Curiosity eliminates the exploration-exploitation dilemma

Erik J Peterson[a,b,1] **and Timothy D Verstynen**[a,b,c,d]

[a]Department of Psychology; [b]Center for the Neural Basis of Cognition; [c]Carnegie Mellon Neuroscience Institute; [d]Biomedical Engineering, Carnegie Mellon University, Pittsburgh PA

Balancing exploration with exploitation is a mathematically intractable problem during reward collection. In this paper, we provide an alternative view that does not depend on exploring to optimize for reward. We argue that the goal of exploration ought to be learning for learning's sake, which we equate to curiosity. Through theory and experiments we show exploration-as-curiosity gives a tractable answer to explore-exploit problems. Simply put, when information is more valuable than rewards, be curious, otherwise seek rewards. We show how this rule succeeds in simulations with naturalistic constraints, including when rewards are sparse, deceptive, high-dimensional, or non-stationary. We suggest four criteria can be used to experimentally distinguish our approach from other theories.

Introduction

In this paper we offer an unconventional perspective on the exploration-exploitation dilemma. We offer this because exploratory behavior in animals can have two different explanations. If environmental rewards are expected by the animal, then exploration is cast as a search for those rewards (1–6). Rewards are how most studies of exploration are framed (2). If instead, there is no reason to expect rewards from the environment, then exploration is a search for information, or curiosity (7–13).

Exploitation has one definition, taking the most rewarding action. It is always the opposite of exploration. An open problem is to unify these opposites, exploration with exploitation. But past attempts at such a union have focused on exploration-for-rewards which in turn leads to the exploration-exploitation dilemma, which is an intractable mathematical problem (14–19). We illustrate this in Figure 1a. Here we offer an alternative approach by unifying exploitation with pure curiosity instead. We illustrate this in Figure 1b. We will prove this take on a union is tractable.

We are interested in curiosity because it is a universal drive, though the definition of what curiosity is remains open. Curiosity can be triggered by an incongruity, or novelty and surprise, or by a gap in knowledge, a prediction error, or by the potential usefulness of some knowledge (9, 20–28). Here we will take an even more general position than these definitions. We argue that curiosity can be recognized by an animal pursuing learning for learning's sake. Even though this definition is both abstract and recursive, we show it lends itself to a precise mathematical form.

Curiosity the drive is as strong as the drive for collecting environmental rewards, if not stronger (7, 9, 11, 12, 25, 29–32). Furthermore, in studies with artificial agents curiosity-like algorithms are highly effective at solving difficult optimization problems (8, 33–40). In other words, we are interested because curiosity is universal and because it works.

This paper consists of a mathematical study of information value, and an optimal value union with reward collection. This is followed by simulations where the performance of our union often exceeds more standard approaches to solving the exploration-exploitation dilemma. We conclude with some questions and answers about our approach.

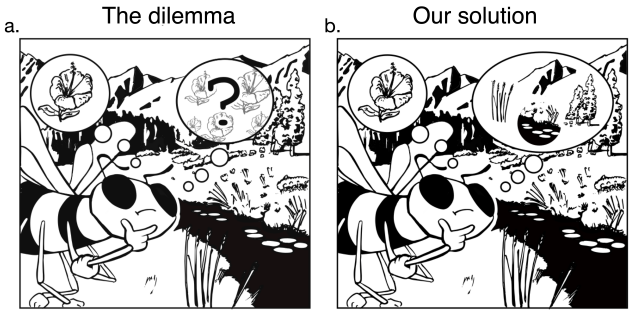

Fig. 1. Two views of exploration and exploitation. **a**. The dilemma: either exploit an action with a known reward (e.g., return to the previous flower) or explore other actions on the chance they will return a better outcome. The central challenge here is that the outcome of exploration is uncertain, intractable, and filled with questions. **b**. Our alternative view of the dilemma, with two goals: either exploit rewards *or* explore to maximize information value, with a curious search of the environment. The solution we propose to this alternative view is to pursue either information value, or reward value, depending on which is expected to be larger. In other words, we set information and reward on equal terms and propose a greedy solution to choose between them. *Artist credit*: Richard Grant.

Results

The central difference between our approach to exploration-exploitation and more standard views in reinforcement learning is that we recognize learning about the environment by curious exploration is just as important for survival as reward collection. To formalize this idea we first introduce reinforcement learning, and its basis in the Bellman equation.

Reward collection - theory. We consider an animal who interacts with an environment over a sequence of states, actions and rewards. The animal's goal is to select actions in a way that maximizes the total reward collected. How should an animal do this?

When the environment is unknown, an animal needs to trade-off exploration and exploitation. The standard way of

Significance Statement

It's optimal to be curious.

The authors have no conflicts of interest to declare.

[1]To whom correspondence should be addressed. E-mail: Erik.Exists@gmail.com

Table 1. Symbols.

Notation	Description
$\mathcal{S}$	The state space of the environment.
$\mathcal{A}$	The action space of the agent.
S	A state.
S'	A next state in a sequence.
$\mathcal{T}(S\|A,t)$	The transition function. Generates S'.
$\mathcal{X}$	The observation set, $\{S, A, S', R\}$
A	An action.
$\mathcal{R}(S,A,t)$	The reward function. Generates R.
R	Reward value from the environment.
$\mathbf{M}$	Memory (some real prameters; a vector space).
$f(\mathcal{X},\mathbf{M})$	A learning function. Generates $\mathbf{M}'$.
E	Subjective information value.
$=$	Denotes assignment.
$\sim$	Denotes a random sample/draw.
$\mathbb{E}$	The expectation operator.
$d(\mathbf{M},\mathbf{M}')$	Information value function. Generates E.
π	An action policy sequence. Generates A.
π^*	An optimal policy sequence.
π_R	A policy sequence to maximize reward (R, H).
π_E	A policy sequence to maximize information (E).
$\pi(S)$	A deterministic action policy sequence.
$\pi(S\|A)$	A stochastic action policy sequence.
$\|$	Denotes any conditional.
$\|.\|$	Denotes the absolute value.
$\|\|.\|\|$	Denotes a vector norm.
δ	Denotes a difference between two scalars or vectors.
∇	Denotes a gradient.
α	Step-size (0,1].
β	Softness parameter for Boltzman/Softmax sampling (0, ∞).
η	Boredom threshold [0,∞).
ρ	Reward bias [0,∞).

approaching this is to assume, "The agent must try a variety of actions and progressively favor those that appear to be best (2)".

We require states, actions, and rewards to be embedded in a discrete time Markov decision process, a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$. Specifically, states are real valued vectors S from a finite set $\mathcal{S}$ of size n. Actions A are from the finite real set $\mathcal{A}$ of size k. We consider a set of policies that consist of a sequence of functions, $\pi = \{\pi_t, \pi_{t+1}, \ldots, \pi_{T-1}\}$. For any step t, actions are generated from states by policies, either deterministic $A = \pi(S)$ or random $A \sim \pi(S|A)$. In our presentation we drop the indexing notation on our policies, using simply π to refer to the sequence as a whole. Rewards are single valued real numbers, R, generated by a reward function $R \sim \mathcal{R}(S|A,t)$. Transitions from state S to new state $S\grave{t}$ are caused by a stochastic transition function, $S' \sim \mathcal{T}(S|A,t)$. We leave open the possibility both $\mathcal{T}$ and $\mathcal{R}$ may be time-varying. A complete observation set is the set $\mathcal{X}_t = \{S, A, S', R\}$. In general we use the $=$ to denoate assignment, in contrast to $\sim$ which we use to denote random samples. The standlone $|$ operator is used to denote conditional probability. An asterix is used to denote optimal value policies, π^*. All the symbols used are summarized in Table 1.

We focus on discrete time systems and finite time horizons. We do not discount the rewards, for simplicity. What we do here should generalize to continuous spaces, and discounted but infinite time horizons (2, 41)

So given a policy π_R, the state value function for reward is given by,

$$V_{\pi_R}(S) = \mathbb{E}_{\pi_R}\Big[\sum_{t=0}^{T} R_t \mid S_t = S\Big] \qquad [1]$$

And its Bellman equation (42) is,

$$\begin{aligned} V_{\pi_R^*}(S) &= \max_{\pi_R} V_{\pi_R}(S) \\ &= \max_{\pi_R} \mathbb{E}_{\pi_R}\Big[\sum_{t=0}^{T} R_t \mid S_t = S\Big] \\ &= \max_{A\in\mathbb{A}} \mathbb{E}\Big[R_t + \sum_{t=0}^{T} R_{t+1} \mid S_t = S,\ A_t = A\Big] \\ &= \max_{A\in\mathbb{A}} \mathbb{E}\Big[R_t + V_{\pi_R^*}(S_{t+1}) \mid S_t = S,\ A_t = A\Big] \end{aligned} \qquad [2]$$

The Bellman equation is desirable because it simplifies the entire action sequence in Eq. 1 into two steps, t and $t+1$. This is what is called an optimal subproblem, because Bellman's insight was that the entire optimal sequence can be reduced to Eq. 2, under certain conditions. The practical problem in this simplification is finding a reliable estimate of $V_{\pi_R^*}(S_{t+1})$ (Eq. 2). A problem we return to further on.

Information collection - theory. We now consider an animal who interacts with an environment over the same space as in reward collection, but whose goal is to select actions in a way that maximizes the total information learned (39, 43–51). In this, we face two questions. How should this information and learning be valued? And like for reward collection above, is there a Bellman solution to our information maximization problem?

Let's say E represents the information value of an observation. The Bellman solution we wish we had is,

$$V_{\pi_E^*}(S) = \max_{A\in\mathbb{A}} \mathbb{E}\Big[E_t + V_{\pi_E^*}(S_{t+1}) \mid S_t = S,\ A_t = A\Big] \qquad [3]$$

In the following paragraphs we will define E and prove this definition has the Bellman solution shown in Eq.3. We will *not* limit ourselves to Markovian ideas of learning and memory in doing this.

We define memory $\mathbf{M}$ as a vector of size p embedded in a real-valued space that is closed and bounded. This idea maps to a range of physical examples of memory, including firing rates of neurons in a population (52), strength differences between synapses (53), or Calcium dynamics (54).

A learning function is then any function f that maps observations $\mathcal{X}$ into memory $\mathbf{M}$. This idea can be expressed using several identical mathematical forms. Using recursive notation, learning can be denoted by $\mathbf{M} \leftarrow f(\mathcal{X}, \mathbf{M})$. Though we will sometimes use $\mathbf{M}'$ to denote the updated memory instead. Other times we will add subscripts like $(\mathbf{M}_t, \mathbf{M}_{t+1}, \ldots)$ to show a temporal path through memory.

We also define a forgetting function, $f^{-1}f(\mathcal{X},\ \mathbf{M}') \rightarrow \mathbf{M}$ which is important later.

A more traditional kind of learning function could be something like $f_{\mathbf{M}}(\mathcal{X}) = \hat{y}$ and we would be concerned with minimizing the loss between a target objective y and the estimate $\hat{y}$, based on *parameters* $\mathbf{M}$. And indeed such a function will

be in operation "behind the scenes". In this paper we are not concerned with the objective y and the absolute error of the loss function, but only with the memory dynamics themselves and their equilibrium points. This approach lets us provide a *technical definition* for information value (55).

As an auxiliary assumption we presume f has been selected by evolution to be formally learnable (56, 57), as this assumption is a minimal requirement for any sensible learning function.

Axioms for information value. We reason the value of any information depends entirely on what is learned by making that observation, and that uniformly distributed learning provides a floor or minimum value for learning. This is in direct contrast to the maximum value expected when estimating entropy/information with uniform distributions. This inversion is the key idea in our metric. Though, we introduce other Axioms to ensure E is sensible, useful, and well-behaved.

If we have a memory $\mathbf{M}$ which has been learned by f over a path of observations, $(\mathbf{X_0}, \mathbf{X_1}, ...)$, can we measure how much value the next observation $\mathbf{X_t}$ should have? We think the answer to this question should ensure,

Axiom 1 (Axiom of Memory)**.** *E is continuous with continuous changes in memory, $\Delta\mathbf{M}$, between $\mathbf{M}'$ and $f(\mathcal{X}, \mathbf{M})$.*

See Figure 2a for an example of E defined only in memory space. A corollary of Axiom 1 is that an observation that doesn't change memory, has no value. So if $\sum_{i=1}^{p} |\Delta M_i| = 0$ then $E = 0$.

Axiom 2 (Axiom of Specificity)**.** *If all p elements $|\Delta M_i|$ in $\mathbf{M}$ are equal, then E is minimized.*

That is, information that is learned evenly in memory cannot induce any inductive bias, and is therefore the least useful learning possible (58, 59). It should therefore be the least valuable information.

Axiom 3 (Axiom of Scholarship)**.** *E is greater than or equal to 0, $E \geq 0$.*

That is, learning something always has a positive value in principle even if later on it has some negative consequences.

Axiom 4 (Axiom of Equilibrium)**.** *Given the same observation $\mathcal{X}_i$, E will decrease to below some finite threshold $\eta > 0$ in finite time.*

That is, learning on $\mathbf{M}$ should eventually find an equilibrium point for each observation. We show examples of this (self-)consistent and inconsistent learning in Figure 2b-c.

A Bellman solution. A common way to arrive at the Bellman solution is to assume a Markov decision space. This is a problem for our definition of memory which we must assume has a natural path dependence. This in turn breaks the Markov assumption, and the needed claim of optimal substructure, the ability to break a sequential learning problem up into smaller pieces (42). To get around this we will prove that exact forgetting of the last observation is another way to achieve optimal substructure, and so to derive a Bellman solution.

In the theorem below which shows the optimal substructure of $\mathbf{M}$ we assume $\mathcal{X}$, and f are given,

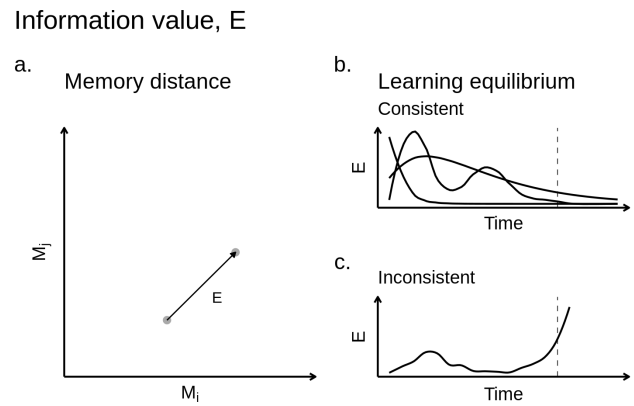

Fig. 2. Information value. **a**. This panel illustrates a two dimensional memory. The information value of an observation depends on the distance memory "moves" during learning, denoted here by E We depict memory distance as euclidean space, but this is one of many possible ways to realize E. **b-c** This panel illustrates some examples of learning dynamics in time, made over a series of observations that are not shown. If information value becomes decelerating in finite time bound, then we say learning is consistent with Def. 2. This is what is shown in panel b. If learning does not decelerate, then it is said to be inconsistent (Panel c). The finite bound is represented here by a dotted line.

Theorem 1 (Optimal substructure)**.** *If $V^*_{\pi_E}$ is the optimal information value given by policy π_E, a memory $\mathbf{M}_t$ has optimal substructure if the last observation $\mathcal{X}$ can be removed from $\mathbf{M}$, by $\mathbf{M}_{t-1} = f^{-1}(\mathcal{X}, \mathbf{M}_t)$ such that the resulting value $V^*_{t-1} = V^*_t - E_t$ is also optimal.*

Forgetting here is of a very particular kind. It is more of a theoretical conconceit than an hard requirement an animal would need to physiologically implement in practice. For the proof to hold it must only be possible to construct such a function f^{-1} in theory. And given the definition includes the observation to be forgotten, we can be sure to be able to construct such a function even if the memory is imperfect, missing information, or otherwise compressed (27, 60, 61).

Having proven the optimal substructure of $\mathbf{M}$, and assuming some arbitrary starting value $E_0 > 0$, it naturally follows that the Bellman solution to information optimization is given by Equation 3. We finally assume that maximizing information value is identical to maximizing curiosity, and refer to our deterministic approach to curiosity as E-exploration, or E-explore for short.

To limit curiosity, both to avoid "useless" minutia (33) and stop exploration behavior overall, we rely on the threshold term, $\eta \geq 0$ (Ax. 4) which we treat as synonymous with boredom (8, 62, 63). We hypothesize boredom is an adaptive parameter that can be tuned by an animal to fit their environment (64). Specifically, we use this boredom threshold in E-explore to ignore "small" amounts of information value by requiring exploration of some observation $\mathcal{X}$ to cease once $E \leq \eta$ for that $\mathcal{X}$.

Reward and information - theory. Finally, we consider an animal who interacts with an environment who wishes to maximize both information *and* reward value. How can an animal do this optimally? To answer this we make two conjectures:

Conjecture 1 (The hypothesis of equal importance)**.** *Reward and information collection are equally important for survival.*

Conjecture 2 (The curiosity trick)**.** *Curiosity is a sufficient solution for all exploration problems (where learning is possible).*

We again wish we had a Bellman solution. We can use these conjectures to arrive at one. If reward and information value are equally important then the Bellman answer is to choose which value is larger for any given state and time. If curiosity is sufficient, then we can use this Bellman rule and pursue information value greedily during exploration confident that we will also arrive at a reward collection policy. The info-reward value function $V_{\pi_{ER}}$ for a given policy is defined by a series of local policy maxima,

$$V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\left[\sum_{t=0}^{T} \max[E_t, R_t] \;\middle|\; S_t = S\right] \qquad [4]$$

Having already established we have a Bellman policy for E and R then we can directly define the Bellman equation for $V_{\pi_{ER}}$,

$$V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\left[\max[E_t, R_t] + V_{\pi_{ER}}(S_{t+1}) \;\middle|\; S_t = S, A_t = A\right] \qquad [5]$$

Because we wish to select actions that also maximize total we can substitute in our respective value functions for reward and information value. This gives us Eq.6, which in turn translates to the Bellman optimal decision policy shown in Eq.7.

$$V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\left[\max[V_{\pi_E}(S), V_{\pi_R}(S)] + V_{\pi_{ER}}(S_{t+1}) \;\middle|\; S_t = S, A_t = A\right] \qquad [6]$$

$$\pi_{ER}(S) = \begin{cases} \pi_R(S) & \text{if } V_{\pi_R}(S) \geq V_{\pi_E}(S) \\ \pi_E(S) & \text{otherwise} \end{cases} \qquad [7]$$

In our analysis of decision making we have assumed value for the next state, $V(S_{t+1})$ is readily and accurately available. In practice for most animals this is often not so, and is a key distinction between the field of dynamic programming and the more biologically sound reinforcement learning (2). If we no longer assume V_{π_E} and V_{π_R} are available but must be estimated from the environment, then we are left with something of a paradox.

We wish to use value functions to decide between information and reward value but are simultaneously estimating those values. There are a range of methods to handle this (2, 65), but we opted to further simplify π_{ER} in two ways. We shifted from using the full value functions to using only the last payout, E_{t-1} and R_{t-1}. We removed all state-dependence leaving only time dependence. We also included η to control the duration of exploration. This gives our final equation, a simple win-stay lose-switch (WSLS) rule,

$$\tilde{\pi}_{ER}(S) = \begin{cases} \pi_R(S) & \text{if } R_{t-1} \geq E_{t-1} - \eta \\ \pi_E(S) & \text{otherwise} \end{cases} \qquad [8]$$

To solve the dilemma we think of π_R and π_E as two "players" in a game that is played for behavioral control (66). We feel this approach has several advantages. It is myopic and therefore can optimally handle nonlinear changes in either the reward function or environmental dynamics (67). While in stationary settings, its regret is bounded (68). It has a long history as a useful sampling algorithm and can approximate Bayesian inference (69). In game theory settings, or other settings with multiple agents, WSLS can lead to cooperative behavior (70). Further, WSLS has a long history in psychological research where it predicts and describes human behavior (71). But most of all, this heuristic is simple to implement and should therefore scale well to the complex natural world.

The remainder of this paper is devoted to studying Eq. 8 in simulated bandit and foraging tasks.

Information collection - simulations. The uncertainty of carrying out exploration is often met by introducing uncertainty in the exploration algorithm via noisy behavior. Our Bellman solution (Eq 3) has instead argued that the most valuable curious search comes from a deterministic greedy algorithm. This is at odds with how exploration is generally viewed, as we have said. To confirm our deterministic method is best therefore we examine a multi-armed information task (Task 1; Figure 3a). This variation of a bandit task (2) replaces rewards with information. In this case, colors. Task 1 returns no environmental rewards, in fact. On each selection, the agent saw one of two colors returned according to specific probabilities (Figure 3a). Colors returned with a uniform probability, as in action 3 , maximize entropy and so offer the most potential information value (Figure 3a).

In Figure 4 we compare ours to a noisy version of curious exploration. As predicted, deterministic exploration generated more information value, in less time when compared against a (slightly) random agent using a curiosity-directed stochastic search.

It also follows from Ax. 3 and Ax. 4 that E-exploration will visit every state at least once in a finite time, assuming $E_0 > 0$. As we stop exploration when $E < \eta$ only states in which there is more to be learned will be revisited. This ensures efficient exploration.

Reward collection - simulations.

Bandits. Can our method solve reward collection problems better than other approaches? To find out we measured the total reward collected in several bandit tasks (Figure 3b-g. We considered eight agents including ours (Table 2). Each agent differed only in their exploration strategy, and examples of these strategies are shown in Figure. 5. In these examples our method has two distinguishing properties. First, exploration is highly structured and often sequential. This follows from our deterministic approach. Second, periods of exploration and exploitation are more clearly defined than the other agents.

Note our algorithm is denoted in green throughout, where E is defined using a Bayesian information gain (72) which we abbreviate as "IG". (We consider other memory definitions later on.)

Example reward time courses for Task 2 are shown in Figure. 6. During successful learning, reward time courses are more tightly clustered than the other agents we considered. Though

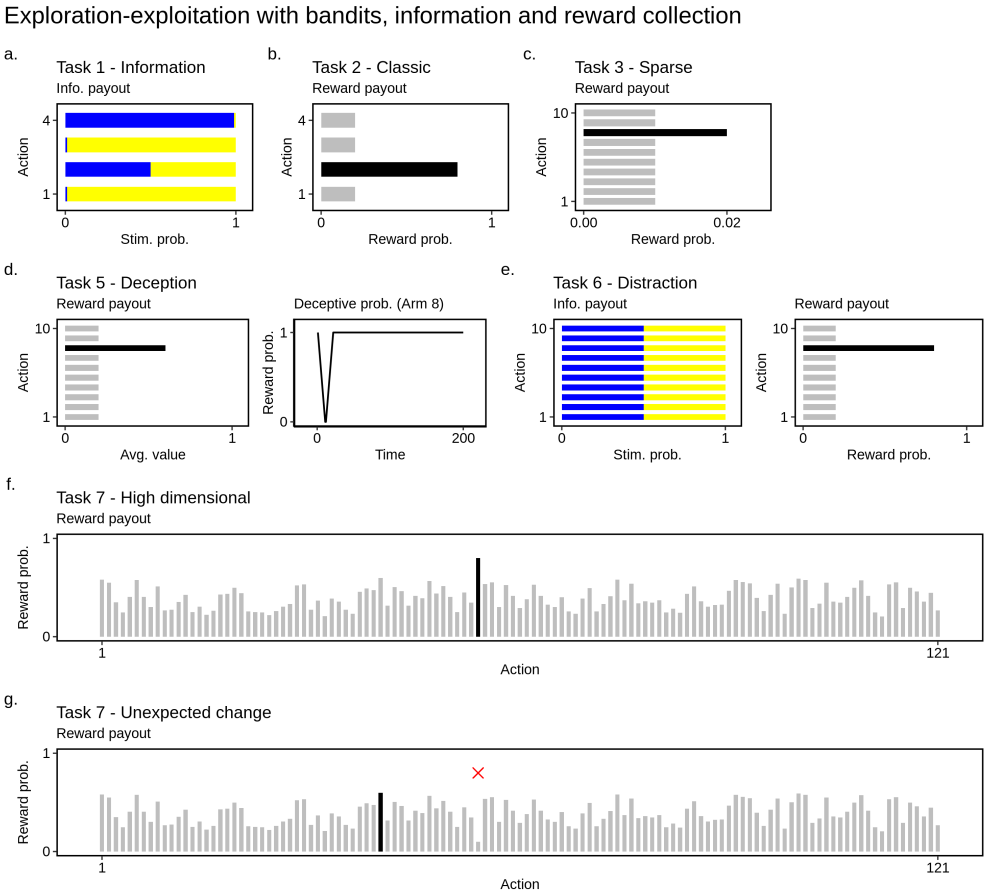

Fig. 3. Multi-armed bandits, illustration and payouts. On each trial the agent must take on n actions. Each action generates a payout. Payouts can be information, reward, or both. For comments on general task design. **a.** A 4 choice bandit for information collection. In this task the payout is information, a yellow or blue "stimulus". A good agent should visit each arm, but quickly discover that only arm two is information bearing. **b.** A 4 choice design for reward collection. The agent is presented with four actions and it must discover which choice yields the highest average reward. In this task that is Choice 2. **c.** A 10 choice sparse reward task. Note the very low overall rate of rewards. Solving this task with consistency means consistent exploration. **d.** A 10 choice deceptive reward task. The agent is presented with 10 choices but the action which is the best in the long-term (>30 trials) has a lower value in the short term. This value first declines, then rises (see column 2). **e.** A 10 choice information distraction task. The agent is presented with both information and rewards. A good agent will realize the information does not predict reward collection, and so will ignore it. **f.** A 121 choice task with a complex payout structure. This task is thought to be at the limit of human performance. A good agent will eventually discover choice number 57 has the highest payout. **g.** This task is identical to *a.*, except for the high payout choice being changed to be the lowest possible payout. This task tests how well different exploration strategies adjust to simple but sudden change in the environment.

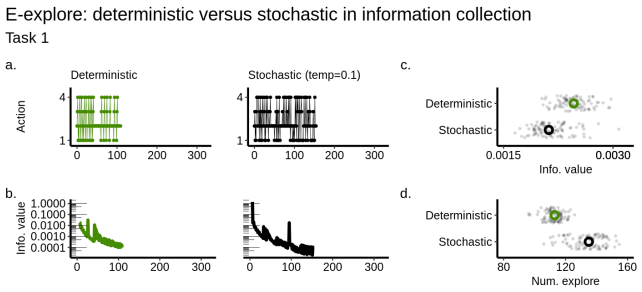

Fig. 4. Comparing deterministic versus stochastic variations of the same curiosity algorithm in Task 1. Deterministic results are shown in the left column and stochastic results are shown in the right. **a**. Examples of exploration behavior. **b**. Information value plotted with time. Matches the behavior shown in a. Large values are prefered. **c**. Average information value for 100 independent simulations. er of steps it took to reach the stopping criterion in c., e.g. the boredom threshold eta (described below). Smaller numbers of steps imply a faster search.

Table 2. Exploration strategies for bandit agents.

Name	Class	Exploration strategy
E-explore	Info. val.	Deterministic max. of information value
Random	Random	Random exploration
Reward	Mixed	Softmax sampling of reward
Reward+Info.	Mixed	Random sampling of reward + β information value
Reward+Novelty	Mixed	Random sampling of reward + β novelty signal
Reward+Entropy	Mixed	Random sampling of reward + β action entropy
Reward+EB	Mixed	Random sampling of reward + β visit counts
Reward+UCB	Mixed	Random sampling of reward + β visit counts

in ours and all other strategies we see a few experiments whose reward traces which trend downward in time. (Figure6).

When quantifying the total reward we noticed three critical results. First, despite exploration in our method never optimizing for reward value, we matched, and often outperformed, standard approaches. These all do rely on reward value at least in part. Second, when properly tuned our method was robust to irrelevant information, and robust to deceptive reward value. Third, our method seems extremely robust towards model (mis)tuning. To highlight these we divided our bandit results into three figures.

In Figure 7 present overall performance in what we considered standard bandit tasks–simple, sparse, high dimensional and non-stationary. Overall, our approach matches the second best strategy and exceeds all the others. This overall pattern was reflected in individual task performance as well (Figure 7b-f).

In Figure 8 we focus on nonstandard bandits designed to highlight and test the limits of curious exploration. In Figure 8a we considered deception. Reinforcement learning algorithms struggle when reward value is deceptive (35, 50), defined as a temporary decrease in reward value in what will ultimately be the optimal choice. Curiosity or novelty search does not struggle here because the short-term loss of value is novel, and so information filled. As such we created a simple deceptive reward bandit, where the best choice had a brief 20 step decline (Figure 3d). All other exploration strategies produced little better than chance performance. While in Figure 8b we consider distraction because curious behavior is prone to distraction. In this task the information payout for all arms was the same, and set to with even probability. This arrangement serves as the strongest possible distraction, and tests our method's ability to avoid excess exploration when the information is not predictive of reward value. Our method collected identical amounts of reward as other agents, which is the best outcome possible.

In Figure 9 we tested the robustness of our algorithm and all other exploration strategies. We re-examined total rewards collected across all possible model tunings or hyperparameters (Methods). Here E-explore was markedly better with substantially higher total reward when we integrate over all parameters (Figure 9a) and when we consider the tasks individually (Figure 9b-f).

Alternate memory. The axioms and derivation for E go to some length to define value independent of the learning and memory model. It is important therefore to demonstrate this generality in action. In the above simulations we used a single kind of Bayesian memory. We measured E using a single distance, the KL divergence (see *Methods*). This is the "info gain", or IG formulation, which has proven so useful in predicting human and other animal's behavior (6, 72). However there is no reason to prefer this overall approach *for all possible problems* (73).

In Figure 10 we compared reward collection performance on two example bandits across five different kinds of learning, memory and distance. In both simple (Task 2) and high-dimensional problems (Task 6) this variety of models was nearly interchangeable. Based on this we do not however argue the choice of memory and distance is arbitrary. No free lunch arguments prevent that (73, 74). Instead, our takeaway for Figure 10 is that a broad range of E definitions may work. In describing animal behavior the theoretical practitioner should choose definitions of learning and memory that seem most practical for the problem at hand knowing there may be several good options.

Foraging. Bandit tasks are a simple means to study the explore-exploit problem space but in an abstract form. They do not adequately describe more natural settings where there is a physical space to traverse and explore. Success of E-exploration on bandits is not an assurance of success in spatial domains.

In Figure 11 we consider a foraging task defined on a 2d space or grid. In our foraging task for any position (x, y) there is a noisy "scent" which is "emitted" by each of 20 randomly placed reward targets. A map of example targets is shown in Figure 11a. The scent is a 2d Gaussian function centered at each target, corrupted by 1 standard deviation of noise (not shown).

In Figure 11b-d we show examples of foraging behavior by 4 agents (Table 3). Two of them are learning agents, ours and a mixed value reinforcement learning model. One is a chemotaxis agent. The other is a random search agent. Figure 11b shows foraging behavior before any learning. Figure 11c shows behavior halfway through the 100 experiments we considered. Figure 11d represents an example of the final behaviors. Notice how over this series ours, and the mixed value agent, learn to

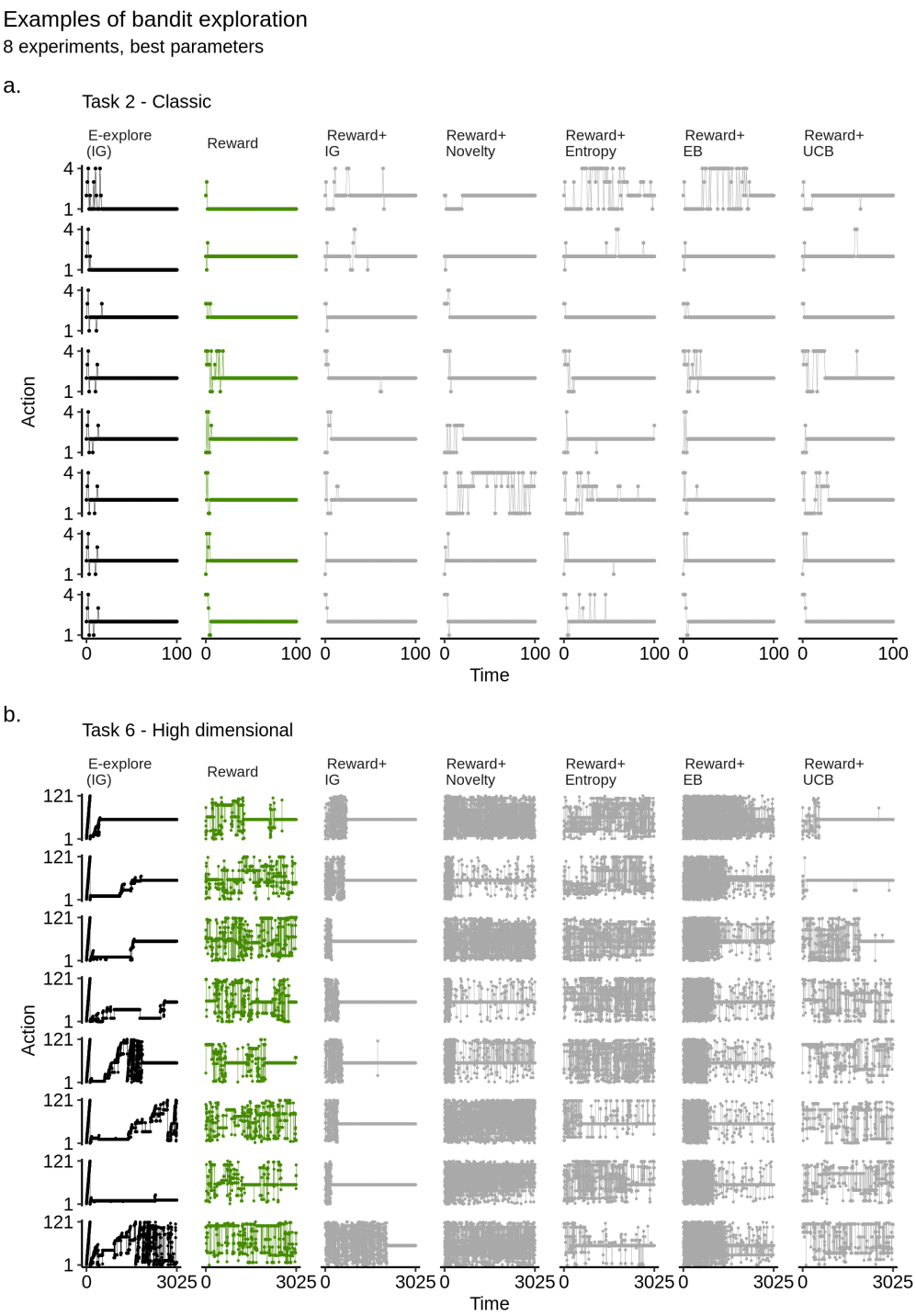

Fig. 5. Exploration behavior on two example tasks. **a**. Task 2. **b**. Task 6.

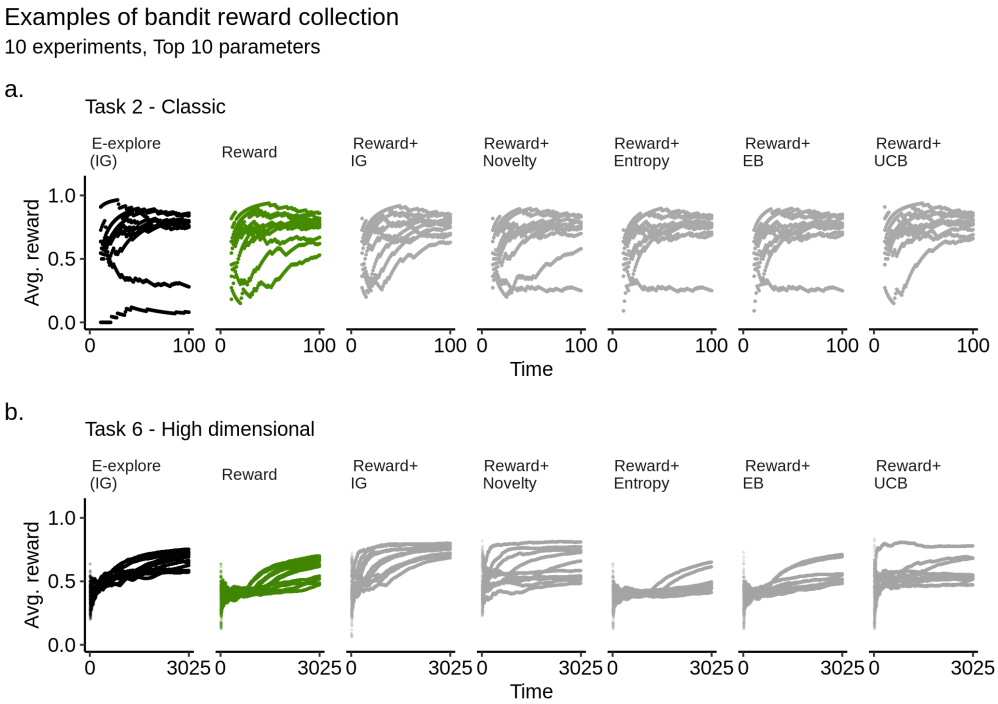

Fig. 6. Reward collection timecourses on two example tasks. **a**. Task 2. **b**. Task 6.

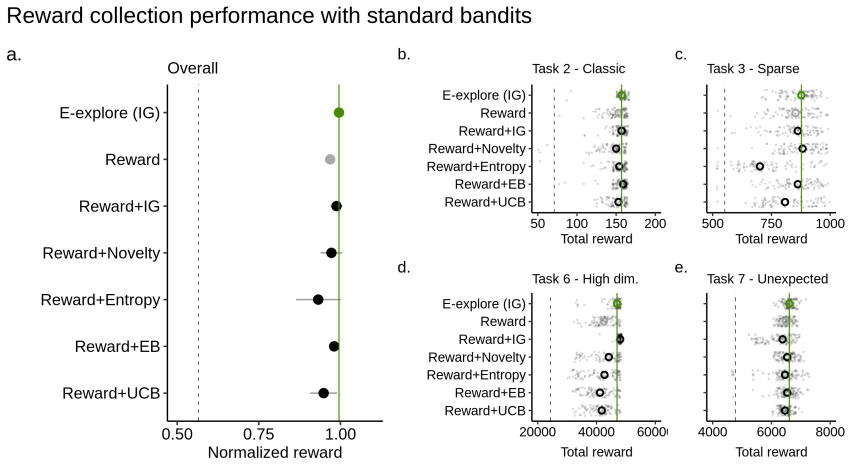

Fig. 7. Reward collection in standard bandits. **b.** Overall results. Total reward collected for each of the four tasks was normalized. Dot represents the median value, error bars represent the median absolute deviation between tasks (MAD). **b.** Results for Task 2, which has four choices and one clear best choice. **c.** Results for Task 3, which has 10 choices and very sparse positive returns. **d.** Results for Task 6, which has 121 choices and a quite heterogeneous set of payouts but still with one best choice. **e.** Results for Task 7, which is identical to Task 6 except the best choice was changed the worst. Agents were pre-trained on Task 6, then tested on 7. In panels b-e, grey dots represent total reward collected for individual experiments while the large circles represent the experimental median.

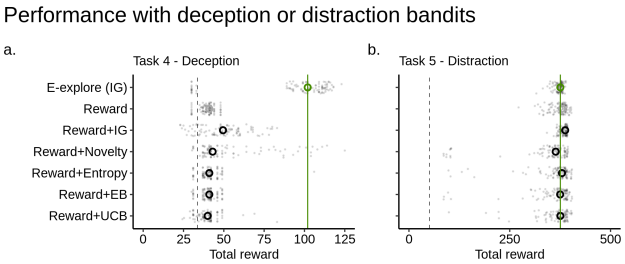

Fig. 8. Reward collection in bandits with deception or distraction. **a.** Results for Task 4, whose best choice is initially "deceptive" in that it returns suboptimal reward value over the first 20 trials. **b.** Results for Task 5, which blends the information foraging Task 1 with a larger version of Task 2 however the information is not predictive of reward payout, and so acts as a curious distraction. Grey dots represent total reward collected for individual experiments while the large circles represent the experimental median.

Table 3. Exploration strategies for foraging agents.

Name	Class	Exploration strategy
E-explore	Info. val.	Deterministic max. of information value
Diffusion	Random	Random exploration
Chemotaxis	Obs.	Sense gradient + Random exploration
Reward+Info.	Mixed	Random sampling of reward + β information

focus their foraging to a small targeted area as is best for this task.

In Figure 11e we show the reward acquisition time course averaged across all learning episodes. This measures both a learning progress and final task performance. Both reward learning agents plateau to the max value of 200, but ours (green) does so more quickly. In contrast, neither the chemotaxis or random agents show a change in profile. The takeaway for Task 8 is our method does scale well to 2d environments and can, at least in this one task, outperform standard approaches here as well.

Discussion

When information is more valuable than rewards, be curious, otherwise seek rewards. Arguing the dilemma can be resolved by this strategy is a controversial claim. We will address potential criticisms as a series of questions and answers.

Why curiosity? Our use of curiosity rests on three well established facts. 1. Curiosity is a primary drive in most, if not all, animal behavior (29). 2. Curiosity is as strong, if not sometimes stronger, than the drive for reward (9, 25, 30). 3. Curiosity as an algorithm is highly effective at solving difficult optimization problems (8, 33–40).

What about information theory? In short, the problems of communication and value are different issues that require different theories.

Weaver (75) in his classic introduction to the topic, describes information as a communication problem with three levels. **A.** The technical problem of transmitting accurately. **B.** The semantic problem of meaning. **C.** The effectiveness problem of using information to change behavior. He then describes how Shannon's work addresses only the problem **A.**, which let's information theory find broad application.

Valuing information is not, at its core, a communications problem. It is a problem in personal history. Consider Bob and Alice who are having a phone conversation, and then Tom and Bob who have the exact same phone conversation. The personal history of Bob and Alice will determine what Bob learns in their phone call, and so determines what he values in that phone call. Ro so we argue. This might be different from what Bob learns when talking to Tom, even when the conversations was identical. What we mean to show by this example is that the personal history, also known as past memory, defines what is valuable on a channel. We have summarized this diagrammatically, in Figure 12.

Finally, we argue there is an analogous set of levels for information value as those Weaver describes for information. There is the technical problem of judging how much was learned. This is the one we address, like Shannon. There is the semantic problem of what this learning "means" and also what its consequences are. There is also the effectiveness problem of using what was learned to some other effect.

Is this too complex? Perhaps turning a single objective into two, as we have, is too complex an answer. If this is true, then it would mean our strategy is not a parsimonious solution to the dilemma. We could or should reject it on that alone?

Questions about parsimony can be sometimes resolved by considering the benefits versus the costs of added complexity. The benefits to E-explore is that it leads to an optimal value solution to the exploration versus exploitation trade-off. At the same time curiosity-as-exploration can build a model of the environment (76), useful for later planning (77, 78), creativity, imagination (79), while also building diverse action strategies (36, 37, 80, 81).

Is this too simple? So have we "cheated" by changing the dilemma problem in the way we have?

The truth is we might have cheated, in this sense: the dilemma might have to be as hard as it has seemed in the past. But the general case for curiosity is clear and backed up by the brute fact of its widespread presence in the animal kingdom. The question is: is curiosity so useful and so robust that it is sufficient for all exploration with learning (35).

The answer to this question is empirical. If our account does well in describing and predicting animal behavior, that would be some evidence for it (7, 9–12, 15, 30, 82–84). If it predicts neural structures (19, 85), that would be some evidence for it. If this theory proves useful in machine learning, that would be some evidence for it (8, 35, 37, 38, 45, 47, 80, 84, 86, 87). In other words how simple, complex, or parsimonious an theoretical idea comes down to usefulness.

Is this a slight of hand, theoretically? Yes. We have taken one problem that cannot be solved, and replaced it with another related problem that can be. In this replacement we swap one animal behavior, reward seeking, for another, information seeking.

Does value as a technical problem even make sense? Having a technical definition of information value, free from any meaning, might seem counterintuitive for any value measure. We suggest it is not any more or less counterintuitive than stripping information of meaning.

What about mutual information, or truth? In other prototypical examples, information value comes from errors or how well what learning corresponds to the environment (88–90). Colloquially, one might call this "truth seeking".

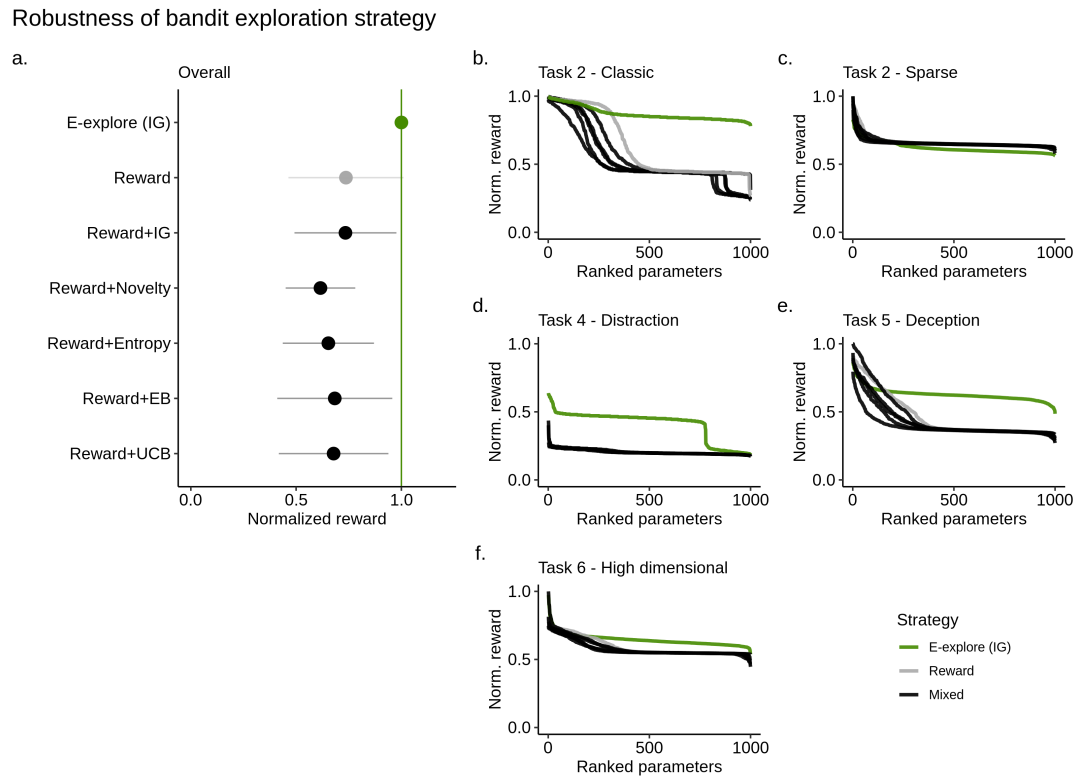

Fig. 9. Exploration parameter sensitivity. **a**. Integrated total reward (normalized) across 1000 randomly chosen exploration hyperparameters. Dots represent the median. Error bars are the median absolute deviation. **b-f**. Normalized total reward for exploration 1000 randomly chosen parameters, ranked according to performance. Each subpanel shows performance on a different task. Lines are colored according to overall exploration strategy - E-explore (ours), reward only, or a mixed value approach blending reward and an exploration bonus).

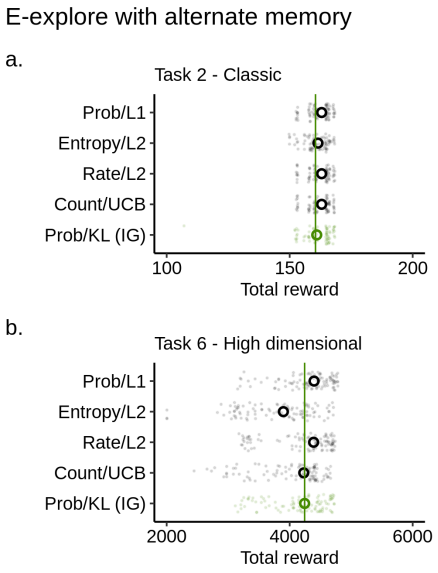

Fig. 10. Reward collection with alternative memory models and distances, akin to other kinds of curiosity. See *Methods* for full details. **a.** Results for Task 2. **b.** Results for Task 6. Grey dots represent total reward collected for individual experiments while the large circles represent the experimental median.

As a people we pursue information which is fictional (91), or based on analogy (92), or outright wrong (93). Conspiracy theories, disinformation, and more mundane errors, are far too commonplace for information value to rest on mutual information or error. This does not mean that holding false beliefs cannot harm survival but this is a second order question as far as information value goes. The consequences of error on information value come after the fact.

So you suppose there is always a positive value for learning of all fictions, disinformation, and deceptions? We must. This is a most unexpected prediction. Yet sit seems consistent with the behavior of humans, and animals alike. Humans do consistently seek to learn falsehoods. Our axioms suggest why this is rational.

Was it necessary to build a general theory for information value to describe curiosity? No. We made an idealistic choice that worked out. The field of curiosity studies has shown that there are many kinds of curiosity (7–9, 23, 25, 27, 28, 30, 33, 34, 39, 47, 80, 94, 95). At the extreme limit of this diversity is a notion of curiosity defined for any kind of observation, and any kind of learning. This is what we offer. At this limit we can be sure to span fields, addressing the problem of information value as it appears in computer science, machine learning, game theory, psychology, neuroscience, biology, economics, among others.

What about other models of curiosity? Curiosity has found many specific definitions (30, 44). Curiosity has been described as a prediction error, by learning progress (8, 96). Schmidhuber (8) noted the advantage of looking at the derivative of errors, rather than errors directly. Itti (97) and others (27, 72, 98, 99) have taken a statistical and Bayesian view often using the KL divergence to estimate information gain or Bayesian surprise. Other approaches have

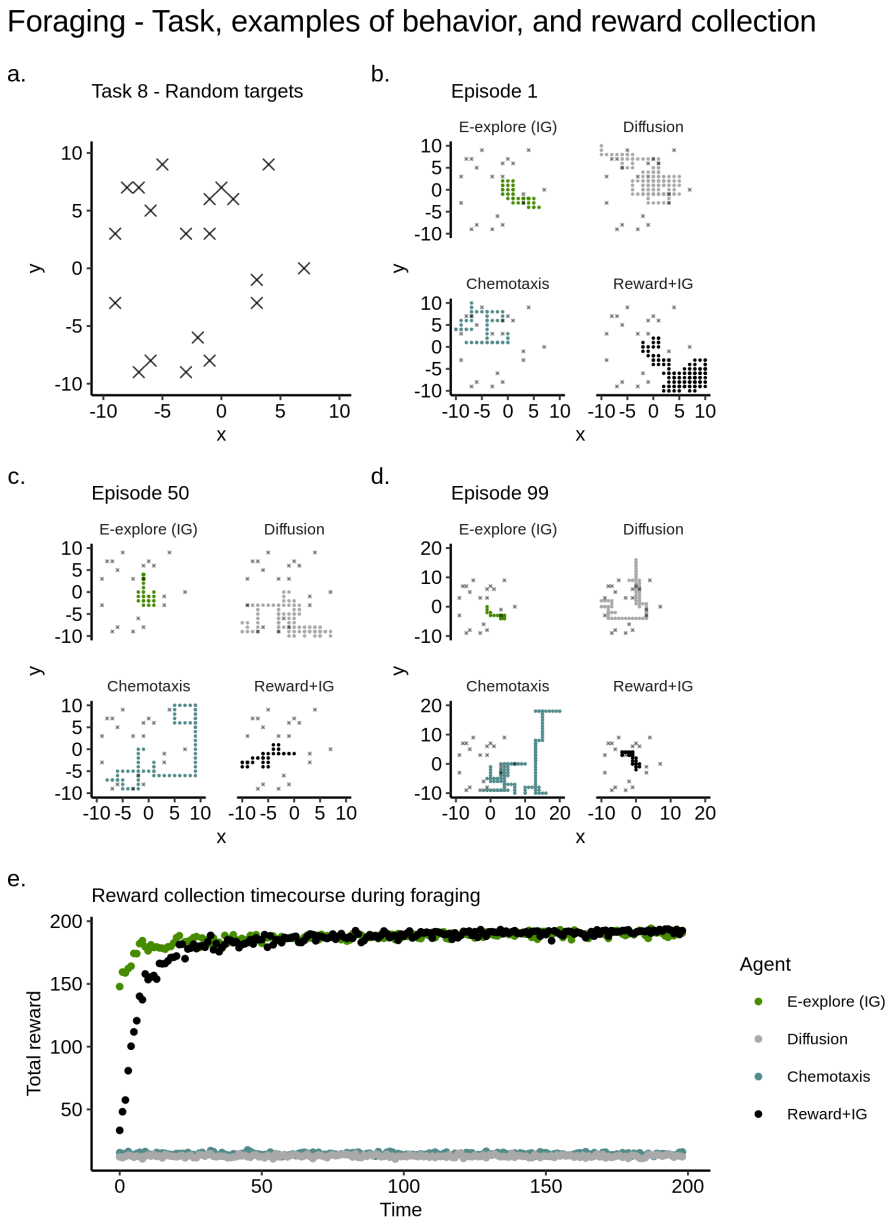

Fig. 11. Reward collection during foraging (Task 8). See *Methods* for full details. **a.** Illustration of a 2 dimensional reward foraging task. **b-d.** Examples of agent behavior, observed over 200 time steps. a. Without any training. b. After 50 training episodes. c. After 99 training episodes. **e.** Total rewards collected over time, median value of 100 independent simulations.

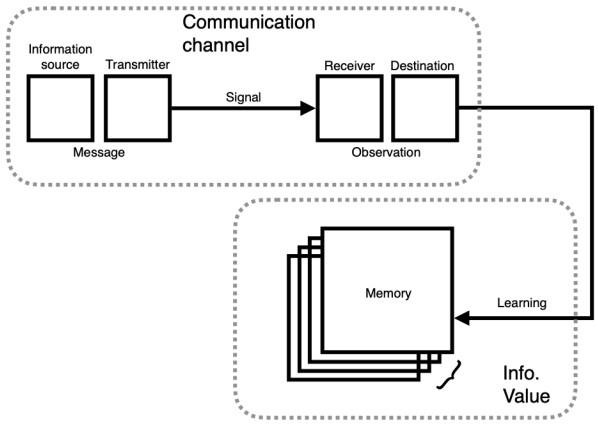

Fig. 12. The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel directly. Value comes from learning about observations taken from the channel, which in turn depends on memory.

been based on adversarial learning, model disagreement (33), or model compression (27). Some measures focused on past experience (26, 76, 100). Others focused on future planning and imagination (101–103). Graphical models of memory have been useful (104).

What distinguishes our approach to information value is we focus on memory dynamics, and base value on some general axioms. We try to embody the idea of curiosity "as learning for learning's sake", not done for the sake of some particular goal (105), or for future utility (106). Our axioms were however designed with all these other definitions in mind. In other words, we do not aim to add a new metric. We aim to generalize the others.

Is information a reward? If reward is any quantity that motivates behavior, then our definition of information value is a reward. This last point does not mean that information value and environmental rewards are interchangeable. Rewards from the environment are a conserved resource, information is not. For example, if a rat shares a potato chip with a cage-mate, it must break the chip up leaving it less food for itself. While if a student shares an idea with a classmate, that idea is not divided up. It depends, in other words.

But isn't curiosity impractical? It does seem curiosity can lead away from a needed solution as towards it. Consider children, who are the prototypical curious explorers (9, 11). This is why we focus on the derivative of memory, and limit curiosity with boredom, and counter curiosity with a drive for reward collecting (i.e., exploitation).

Consider how science and engineering can interact. Science is an open-ended inquiry, and engineering often is a purpose driven enterprise. They each have their own pursuits, but they also learn from each other, often in alternating iterations (1). Their different objectives though is what makes them such good long-term collaborators.

What is boredom, besides being a tunable parameter? A more complete version of the theory would let us derive a useful or even optimal value for boredom if given a learning problem or environment. We cannot do this. It is the next problem we will work on, and it is important.

Does this mean you are hypothesizing that boredom is actively tuned? Yes we are predicting that.

But can animals tune boredom? Geana and Daw (64) showed this in a series of experiments. They reported that altering the expectations of future learning in turn alters self-reports of boredom. Others have shown how boredom and self-control interact to drive exploration (62, 63, 107).

Do you have any evidence in animal behavior and neural circuits? There is some evidence for our theory of curiosity in psychology and neuroscience, however, in these fields curiosity and reinforcement learning have developed as separate disciplines (2, 7, 9). Indeed, we have highlighted how they are separate problems, with links to different basic needs: gathering resources to maintain physiological homeostasis (108, 109) and gathering information to decide what to learn and to plan for the future (2, 56). Here we suggest that though they are separate problems, they are problems that can, in large part, solve one another. This insight is the central idea to our view of the explore-exploit decisions.

Yet there are hints of this independent cooperation of curiosity and reinforcement learning out there. Cisek (2019) has traced the evolution of perception, cognition, and action circuits from the Metazoan to the modern age (85). The circuits for reward exploitation and observation-driven exploration appear to have evolved separately, and act competitively, exactly the model we suggest. In particular he notes that exploration circuits in early animals were closely tied to the primary sense organs (i.e. information) and had no input from the homeostatic circuits (85, 108, 109). This neural separation for independent circuits has been observed in some animals, including zebrafish (110) and monkeys (12, 111).

Is the algorithmic run time practical? Computer scientists often study the run time of an algorithm as a measure of its efficiency. The worst case algorithmic run time of our method is linear and additive in the independent policies. If it takes T_E steps for π_E to converge, and T_R steps for π_R, then the worst case run time for π_{ER} is $T_E + T_R$.

What predictions does this theory make? We have four strong predictions to make:

1. **Invariant exploration strategy**. The presence of future reward should not change the exploration strategy. As shown in, (112). *Auxiliary assumptions*. Learning is possible because there is some novelty ($E > \eta$), and some observations available to learn from.

2. **Information preference following reward absence**. Animals are always maximizing the sum of reward and information. When reward is absent, an animal should immediately choose to do an information collection with it's next action. *Auxiliary assumptions*: Learning is possible, because there is some novelty ($E > \eta$), and some observations available to learn from. We ignore "maintenance" behaviors like grooming, consummatory, etc.

3. **Deterministic search**. It is possible to predict exactly exploratory choices made during a curious search, because the optimal search policy is strictly deterministic. To evaluate this we must move from population-level distribution analysis of behavior in foraging and bandit tasks, to moment-by-moment prediction of animal behavior (113, 114). *Auxiliary assumptions*: The noise in the neural circuits which implements behavior is sufficiently

weak so it does add unaccounted for variation in the behavior itself.

4. **Well defined pure-periods**. There should be well-defined periods of exploration and exploitation and the pattern of exploration should not favor reward value over learning value. In other words, fitting a hidden markov model to the action space should yield only two hidden states–one for all exploration and one for exploitation. For example, (115).

Summary. We argue here that one should put aside any intuitions that suggest a curious search is too inefficient to be practical in solving reward collection problems. We have shown it can be practical, and there is an optimal value way to solve the dilemma. Be greedy and curious.

Methods and Materials

Tasks. Bandit tasks did not distinguish between states and actions, and the number of actions is equal to the number of arms. Each arm/action is a distinct payout function. Payouts for the foraging task are described below.

We studied seven tasks, each designed to test exploration in a common setting or test the limits of curiosity. Despite their differences each task had the same basic structure. On each trial there were a set of n choices, and the agent should try and learn the best one. Each choice action returns a "payout" according to a predetermined probability. Payouts are information, reward, or both (Figure 3). Note that information was symbolic, denoted by a color code, "yellow" and "blue" and as is custom reward was a positive real number.

Task 1 was designed to examine information foraging. At no point does this task generate rewards. There were four choices. Three of these generated either a "yellow" or "blue" symbol, with a probability of $p = 0.99$. These have near zero entropy, and as result are an information poor foraging choice. The final choice had an even probability of generating either "yellow" of "blue". This was the maximum entropy choice, The ideal choice for information foraging. See Figure 3**a**.

Task 2 was designed to examine reward collection, in a setting very common in the decision making literature. At no point does the task generate symbolic information. Rewards were 0 or 1. There were again four choices. The best choice had a payout of $p(R = 1) = 0.8$. This is a much higher average payout than the others ($p(R = 1) = 0.2$). See Figure 3**b**.

Task 3 was designed with very sparse rewards (116, 117). There were 10 choices. The best choice had a payout of $p(R = 1) = 0.02$. The other nine had, $p(R = 1) = 0.01$ for all others. See Figure 3c.

Task 4 was designed with deceptive rewards. By deceptive we mean that the best long-term option presents itself initially with lower value. The best choice had a payout of $p(R > 0) = 0.6$. The others had $p(R > 0) = 0.4$. Value for the best arm dips, then recovers. This is the "deception" It happens over the first 20 trials. Rewards were real numbers, between 0-1. See Figure 3d.

Task 5 designed to distract our curiosity algorithm with irrelevant information. Like task 1, there is one best option ($p(R = 1) = 0.8$) and 9 others ($p(R = 1) = 0.2$). In addition to reward, the task generated information. Information was max entropy and uncorrelated with rewarding payouts. Every choice had an even probability of generating either "yellow" of "blue". See Figure 3e.

Tasks 6-7 were designed with 121 choices, and a complex payout structure. Tasks of this size are at the limit of human performance (118). We first trained all agents on *Task 6*, whose payout can be seen in Figure3f-g. This task, like the others, had a single best payout $p(R = 1) = 0.8$. After training for this was complete, final scores were recorded as reported, and the agents were then challenged by *Task 7*. *Task 7* was identical except that the best option was changed to be the worst $p(R = 0) = 0.2$ (Figure 3f).

Task 7 was designed to examine how robust reward collection is to unexpected changes in the environment. It has a payout structure that was identical to *Task 6*, except the best arm was changed to be the worst (($p(R = 1) = 0.8$ became $p(R = 1) = 0.1$). Learners which have a good model of the other choices' values should recover more quickly leading to a greater degree of reward collected. Figure 3g).

Task 8 was a 2d foraging task (Figure 11a) where 20 renewing targets were randomly placed in a (20, 20) unit grid. Each target "emitted" a 2d Gaussian "scent" signal with a 1 standard deviation width. Agents (described below) began at the (0,0) center position and had 4 possible actions. They could move *(up, down, left, right)*. If the agent reached a target it would receive reward value. All other positions generate 0 rewards. Each reward position would continue to generate reward as long as the agent remained in that position. Odors from different targets that happened to overlap were additively combined. The simulated sensing machinery for each agent was designed by us to be noisy and so added a 1 standard deviation or random Gaussian noise to each measurement. Each foraging simulation was run for 200 steps, which as a result sets the upper bound for total rewards collected in this task at 200.

Agents. We considered two kinds of specialized agents. Those suited to solve bandit tasks and those suited to solve foraging tasks. Though, the base exploration algorithms at work were in part shared between these two classes of tasks.

Bandit: E-explore - Our algorithm. It pursued either pure exploration based on E maximization or pure exploitation based on pure R maximization. All its actions are deterministic and greedy. Both E and R maximization was implemented as an actor-critic architecture (2) where value updates were made in crtic according to Eq. 22. Actor action selection was governed by,

$$A_t = \text{argmax}\, A \in \mathbb{A} = Q(., A_t) \qquad [9]$$

where we use the "." to denote the fact our bandits have no meaningful state.

This agent has two parameters, the learning rate α (Eq. 22) and the boredom threshold η (Eq. 8). It's payout function was simply, $G = R_t$ (Eq. 22).

Note that no matter which policy is in control in Eq. 8 each policy can learn from the other policies' observations. That is, while the two policies compete for control they also learn from each other.

Also note when we compare reward R and information value E as above we have not ensured they have a common scale. We use some auxiliary assumptions to sidestep this potential problem. We limit the probability of having zero reward to

be, itself, non-zero. That is, $p(R_t = 0) > 0$. We also limit E_0 to be positive and non-zero, $(E_0 - \eta) \geq 0$. These constraints are enough to ensure exploration will dominate exploitation at a rate inversely proportional to rewards received.

Bandit: Reward. An algorithm whose objective was to estimate the reward value of each action, and stochastically select the most valuable action in an effort to maximize total reward collected. It's payout function was simply, $G = R_t$ (Eq. 22). Similar to the E-explore agent, this agent used actor-critic, but its actions were sampled from a softmax / Boltzman distribution,

$$p(A_t) = \frac{e^{\gamma Q_R(.,A_t)}}{\sum_{A \in \mathbb{A}} e^{\gamma Q_R(.,A)}} \quad [10]$$

Where γ is the "temperature" parameter. Large γ generate more random actions. This agent has two parameters, the learning rate α (Eq. 22) and the temperature $\gamma > 0$ (Eq. 10).

Bandit: Reward+Info. This algorithm was based on the Reward algorithm defined previously but it's reward value was augmented or mixed with E. Specifically, the information gain formulation of E described below. It's payout is, $G_t = R_t + \beta E_t$ (Eq. 22). Critic values were updated using this G, and actions were selected according to Eq. 10. This agent has three parameters, the learning rate α (Eq. 22) the temperature γ (Eq. 10) and the exploitation weight $\beta > 0$. Larger values of β will tend to favor exploration.

Bandit: Reward+Novelty. This algorithm was based on the Reward algorithm defined previously but it's reward value wass augmented or mixed with a novelty/exploration bonus (Eq. 11).

$$B_{A_t} = \begin{cases} B & \text{if } A_t \text{ not it } \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \quad [11]$$

Where $\mathbb{Z}$ is the set of all actions that agent has taken so far in a simulation. Once all actions have been tried, $B = 0$.

This agent's payout is, $G_t = R_t + B_t$ (Eq. 22). Critic values were updated using this G and actions were selected according to Eq. 10. This agent has three parameters, the learning rate α (Eq. 22), the temperature γ (Eq. 10), and the bonus size $B > 0$.

Bandit: Reward+Entropy. This algorithm was based on the Reward algorithm defined previously but it's reward value was augmented or mixed with an entropy bonus (Eq. 12). This bonus was inspired by the "softactor" method common to current agents in the related field of deep reinforcement learning (119).

$$H(A_t) = \sum_{A \in \mathbb{A}} p(A) \log p(A) \quad [12]$$

Where $p(A)$ was estimated by a simple normalized action count.

This agent's payout is, $G_t = R_t + \beta H_t$ (Eq. 22). Critic values were updated using this G, and actions were selected according to Eq. 10. It has three free parameters, the learning rate α (Eq. 22), the temperature γ (Eq. 10), and the exploitation weight $\beta > 0$.

Bandit: Reward+EB. This algorithm was based on the Reward algorithm defined previously but it's reward value was augmented or mixed with an evidence bound (EB) statistic (120) (Eq 13).

$$\text{EB}(A) = \sqrt{C(A)} \quad [13]$$

Where $C(A)$ is a running count of each action, $A \in \mathbb{A}$. This agent's payout is, $G_t = R_t + \beta$ EB (Eq. 22). Critic values were updated using this G, and actions were selected according to Eq. 14.

$$A_t = \text{argmax}\, A \in \mathbb{A} = Q(., A_t) + \beta \text{ EB}(A) \quad [14]$$

It has three two parameters, the learning rate α (Eq. 22) and the exploitation weight $\beta > 0$.

Bandit: Reward+UCB. This algorithm was based on the Reward algorithm defined previously, but it's reward value was augmented or mixed with an upper confidence bound (UCB) statistic (120) (Eq 15).

$$\text{UCB}(A) = \frac{2 \log(N_A + 1)}{\sqrt{C(A)}} \quad [15]$$

Where $C(A)$ is a running count of each action, $A \in \mathbb{A}$ and N_A is the total count of all actions taken. This agent's payout is, $G_t = R_t + \beta$ UCB (Eq. 22). Critic values were updated using this G, and actions were selected according to Eq. 16.

$$A_t = \text{argmax}\, A \in \mathbb{A} = Q(., A_t) + \beta \text{ UCB}(A) \quad [16]$$

It has three two parameters, the learning rate α (Eq. 22) and the exploitation weight $\beta > 0$.

Forage: E-explore. This algorithm was identical to *Bandit: E-explore* except for the change in action space this task required (described above).

Forage: Reward+Info. This algorithm was identical to *Bandit: Reward+Info* except for change in action space the forage task required (as described in the *Task 8* section above).

Forage: Diffusion This algorithm was designed to ape 2d brownian motion, and therefore to implement a form of random search (121). At every timestep this agent either continued in motion, if it had not traversed a previously sampled length l, or it drew a new l and random uniform direction, $d_f = U(\text{up}, \text{down}, \text{left}, \text{right})$. Movement lengths were sampled from an exponential distribution (Eq. 17).

$$p(l) = e^{\lambda} \quad [17]$$

Where λ is the length scale which was set so $\lambda = 1$ in all experiments.

Forage: Chemotaxis Movement in this algorithm was based on the Diffusion model, except it used a biased approach to decide when to turn. Specifically, it used the scent concentration gradient to bias its change in direction (Eq. 18).

$$\nabla(A) = \text{sign } (O - O') \quad [18]$$

Where O is the sense or scent observation made by the agent at its current grid position, and sign returns 1 if the argument is positive, and -1 otherwise.

Turning decisions were based on a biased random walk model, often used in -taxis behaviors (122). In this there are two independent probability thresholds, p_f and p_r. If the gradient is positive, then the p_f is "selected" and a coin flip happens, generating a random value $0 \leq x_i \leq 1$. If $p_f \geq x_i$, then the current direction d_f is maintained, otherwise a new direction is randomly chosen (see Diffusion above for details). If the gradient was negative, then the criterion $p_r \geq x_i$ is used.

In preliminary simulations we found performance was always best when, $p_f = 1$ and $p_r = 0$. As a result agents always continue in the same direction when the gradient is positive, and always turn when the gradient is negative.

Task memory for E-explore. In the majority of simulations we have used a Bayesian or Info Gain formulation for curiosity and E. Each task's observations fit a simple discrete probabilistic model, with a memory "slot" for each action for each state. Specifically, probabilities were tabulated on state reward tuples, (S, R). To measure distances in this memory space we used the Kullback–Leibler divergence (described below). This model was sometimes denoted as simple "Info." in most figures, with the exception of Figure 10. There it was denoted as "Prob/KL", for clarity.

The Kullback–Leibler divergence (KL) is a widely used information theory, and Bayesian modelling. It measures the information gained or lost by replacing one distribution with another. It used in machine learning (123), Bayesian reasoning (97), visual neuroscience (97), experimental design (124), compression (27, 60) and information geometry (125).

In our implementation $\mathbf{M}'$ corresponded to the updated memory $\mathbf{M}_{t_1}$ and $\mathbf{M}$ was a "copy" of the older memory.

$$KL(\mathbf{M}', \mathbf{M}) = \sum_{x \in X} M'(s) \log \frac{\mathbf{M}'(x)}{\mathbf{M}(x)} \qquad [19]$$

In Figure 10 we explored alternative alternative models and/or distance metrics to define E. All consistent with our axioms. Their details follow.

Prob/L1. This model used the same memory model as *Prob/KL*, but measured distances using the L1 norm,

$$L^1(M, M') = \sum_{x \in X} |\mathbf{M}(x) - \mathbf{M}'(x)| \qquad [20]$$

Entropy/L2. This model used the same memory model as *Prob/KL* except it considered the entropy (Eq. 12) of each action. Specifically, it measure differences in entropy between two memories, $\mathbf{M}'$ and $\mathbf{M}$ using a L2 norm,

$$L^2(M, M') = \sqrt{\sum_{x \in X} [\mathbf{M}(x) - \mathbf{M}'(x)]^2} \qquad [21]$$

Rate/L2. This model used the absolute value of the derivative of the reward rate, measured by a rolling average ($n = 10$). Specifically, each action had its own running average and differences between sequential averages between M' and $\mathbf{M}$ were measured using the L2 norm (Eq. 21).

Count/UCB. This model used the UCB statistic, as defined in Eq. 15.

Initializing E_0. In all simulations we assume an animal will have a uniform prior over the possible actions, $\mathbb{A}$. Thus the maximum possible information value is the maximum entropy of $\mathbb{A}$, thus $E_0 = \sum_K p(A_k) \log p(A_k)$.

Learning equations. Reward and information value learning for all agents on the bandit tasks were made using update rule below,

$$V(S) = V(S) + \alpha(G_t - V(S_t)) \qquad [22]$$

Table 4. Hyperparameter tuning - parameters and ranges.

Task	Agent	Parameter	Range
Bandit	E-explore	η	(1e-9, 1e-2)
Bandit	E-explore	α	(0.001, 0.5)
Forage	E-explore	η	(1e-9, 1e-2)
Forage	E-explore	α	(0.001, 0.5)
Bandit	Reward	γ	(0.001, 1000)
Bandit	Reward	α	(0.001, 0.5)
Bandit	Reward+Info.	γ	(0.001, 1000)
Bandit	Reward+Info.	β	(0.001, 10)
Bandit	Reward+Info.	α	(0.001, 0.5)
Forage	Reward+Info.	γ	(0.001, 1000)
Forage	Reward+Info.	β	(0.001, 10)
Forage	Reward+Info.	α	(0.001, 0.5)
Bandit	Reward+Novelty	γ	(0.001, 1000)
Bandit	Reward+Novelty	B	(1, 100)
Bandit	Reward+Novelty	α	(0.001, 0.5)
Bandit	Reward+Entropy	γ	(0.001, 1000)
Bandit	Reward+Entropy	β	(0.001, 10)
Bandit	Reward+Entropy	α	(0.001, 0.5)
Bandit	Reward+EB	β	(0.001, 10)
Bandit	Reward+EB	α	(0.001, 0.5)
Bandit	Reward+UCB	β	(0.001, 10)
Bandit	Reward+UCB	α	(0.001, 0.5)

Where $V(S)$ is the value for each state, G_t is the *return* for the current trial, either R_t or E_t, and α is the learning rate $(0 - 1]$. See the *Hyperparameter optimization* section for information on how α is chosen for each agent and task.

Value learning updates for all relevant agents in the foraging task were made using the TD(0) learning rule (2),

$$V(S_t) = V(S_t) + \alpha(G_t - V(S_{t+1}) - V(S_t) \qquad [23]$$

Hyperparameter optimization. The hyperparameters for each agent were tuned independently for each task by random search (126). Generally, we reported results for the top 10 values, sampled from 1000 possibilities. Top 10 parameters and search ranges for all agents are shown in Table 4. Parameter ranges were held fixed for each agent across all the bandit tasks.

Acknowledgments

We wish to thank Jack Burgess, Matt Clapp, Kyle "Donovank" Dunovan, Richard Gao, Roberta Klatzky, Jayanth Koushik, Alp Muyesser, Jonathan Rubin, and Rachel Storer for their comments on earlier drafts. We also wishes to thank Richard Grant for his illustration work in Figure 1.

The research was sponsored by the Air Force Research Laboratory (AFRL/AFOSR) award FA9550-18-1-0251. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. government. TV was supported by the Pennsylvania Department of Health Formula Award SAP4100062201, and National Science Foundation CAREER Award 1351748.

1. Gupta AA, Smith K, Shalley C (2006) The Interplay between Exploration and Exploitation. *The Academy of Management Journal* 49(4):693–706.
2. Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning Series. (The MIT Press, Cambridge, Massachusetts), Second edition edition.
3. Woodgate JL, Makinson JC, Lim KS, Reynolds AM, Chittka L (2017) Continuous Radar Tracking Illustrates the Development of Multi-destination Routes of Bumblebees. *Scientific Reports* 7(1).
4. Lee MD, Zhang S, Munro M, Steyvers M (2011) Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research* 12(2):164–174.

5. Schulz E, et al. (2018) Exploration in the wild, (Animal Behavior and Cognition), Preprint.
6. Calhoun AJ, Chalasani SH, Sharpee TO (2014) Maximally informative foraging by Caenorhabditis elegans. *eLife* 3:e04220.
7. Berlyne D (1950) Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology* 41(1):68–80.
8. Schmidhuber (1991) A possibility for implementing curiosity and boredom in model-building neural controllers. *Proc. of the international conference on simulation of adaptive behavior: From animals to animats* pp. 222–227.
9. Kidd C, Hayden BY (2015) The Psychology and Neuroscience of Curiosity. *Neuron* 88(3):449–460.
10. Jaegle A, Mehrpour V, Rust N (2019) Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Arxiv* 1901.02478:13.
11. Sumner ES, et al. (2019) The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss. *PsyArxiv* h437v:11.
12. Wang MZ, Hayden BY (2019) Monkeys are curious about counterfactual outcomes. *Cognition* 189:1–10.
13. Auersperg AM (2015) Exploration Technique and Technical Innovations in Corvids and Parrots in *Animal Creativity and Innovation*. (Elsevier), pp. 45–72.
14. Kelly JL (1956) A New Interpretation of Information Rate. *the bell system technical journal* p. 10.
15. Berger-Tal O, Nathan J, Meron E, Saltz D (2014) The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE* 9(4):e95693.
16. Dayan P, Sejnowski TJ (1996) Exploration bonuses and dual control. *Mach Learn* 25(1):5–22.
17. Thrun SB (1992) Eficient Exploration In Reinforcement Learning. *NIPS* p. 44.
18. Mehlhorn K, et al. (2015) Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* 2(3):191–215.
19. Kobayashi K, Hsu M (2019) Common neural code for reward and information value. *Proc Natl Acad Sci USA* 116(26):13061–13066.
20. Berlyne DE (1954) A theory of human curiosity. *British Journal of Psychology. General Section* 45(3):180–191.
21. Dubey R, Mehta H, Lombrozo T (2020) Curiosity is Contagious: A Social Influence Intervention to Induce Curiosity. *PsyArxiv*.
22. Brändle F, Wu CM, Schulz E (2020) What Are We Curious about? *Trends in Cognitive Sciences* 24(9):685–687.
23. Wang MZ, Hayden BY (2020) Curiosity, latent learning, and cognitive maps, (Neuroscience), Preprint.
24. Lydon-Staley DM, Zhou D, Blevins AS, Zurn P, Bassett DS (2021) Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nat Hum Behav* 5(3):327–336.
25. Loewenstein G (1994) The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116(1):75.
26. Schmidhuber J (1991) Curious model-building control systems in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. (IEEE, Singapore), pp. 1458–1463 vol.2.
27. Schmidhuber J (2008) Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes. *arXiv:0812.4360 [cs]*.
28. Zhou D, Lydon-Staley DM, Zurn P, Bassett DS (2020) The growth and form of knowledge networks by kinesthetic curiosity. *arXiv:2006.02949 [cs, q-bio]*.
29. Inglis IR, Langton S, Forkman B, Lazarus J (2001) An information primacy model of exploratory and foraging behaviour. *Animal Behaviour* 62(3):543–557.
30. Gottlieb J, Oudeyer PY (2018) Towards a neuroscience of active sampling and curiosity. *Nat Rev Neurosci* 19(12):758–770.
31. Gopnik A (2020) Childhood as a solution to explore–exploit tensions. *Phil. Trans. R. Soc. B* 375(1803):20190502.
32. Song M, Bnaya Z, Ma WJ (2019) Sources of suboptimality in a minimalistic explore–exploit task. *Nature Human Behaviour* 3(4):361–368.
33. Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-Driven Exploration by Self-Supervised Prediction in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. (IEEE, Honolulu, HI, USA), pp. 488–489.
34. Stanton C, Clune J (2018) Deep Curiosity Search: Intra-Life Exploration Can Improve Performance on Challenging Deep Reinforcement Learning Problems. *arXiv:1806.00553 [cs]*.
35. Fister I, et al. (2019) Novelty search for global optimization. *Applied Mathematics and Computation* 347:865–881.
36. Mouret JB, Clune J (2015) Illuminating search spaces by mapping elites. *arXiv:1504.04909 [cs, q-bio]*.
37. Colas C, Huizinga J, Madhavan V, Clune J (2020) Scaling MAP-Elites to Deep Neuroevolution. *Proceedings of the 2020 Genetic and Evolutionary Computation Conference* pp. 67–75.
38. Cully A, Clune J, Tarapore D, Mouret JB (2015) Robots that can adapt like animals. *Nature* 521(7553):503–507.
39. Schwartenbeck P, et al. (2019) Computational mechanisms of curiosity and goal-directed exploration. *eLife* (e41703):45.
40. Laversanne-Finot A, Péré A, Oudeyer PY (2018) Curiosity Driven Exploration of Learned Disentangled Goal Spaces. *arXiv:1807.01521 [cs, stat]*.
41. Bertsekas D (2017) *Dynamic Programming and Optimal Control, Vol. I*. (Athena Scientific), Fourth edition.
42. Bellmann R (1954) The theory of dynamic programming. *Bull. Amer. Math. Soc* 60(6):503–515.
43. Schmidhuber J (1991) Curious model-building control systems in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. (IEEE, Singapore), pp. 1458–1463 vol.2.
44. Oudeyer PY (2018) Computational Theories of Curiosity-Driven Learning. *arXiv:1802.10546 [cs]*.
45. Burda Y, et al. (2018) Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat]*.
46. Zhang S, Yu AJ (2013) Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. *NeurIPS* 26.
47. de Abril IM, Kanai R (2018) Curiosity-Driven Reinforcement Learning with Homeostatic Regulation in *2018 International Joint Conference on Neural Networks (IJCNN)*. (IEEE, Rio de Janeiro), pp. 1–6.
48. Zhou D, Lydon-Staley DM, Zurn P, Bassett DS (2020) The growth and form of knowledge networks by kinesthetic curiosity. *arXiv:2006.02949 [cs, q-bio]*.
49. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* 143(6):2074–2081.
50. Lehman J, Stanley KO (2011) Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19(2):189–223.
51. Velez R, Clune J (2014) Novelty search creates robots with general skills for exploration in *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation - GECCO '14*. (ACM Press, Vancouver, BC, Canada), pp. 737–744.
52. Wang XJ (2021) 50 years of mnemonic persistent activity: Quo vadis? *Trends in Neurosciences* p. S0166223621001685.
53. Stokes MG (2015) 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences* 19(7):394–405.
54. Higgins D, Graupner M, Brunel N (2014) Memory Maintenance in Synapses with Calcium-Based Plasticity in the Presence of Background Activity. *PLoS Computational Biology* 10(10):e1003834.
55. Shannon CE (1948) A Mathematical Theory of Communication. *The Bell system technical journal* 27(3):55.
56. Valiant LG (1984) A theory of the learnable. *Commun. ACM* 27(11):1134–1142.
57. Rich P, de Haan R, Wareham T, van Rooij I (2021) How hard is cognitive science?, (PsyArXiv), Preprint.
58. Mitchell TM (1980) The Need for Biases in Learning Generalizations. *New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.*. pp. 184–191.
59. Thrun S (1998) Lifelong Learning Algorithms in *Learning to Learn*, eds. Thrun S, Pratt L. (Springer US, Boston, MA), pp. 181–209.
60. Ganguli S, Sompolinsky H (2010) Short-term memory in neuronal networks through dynamical compressed sensing. p. 9.
61. Kingma DP, Welling M (2013) Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*.
62. Hill AB, Perkins RE (1985) Towards a model of boredom. *British Journal of Psychology* 76(2):235–240.
63. Bench S, Lench H (2013) On the Function of Boredom. *Behavioral Sciences* 3(3):459–472.
64. Geana A, Daw N (2016) Boredom, Information-Seeking and Exploration. *CogSci* p. 6.
65. Schultz W (1998) Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology* 80(1):1–27.
66. Estes W (1994) Toward a statistical theory of learning. *Psychological Review* 101:94–107.
67. Hocker D, Park IM (2019) Myopic control of neural dynamics. *PLOS Computational Biology* 15(3):24.
68. Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.
69. Bonawitz E, Denison S, Gopnik A, Griffiths TL (2014) Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology* 74:35–65.
70. Nowak M, Sigmund K (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364(6432):56–58.
71. Worthy DA, Todd Maddox W (2014) A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology* 59:41–49.
72. Reddy G, Celani A, Vergassola M (2016) Infomax strategies for an optimal balance between exploration and exploitation. *Journal of Statistical Physics* 163(6):1454–1476.
73. Wolpert D, Macready W (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Computat.* 1(1):67–82.
74. Whitley D, Rowe J (2011) *A "No Free Lunch" Tutorial: Sharpened and Focused No Free Lunch*. (WORLD SCIENTIFIC) Vol. 1, pp. 255–287.
75. Shannon C, Weaver W (1964) *The Mathematical Theory of Communication*. (The university of Illinois Press).
76. Ha D, Schmidhuber J (2018) World Models. *arXiv:1803.10122 [cs, stat]*.
77. Ahilan S, et al. (2019) Learning to use past evidence in a sophisticated world model. *PLoS Comput Biol* 15(6):e1007093.
78. Poucet B (1993) Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psycholocial Review* 100(1):162–183.
79. Schmidhuber J (2010) Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2(3):230–247.
80. Lehman J, Stanley KO (2011) Novelty Search and the Problem with Objectives in *Genetic Programming Theory and Practice IX*, eds. Riolo R, Vladislavleva E, Moore JH. (Springer New York, New York, NY), pp. 37–56.
81. Lehman J, Stanley KO, Miikkulainen R (2013) Effective diversity maintenance in deceptive domains in *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference - GECCO '13*. (ACM Press, Amsterdam, The Netherlands), p. 215.
82. Colas C, et al. (2020) Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration. *arXiv:2002.09253 [cs]*.
83. Rahnev D, Denison RN (2018) Suboptimality in perceptual decision making. *Behavioral and Brain Sciences* 41:e223.
84. Wilson RC, Bonawitz E, Costa V, Ebitz B (2020) Balancing exploration and exploitation with

information and randomization, (PsyArXiv), Preprint.
85. Cisek P (2019) Resynthesizing behavior through phylogenetic refinement. *Atten Percept Psychophys*.
86. Stanley KO, Miikkulainen R (2004) Evolving a Roving Eye for Go in *Genetic and Evolutionary Computation – GECCO 2004*, eds. Kanade T, et al. (Springer Berlin Heidelberg, Berlin, Heidelberg) Vol. 3103, pp. 1226–1238.
87. Pathak D, Gandhi D, Gupta A (2019) Self-Supervised Exploration via Disagreement. *Proceedings of the 36th International Conference on Machine Learning* p. 10.
88. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10(9):1214–1221.
89. Kolchinsky A, Wolpert DH (2018) Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8(6):20180041.
90. Tishby N, Pereira FC, Bialek W (2000) The Information Bottleneck Method. *Arxiv* 0004057:11.
91. Sternisko A, Cichocka A, Van Bavel JJ (2020) The dark side of social movements: Social identity, non-conformity, and the lure of conspiracy theories. *Current opinion in psychology* 35:1–6.
92. Gentner D, Holyoak KJ (1997) Reasoning and learning by analogy: Introduction. *American psychologist* 52(1):32.
93. Loftus EF, Hoffman HG (1989) Misinformation and memory: The creation of new memories. *Journal of experimental psychology: General* 118(1):100.
94. Kashdan TB, Disabato D, Goodman FR, McKnight P (2019) The Five-Dimensional Curiosity Scale Revised (5DCR): Briefer subscales while separating general overt and covert social curiosity, (Open Science Framework), Preprint.
95. Keller H, Schneider K, Henderson B, eds. (1994) *Curiosity and Exploration*. (Springer Berlin Heidelberg, Berlin, Heidelberg).
96. Kaplan F, Oudeyer PY (2007) The progress drive hypothesis: An interpretation of early imitation in *Imitation and Social Learning in Robots, Humans and Animals*, eds. Nehaniv CL, Dautenhahn K. (Cambridge University Press, Cambridge), pp. 361–378.
97. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. *Vision Research* 49(10):1295–1306.
98. Friston KJ, et al. (2017) Active Inference, Curiosity and Insight. *Neural Computation* 29(10):2633–2683.
99. Calhoun AJ, et al. (2015) Neural Mechanisms for Evaluating Environmental Variability in Caenorhabditis elegans. *Neuron* 86(2):428–441.
100. Kim K, Sano M, De Freitas J, Haber N, Yamins D (2020) Active World Model Learning with Progress Curiosity. *arXiv:2007.07853 [cs, stat]*.
101. Savinov N, et al. (2019) Episodic Curiosity through Reachability. *arXiv:1810.02274 [cs, stat]*.
102. Colas C, et al. (2020) Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration. *arXiv:2002.09253 [cs]*.
103. Mendonca R, Rybkin O, Daniilidis K, Hafner D, Pathak D (2021) Discovering and Achieving Goals via World Models. *arXiv:2110.09514 [cs, stat]*.
104. Colas C, et al. (2020) Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration. *arXiv:2002.09253 [cs]*.
105. Gabaix X, Laibson D, Moloche G, Weinberg S (2006) Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *THE AMERICAN ECONOMIC REVIEW* 96(4):26.
106. Dubey R, Griffiths TL (2020) A rational analysis of curiosity. *arXiv:1705.04351 [cs]*.
107. Wolff W, Martarelli CS (2020) Bored Into Depletion? toward a Tentative Integration of Perceived Self-Control Exertion and Boredom as Guiding Signals for Goal-Directed Behavior. *Perspect Psychol Sci* 15(5):1272–1283.
108. Keramati M, Gutkin B (2014) Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife* 3:e04811.
109. Juechems K, Summerfield C (2019) Where does value come from?, (PsyArXiv), Preprint.
110. Marques J, Meng L, Schaak D, Robson D, Li J (2019) Internal state dynamics shape brainwide activity and foraging behaviour. *Nature* p. 27.
111. White JK, et al. (2019) A neural network for information seeking, (Neuroscience), Preprint.
112. Rosenberg M, Zhang T, Perona P, Meister M (2021) Mice in a labyrinth: Rapid learning, sudden insight, and efficient exploration. *bioRxiv* 426746:36.
113. Peters O, Gell-Mann M (2016) Evaluating gambles using dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26(2):023103.
114. Mangalam M, Kelty-Stephen DG (2021) Point estimates, Simpson's paradox, and nonergodicity in biological sciences. *Neuroscience & Biobehavioral Reviews* 125:98–107.
115. Song M, Bnaya Z, Ma WJ (2019) Sources of suboptimality in a minimalistic explore–exploit task. *Nature Human Behaviour* 3(4):361–368.
116. Silver D, et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.
117. Silver D, et al. (2018) Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *Science* 362(6419):1140–1144.
118. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B (2018) Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour* 2(12):915–924.
119. Haarnoja T, et al. (2018) Soft Actor-Critic Algorithms and Applications. *arXiv:1812.05905 [cs, stat]*.
120. Bellemare MG, et al. (2016) Unifying Count-Based Exploration and Intrinsic Motivation. *arXiv:1606.01868 [cs, stat]*.
121. Bartumeus F, Catalan J, Fulco UL, Lyra ML, Viswanathan GM (2002) Optimizing the Encounter Rate in Biological Interactions: Lévy versus Brownian Strategies. *Phys. Rev. Lett.* 88(9):097901.
122. Bi S, Sourjik V (2018) Stimulus sensing and signal processing in bacterial chemotaxis. *Current Opinion in Microbiology* 45:22–29.
123. Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. (MIT Press).
124. López-Fidalgo J, Tommasi C, Trandafir PC (2007) An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2):231–242.
125. Ay N (2015) Information Geometry on Complexity and Stochastic Interaction. *Entropy* 17(4):2432–2458.
126. Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13:281–305.
127. Gershman SJ (2018) Deconstructing the human algorithms for exploration. *Cognition* 173:34–42.

Mathematical Appendix.

Optimal substructure in memory. To find an optimal value solution for E using the Bellman equation we must prove our memory $\mathbf{M}$ has optimal substructure. This is because the normal route to a Bellman solution, which assumes the problem rests in a Markov Space, is closed to us. To understand why it is closed to us let's consider a normal Markov case.

In Markov spaces there are a set of states where the transition to the next S_t depends only on the previous state S_{t-1}. If we were to optimize over these states, as in typical reinforcement learning, we can know each transition as its own "subproblem", dynamics programming approaches apply.

The problem for our definition of information value is that it relies on a memory definition which is necessarily composed of many past observations, arbitrarily, and so it cannot for certain be a Markov space. In fact, it is probably now. So if we wish to use a dynamics programming approach we need to find another way to establish optimal substructure, which is the ability to make the simpler independent "subproblems" dynamic programming and the Bellman equation rely on. Establishing these subproblems is the focus of Theorem 1, which proceeds by contradiction.

The "heavy lifting" in the proof is done by assuming we have a perfect forgetting function for the most recent past observation. This is what we denote asas f^{-1}. In this theorem we assume $\mathcal{X}$, A, $\mathbf{M}$, f, and $\mathcal{T}$ are implicitly given.

Theorem 1 (Optimal substructure)**.** *If $V^*_{\pi_E}$ is the optimal information value given by policy π_E, a memory $\mathbf{M}_t$ has optimal substructure if the last observation S can be removed from $\mathbf{M}$, by $\mathbf{M}_{t-1} = f^{-1}f(\mathcal{X}, \mathbf{M}_t)$ such that the resulting value $V^*_{t-1} = V^*_t - E_t$ is also optimal.*

Proof. Given a known optimal value V^* given by π_E we assume for the sake of contradiction there also exists an alternative policy $\hat{\pi}_E \neq \pi_E$ at time t that gives a memory $\hat{\mathbf{M}}_{t-1} \neq \mathbf{M}_{t-1}$ and for which $\hat{V}^*_{t-1} > V^*_{t-1}$.
To recover the known optimal memory $\mathbf{M}_t$ we lift $\hat{\mathbf{M}}_{t-1}$ to $\mathbf{M}_t = f(\hat{\mathbf{M}}_{t-1}, S_t)$. This implies $\hat{V}^* > V^*$ which contradicts the purported optimality of V^* and therefore π_E. □

The assumption of deterministic dynamics might seem limiting, especially when viewed from classical Markov decision problems. In settings like bandit problems, for example, (127). But in the natural world this limit is not so terrible. Determinism is consistent with the dynamics of the natural world en large, at the scale of animal behavior. We addressed this inconsistency in the main text, using numerical simulations. Though our proofs hold for deterministic problems in practice the resulting algorithms we propose do very well on many traditional test problems for reinforcement learning.