The curiosity trick

Erik J Peterson[1,2*] **and Timothy D Verstynen**[1,2,3,4]

***For correspondence:**
Erik.Exists@gmail.com (EJP)

[1]Department of Psychology; [2]Center for the Neural Basis of Cognition; [3]Carnegie Mellon Neuroscience Institute; [4]Biomedical Engineering, Carnegie Mellon University, Pittsburgh PA

Abstract The exploration-exploitation dilemma is commonly thought to be an intractable but problem in reward learning and therefore in the biological sciences. In this paper, we provide an alternative view that does not depend on exploring to optimize for reward value. We redefine exploration as having no objective but learning itself, which we equate to curiosity. Through theory and experiments we show this alternative leads to an optimal value solution to exploration-exploitation problems, based on the famous win-stay, lose-switch rule. This solution rests on our conjecture that information and reward are equally valuable for survival, but independently valued. We show this "curiosity trick" succeeds in naturalistic conditions, when rewards are sparse, deceptive, and non-stationary.

Exploration behavior can have two very different explanations depending on whether an animal might receive a reward. If there is no reason to expect a reward, exploration is treated as a search for information. For example, when a rat is placed in a novel environment it will explore even if no food and water is expected **?**. We equate this with curiosity **???????**. If reward is expected however exploration is interpreted as a search for reward **??????**.

An open problem in the decision sciences is to unify exploration with exploitation, which we define as the policy of choosing the most rewarding action. This union is known to be difficult and the conflict between them leads to the famous exploration-exploitation dilemma (**??????**). We illustrate this classic paradox in Fig. 1a.

In this paper, we show that exploration is better handled by an information search, even when the goal is to collect the most reward. We can justify our use of curiosity because it is already known to be a primary drive in most, if not all, animals **???**. It is as strong, if not sometimes stronger, than the drive for reward **???????**. We illustrate our alternative in Fig. 1b.

If exploration is just a search for information, then what was the dilemma can be divided into two problems. One problem is exploration (recast as a curious search). The other is exploitation, which still means a search for max reward. The focus of this paper is deriving a greedy and deterministic algorithm for solving both problems, simultaneously. We first develop a new mathematical view of information value, curiosity, and a new understanding of ideal curious search. The second half uses these first results to prove our curiosity trick leads to an optimal value solution for nearly all explore-exploit decisions.

Results

We consider an animal who interacts with an environment over a sequence of states, actions and rewards. Their reward learning is embedded in the now standard Markov decision process. It's goal is to select actions in a way that maximizes the total reward collected from the environment for

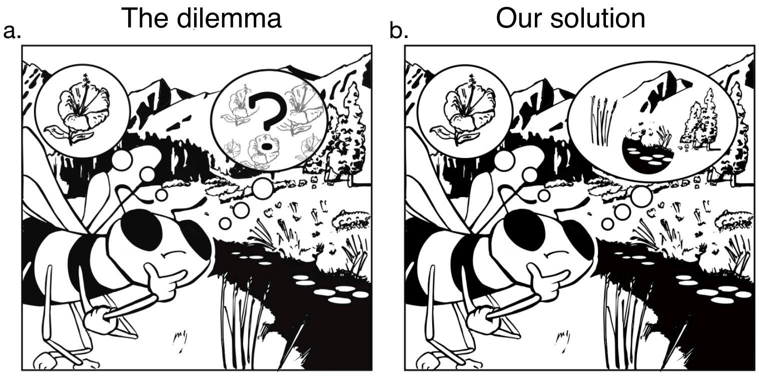

Figure 1. Two views of exploration and exploitation. **a**. The dilemma: either exploit an action with a known reward (e.g., return to the previous flower) or explore other actions on the chance they will return a better outcome. The central challenge here is that the outcome of exploration is uncertain, and filled with questions. **b**. An alternative view of the dilemma, with two goals: either maximize rewards *or* maximize information value,s with a curious search of the environment. *Artist credit*: Richard Grant.

every state $\mathbf{S}$ (Eq 1).

$$V_R(\mathbf{S}) = \operatorname*{argmax}_{\mathbf{A}} \left[\sum_T R \right] \tag{1}$$

How should an animal do this? To maximize reward an animal must tradeoff between exploration, and exploitation. The standard way of approaching this tradeoff is to assume, "The agent must try a variety of actions and progressively favor those that appear to be best" **?**. Here we suggest two fairly radical changes to this standard formula. Don't stop searching progressively, do so suddenly (greedily). Explore only for learnings sake, out of curiosity (**?**). To demonstrate the advantages of this first we need to establish some shared formalities.

Markovian formalities

Our Markov decision process (MDP) consists of states of real valued vectors $\mathbf{S}$ from a finite set $\mathcal{S}$ of size n. As are actions, $\mathbf{A}$ from a finite set $\mathcal{A}$ of size k. Rewards are non-negative real numbers, R. Transitions from state to state are handled by the transition function $\Lambda(\mathbf{S})$. In our mathmatics we assume Λ is a deterministic function of its inputs. In our simulations we assume it is a stochastic function.

To study learning, memory and curiosity hollistically we assume animals can "fuse" the elements of the MDP into single observations, $\mathbf{X}$. It is convenient to assume there observations are embedded in a finite real space, $\mathbf{X} \in \mathcal{X}$ of size l. This fusing is imagined to happen via a communication channel, T, which is identical to out observation function, $T(\mathbf{S}, \mathbf{A}, \mathbf{S}', R, \mathbf{M})$. $\mathbf{M}$ is the animal's memory, that we define below.

Observations made by T can be internally generated from memory $\mathbf{M}$, or externally generated from the environment ($\mathbf{S}$,$\mathbf{S}'$). Or both in combination, as happens in recurrent neural circuits. Observations may contain action information $\mathbf{A}$ but this optional. Observations may contain reward information R or not. Though ommiting reward would hamper solving dilemma.

A general kind of curiosity

To make our eventual solution general, we first need to separate reward value from information value in a principled and general way. We do this axiomatically. Working models of curiosity tend to conflate the learning rule, with the objective itself. We will now describe an optimal value, but deterministic model, of curiosity. It is defined by some easily satisfiable requirements (axioms). This basis makes curiosity domain general, and learning rule independent. We assume curiosity is value driven, and should be as time-efficient as possible.

We reason that the value of any observation $\mathbf{X}$ made by an animal depends entirely on what an animal learns by making that observation—how it changes an animal's memory. But how can we define memory in a way that is general, but practically useful?

Sustained firing, the strength between two synapses, elevated Calcium concentration are three simple examples of learning and memory in the nervous system. Each of these examples though can be represented as a vector space. In fact, most any memory system can be so defined.

Memory formalities

We define memory as a set of real valued numbers, embedded in a finite space that is closed and bounded with p dimensions. In practice we regard memory $\mathbf{M}$ as learning something about a sequence of observations ($\mathbf{X}_0$ $\mathbf{X}_1$, $\mathbf{X}_2$, ...) which have arrived via a the communications channel T.

A learning function is then any function f that maps observations $\mathbf{X}$ into memory $\mathbf{M}$. This f is considered to be a valid learning function as long as the mapping changes $\mathbf{M}$ some small amount. A non-constant mapping, in other words. Using recursive notation, this kind of learning is denoted by $\mathbf{M} \leftarrow f(\mathbf{X}, \mathbf{M})$. Though we will sometimes use $\mathbf{M}'$ to denote the updated memory. Other times we will add subscripts, like $\mathbf{X}_0, \mathbf{X}_1, ldots$, to express the history of a memory, or the path is has followed.

We can also define a forgetting function that *can be any function* that inverts the memory, $f^{-1}(\mathbf{X};\ \mathbf{M}') \rightarrow \mathbf{M}$. Forgetting might not seem crucial, here. But it will prove essential later on in developing exploration algorithms that provably maximize $\hat{E}$.

For individual animals we assume f and f^{-1} have been chosen by evolution and experience to be suitable and efficient for the environment **????????**.

Axiomatic information value

Our definition of information value closely follows that of Shannon and his definition of a communication channel. If the communication channel is agnostic to semantic considerations, as in Shannon's definition, we reason its information value should be agnostic as well.

If we have a memory $\mathbf{M}$, which has been learned by f over a history of observations, $(\mathbf{X_0}, \mathbf{X_1}, ...)$, Can we measure how much value the next observation $\mathbf{X_t}$ should have? We think this measure $\hat{E}$ should have certain intuitive properties.

Axiom 1 (Axiom of Memory). $\hat{E}$ depends only on difference $\delta\mathbf{M}$ between $\mathbf{M}$ and $\mathbf{M}'$.

That is, the value of an observation $\mathbf{X}|$ depends only on how the current memory changes.

Axiom 2 (Axiom of Change). $\hat{E} = 0$ if and only if $\delta M = 0$.

That is, an observation that doesn't change the memory has no value.

Axiom 3 (Axiom of Scholarship). $\hat{E} \geq 0$.

That is, all (new) information is in principle valuable even its consequences are later found to be negative.

Axiom 4 (Axiom of Completeness) $\hat{E}$ should increase monotonically with the total change in memory.

That is, there should be a one-to-one relationship how memory changes and how information is valued.

Axiom 4 (Axiom of Equilibrium) For the same observation $\hat{E}$ should approach 0 in finite time.

That is, learning on $\mathbf{M}$ makes continual progress toward equilibrium (i.e. self-consistency) with each observation, after a finite time period has passed. After this it will approach a steady-state value of $\delta\mathbf{M} \rightarrow 0$ and $\hat{E} \rightarrow 0$.

All we have really done in these axioms is remove any mention of a learning rule, its properties, or learning target, its propertiess **??????**. This is important not for its own sake but because, we it lets use consider information value independent of any semantics. Anything we prove in this general setting will be therefore be true for any learning algorithm, who meets our requirements.

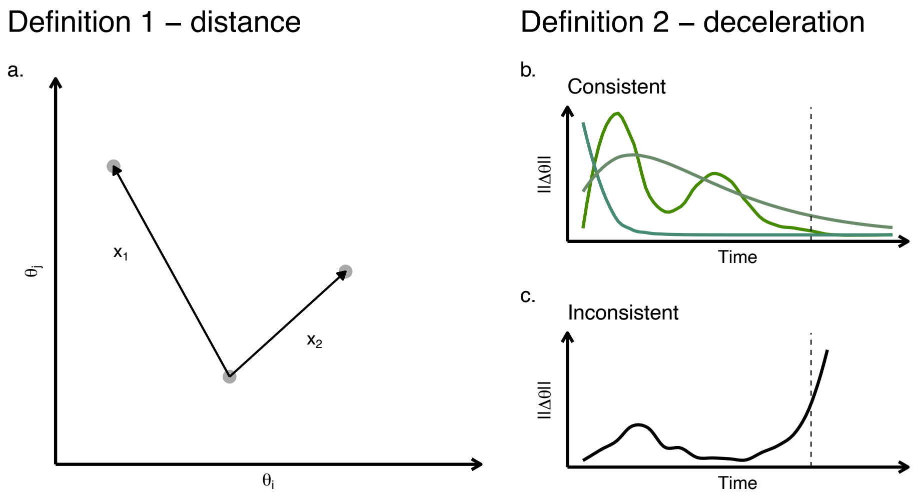

Figure 2. Geometric information value, and our constraints on its dynamics. **a**. This panel illustrates a two dimensional memory. The information value of two observations $\mathbf{X}_1$ and $\mathbf{X}_2$ depends on the norm of memory with learning. Here we show this distance as a euclidean norm, denoted as a black arrow. **b-c** This panel illustrates learning dynamics with time (over a series of observations that are not shown). If information value becomes decelerating in finite time bound, then we say that learning is consistent with Def. 2. This is shown in panel b. The bound is depicted as a dotted line. If learning does not decelerate, then it is said to be inconsistent (Panel c). *It is important to note:* our account of information value and curiosity does not work when learning is inconsistent.

Practical (geometric) information value

Meeting our requirements is not difficult. The practical example of axiomatic value we use throughout this paper is based on the geometric norm. Though not necessaailry a unique solution this is quite a natural way, based only the dynamics and "shape" of the memory space itself (Eq 2).

$$\hat{E} = ||\ f(\mathbf{M}, \mathbf{X}) - \mathbf{M}\ || \tag{2}$$

To ensure equilibrium in Eq 2, we require $\nabla^2 \mathbf{M} < 0$ for all $t \geq T^*$. Informally, the idea here is that any notion of sensible and useful learning must in time converge, which we take to mean that in finite time $\hat{E}$ must approach 0 (Fig. 2). We term this consistent learning.

Deterministic curiosity

A nice solution to maximizing $\hat{E}$, and so ideal curiosity, would be a Bellman solution (**?**) because it guarantees optimal value using just recursion, a very biological idea. It is convenient also because the Bellman equation is the basis for the theory of reinforcement learning (**?**).

The common way to arrive at Bellman solution is to assume a Markov decision space. This is problem for memory because it naturally has a long-term path dependence, which breaks the Markov assumption and its neeeded claim of optimal substructure. To get around this we prove that forgetting is another way to achieve optimal substructure.

Given an arbitrary starting value $E_0 > 0$, the best solution to maximizing $\hat{E}$ is given by Eq. 3. A full derivation for this fact is provided in the Appendix.

$$V^{\pi_E}(\mathbf{X}_0) = \underset{\mathbf{A}}{\operatorname{argmax}} \left[E_0 + V_E(\mathbf{X}_1) \right] \tag{3}$$

As long learning about the different observations $\mathbf{X}$ is independent, and as long as learning continues until steady-state, defined as $E_t < \eta$, there are many equally good solutions. Under these conditions we can simplify Eq. 3 further, giving Eq. 4.

$$V^{\pi_E}(\mathbf{X}_0) \approxeq \underset{\mathbf{A}}{\operatorname{argmax}} \left[E_0 + E_1 \right] \tag{4}$$

A note on exploration quality

Having a policy π_E^* that maximizes $\hat{E}$ does not necessarily ensure that exploration is of good quality. We think it is reasonable to call an exploration good if it meets the three criteria below. In Theorem **??** we prove that our Bellman solution π_E satisfies these, when $\eta > 0$.

1. Exploration should visit all available states of the environment at least once.
2. Exploration should cease when learning has plateaued.
3. Exploration should take as few steps as possible to achieve 1 and 2.

A note on boredom

The central failure mode of curiosity is what we'll call minutia, defined as those learning that has little to no use, ever. A good example of this is to, "Imagine a scenario where the agent is observing the movement of tree leaves in a breeze. Since it is inherently hard to model breeze, it is even harder to predict the location of each leaf" **?**. This will imply, they note, a curious agent will always remain curious about chaos in the leaves even though they have no hope of successful learning of them.

To limit curiosity we will use boredom which we consider an adaptive trait. Others have considered boredom constructilvy arguing it is a useful way to motivate aversive tasks (**?**), or curiosity **?**. We take a slightly different view. We treat boredom as a tunable free parameter, $\eta \geq 0$. We consider boredom as a means to ignore marginal value, by requiring curious exploration to stop once $E \leq \eta$.

Of equal importance

How can we justify using curiosity for the dilemma when the goal of exploitation is reward collection? Curiousity is just as important for survival as reward collection **?**. Curiosity is a primary drive in most, if not all, animals **?**. It is as strong, if not sometimes stronger, than the drive for reward **???**.

Still, intuition suggests that curiosity is wrong-headed. Searching in an open-ended way for information must be less efficient than a direct search for reward? In answer to this we make two conjectures.

A hypothesis of equal important (Conjecture 1): Reward value and information value are equally important.

If both reward and information are equally important, then time is not wasted in a curious search, and so the process is not *necessarily* inefficient or even indirect. The question becomes instead, is curiosity practical enough to serve as a usseful trick to solve reward collection search problems? This leads us to the next conjecture.

The curiosity trick (Conjecture 2): Curiosity is a sufficient solution to exploration (when learning is the goal)

It's worth noting here that curiosity, as an algorithm, is highly effective at solving optimization problems **???????????**.

A win-stay, lose-switch solution

If reward and information are equally important, how can an animal go about balancing them? If you accept our conjectures then answering this question is the same as answering the dilemma. To solve this dual value learning problem we've posed—information and reward maximization—we need an algorithm that can maximize the total value of both objectives. We found a simple answer in game theory and in the computer science sub-field of optimal scheduling.

Win-stay lose-shift is a strategy famous from game theory (**?**), where a player will keep its current strategy under winning conditions, and shift otherwise. We see the same kind of patterning as useful here, except the "players" are two different policies inside the animal. They are competing for behavioral control. Imagine that in the previous action our bee from Fig. 1 observed 0.7 units of information value, $\hat{E}$, and no reward (i.e. 0). Using our rule on the next action, the bee should then choose its exploration policy. Let us say that this time $\hat{E}$ decreases to 0.6, and a reward value of 1.0 is observed. Now the bee should change its strategy for the next round, to exploit instead. In general then if there was more reward value then information value last round ($R_{t-1} > E_{t-1}$), our rule dictates an animal should choose the reward policy. If information value dominated, $E_{t-1} >\geq R_{t-1}$, then it should choose the information gathering policy.

Here we propose that a win-stay, lose-switch strategy is in fact an optimal and tractable solution to the scheduling problem of exploration (by curiosity) and exploitation (for reward).

To prove the optimality of our rule we will use regret minimization, which is a standard metric of optimality in reinforcement learning (**?**). Regret is a numerical quantity of the difference between the most valuable choice and the value of the choice that was made. To approximate regret G we use Eq. 5, where V is the value of the chosen action, and V^* is the maximum value.

$$G = V^* - V \tag{5}$$

An optimal value algorithm always makes the most valuable choice. It therefore will have zero regret, by Eq. 5. This is impossible to achieve, of course, using stochastic search. To find the algorithm that maximizes both information and reward value with zero regret, we therefore need a deterministic method.

To prove things about our solution we will move from game theory, which gave us our inspiration, to the field of optimal scheduling (**??**). Put another way, we will now answer the question, when does win-stay, lose-switch strategy work as an optimal scheduler?

We imagine the policies for exploration and exploitation acting as two possible "jobs" competing for control of behavior, a fixed resource. We know by definition that each of these jobs produces non-negative values which an optimal job scheduler could use: $\hat{E}$ for information or R for reward/reinforcement learning. We can also ensure, again by definition, that each of the jobs takes a constant amount of time and that each policy can only take one action at a time. These two properties, and one further assumption, are sufficient.

We believe it is also reasonable to assume that there will be no model of the environment available to the learner at the start of a problem, and that its environment is nonlinear but deterministic.

We arrived at our solution by taking the simplest possible path available. We just wrote down the simplest equation that could have zero regret answer, and began to study that. This is Eq. 7. We refer to it as a meta-greedy algorithm. It decides in greedy fashion which of our independent policies to use.

$$\Pi^{\pi} = \left\{ \pi_E, \; \pi_R \right\} \tag{6}$$

$$\underset{\Pi^{\pi}}{\operatorname{argmax}} \; \left[\hat{E} - \eta, \; R \right]_{\mathbf{(2,\,1)}} \tag{7}$$

In Eq 7 we modified argmax notation to include a subscript vector to denote a ranking for which value/policy should be preferred in the event of a tie. In this case its $(2, 1)$. We did this because an exact rule for tie breaking is it turns out critical.

In Theorem 5 we prove Eq. 7 is Bellman optimal, and so it will always find a maximum value sequence of $\hat{E}$ and R values. Given, that is, that the environment is deterministic.

In Eq 7 we compare reward R and information value E but have not ensured they have same "units", or are comparible valuables. We sidestep this isssue by limiting the probability of having zero reward to be, itself, non-zero. That is, $p(R_t = 0) > 0$. We also limit E_0, the first set of values, to be positive and non-zero when used with boredom, $(E_0 - \eta) \geq 0$. These constraints ensure complete "good" exploration (defined above) which in turn ensures optimal reward learning is possible.

A note about subjective reward value

We have been studying a kind of reinforcement learning where the value of a reward is fixed. That is, where environment reward value R is always equal to it subjective value to the animal $\hat{R}$. It is the most common style of reinforcement learning (**?**). In natural behavior though the motivational value of reward often declines as the animal reaches satiety. We'll call reward learning like this the homeostatic style (**???**). Fortunately, it has already been proven that a reward collection policy designed for one style will work for the other, at least this is so an infinite time-horizon (**?**). However to use Eq. 7 for reward homeostasis we need to make one modification. Under homeostasis ties between values in Eq. 7 should be broken in favor of exploration, and information seeking. This modification leads to Eq. 8. This equation implies under reward homeostasis an animal's default policy is information seeking, and therefore environmental exploration.

$$\underset{\Pi^\pi}{\text{argmax}} \left[\hat{E} - \eta,\ \hat{R}\right]_{\mathbf{(1,\,2)}} \tag{8}$$

Information collection in practice

The work so far has built up the idea that the most valuable, and most efficient, curious search will come from a deterministic algorithm. That is, every step strictly maximizes $\hat{E}$. It is this determinism which will let us resolve the dilemma, later on. A deterministic view of exploration seems at odds with how the problem is generally viewed today, which involves searching with some amount of randomness. Whereas if our analysis is correct, randomness is not needed or desirable, as it must lead to less value and a longer search.

We confirm and illustrate this result using an example of a simple information foraging task (Task 1; Fig.). This variation of the bandit task (**?**) replaces rewards with information, in this case colors. On each selection, the agent sees one of two colors according to a specific probability shown in Fig. 6a. When the relative color probabilities are more similar that arm has more entropy and so more to learn and more information value. Arms that are certain to present only one color lose their informative value quickly.

The results of this simple information seeking task are shown in Figure 4. Deterministic curiosity in this task generated more information value, in less time, and with zero regret when compared against a stochastic agent using more directed random search. As our work here predicts, noise only made the search happen slower and lead to regret. Besides offering a concrete example, performance in Task 1 is a first step in showing that even though π_E's optimality is proven for a deterministic environment, it can perform well in other settings.

Reward collection in practice

The general form of the 7 tasks we will study is depicted in Fig 5. As with our first task (Fig. 3) they are all variations of the classic multi-armed bandit (**?**). The payouts for each of the tasks are shown separately in Fig. 6. Each was designed to either test exploration performance in a way that matches recent experimental study(s), or to test the limits of curiosity. Every trial has a set of n choices. Each choice returns a "payout", according to a predetermined probability. Payouts are information, a reward, or both. Note that, as in Task 1, information was symbolic, denoted by a color code, "yellow" and "blue" and as is custom reward was a positive real number.

To evaluate the performance of our curiosity algorithm, we compare it against a range of other exploration strategies found in the literature. For a brief description of each, see Table 1. We feel this set is a broad and fair representation of today's state-of-the-art approaches.

They all have in common that their central goal is to maximize reward value, though this is often supplemented by some other goal. The results in full for all tasks and exploration strategies

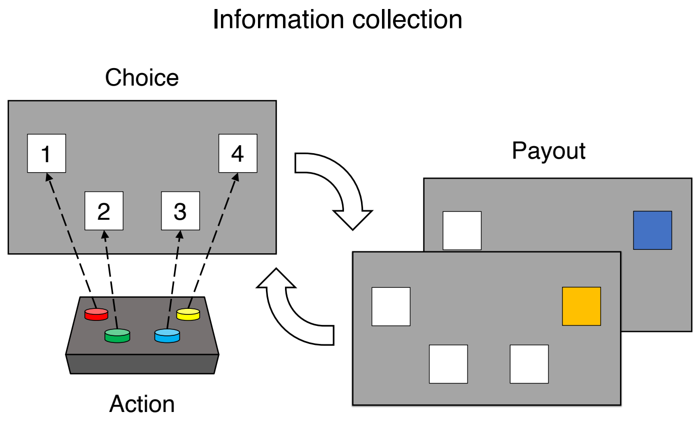

Figure 3. A simple four choice information foraging task. The information in this task is a yellow or blue stimulus, which can change from trial to trial. A good learner in this task is one who tries to learn the probabilities of all the symbols in each of the four choices. The more random the stimuli are in a choice, the more potential information/entropy there is to learn. *Note*: the information payout structure is shown below (Fig. 6a).

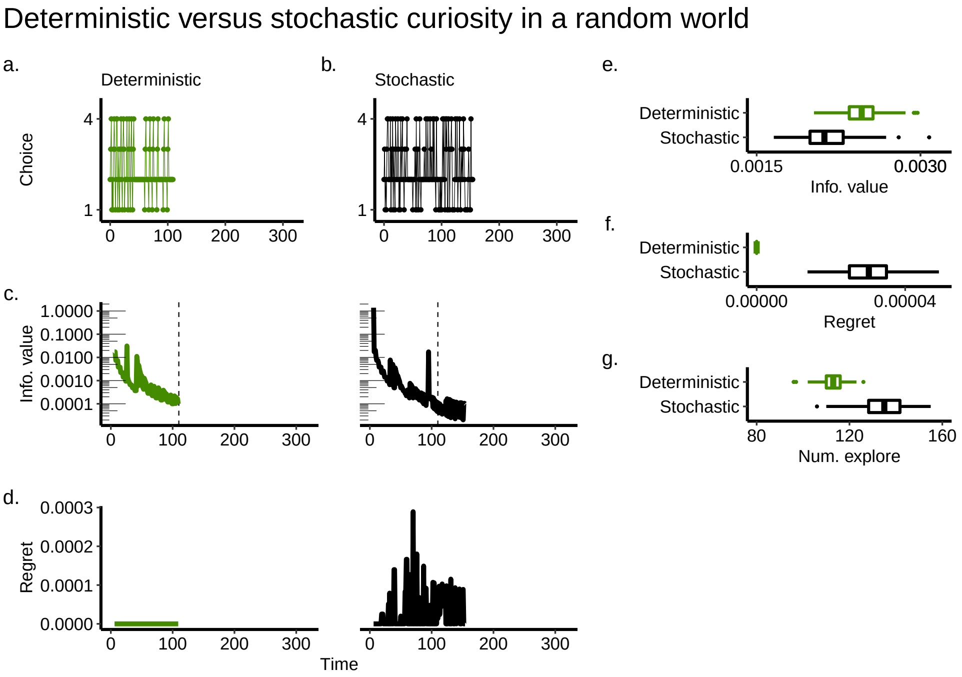

Figure 4. Comparing deterministic versus stochastic variations of the same curiosity algorithm, in a simple information foraging task (Task 1). Deterministic results are shown in the left column, and stochastic are shown in the right. The hyperparameters for both models were found by random search, which is described in the Methods. **a-b**. Examples of choice behavior. **c-d**. Information value plotted with time for the behavior shown in a-b. **e-f**. Regret plotted with time for the behavior shown in a-b. Note how our only deterministic curiosity generates zero regret, inline with theoretical predictions. **e**. Average information value for 100 independent simulations. Large values mean a more efficient search. **f**. Average regret for 100 independent simulations. Ideal exploration should have no regret. **g**. Number of steps it took to reach the boredom threshold *eta*. Smaller values imply a faster search.

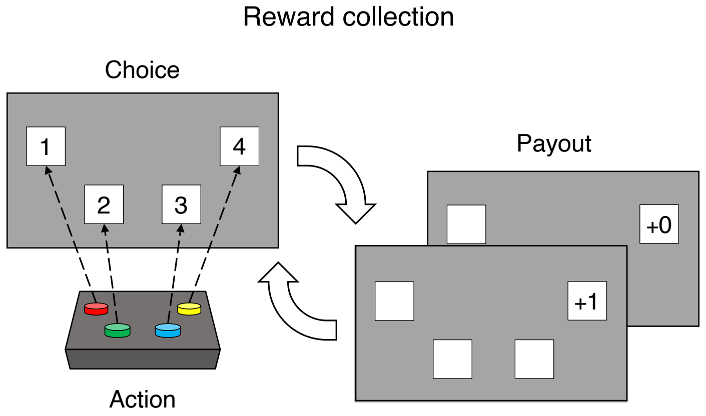

Figure 5. A 4 choice reward collection task. The reward is numerical value, here a 1 or 0. A good learner in this task is one who collects the most rewards. The task depicted here matches that in Fig 6b. Every other reward collection task we study has the same basic form, only the number of choices increases and the reward spaces are more complex.

Table 1. Exploration strategies.

Name	**Class**	**Exploration strategy**
Curiosity	Deterministic	Maximize information value
Random/Greedy	Random	Alternates between random exploration and greedy with probability ϵ.
Decay/Greedy	Random	The ϵ parameter decays with a half-life τ
Random	Random	Pure random exploration
Reward	Extrinsic	Softmax sampling of reward value
Bayesian	Extrinsic + Intrinsic	Sampling of reward value + information value
Novelty	Extrinsic + Intrinsic	Sampling of reward value + novelty signal
Entropy	Extrinsic + Intrinsic	Sampling of reward value + action entropy
Count (EB)	Extrinsic + Intrinsic	Sampling of reward value + visit counts
Count (UCB)	Extrinsic + Intrinsic	Sampling of reward value + visit counts

are shown in Fig. ??. All learners, including ours, used the same exploitation policy, based on the temporal difference learning rule (**?**).

Task 1. is an information gathering task, which we discussed already above. *Task 2* was designed to examine reward collection, in probabilistic reward setting very common in the decision making literature (**???????**). Rewards were 0 or 1. The best choice had a payout of $p(R = 1) = 0.8$. This is a much higher average payout than the others ($p(R = 1) = 0.2$). At no point does the task generate symbolic information. See, Fig. **??b**.

The results for Task 1 were discussed above. As for Task 2, we expected all the exploration strategies to succeed. While this was indeed the case, our deterministic curiosity was the top-performer in terms of median rewards collected, though by a small margin (Fig **??a**).

Task 3. was designed with very sparse rewards (**???**) and there were 10 choices, making this a more difficult task (Fig. **??c**). Sparse rewards are a common problem in the natural world, where an animal may have to travel and search between each meal This is a difficult but not impossible task for vertebrates (**?**) and invertebrates (**?**). That being said, most reinforcement learning algorithms will struggle in this setting because the thing they need to learn, that is rewards, are often absent. In this task we saw quite a bit more variation in performance, with the novelty-bonus strategy taking the top slot (this is the only time it does so).

Task 4. was designed with deceptive rewards. By deceptive we mean that the best long-term

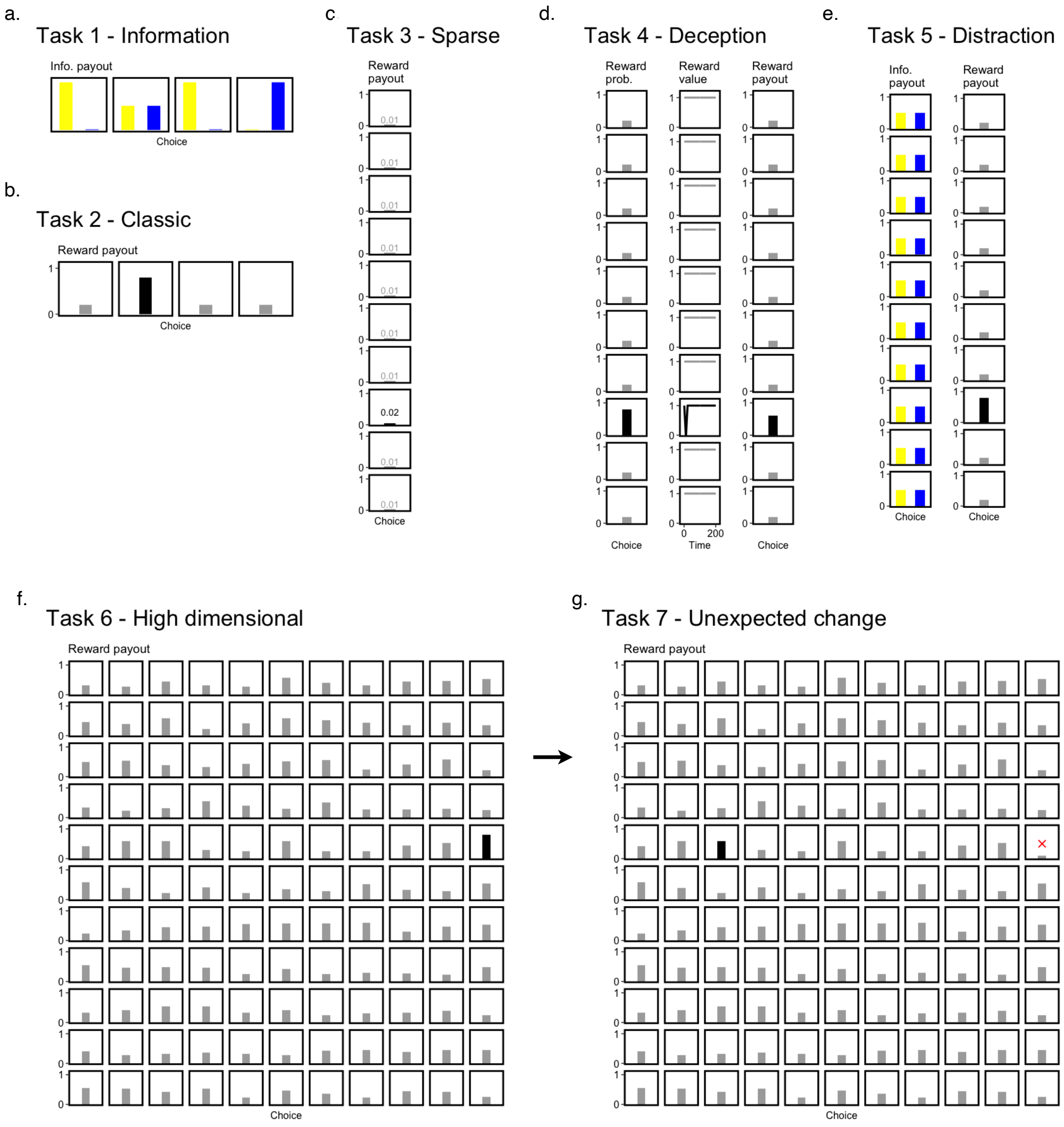

Figure 6. Payouts for *Tasks 1 - 7*. Payouts can be information, reward, or both. For comments on general task design, see Fig 5. **a.** A classic four-choice design for information collection. A good learner should visit each arm, but quickly discover that only arm two is information bearing. **b.** A classic four-choice design for testing exploration under reward collection. The learner is presented with four choices and it must discover which choice yields the highest average reward. In this task that is Choice 2. **c.** A ten choice sparse reward task. The learner is presented with four choices and it must discover which choice yields the highest average reward. In this task that is Choice 8 but the very low overall rate of rewards makes this difficult to discover. Solving this task with consistency means consistent exploration. **d.** A ten choice deceptive reward task. The learner is presented with 10 choices, but the choice which is the best on the long-term (>30 trials) has a lower value in the short term. This value first declines, then rises (see column 2). **e.** A ten choice information distraction task. The learner is presented with both information and rewards. A good learner though 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 learner 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.

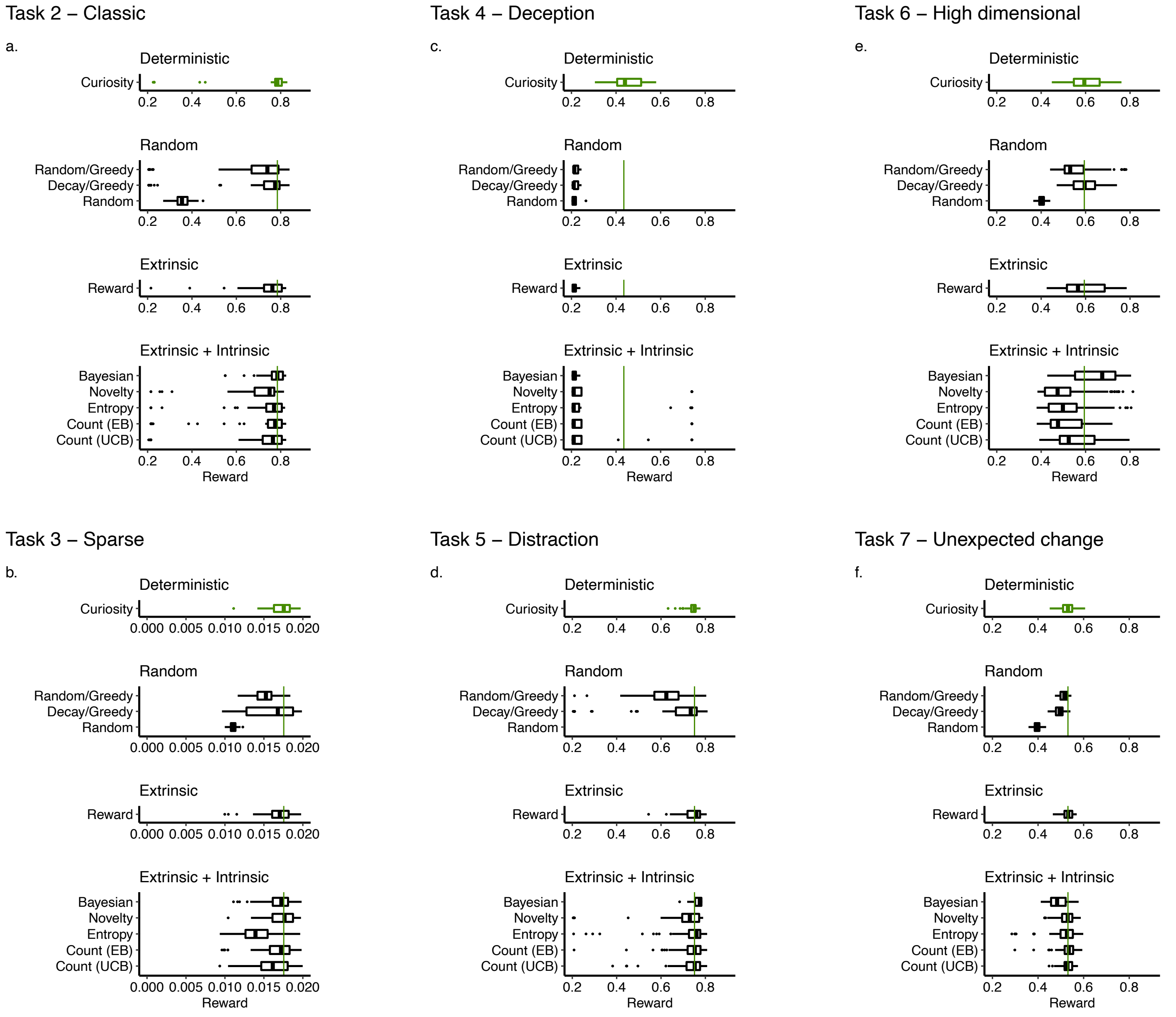

Figure 7. Summary of reward collection (*Tasks 2-7*). The strategies in each panel are grouped according to the class of search they employed (Curiosity. Random, Extrinsic reward or Extrinsic + Intrinsic rewards). **a.** Results for Task 2, which has four choices and one clear best choice. **b.** Results for Task 3, which has 10 choices and very sparse positive returns. **c.** Results for Task 4, whose best choice is initially "deceptive" in that it returns suboptimal reward value over the first 20 trials. **d.** Results for Task 5, which blends the information foraging task 1 with a larger version of Task 2. The yellow/blue stimuli are a max entropy distraction which do not predict the reward payout of each arm. **e.** Results for Task 6, which has 121 choices and a quite heterogeneous set of payouts but still with one best choice. **f.** Results for Task 7, which is identical to Task 6 except the best choice was changed to be the worst. The learners from Task 6 were trained on this Task beginning with the learned values from their prior experience – a test of robustness to sudden change in the environment. *Note*: To accommodate the fact that different tasks were run different numbers of trials, we normalized total reward by trial number. This lets us plot reward collection results for all tasks on the same scale.

Figure 7–Figure supplement 1. Summary of regret (*Tasks 2-7*)

Figure 7–Figure supplement 2. Summary of information value (*Tasks 2-7*)

Figure 7–Figure supplement 3. Examples of exploration (Task 1). Here we simulated different sets of hyper-parameters on the same task and random seed. We view each parameter set as a unique "animal" (**?**). So this figure estimates the degree and kind of variability we might expect in a natural experiment.

option presents itself initially with a decreasing reward value (Fig. **??d**). Such small deceptions abound in many natural contexts where one must often make short-term sacrifices (**?**). It is well known that classic reinforcement learning will often struggle in this kind of task (**??**). Here our deterministic curiosity is the only strategy that reaches above chance performance. Median performance of all other strategies are similar to the random control (Fig **??c**).

Task 5 was designed to fool curiosity, our algorithm, by presenting information that was utterly irrelevant to reward collection, but had very high entropy and so "interesting" to our algorithm. We fully anticipated that this context would fool just our algorithm, with all other strategies performing well since they are largely agnostic to entropy. However, despite being designed to fail, deterministic curiosity still produced a competitive performance to the other strategies (Fig **??d**).

Tasks 6-7 were designed as a pair, with both having 121 choices, and a complex payout structure. Tasks of this size are at the limit of human performance (**?**). We first trained all learners on *Task 6*, then tested them in Task 7 which identical to 6, except the best payout arm is reset to be worst (Fig. **??e**-**f**). In other words Tasks 6 and 7 were joined to measure learning in a high dimensional, but shifting, environments.

In Task 6 deterministic curiosity performed well, securing a second place finish. We note the Bayesian strategy outperformed our approach. However, under the sudden non-stationarity in reward when switching tasks, the top Bayesian model became the worst on Task 7, and deterministic curiosity took the top spot. Compare Fig. **??e** to **f**. This is the robustness that we'd expect for any curiosity algorithm, whose main goal is to learn everything unbiased by other objectives. The environment and the objectives do change in real life, and so we must be prepared for this. Note how the other less-biased-toward-reward-value exploration models (count-based, novelty, and entropy models) also saw gains, to a lesser degree.

Robustness

A good choice of boredom eta is critical for our definition of curiosity, and was optimized carefully. We show an example of this importance in Fig. 8. On the left hand side we did a random search over 10000 possible parameters to find a value for η that produced excellent results. (This was the same kind of search we did to achieve the results shown in Fig. **??**. On the right hand side we choose a value for boredom we knew would lead to a poor relative result. We then ran the 100 different randomly generated simulations which gives us the results seen in Fig. 8).

Carefully optimized boredom converged far more quickly (Fig. 8a-b), generating more reward over time (Fig. 8c-d), and showed far more in consistency during exploration (Fig. 8e-f).

These experiments in setting η are also useful in demonstrating the wide variance in behavior that is possible with a deterministic search. Exploration in these figures is deterministic, as prescribed by Eq. 7. The environment was stochastic however and so the variations come from the environment changing what is learned, which then changes what the best learner should choose to learn next.

Unlike idealised simulations, animals cannot pre-optimize their search parameters for every task. We therefore explored reward collection as a function of 1000 randomly chosen model parameters, and reexamined performance on Tasks 1 and 7=6. These were chosen to represent an "easy" task, and a "hard one". These results are shown in Fig. 2.

As we observed in our other model comparisons, exploitation by deterministic curiosity produced top-3 performance on both tasks (Fig. 2a-b). Most interesting is the performance for the worst parameters. Here deterministic curiosity was markedly better, with substantially less variability across different parameter schemes. We can explain this in the following way. All the other exploration strategies use parameters to tune the degree of exploration. With deterministic curiosity, however, we tune only when exploration should stop. The degree of exploration, the measure by how close it is to maximum entropy, is fixed by the model itself. It seems that here at least it is far more robust to tune the stopping point, rather than the degree of exploration.

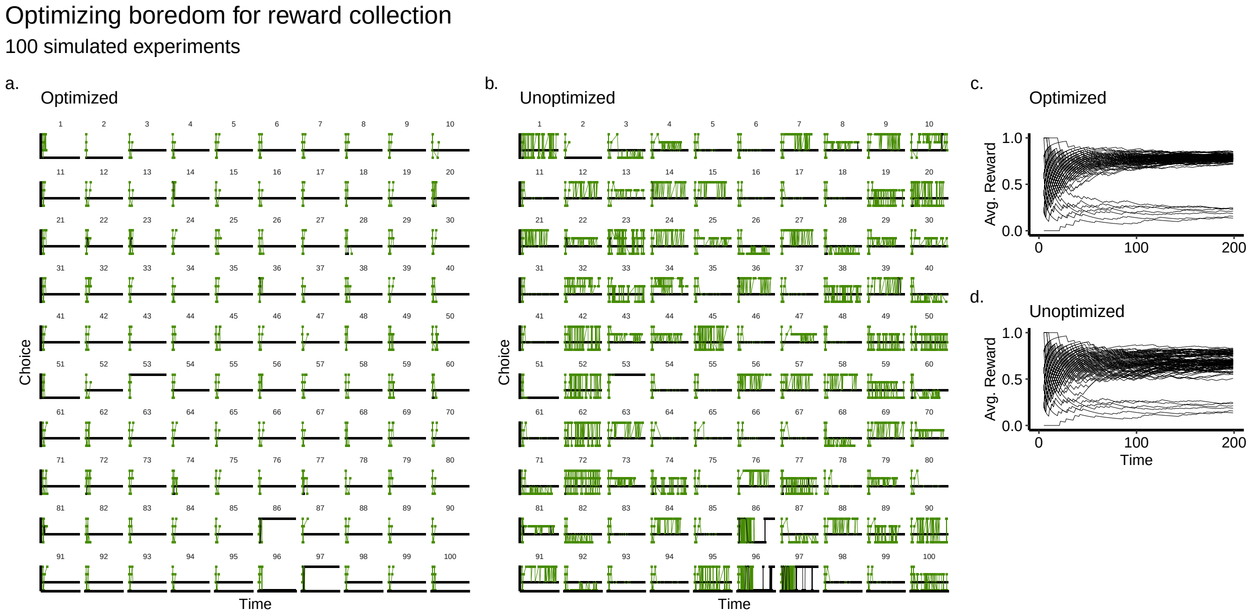

Figure 8. The importance of boredom during reward collection (Task 2). One hundred example experiments, each began with a different random seed. **a**. Behavioral choices with optimized boredom. **b**. Behavioral choices with unoptimized boredom. **c,d**. Average reward as a function of time for optimized (c) and unoptimized (d) boredom.

Discussion

We have offered a way around the exploration-exploitation dilemma. We proved the classic dilemma, found when optimizing exploration for reward value, can be resolved by exploring instead for curiosity. And that in this form the rational trade-off we are left with is solved with a classic strategy taken from game theory. The win-stay lose-shift strategy, that is.

This is a controversial claim. So we will address some of its criticisms as a series of questions.

Why curiosity?

Curiosity is an ecologically rational behavoir **?**. This arguemnt rests on what we feel are three well established facts. Curiousity is just as important for survival as reward collection **?**. Curiosity is a primary drive in most, if not all, animal behavoir **?**. It is as strong, if not sometimes stronger, than the drive for reward **???**. Curiosity as an algorithm is highly effective at solving difficult optimization problems **???????????**.

In other words, curiosity is not irrational , nor a paradox , nor a heuristic , nor even quitesssentially cognitive **?**. It is a basic homeostatic drive **?**.

What about information theory?

In short, the problems of communication and value are different problems that require different theories.

Weaver (**?**) in his classic introduction to the topic, describes information as a communication problem with three levels. 1. The technical problem of transmitting accurately. 2. The semantic problem of meaning. 3. The effectiveness problem of changing behavior. He then describes how Shannon's work addresses only problem A. It has turned out that Shannon's separation of the problem allowed the information theory to have an incredibly broad application.

Unfortunately valuing information is not, at its core, a communications problem. Consider the very simple example Bob and Alice, having a phone conversation and then Tom and Bob having the exact same 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. These might be completely

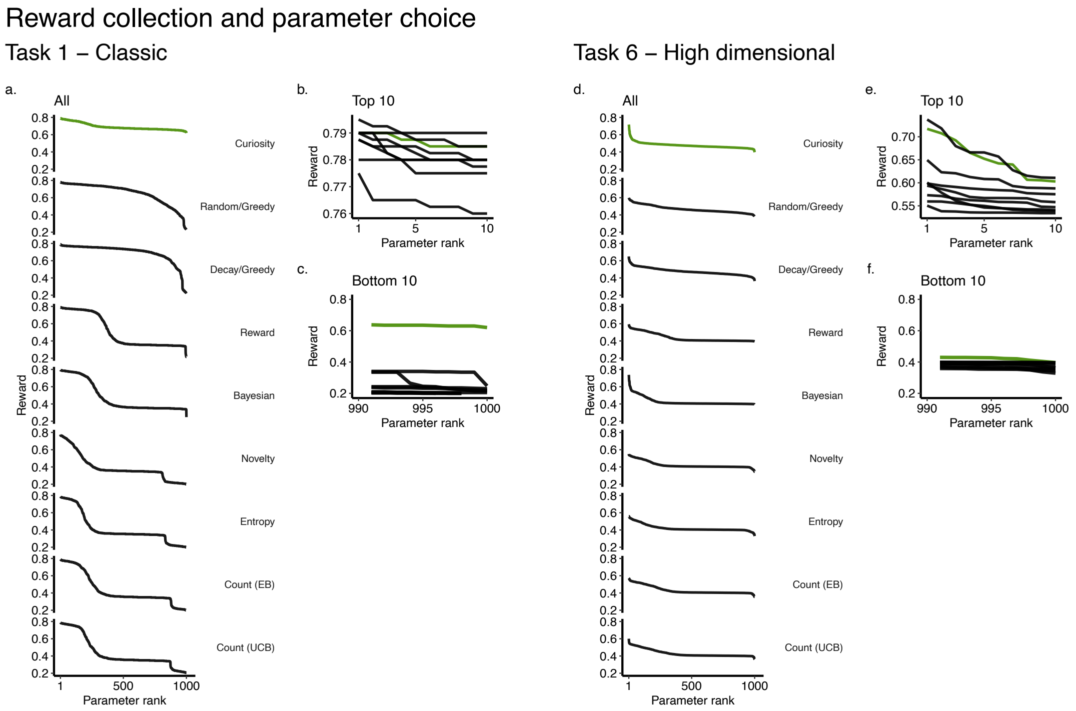

Figure 9. Parameter selection and search performance (*Task 2* and *6*). We wanted to evaluate how robust performance was to poor and random hyperparameters. A very robustness exploration algorithm would produce strong performance with both the best parameter choices, and the worst parameter choices. **a,d** Total reward for 1000 search hyperparameters, for each of strategies we considered. **b,e** Performance with the top 10 hyperparameters, curiosity (green) compared to all the others. **c.f** Performance with the bottom 10 hyperparameters.

Figure 9–Figure supplement 1. Adding noise to deterministic curiosity does degrade its performance. In this example control we added two levels of random noise. This only served to increase variability of convergence (a-c) and reduce the total rewards collected (d). Note: decision noise was modelled by sampling a Boltzmann distribution, as described in the *Methods*.

Figure 9–Figure supplement 2. Noise across strategies. Adding noise to degrades performance to nearly the same degree (Task 2).

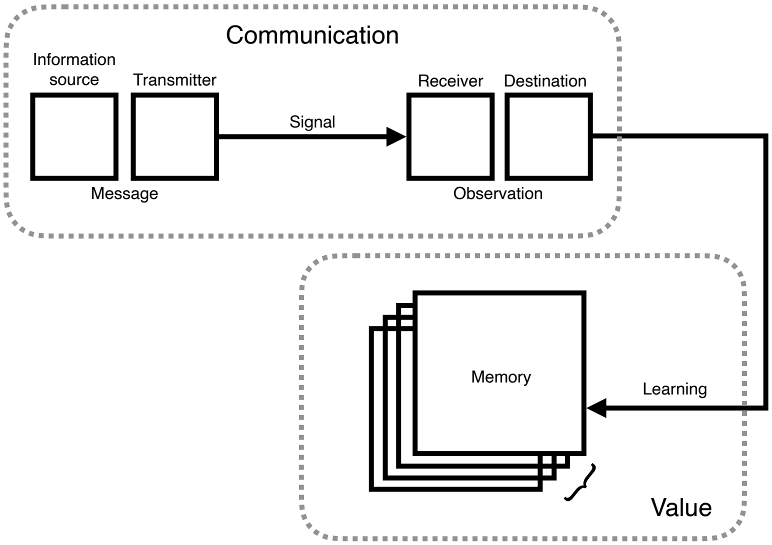

Figure 10. The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel directly. Value is derived from learning about observations from the channel, which is in turn dependent on memory and its past history.

different from what he learns when talking to Tom, even if the conversations are identical. What we mean by this example is that the personal history defines what is valuable on a channel, and this is independent of the channel and the messages sent. We have summarized this diagrammatically, in Fig. 10.

There is, we argue, an analogous set of levels for information value as those Weaver describes. a. There is the technical problem of judging how much was learned. b. There is the semantic problem of both what this learning "means", and also what its consequences are. c. There is an effectiveness problem of using what was learned to some other effect. In many ways b and c are similar to those in the information theory. It is a. which is most different.

For the same reasons Shannon avoided b and c in his theory, we have avoided them also. Both are very hard to measure, which makes them very hard to study. We feel it necessary to first build a theory of information value that can act as a companion for technical problem of information theory. That is to say, Shannon's account. This is why we took an axiomatic approach, and why we have tried to ape Shannon's axioms, as it made sense to do so.

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, and so we could or should reject it on that alone.

Questions about parsimony are resolved by considering the benefits versus the costs. The benefits to curiosity-based search is that it leads to no regret solutions to exploration, to exploration-exploitation, and that it so far seems to do very well in practice, at reward collection. At the same time curiosity-as-exploration can also be building a model of the environment, useful for later planning (**??**), creativity, imagination (**?**), while also building, in some cases, diverse action strategies (**????**). In other words, we argue it is a necessary complexity.

Is this too simple?

Solutions to the regular dilemma are complex and require sophisticated mathematical efforts and complex computations, when compared to our solution. Put another way, our solution may seem too simple to some. So have we “cheated” by changing the problem in the way that we have?

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

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 (**???????????**). If it predicts neural structures (**??**), that would be some evidence for it (we discuss this more below). If this theory proves useful in machine learning and artificial intelligence research, that would be some evidence for it (**??????????**). These are empirical predictions that require further investigation.

Is this a slight of hand, theoretically?

Yes. It is often useful in mathematical studies to take one problem that cannot be solved, and replace it with another related problem that can be. This is what we have done for exploration and the dilemma. The regular view of the problem has no tractable solution, without regret. Ours has a solution, with no regret.

In this case, we swap one highly motivated animal behavoir (reward seeking) for another (information seeking).

Does value as a technical problem even make sense?

Information value has been taken up as a problem in meaning (**?**), or usefulness (**?**). These are based on what we will call *b-type* questions (a notion we defined in the section above). It is not that b questions are not important or are not valuable. It is that they are second order questions. The first order questions are the ’technical’ or a-type questions (also defined above). Having a technical definition of value, free from any meaning, might seem counterintuitive for a value measure. We argue that it is not more or less counterintuitive than stripping information of meaning in the first place. Indeed, this is how Shannon got his information theory. The central question for our account of value is, is it useful? It is enough to ensure a complete but perfectly efficient search for information/learning in any finite space. It is enough to play a critical part in solving/replacing the dilemma, which has for many years looked intractable.

What about mutual information, or truth?

In other prototypical examples, information value is based on how well what is learned relates to the environment (**???**), that is the mutual information the internal memory and the external environment. Colloquially, one might call this “truth". The larger the mutual information between the environment and the memory, the more value it is said that we should then expect. This view seems to us at odds with actual learning behavior, especially in humans.

As a people we pursue information which is fictional (**?**), or based on analogy (**?**), or outright wrong (**?**). Conspiracy theories, disinformation, and our many more mundane errors, are far too commonplace for value to be based on fidelity alone. This does not mean that holding false beliefs cannot harm survival, but this is a second order question as far as learning goes. The fact that humans consistently seek to learn false things calls out us to accept that learning alone is a motivation for behavior.

So you suppose there is always positive value for learning of all fictions, disinformation, and deceptions?

We must. This is our most unexpected prediction. Yet, logically, it is internally consistent with our work we believe qualitatively consistent with the behavior of humans and animals alike.

What about information foraging?

Our work is heavily inspired by information foraging (**??**). Our motivation in developing our novel strategy was that information value should not depend on only a probabilistic reasoning, as it does in information foraging. This is for the simple reason that many of the problems animals need to learn are not probabilistic problems *from their point of view*. Of course, most any learning problem can be described as probabilistic from an outsider's perspective; the one scientists' rely on. But that means probabilistic approaches are descriptive approximations. We also feel curiosity is so important it needed a theoretical account that is both mechanistic and descriptive at the same time. This is why we developed an axiomatic approach that ended up addressing the technical problem of information value.

Was it necessary to build a general theory for information value just to describe curiosity?

No. It was an idealistic choice that worked out. The field of curiosity studies has shown that there are many kinds of curiosity. At the extreme limit of this diversity is a notion of curiosity defined for any kind of observation, and any kind of learning. If we were able to develop a theory of value driven search at this limit, we can be sure it will apply in all possible examples of curiosity that we seem sure to discover in future. At this limit we can also be sure that the ideas will span fields, addressing the problem as it appears in computer science, psychology, neuroscience, biology, and perhaps economics.

Doesn't your definition of regret ignore lost rewards when the curiosity policy is in control?

Yes. Regret is calculated for the policy which is in control of behavior, not the policy which *could* have been in control of behavior.

Is information a reward?

It depends. If reward is defined as more-or-less any quantity that motivates behavior, then our definition of information value is a reward. This does not mean, however, that information value and environmental rewards are interchangeable. In particular, if you add reward value and information value to motivate search you can drive exploration in practical and useful ways. But this doesn't change the intractability of the dilemma proper (**????**). So exploration will still generate substantial regret, as we show in Fig ??. But perhaps more important is that these kinds of intrinsic rewards (**????**) introduce a bias that will drive behavior away from what, could otherwise be, ideal skill learning (**??**).

But isn't curiosity impractical?

It does seem curiosity is just as likely to lead away from a needed solution as towards it. Especially so if one watches children who are the prototypical curious explorers (**??**). This is why we limit curiosity with boredom, and counter it with a competing drive for reward collecting (i.e., exploitation).

We believe there is a useful analogy between science and engineering. Science can be considered an open-ended inquiry, and engineering a purpose driven enterprise. They each have their own pursuits but they also learn from each other, often in alternating iterations (**?**). It is their different objectives though which make them such good long-term collaborators.

But what about curiosity on big problems?

A significant drawback to curiosity, as an algorithm, is that it will struggle to deliver timely results when the problem is very large, or highly detailed. First, it is worth noting that most learning algorithms struggle with this setting (**??**), that is unless significant prior knowledge can be brought to bear (**??**) or the problem can be itself compressed (**??**). Because "size" is a widespread problem in learning, there are a number of possible solutions and because our definition of learning, memory and curiosity is so general, we can take advantage of most. This does not guarantee curiosity will work out for all problems; all learning algorithms have counterexamples (**?**).

But what about other forms of curiosity?

We did not intend to dismiss the niceties of curiosity that others have revealed. It is that these prior divisions parcel out the idea too soon. Philosophy and psychology consider whether curiosity is internal or external, perceptual versus epistemic, or specific versus diversive, or kinesthetic (**???**). Computer science tends to define it as needed, for the task at hand (**?????**). Before creating a taxonomy of curiosity it is important to first define it, mathematically. Something we have done here.

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. This would also answer the question of what boredom is computationally. We cannot do this. It is a next problem we will work on, and it is very important. The central challenge is deriving boredom for any kind of learning algorithm.

Does this mean you are hypothesizing that boredom is actively tuned?

Yes we are predicting exactly that.

But do people subjectively feel they can tune their boredom?

From our experience, no. Perhaps it happens unconsciously? For example we seem to alter physiological learning rates without subjectively experiencing it (**?**).

How can an animal tune boredom in a new setting?

The short answer is that one would need to guess and check. Our work in Fig 2 suggests that this can be somewhat robust.

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 largely developed as separate disciplines (**???**). Indeed, we have highlighted how they are separate problems, with links to different basic needs: gathering resources to maintain physiological homeostasis (**??**) and gathering information to decide what to learn and to plan for the future (**??**). 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 (**?**). 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 (**???**). This neural separation for independent circuits has been observed in "modern" high animals, including zebrafish (**?**) and monkeys (**??**).

What about aversive values?

We do not believe this is complete theory. For example, we do not account for any consequences of learning, or any other aversive events. Accounting for this is another important next step.

The need to add other values fits well within our notion of competitive simple policies, in a win-stay loose-shift game. The need for taking other values into account is why we choose the notation we have. For example, if we added a fear value F our method could be expanded to read,

$$\Pi^{\pi} = \left[\pi_E,\ \pi_R,\ \pi_F\right] \tag{9}$$

$$\underset{\Pi^{\pi}}{\operatorname{argmax}} \left[E - \eta,\ \hat{R},\ F\right]_{\mathbf{(2,\,3,\,1)}} \tag{10}$$

Where we assume the fear/aversive term should take precedence over all other options, whe there is a tie.

What about other values?

Another more mundane example would be to add a policy for a grooming drive, C.

$$\Pi^{\pi} = \left[\pi_E,\ \pi_R,\ \pi_C\right] \tag{11}$$

$$\underset{\pi^{\pi}}{\operatorname{argmax}} \left[\hat{E} - \eta,\ \hat{R},\ \hat{C}\right]_{\mathbf{(1,\,2,\,3)}} \tag{12}$$

Where we assume the grooming term should take precedence over all other options, when there is a tie.

What about fitting deterministic models?

Exploration in the lab is often described as probabilistic (**????**). By definition probabilistic models do not make exact predictions of behavior, only statistical ones. Our approach is deterministic and so can make exact predictions. If, that is, we can fit it and it turns out to be correct. In species ranging from human to *Caenorhabditis elegans*, there are hundreds of exploration-exploitation experiments. A deterministic theory can, in principle, open up entirely new avenues for reanalysis. Testing and exploring these is a critical next step.

Does regret matter?

Another way around the dilemma is to ignore regret altogether. This is the domain of pure exploration methods, like the Gitten's Index (**?**), or probably-approximately correct approaches (**?**), or other bounded forms of reinforcement learning (**?**). Such methods can guarantee that you find the best value, eventually. These methods do not consider regret. But regret is critical for making sure that an animal's actions are efficient when resources and time are scarce, as they are in almost every natural setting.

Is the algorithmic run time practical?

It is common in computer science to study the run time of an algorithm as a measure of its efficiency. So what is the run time of our win stay, loose switch solution? There is unfortunately no fixed answer for the average or best case. The worst case algorithmic run time 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_π is $T_E + T_R$. Of course this is the worst case run time and would still be within reasonable ranges for learning in most naturalistic contexts.

Summary

There will certainly be cases in an animal's life when their exploration must focus on tangible physical rewards. Even here, searching for information about physical rewards, rather than searching for their value, will make the search more robust to deception, and more reliable in the future. We think of this as curiosity about reward. There are many more cases though where curiosity can include observations of the environment and the self. We have argued one should put aside intuitions

Box 1. Norms.

A **norm** is a mathematical way to measure the overall size or extent of an "object" in a space. The object we are concerned with is a mathematical object with a form given by, $\Delta\mathbf{M} = f_{\mathbf{M}}(x) - \mathbf{M}$. But to define the norm let us as the first step consider a generic vector v. Keeping with convention will have its norm denoted by $||v||$, and defined by three properties:

1. $||v|| > 0$ when $v \neq 0$ and $||v|| = 0$ only if $v = 0$ (point separating)
2. $||kv|| = ||k||\ ||v||$ (absolute scalability)
3. $||v + w|| \leq ||v|| + ||w||$ (the triangle inequality)

What does this mean for $\Delta\mathbf{M}$? A norm over a changing memory provides most the properties we require for our definition of $\hat{E}$. It depends only on learning (and forgetting), that dependence must be monotonic, a norm is zero only when nothing in the space/memory changes and is positive definite. These satisfy the first four properties. For the fifth, see Box 2.

Box 2. The Laplacian.

Here we use ∇^2 to represent the **Laplacian**, which provides the second derivative of vector valued functions, giving us the remaining property we need. (See Box 1 for the others). To explain why, suppose we are working with a 1- dimensional memory. In this case the Laplacian reduces to the second derivative on our memory, $\frac{d^2\mathbf{M}}{dx^2} < 0$. A negative second derivative will force the changes in memory to be decelerating, by definition. In higher dimensions of memory then, a reader can think of the Laplacian as the "second derivative" of vector-valued functions like f. That is, negative Laplacian imply decelerating by definition.

that suggest an open-ended curious search is too inefficient to be practical. We have shown it can be made very practical. We have argued one should put aside traditional intuitions for what exploratory behavior in animals represents. We have shown how there is a zero-regret way to solve the dilemma, a problem that has seemed unsolvable. Simply be curious.

Mathematical Appendix.

Information value as a dynamic programming problem

To find a dynamic programming solution based on the Bellman equation (Eq **??**) we must prove our memory $\mathbf{M}$ has optimal substructure. This is because the normal route, which assumes the problem rests in a Markov Space, is closed to us. By optimal substructure we mean that the process of learning in $\mathbf{M}$ can be partitioned into a collection, or series of memories, each of which is itself a solution. That is, it lets us find a kind of recursive structure in memory learning, which is what we need to apply the Bellman. Proving optimal substructure also allows for simple proof by induction (**?**), a property we will take advantage of.

————————————————————————————-

————————————————————————————-

Theorem 1 (Optimal substructure)**.** *Let X, A, $\mathbf{M}$, $f_{\mathbf{M}}$, (Def. 1), π_E and δ be given. Assuming transition function δ is deterministic, if $V^*_{\pi_E}$ is the optimal information value given by policy π_E, a memory $\mathbf{M}_{t+1}$ has optimal substructure if the the last observation x_t can be removed from $\mathbf{M}_t$, by $\mathbf{M}_t = f^{-1}(\mathbf{M}_{t+1}, x_t)$ 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 $\hat{\pi}_E \neq \pi_E$ that gives a memory $\hat{M}_{t-1} \neq M_{t-1}$ and for which $\hat{V}^*_{t-1} > V^*_{t-1}$.

Box 3. Optimal substructure

To use Bellman's principle of optimality (Box 4) the problem we want to solve needs to have **optimal substructure**. This opaque term can be understood by looking in turn at another theoretical construct, Markov spaces.

In Markov spaces there are a set of states or observations $(x_0, x_1, ..., x_T)$, where the transition to the next x_t depends only on the previous state x_{t-1}. This limit means that if we were to optimize over these states, as we do in reinforcement learning, we know that we can treat each transition as its own "subproblem" and therefore the overall situation has "optimal substructure".

In reinforcement learning the problem of reward collection is defined by fiat as happening in a Markov space (**?**). The problem for curiosity is it relies on memory which is necessarily composed of many past observations, in many orders, and so it cannot be a Markov space. So if we wish to use dynamic programming to maximize $\hat{E}$, we need to find another way to establish optimal substructure. This is the focus of Theorem 1.

To recover the known optimal memory M_t we lift $\hat{M}_{t-1}$ to $M_t = f(\hat{M}_{t-1}, x_t)$. This implies $\hat{V}^* > V^*$ which in turn contradicts the purported original optimality of V^* and therefore $\hat{\pi}_E$. □

This proof required two things. First, it was required we extend the idea of memory. We need to introduce a mechanism for forgetting of a very particular kind. We must assume that any last learning step $f(x, \theta) \rightarrow \theta'$ can be undone by a new vector valued function f^{-1}, such that $f(x, \theta') \rightarrow \theta$. In other words we must assume what was last remembered, can always be forgotten.

Second we must also assume the environment δ is deterministic. Determinism is consistent with the natural world, which does evolve in a deterministic way, at the scales we concerned with. This assumption is however at odds with much of reinforcement learning theory (**?**) and past experimental work (**?**). Both tend to study stochastic environments. We'll address this discrepancy later on using numerical simulations.

Bellman solution

Knowing the optimal substructure of $\mathbf{M}$ and given an arbitrary starting value E_0, the Bellman solution to curiosity optimization of $\hat{E}$ is given by,

$$V_E(x) = E_0 + \underset{a}{\text{argmax}} \left[E_t \right] \tag{14}$$

To explain this derivation we'll begin simply with the definition of a value function and work from there. A value function $V(x)$ is defined as the best possible value of the objective for x. Finding a Bellman solution is discovering what actions a to take to arrive at $V(x)$. Bellman's insight was to break the problem down into a series of smaller problems, leading a recursive solution. Problems that can be broken down this way have optimal substructure. The value function for a finite time horizon T and some payout function F is given by Eq 15. It's Bellman form is given by Eq. 16

$$V(x) = \underset{a}{\text{argmax}} \left[\sum_T F(x, a) \right] \tag{15}$$

$$V(x) = \underset{a}{\text{argmax}} \left[F(x_0, a_0) + V(x_1) \right] \tag{16}$$

In reinforcement learning the value function is based on the sum of future rewards giving Eq. 17-18,

$$V_R(x) = \underset{a}{\text{argmax}} \left[\sum_T R_t \right] \tag{17}$$

Box 4. Dynamic programming and Bellman optimality.

The goal of dynamic programming is to find a series of actions $(a_1, a_2, ..a_T)$, drawn from a set A, that maximize each payout $(r_1, r_2, ..., r_T)$ so the total payout received is as large as possible. If there is a policy $a = \pi(x)$ to take actions, based on a series of observations $(x_0, x_1, ..x_T)$., then an optimal policy π^* will always find the maximum total value $V^* = \text{argmax}_A \sum_T r_t$. The question that Bellman and dynamic programming solve then is how to make the search for finding $\pi*$ tractable. In the form above one would need to reconsider the entire sequence of actions for any one change to that sequence, leading to a combinatorial explosion.
Bellman's insight was a way to make the problem simpler, and break it down into a small set of problems that we can solve in a tractable way without an explosion in complexity. This is his principle of optimality, which reads:

> An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. (**?**)

Mathematically this allows us to translate the full problem, $V^* = \text{argmax}_\pi \sum_T r_1, r_2, ..., r_T$ to a recursive one, which is given below.

$$V^* = \underset{\pi}{\text{argmax}} \left[r_0 + V(x_1) \right] \tag{13}$$

Using this new form we can find an optimal solution by moving though the sequence of action iteratively, and solving each step independently. The question which remains though is how can we ensure our problem can be broken down this way? This is addressed in Box 3.

$$V_R(x) = \underset{a}{\text{argmax}} \left[R_0 + V(x_1) \right] \tag{18}$$

In our objective the best possible value is not found by temporal summation as it is during reinforcement learning. $\hat{E}$ is necessarily a temporally unstable function. We expect it to vary substantially before contracting to approach 0. It's best possible value is well approximated then by only looking at the maximum value for the last time step, making our optimization myopic, that is based on only last values encountered (**?**).

$$V_E(x) = \underset{a}{\text{argmax}} \left[E_t \right] \tag{19}$$

$$\begin{aligned} V_E(x) &= \underset{a}{\text{argmax}} \left[E_0 + V(x_1) \right] \\ &= E_0 + \underset{a}{\text{argmax}} \left[E_t \right] \end{aligned} \tag{20}$$

Good exploration by curiosity optimization

Recall from the main text we consider that a good exploration should,

1. Visit all available states of the environment at least once.
2. Should cease only once learning about the environment has plateaued.
3. Should take as few steps as possible to achieve criterion 1 and 2.

Limiting π_E to a deterministic policy makes proving these three properties amounts to solving sorting problems on $\hat{E}$. If a state is visited by our algorithm it must have the highest value. So if every state must be visited under a deterministic greedy policy every state must, at one time or another, generate maximum value. This certain if we know that all values will begin contracting towards zero. *Violations of this assumption are catastrophic*. We discuss this limit in the Discussion but note that things are not as dire as they may seem.

Definitions. Let Z be the set of all visited states, where Z_0 is the empty set $\{\}$ and Z is built iteratively over a path P, such that $Z \rightarrow \{x | x \in X \text{ and } x \notin Z\}$.

To formalize the idea of ranking we take an algebraic approach. Give any three real numbers (a, b, c),

$$a \leq b \Leftrightarrow \exists\ c;\ b = a + c \tag{21}$$

$$a > b \Leftrightarrow (a \neq b) \wedge (b \leq a) \tag{22}$$

Theorem 2 (State search: breadth)**.** *Given some arbitrary value E_0, an exploration policy governed by π^*_E will visit all states $x \in X$ in finite number of steps T.*

Proof. Let $\mathbf{E} = (E_1, E_2, ...)$ be ranked series of $\hat{E}$ values for all states X, such that $(E_1 \geq E_2, \geq ...)$. To swap any pair of values $(E_i \geq E_j)$ so $(E_i \leq E_j)$ by Eq. 21 $E_i - c = E_j$.

Therefore, again by Eq. 21, $\exists \int \delta E(s) \rightarrow -c$.

Recall: $\nabla^2 f_{\mathbf{M}}(x) < 0$ after a finite time T.

However if we wished to instead swap $(E_i \leq E_j)$ so $(E_i \geq E_j)$ by definition $\nexists c; E_i + c = E_j$, as $\nexists \int \delta \rightarrow c$.

To complete the proof, assume that some policy $\hat{\pi}_E \neq \pi^*_E$. By definition policy $\hat{\pi}_E$ can be any action but the maximum, leaving $k - 1$ options. Eventually as $t \rightarrow T$ the only possible swap is between the max option and the kth, but as we have already proven this is impossible as long as Axiom 2 holds. Therefore, the policy $\hat{\pi}_E$ will leave at least 1 option unexplored and $S \neq Z$. □

Theorem 3 (State search: depth)**.** *An exploration policy governed by π^*_E will only revisit states for where learning is possible.*

Proof. Recall: Axiom 2. Each time π^*_E visits a state s, so $\mathbf{M} \rightarrow \mathbf{M}'$, $F(\mathbf{M}', a_{t+dt}) < F(\mathbf{M}, a_t)$

In Theorem 2 we proved only a deterministic greedy policy will visit each state in X in T trials.

By induction, if $\pi^* E$ will visit all $x \in X$ in T trials, it will revisit them at most $2T$, therefore as $T \rightarrow \infty$, $E \rightarrow \eta$. □

We assume in the above the action policy π_E can visit all possible states in X. If for some reason π_E can only visit a subset of X then proofs above apply only to that subset.

These proofs come with some fine print. E_0 can be any positive and finite real number, $E_0 > 0$. Different choices for E_0 will not change the proofs, especially their convergence. So in that sense one can chosen it in an arbitrary way. Different choices for E_0 can however change individual choices, and their order. This can be quite important in practice, especially when trying to describe some real data. This choice will not change the algorithm's long-term behavior *but* different choices for E_0 will strongly change the algorithm's short-term behavior. This can be quite important in practice.

Optimality of π_π

Recall that in the main text we introduce the equation below as a candidate with zero regret solution to the exploration-exploitation dilemma.

$$\pi^\pi = \left[\pi_E,\ \pi_R\right] \tag{23}$$

$$\underset{\pi^\pi}{\text{argmax}}\ \left[E_{t-1} - \eta,\ R_{t-1}\right]_{\mathbf{(2,\,1)}} \tag{24}$$

In the following section we prove two things about the optimality of π_π. First, if π_R had any optimal asymptotic property for value learning before their inclusion into our scheduler, they retain that optimal property under π_π when $\eta = 0$, or is otherwise sufficiently small. Second, show that if both π_R and π_E are greedy, and π_π is greedy in its definition, then Eq 23 is certain to maximize total value. The total value of R and $\hat{E}$ is the exact quantity to maximize if information seeking and reward seeking are equally important, overall. This is, as the reader may recall, one of our key assumptions. Proving this optimality is analogous to the classic activity selection problem from the job scheduling literature (**??**).

Theorem 4 (π_π is unbiased)**.** *Let any S, A, $\mathbf{M}$, π_R, π_E, and δ be given. Assuming an infinite time horizon, if π_E is optimal and π_R is optimal, then π_π is also optimal in the same sense as π_E and π_R.*

Proof. The optimality of π_π can be seen by direct inspection. If $p(R = 0) > 0$ we are given an infinite horizon, then π_E will have a unbounded number of trials meaning the optimally of P^* holds. Likewise, $\sum E < \eta$ as $T \rightarrow \infty$, ensuring pi_R will dominate π_π therefore π_R will asymptotically converge to optimal behavior. □

In proving this optimality of π_π we limit the probability of a positive reward to less than one, denoted by $p(R_t = 1) < 1$. Without this constraint the reward policy π_R would always dominate π_π when rewards are certain. While this might be useful in some circumstances, from the point of view π_E it is extremely suboptimal. The model would never explore. Limiting $p(R_t = 1) < 1$ is a reasonable constraint, as rewards in the real world are rarely certain. A more naturalistic way to handle this edge case is to introduce reward satiety, or a model physiological homeostasis (**??**).

Optimal scheduling for dual value learning problems

In classic scheduling problems the value of any job is known ahead of time (**??**). In our setting, this is not true. Reward value is generated by the environment, *after* taking an action. In a similar vein, information value can only be calculated *after* observing a new state. Yet Eq. 23 must make decisions *before* taking an action. If we had a perfect model of the environment, then we could predict these future values accurately with model-based control. In the general case though we don't know what environment to expect, let alone having a perfect model of it. As a result, we make a worst-case assumption: the environment can arbitrarily change–bifurcate–at any time. This is a highly nonlinear dynamical system (**?**). In such systems, myopic control–using only the most recent value to predict the next value– is known to be an robust and efficient form of control (**?**). We therefore assume that last value is the best predictor of the next value, and use this assumption along with Theorem 5 to complete a trivial proof that Eq. 23 maximizes total value.

Optimal total value

If we prove π_π has optimal substructure, then using the same replacement argument (**?**) as in Theorem 5, a greedy policy for π_π will maximize total value.

Theorem 5 (Total value maximization of π_π)**.** *Let any S, A, $\mathbf{M}$, π_R, and δ be given. If π_R is defined on a Markov Decisions, then π_π is Bellman optimal and will maximize total value.*

Proof. We assume Reinforcement learning algorithms are embedded in Markov Decisions space, which by definition has the same decomposition properties as that found in optimal substructure.

Recall: The memory $\mathbf{M}$ has optimal substructure (Theorem 1.

Recall: The asymptotic behavior of π_R and π_E are independent under π_π (Theorem 5

Recall: The controller π_π is deterministic and greedy.

If both π_R and π_E have optimal substructure and are independent, then π_π must also have optimal substructure. If π_π has optimal substructure, and is greedy-deterministic then it is Bellman optimal. □

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, and Jack Burgess for his assitance with Figure **??**.

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.

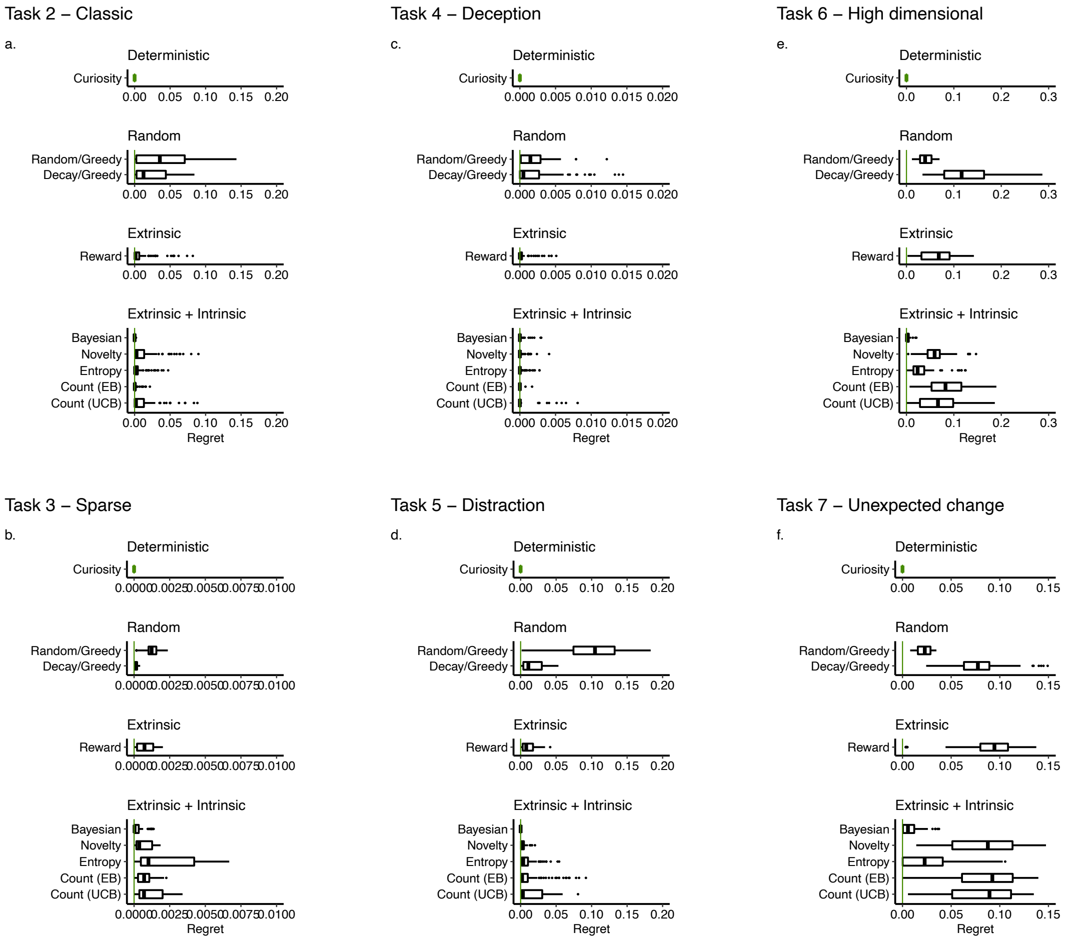

Figure 7–Figure supplement 1. Summary of regret (*Tasks 2-7*)

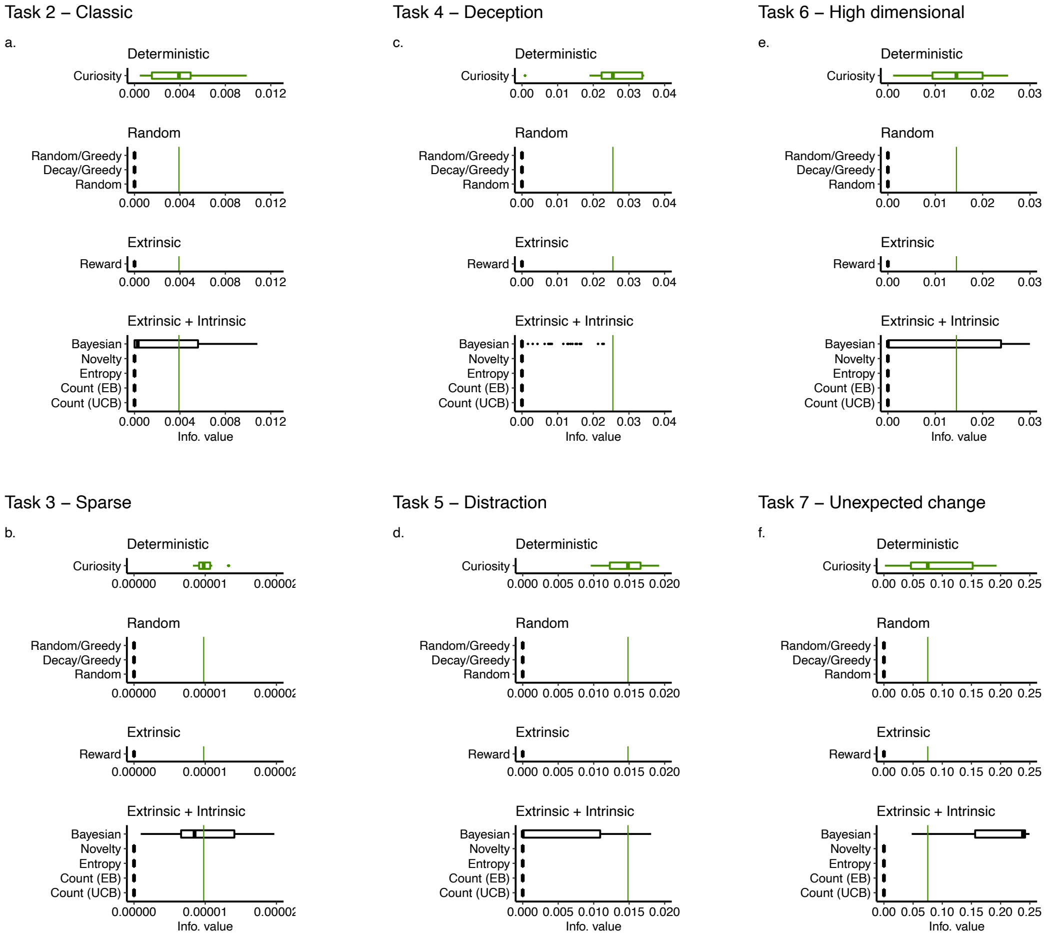

Figure 7–Figure supplement 2. Summary of information value (*Tasks 2-7*)

764

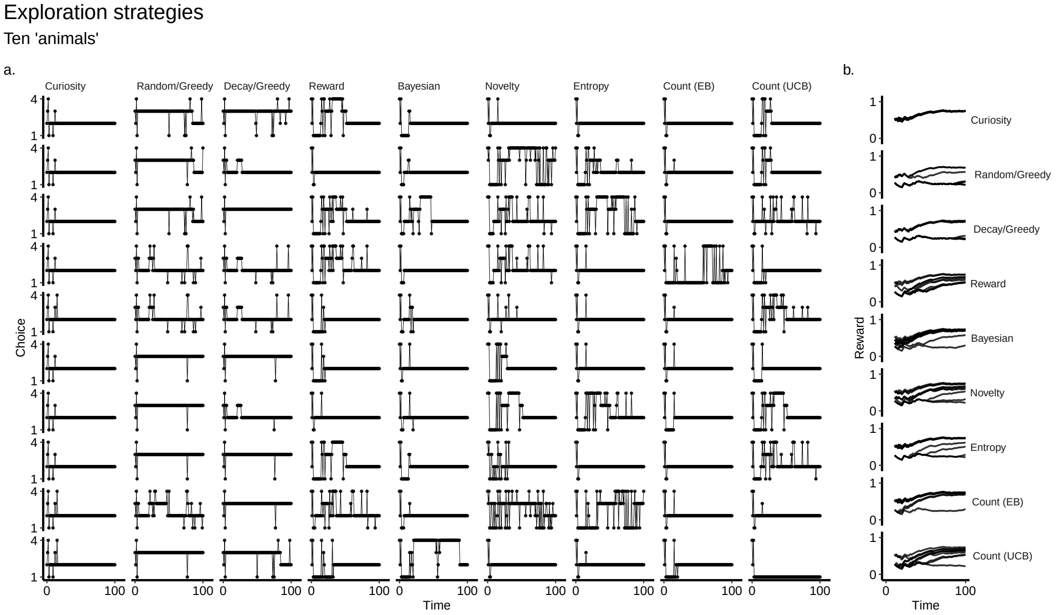

Figure 7–Figure supplement 3. Examples of exploration (Task 1). Here we simulated different sets of hyper-parameters on the same task and random seed. We view each parameter set as a unique “animal” (**?**). So this figure estimates the degree and kind of variability we might expect in a natural experiment.

765

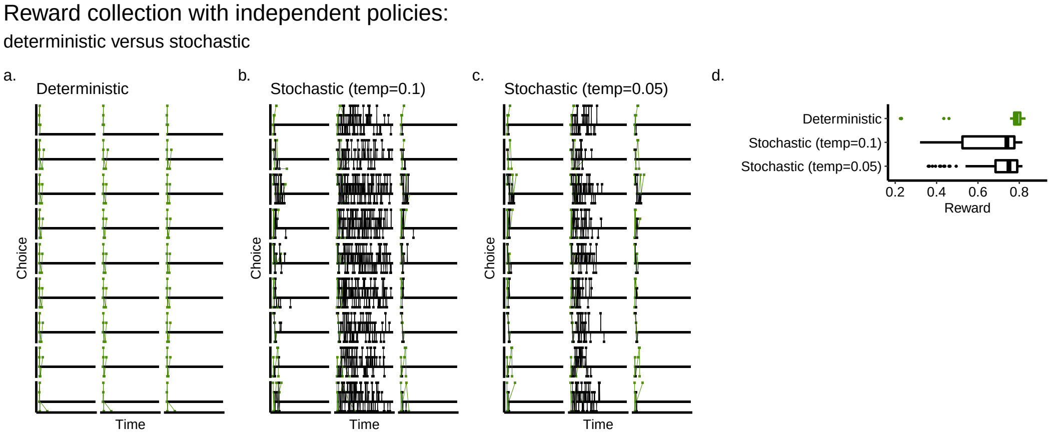

Figure 9–Figure supplement 1. Adding noise to deterministic curiosity does degrade its performance. In this example control we added two levels of random noise. This only served to increase variability of convergence (a-c) and reduce the total rewards collected (d). Note: decision noise was modelled by sampling a Boltzmann distribution, as described in the *Methods*.

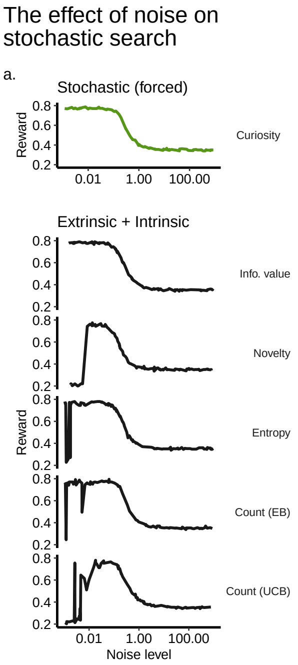

Figure 9–Figure supplement 2. Noise across strategies. Adding noise to degrades performance to nearly the same degree (Task 2).