the reward obtained, on average, but often with more variability than the infomax strategy across repeated instances of the same task conditions. Infomax thus guards against precipitously low reward returns by producing consistently adequate rewards. (C) Robustness differential $\Delta\kappa = \kappa_{rm} - \kappa_{im}$, with a model's robustness defined as $\kappa = \frac{\langle\tilde{\rho}\rangle}{\mathrm{std}(\tilde{\rho})}$. Even as rewardmax yields larger average reward rates, infomax delivers a narrower distribution of reward rates, making it a more robust strategy. This robustness advantage is enhanced as discounting on future utility increases.

As expected, the rewardmax model tends to outperform the infomax model in terms of average reward obtained across task conditions (by approx. 25–33% of a correct response; figure 4A). However, this performance advantage is more nuanced than one might expect. First, the advantages are most prominent in environments with a combination of intermediate-to-high reward reliability (q) and stability (ϵ). In contrast, both strategies lead to nearly identical average reward rates when reliability is low and when purely exploitative behaviours predominate for both models. Second, as future utility is discounted ($\gamma \to 0$), the infomax model increasingly produces larger average reward rates than the rewardmax model in highly volatile and/or unreliable environments. This effect can be explained by the fact that, as $\gamma \to 0$, the rewardmax model transitions to a greedy strategy that makes only commitments, ignoring potential future gains of sampling, whereas the infomax model continues to balance sampling-based and response feedback-based evidence gathering, making it a more robust strategy across objectives. This difference, coupled with a reward structure that includes punishment for incorrect decisions, allows the infomax model to narrow the performance advantage.

In addition to these differences in average reward rates, the two strategies produce striking differences in the reliability with which those rewards are obtained (figure 4B). Broad distributions offer a substantial probability of low performance, even if they have higher average reward rates. In contrast, narrow distributions guard against the probability of low performance, even if their

average reward rates are slightly lower. We defined a robustness metric, κ, as

$$\kappa = \frac{\langle \tilde{\rho} \rangle}{\mathrm{std}(\tilde{\rho})}, \tag{3.2}$$

which favours models that deliver a large average reward rate (placed in the numerator to avoid instability when values approach zero in our task) as well as a consistent reward rate (the standard deviation of which is in the denominator). Robustness is consistently higher for the infomax than the rewardmax strategy, across a broad range of task conditions and particularly as future utility discounting increases (figure 4C). These results demonstrate reasons for an infomax strategy to be preferred over a rewardmax strategy in naturalistic environments that carry risk.

4. Discussion

In this work, we provided a detailed analysis of how reward-maximizing (rewardmax) and information-maximizing (infomax) strategies perform a sequential, dynamic extension of a classic foraging task. We showed that both strategies exhibit a similar explore–exploit trade-off, with more exploring in volatile environments and more exploiting in stable environments. We also showed that these strategies can be distinguished from each other by their action sequences in non-extremal environments. These results complement previous comparisons of normative and heuristic strategies, which found that the largest performance differences between different decision strategies occur in environments with intermediate levels of uncertainty [46,47]. Finally, we showed that a rewardmax model often provides a performance advantage in average reward rate, but an infomax model dominates in producing a consistent distribution of nearly optimal reward rates. These strategy-level results complement eco-evolutionary work showing that the structure and heterogeneity of resource landscapes can shape the very cognitive and perceptual traits that support foraging decisions [48]. The optimal strategies in our model should be viewed as normative benchmarks rather than literal descriptions of animal decision-making. They illustrate the trade-off between gathering information about patch quality and exploiting a patch for immediate reward, and they identify the environmental conditions under which each behaviour is favoured. These benchmarks provide a useful baseline for comparing heuristic or approximate strategies that real foragers may use.

These results demonstrate the practical utility of information seeking in naturalistic foraging settings with risk and provide several testable hypotheses for further study. First, we predict that naturalistic behaviours should exhibit a sharp transition from exploration to exploitation at the boundary between moderate and high environmental stability and feedback reliability. Identifying these abrupt shifts in strategy could clarify how organisms adapt to changing environments. Second, as reward reliability decreases or environmental volatility increases, reliable long-term reward outcomes depend more on favouring information gathering, although potentially at the cost of short-term gains. This contrast underscores the trade-off between immediate gains and long-term stability, a pattern that resembles bet-hedging strategies in which organisms accept lower short-term returns to reduce variability in long-term outcomes [12].

This interpretation parallels empirical work showing that organisms frequently favour consistent outcomes over maximizing expected gains. In behavioural ecology, animals often choose options with lower variance in returns when environmental conditions are harsh or unpredictable, as predicted by risk-sensitive foraging theory [49,50]. Similarly, studies of human decision-making highlight satisficing principles and heuristics that prioritize reliable, good-enough outcomes rather than strict maximization of expected value [51]. This pattern is reflected in our robustness metric κ (mean reward divided by its standard deviation), which quantifies the stability of returns and highlights the consistent advantage of infomax in volatile environments.

Although our work focused on decision-making in the context of a foraging task, our findings probably generalize to other problems. In particular, our task shares features with those modelled by POMDPs in theories of economics, planning and machine learning [52–54]. Classical and contemporary POMDP formulations, including those that allow observation as a distinct action [34,35], typically optimize a single composite objective and therefore do not compare the distinct optimal epistemic and instrumental policies we analyse here. Stationary POMDPs (i.e. those designed to operate in statistically stable environments) have been compared with traditional infotaxis and reward-maximizing strategies [28,55]. Our work shows that those comparisons can miss many of the rich differences in behaviour that these strategies can produce in more dynamic environments. Consequently, based on our results, future utility discounting can dramatically alter the relative merits of reward- versus information-seeking strategies.

Active-inference formulations have also been applied to volatile multi-armed bandit tasks [30,56]. However, these tasks do not allow evidence accumulation independent of reward sampling and therefore cannot characterize the distinct optimal policies for pure epistemic versus pure instrumental objectives. Our decomposition isolates these components explicitly and reveals how their divergences depend on environmental volatility and feedback reliability.

We aimed to construct a simple sequential, dynamic foraging environment and feedback structure, so we could link observed behaviour directly to the agent's model selection. However, there are several natural extensions of our task that could yield further insights. The first would be to consider environments with state changes that can occur within a single deliberative process, so the belief must adapt to state fluctuations between commitments. This type of dynamic environment has been studied for non-sequential decision tasks [46,57,58]. We hypothesize that such environments would probably engender additional differences in rewardmax versus infomax behaviours, given the even more extensive need to learn about the statistics of the within-trial change process to inform the decisions.

In natural foraging settings, animals often choose among more than two patches, and rewards need not be mutually exclusive. In such environments, the informational structure changes substantially. For example, a lack of reward after committing to one option is far less diagnostic: several remaining patches may still be high-value, whereas in a two-state setting, a negative outcome sharply favours the single alternative. This reduced specificity would make sampling more valuable and probably produce

longer exploratory bouts and slower commitment [59–61]. Accordingly, the divergence we observe between rewardmax and infomax in the binary case is probably conservative, and the informational advantage of infomax may become even more pronounced in higher-dimensional settings. A full analysis of this case would require extending the dynamic programming formulation to higher-dimensional belief spaces, which we view as a promising direction for future work.

A second possible extension would be to consider multiple rewarding alternatives beyond the two we considered. For our two-alternative task, the information gained by mutual exclusivity allowed both models to perform well across environmental parametrizations. This mutual exclusivity is broken when there are more than two possible environmental states, and evidence against one alternative increases the probability of multiple other alternatives. We hypothesize that under these conditions, an infomax strategy has even more extensive benefits over a rewardmax strategy, given our finding that infomax tends to perform well in difficult, risky environments. We also predict that an infomax strategy may exhibit complex commitment patterns, because committing to, and obtaining feedback from, different alternatives in sequence can yield significant information, even if feedback indicates an incorrect decision. These nuanced commitment behaviours may prove useful in studying various choice paradoxes observed in cognitive psychology, such as the independence of irrelevant alternatives [7,62,63].

Our model assumes that environmental states evolve as a memoryless Markov process, which allows for the derivation of closed-form optimal policies. Real foraging landscapes often exhibit structured depletion and recovery. Incorporating such dynamics would probably accentuate the differences we observe between reward-maximizing and information-maximizing strategies. A further extension would incorporate environments whose dynamics depend directly on the agent's actions. In natural settings, consumed resources require time to replenish, and foragers must learn not only the reward contingencies but also the recovery dynamics they induce. Introducing such feedback between decisions and environmental state transitions would create a richer class of sequential problems and may reveal additional differences between rewardmax and infomax strategies, particularly in how they track or exploit resource regeneration over time.

Finally, although our focus here was on contrasting optimal reward-seeking and information-seeking strategies, the explicit separation of these objectives provides a natural bridge to computational phenotyping approaches. In such frameworks, differences in how individuals weight epistemic and instrumental components of utility are used to characterize behavioural heterogeneity, often framed in terms of the complete class theorem [64], which links observable policies to implied preferences and beliefs. This perspective underlies recent computational psychiatry work showing how individual differences in priors and utility weightings can explain variation in approach-avoidance conflict and decision uncertainty [65–67]. Extending our model to include inter-individual variability, learning of environmental parameters, or hierarchical inference could therefore allow this task to serve as a compact testbed for such phenotyping efforts.

Overall, our results demonstrate that while reward-seeking strategies maximize average gains, information-seeking strategies provide greater consistency and robustness in uncertain environments. Infomax's ability to generate more stable reward distributions suggests that organisms operating in dynamic settings may benefit from prioritizing information acquisition, even at the cost of occasional missed immediate rewards. These findings highlight the adaptive value of information-seeking behaviours for future gains, particularly in volatile environments in which reward contingencies change unpredictably. In other words, the pursuit of information may be as crucial to survival as the pursuit of resources themselves.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. See [68] for the Python code used to generate all results and figures.

Supplementary material is available online [69].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. N.W.B.: conceptualization, formal analysis, investigation, methodology, software, visualization, writing—original draft; J.I.G.: conceptualization, funding acquisition, supervision, writing—review and editing; K.J.: conceptualization, funding acquisition, supervision, writing—review and editing; Z.P.K.: conceptualization, funding acquisition, project administration, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. At the time of submission and acceptance, K.J. was a member of the editorial board. He was not involved in the editorial process for this manuscript.

Funding. N.W.B., J.I.G., K.J. and Z.P.K. were all supported by a CRCNS grant (NSF DMS-2207700). Z.P.K. was supported by NIH BRAIN 1R01EB029847-01. K.J. was supported by NSF-DBI-1707400 and NIH RF1MH130416.

References

1. Gilmour ME, Castillo-Guerrero JA, Fleishman AB, Hernández-Vázquez S, Young HS, Shaffer SA. 2018 Plasticity of foraging behaviors in response to diverse environmental conditions. *Ecosphere* **9**, e02301. (doi:10.1002/ecs2.2301)
2. Ropert-Coudert Y, Kato A, Chiaradia A. 2009 Impact of small-scale environmental perturbations on local marine food resources: a case study of a predator, the little penguin. *Proc. R. Soc. B Biol. Sci.* **276**, 4105–4109. (doi:10.1098/rspb.2009.1399)
3. Barendregt NW, Gold JI, Josić K, Kilpatrick ZP. 2022 Normative decision rules in changing environments. *eLife* **11**, e79824. (doi:10.7554/eLife.79824)
4. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006 The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765. (doi:10.1037/0033-295X.113.4.700)
5. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. 2012 The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32**, 3612–3628. (doi:10.1523/JNEUROSCI.4010-11.2012)
6. Nguyen KP, Josić K, Kilpatrick ZP. 2019 Optimizing sequential decisions in the drift–diffusion model. *J. Math. Psychol.* **88**, 32–47. (doi:10.1016/j.jmp.2018.11.001)
7. Tajima S, Drugowitsch J, Patel N, Pouget A. 2019 Optimal policy for multi-alternative decisions. *Nat. Neurosci.* **22**, 1503–1511. (doi:10.1038/s41593-019-0453-9)
8. Hayden BY, Walton ME. 2014 Neuroscience of foraging. *Front. Neurosci.* **8**, 81. (doi:10.3389/fnins.2014.00081)

9. Hills TT. 2006 Animal foraging and the evolution of goal-directed cognition. *Cogn. Sci*. **30**, 3–41. (doi:10.1207/s15516709cog0000_50)
10. Mobbs D, Trimmer PC, Blumstein DT, Dayan P. 2018 Foraging for foundations in decision neuroscience: insights from ethology. *Nat. Rev. Neurosci.* **19**, 419–427. (doi:10.1038/s41583-018-0010-7)
11. Charnov EL. 1976 Optimal foraging, the marginal value theorem. *Theor. Popul. Biol*. **9**, 129–136. (doi:10.1016/0040-5809(76)90040-X)
12. Pyke GH. 1984 Optimal foraging theory: a critical review. *Annu. Rev. Ecol. Syst*. **15**, 523–575. (doi:10.1146/annurev.es.15.110184.002515)
13. Green RF. 1984 Stopping rules for optimal foragers. *Am. Nat*. **123**, 30–43. (doi:10.1086/284184)
14. Valone TJ. 2006 Are animals capable of Bayesian updating? An empirical review. *Oikos* **112**, 252–259. (doi:10.1111/j.0030-1299.2006.13465.x)
15. Kilpatrick ZP, Davidson JD, El Hady A. 2021 Uncertainty drives deviations in normative foraging decision strategies. *J. R. Soc. Interface* **18**, 20210337. (doi:10.1098/rsif.2021.0337)
16. McNamara J. 1982 Optimal patch use in a stochastic environment. *Theor. Popul. Biol*. **21**, 269–288. (doi:10.1016/0040-5809(82)90018-1)
17. Olschewski S, Mullett TL, Stewart N. 2025 Optimal allocation of time in risky choices under opportunity costs. *Cogn. Psychol*. **157**, 101716. (doi:10.1016/j.cogpsych.2025.101716)
18. Cisek P, Puskas GA, El-Murr S. 2009 Decisions in changing conditions: the urgency-gating model. *J. Neurosci*. **29**, 11560–11571. (doi:10.1523/JNEUROSCI.1844-09.2009)
19. Ho MK, Abel D, Correa CG, Littman ML, Cohen JD, Griffiths TL. 2022 People construct simplified mental representations to plan. *Nature* **606**, 129–136. (doi:10.1038/s41586-022-04743-9)
20. Calhoun AJ, Chalasani SH, Sharpee TO. 2014 Maximally informative foraging by *Caenorhabditis elegans*. *eLife* **3**, e04220. (doi:10.7554/eLife.04220)
21. Vergassola M, Villermaux E, Shraiman BI. 2007 'Infotaxis' as a strategy for searching without gradients. *Nature* **445**, 406–409. (doi:10.1038/nature05464)
22. Davidson JD, El Hady A. 2019 Foraging as an evidence accumulation process. *PLoS Comput. Biol*. **15**, e1007060. (doi:10.1371/journal.pcbi.1007060)
23. Masson JB. 2013 Olfactory searches with limited space perception. *Proc. Natl Acad. Sci. USA* **110**, 11261–11266. (doi:10.1073/pnas.1221091110)
24. Bartumeus F, Campos D, Ryu WS, Lloret-Cabot R, Méndez V, Catalan J. 2016 Foraging success under uncertainty: search tradeoffs and optimal space use. *Ecol. Lett*. **19**, 1299–1313. (doi:10.1111/ele.12660)
25. Barbieri C, Cocco S, Monasson R. 2011 On the trajectories and performance of Infotaxis, an information-based greedy search algorithm. Europhys. Lett. **94**, 20005. (doi:10.1209/0295-5075/94/20005)
26. Karpas ED, Shklarsh A, Schneidman E. 2017 Information socialtaxis and efficient collective behavior emerging in groups of information-seeking agents. *Proc. Natl Acad. Sci. USA* **114**, 5589–5594. (doi:10.1073/pnas.1618055114)
27. Voges N, Chaffiol A, Lucas P, Martinez D. 2014 Reactive searching and infotaxis in odor source localization. *PLoS Comput. Biol*. **10**, e1003861. (doi:10.1371/journal.pcbi.1003861)
28. Loisy A, Eloy C. 2022 Searching for a source without gradients: how good is infotaxis and how to beat it. *Proc. R. Soc. A* **478**, 20220118. (doi:10.1098/rspa.2022.0118)
29. Friston K, Da Costa L, Hafner D, Hesp C, Parr T. 2021 Sophisticated inference. *Neural Comput.* **33**, 713–763. (doi:10.1162/neco_a_01363)
30. Marković D, Stojic H, Schwöbel S, Kiebel SJ. 2021 An empirical evaluation of active inference in multi-armed bandits. *Neural Netw.* **144**, 229–246. (doi:10.1016/j.neunet.2021.08.027)
31. Parr T, Pezzulo G, Friston KJ. 2022 *Active inference*. Cambridge, MA, USA: MIT Press. See https://direct.mit.edu/books/book/5299/Active-InferenceThe-Free-Energy-Principle-in-Mind.
32. Nguyen KP. 2020 How trial correlations and feedback shape sequential decision-making. PhD thesis, University of Houston. UH Electronic Theses and Dissertations. https://hdl.handle.net/10657/10253.
33. Philippi T, Seger J. 1989 Hedging one's evolutionary bets, revisited. *Trends Ecol. Evol*. **4**, 41–44. (doi:10.1016/0169-5347(89)90138-9)
34. Kaelbling LP, Littman ML, Cassandra AR. 1998 Planning and acting in partially observable stochastic domains. *Artif. Intell*. **101**, 99–134. (doi:10.1016/S0004-3702(98)00023-X)
35. Littman ML. 2009 A tutorial on partially observable Markov decision processes. *J. Math. Psychol*. **53**, 119–125. (doi:10.1016/j.jmp.2009.01.005)
36. Levin SA. 2000 Multiple scales and the maintenance of biodiversity. *Ecosystems* **3**, 498–506. (doi:10.1007/s100210000044)
37. McMahon LA, Rachlow JL, Shipley LA, Forbey JS, Johnson TR. 2017 Habitat selection differs across hierarchical behaviors: selection of patches and intensity of patch use. *Ecosphere* **8**, e01993. (doi:10.1002/ecs2.1993)
38. Gold JI, Shadlen MN. 2007 The neural basis of decision making. *Annu. Rev. Neurosci*. **30**, 535–574. (doi:10.1146/annurev.neuro.29.051605.113038)
39. Wald A. 1945 Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16**, 117–186. (doi:10.1214/aoms/1177731118)
40. Bellman R. 1966 Dynamic programming. *Science* **153**, 34–37. (doi:10.1126/science.153.3731.34)
41. Drugowitsch J. 2015 *Notes on normative solutions to the speed-accuracy trade-off in perceptual decision-making*. FENS-Hertie Winter School 2015 'The neuroscience of decision making'. See https://www.drugowitschlab.org/assets/pdf/fens2015_notes.pdf.
42. Cover TM, Thomas JA. 2006 *Elements of information theory*, 2nd edn. Hoboken, NJ: John Wiley & Sons.
43. Addicott MA, Pearson JM, Sweitzer MM, Barack DL, Platt ML. 2017 A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology* **42**, 1931–1939. (doi:10.1038/npp.2017.108)
44. Ma T, Hermundstad AM. 2024 A vast space of compact strategies for highly efficient decisions. *Sci. Adv.* **10**, eadj4064.
45. Robbins H. 1952 Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc*. **58**, 527–535. (doi:10.1090/S0002-9904-1952-09620-8)
46. Barendregt NW, Josić K, Kilpatrick ZP. 2019 Analyzing dynamic decision-making models using Chapman-Kolmogorov equations. *J. Comput. Neurosci*. **47**, 205–222. (doi:10.1007/s10827-019-00733-5)
47. Tavoni G, Doi T, Pizzica C, Balasubramanian V, Gold JI. 2022 Human inference reflects a normative balance of complexity and accuracy. *Nat. Hum. Behav*. **6**, 1153–1168. (doi:10.1038/s41562-022-01357-z)
48. Gibbs R, Landi P, Hui C. 2024 Heterogeneity in the resource landscape encourages increased cognitive and perceptive capabilities in foragers. *Ecol. Modell*. **492**, 110693. (doi:10.1016/j.ecolmodel.2024.110693)
49. Kacelnik A, Bateson M. 1996 Risky theories—the effects of variance on foraging decisions. *Am. Zool*. **36**, 402–434. (doi:10.1093/icb/36.4.402)
50. Bateson M. 2007 Recent advances in our understanding of risk-sensitive foraging preferences. *Proc. Nutr. Soc*. **66**, 397–404. (doi:10.1017/S0029665107005705)
51. Gigerenzer G, Gaissmaier W. 2011 Heuristic decision making. *Annu. Rev. Psychol*. **62**, 451–482. (doi:10.1146/annurev-psych-120709-145346)
52. Auer P, Cesa-Bianchi N, Fischer P. 2002 Finite-time analysis of the multiarmed bandit problem. *Mach. Learn*. **47**, 235–256. (doi:10.1023/A:1013689704352)
53. Bergemann D, Välimäki J. 2008 Bandit problems. In *The new Palgrave dictionary of economics*, 2nd edn. London, UK: Palgrave Macmillan.
54. Kocsis L, Szepesvari C. 2006 Bandit based Monte-Carlo planning (eds F Fürnkranz, T Scheffer, M Spiliopoulou). In *Machine Learning: ECML 2006*, pp. 282–293. Berlin, Germany: Springer. (doi:10.1007/11871842_29)
55. Heinonen RA, Biferale L, Celani A, Vergassola M. 2023 Optimal policies for Bayesian olfactory search in turbulent flows. *Phys. Rev. E* **107**, 055105. (doi:10.1103/PhysRevE.107.055105)
56. Sajid N, Tigas P, Friston K. 2022 Active inference, preference learning and adaptive behaviour. *IOP Conf. Ser.: Mater. Sci. Eng.* **1261**, 012020. (doi:10.1088/1757-899X/1261/1/012020)
57. Glaze CM, Kable JW, Gold JI. 2015 Normative evidence accumulation in unpredictable environments. *eLife* **4**, e08825. (doi:10.7554/eLife.08825)
58. Veliz-Cuba A, Kilpatrick ZP, Josić K. 2016 Stochastic models of evidence accumulation in changing environments. *SIAM Rev*. **58**, 264–289. (doi:10.1137/15M1028443)

59. McMillen T, Holmes P. 2006 The dynamics of choice among multiple alternatives. *J. Math. Psychol.* **50**, 30–57. (doi:10.1016/j.jmp.2005.10.003)
60. Moreno-Bote R, Ramírez-Ruiz J, Drugowitsch J, Hayden BY. 2020 Heuristics and optimal solutions to the breadth-depth dilemma. *Proc. Natl Acad. Sci. USA* **117**, 19799–19808. (doi:10.1073/pnas.2004929117)
61. Usher M, McClelland JL. 2001 The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.* **108**, 550–592. (doi:10.1037/0033-295x.108.3.550)
62. Luce RD. 1959 *Individual choice behavior: a theoretical analysis*. New York, NY: Wiley.
63. Shafir S, Waite TA, Smith BH. 2002 Context-dependent violations of rational choice in honeybees (*Apis mellifera*) and gray jays (*Perisoreus canadensis*). *Behav. Ecol. Sociobiol.* **51**, 180–187. (doi:10.1007/s00265-001-0420-8)
64. Wald A. 1947 An essentially complete class of admissible decision functions. *Ann. Math. Statist.* **18**, 549–555. (doi:10.1214/aoms/1177730345)
65. Smith R, Khalsa SS, Paulus MP. 2021 An active inference approach to dissecting reasons for nonadherence to antidepressants. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **6**, 919–934. (doi:10.1016/j.bpsc.2019.11.012)
66. Smith R, Kirlic N, Stewart JL, Touthang J, Kuplicki R, Khalsa SS, Feinstein J, Paulus MP, Aupperle RL. 2021 Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: a computational modelling approach. *J. Psychiatry Neurosci.* **46**, E74–E87. (doi:10.1503/jpn.200032)
67. Smith R *et al*. 2021 Long-term stability of computational parameters during approach-avoidance conflict in a transdiagnostic psychiatric patient sample. *Sci. Rep.* **11**, 11783. (doi:10.1038/s41598-021-91308-x)
68. Barendregt N. 2026 Code for: Information-seeking decision strategies mitigate risk in dynamic, uncertain environments. Zenodo. (doi:10.5281/zenodo.18633175)
69. Barendregt N, Gold JI, Josic K, Kilpatrick ZP. 2026 Supplementary material from: Information-Seeking Decision Strategies Mitigate Risk in Dynamic, Uncertain Environments. FigShare. (doi:10.6084/m9.figshare.c.8400720)