# Curiosity is the trick for solving explore-exploit problems

**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**   Explore-exploit problems are a common but intractable problems in reward learning, and therefore in the biological sciences. In this paper, we provide an alternative view of these problems that does not depend on exploring to optimize for reward value, even though this remains the goal of exploitation. Here exploration has no objective but learning itself which we equate to curiosity. Through theory and experiments we show this "curiosity trick" leads to optimal value solutions for potentially all explore-exploit problems. In particular, we show this approach succeeds in naturalistic conditions, when rewards are sparse, deceptive, and non-stationary. We also suggest three experiments that might be used to distinguish our approach from other theories.

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. It's goal is to select actions in a way that maximizes the total reward collected from the environment, for every state.
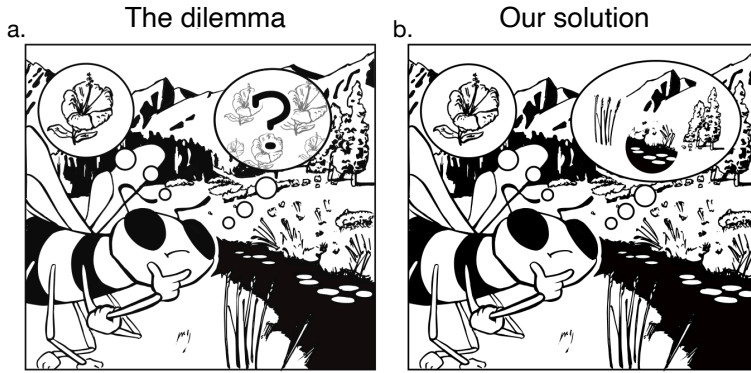
**Figure 1.** Two views of exploration and exploitation. **a**. The dilemma: either exploit an action with a known reward (e.g., return to the previous flower) or explore other actions on the chance they will return a better outcome. The central challenge here is that the outcome of exploration is uncertain, 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.

How should an animal do this? To maximize reward an animal must tradeoff between explo- ration, 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 (**?**). First though we need to establish some shared formalities.

## Markovian formalities - reward collection

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 is handled by the transition function $\Lambda(\mathbf{S}, \mathbf{A})$. In our mathmatics we assume $\Lambda$ is a deterministic function of its inputs. In our simulations we assume it is a stochastic function.

These prelimaries in place we can write down the value function definition, and the standard Bellman equation **?** for reward collection, as given by reinforcement learning theory (**?**). We assume below that the horizon is finite, bounded by $T$, for simplicity.

$$
\begin{aligned}
V_R^*(\mathbf{S}) &= \operatorname*{argmax}_{\mathbf{A} \in \mathbb{A}} \left[ \sum_{k=0}^{T} R \right] \\
&= \operatorname*{argmax}_{\mathbf{A} \in \mathbb{A}} \left[ R_t + \sum_{k=0}^{T} R_{t+1} \mid \mathbf{S_t} = \mathbf{S}, \ \mathbf{A_t} = \mathbf{A} \right] \\
&= \operatorname*{argmax}_{\mathbf{A} \in \mathbb{A}} \left[ R_t + V_R^*(\mathbf{S_{t+1}}) \mid \mathbf{S_t} = \mathbf{S}, \ \mathbf{A_t} = \mathbf{A} \right] \\
&= R_0 + V_R^*(\mathbf{S_t}) + V_R^*(\mathbf{S_{t+1}}), \ \dots
\end{aligned}
\tag{1}
$$

The second to last equation is the optimal Bellman equation for $V_R^*$. This can be restated more compactly as,

$$
V_R^*(\mathbf{S}) = \operatorname*{argmax}_{\mathbf{A} \in \mathbb{A}} \left[ R_t + V_R^*(\Lambda(\mathbf{S}, \mathbf{A})) \right]
\tag{2}
$$

## non-Markovian formalities - information collection

Imagine we wish to arrive an equation to maximize curious behavoir, in terms of information gained, what we'll call information value. Assume, as a stand-in, $\hat{E}$ represents the information value of some observation. What we would like is solution for this term $\hat{E}$ that would let us arrive at an

63 equation like Eq. 2, but for information value in place of reward. This could hypothetical equation
64 could be written as,

$$V_{\hat{E}}^*(\mathbf{S}) = \underset{\mathbf{A} \in \mathbb{A}}{\operatorname{argmax}} \left[ \hat{E}_t + V_{\hat{E}}^*(\Lambda(\mathbf{S}, \mathbf{A})) \right] \tag{3}$$

65 How can we arrive at such an equation that is both general, and therefore suitaable for use
66 across species, and how can we do so when we necessaailry limit ourselves to non-Markovians kind
67 of memory? To setup the formality we need to answer these questions we define observations,
68 then memory and learning. This sets the stage for defining $\hat{E}$ itself.
69 *Obervations.* Animals at different stages in development, and in different envinroments, express
70 curious behvaoir about the enviriniment, their own actions, and their own mental processes. We
71 assume a parsimonious way to handle this is be assuming curious animals often "bind" **?** the
72 elements of environment, the Markov process ($mathbf S, \mathbf{A}, \mathbf{S}', R$), into single observations $\mathbf{X}$. For
73 conveience, we assume observations are embedded in a finite real space, $\mathbf{X} \in \mathcal{X}$ of size $l$. This
74 binding is imagined to happen via a communication channel, $T$, identical to our observation function,
75 $T(\mathbf{S}, \mathbf{A}, \mathbf{S}', R, \mathbf{M})$. ($\mathbf{M}$ is the animal's memory, that we define below.)
76 Observations may contain action information $\mathbf{A}$ but this optional. Observations may contain
77 reward information $R$ or not. Though ommiting reward would hamper solving dilemma. Observa-
78 tions made by $T$ can also be internally generated from memory $\mathbf{M}$, or externally generated from
79 the environment ($\mathbf{S}, \mathbf{S}'$). Or both in combination, as happens in recurrent neural circuits.
80 *Memory.* Sustained firing, the strength between two synapses, elevated Calcium concentration
81 are three simple examples of learning and memory in the nervous system. Each of these examples
82 though can be represented as a vector space. In fact, most any memory system can be so defined.
83 We define memory as a set of real valued numbers, embedded in a finite space that is closed and
84 bounded with $p$ dimensions. A learning function is then any function $f$ that maps observations $\mathbf{X}$
85 into memory $\mathbf{M}$. This $f$ is considered to be a valid learning function as long as the mapping changes
86 $\mathbf{M}$ some small amount. A non-constant mapping, in other words. Using recursive notation, this kind
87 of learning is denoted by $\mathbf{M} \leftarrow f(\mathbf{X}, \mathbf{M})$. Though we will sometimes use $\mathbf{M}'$ to denote the updated
88 memory in place of the recursion. Other times we will add subscripts, like ($\mathbf{M}_t, \mathbf{M}_{t+1}, \ldots$), to express
89 the path of a memory.
90 We can also define a forgetting function that *can be any function* that inverts the memory,
91 $f^{-1}(\mathbf{X}; \mathbf{M}') \rightarrow \mathbf{M}$. Forgetting might not seem crucial, here. But it will prove essential later on in
92 developing exploration algorithms that provably maximize $\hat{E}$.
93 Most important, we do not assume memory is Markovian.
94 For individual animals we assume $f$ and $f^{-1}$ have been chosen by evolution and experience to
95 be learnable, efficient, and have sufficiently useful inductive bias for their particular environment
96 **??**. This assumption is not necessary for the mathematical results which follow.

## Formal regret

98 Regret minimization is a standard metric of optimality in reinforcement learning (**?**). Formally, given
99 *any* any optimal value solution we can define regret $G$ in terms of this, given by $G = V^* - V_t$, where
100 $V_t$ is the value of the choosen action $\mathbf{A}_t$. An optimal value algorithm can then be restated, as a "best"
101 series of best actions ($\mathbf{A}_t, \mathbf{A}_{t+1}, \mathbf{A}_{t+2}, \ldots$) for which $\sum_{k=0}^{T} G_t = 0$.

## A general kind of (axiomatic) curiosity

103 To make our eventual solution to the dilemma general, we first need to separate reward value
104 from information value in a principled and general way. We do this axiomatically because working
105 models of curiosity tend to conflate the learning rule, with the objective itself. We reason instead
106 that the value of any observation made by an animal depends entirely on what an animal learns by
107 making that observation—literally how it changes that animal's memory.
108 We assume curiosity is driven by information value, and should be as time-efficients as possible.
109 We will now describe an optimal value, but deterministic model, of curiosity. It is defined by some

easily satisfiable requirements (axioms), which makes information value and curiosity domain general, and learning rule independent.

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 memory changes, by $f$.

**Axiom 2** (Axiom of Novelty). *$\hat{E} = 0$ 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 5** (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 $\delta\mathbf{M} \to 0$ and $\hat{E} \to 0$.

All we have really done in these axioms is remove any mention of a learning rule, its properties, or learning target, its properties from previous definitions **??????**. This is important not for its on sake but because it with thesse axioms we can consider information value, and therefore curiosity, independent of any learning semantics, or any semantics at all. Anything we prove in this more abstract setting will be therefore be true for any learning algorithm, who meets our requirements.

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, ||.|| (as shown in Fig. 2). Though not necessaailry a unique solution, a norm on the memory gradient is a simple, perhaps natural, way to satisfy Axioms 1-4 (Eq 4).

$$\hat{E} = || \nabla\mathbf{M} || \tag{4}$$

To satify Axiom 5 we further 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 converge. We take this to mean that in finite time bound, $T^*$, $\hat{E}$ must approach 0 (as shown in Fig. 2). We term this consistent learning.

In practice here we will not work with the gradient specifically, but use a discrete time map in its place,

$$\hat{E} = || f(\mathbf{M_t}, \mathbf{X}) - \mathbf{M_t} || \tag{5}$$

## Definition 1 – distance

a.
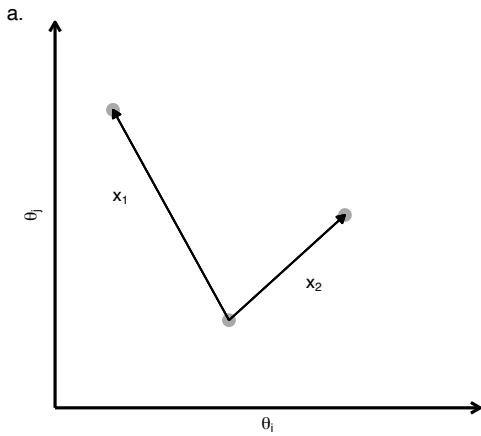


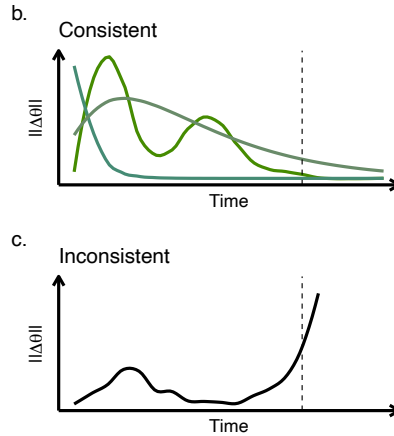## Definition 2 – deceleration

b.



c.



**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.

---

147  ## Bellman curiosity

148  An ideal learning-rule independent definition of curiosity would be a Bellman solution of this type (**?**),
149  as we noted above. It is ideal because it guarantees optimal value using recursion–a very biological
150  idea. It is convenient also because the Bellman equation is the basis for the theory of reinforcement
151  learning (**?**), as we also illustrated above.

152      The common way to arrive at Bellman solution is to assume a Markov decision space. This is
153  problem for our defition of memory, which we assume has a natural long-term path dependence.
154  This breaks the Markov assumption, and its neeeded claim of optimal substructure. To get around
155  this we proved that exact forgetting of the last observation is another way to achieve optimal
156  substructure, and so to derive a Bellman solution.

157      Given an arbitrary starting value $E_0 > 0$, the best solution to maximizing $\hat{E}$ is given by Eq. 6. A
158  full derivation for this fact is provided in the Appendix.

$$V_E^*(\mathbf{X}) = \operatorname*{argmax}_{\mathbf{A}} \left[ E_0 + V_E(\mathbf{X}_1) \right] \tag{6}$$

159      As long learning about the different observations $\mathbf{X}$ is independent, and as long as learning
160  continues until steady-state, defined as $E_t < \eta$, there are many equally good solutions. Under these
161  conditions we can simplify Eq. 6 further, giving Eq. 7.

$$V_E^*(\mathbf{X}) = \operatorname*{argmax}_{\mathbf{A}} \left[ E_t + E_{t+1} \right] \tag{7}$$

162  ## A note on boredom

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

169    To limit curiosity we introduction a penalty term, $\eta$ which we treat as synomoous with boredom.
170 We consider boredom an adaptive trait, in other words. Others have considered boredom arguing
171 it is a useful way to motivate aversive tasks (**?**), or curiosity **?**. We take a slightly different view. We
172 treat boredom as a tunable free parameter, $\eta \geq 0$ and a means to ignore marginal value by requiring
173 curious exploration to stop once $E \leq \eta$.

## A note on exploration quality

175 Having a policy $\pi_E^*$ that optomally maximizes $\hat{E}$ does not necessarily ensure that exploration is of
176 good quality. We think it is reasonable to call an exploration good if it meets the criteria below.

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

180    In Theorems 2 and 3 we prove that our Bellman solution $\pi_E$ satisfies these, when $\eta > 0$ and
181 when the observations $\mathbf{X}$ contain complete state information, $\mathbf{S} \in \mathbf{X}$. See the Appendix for the
182 proofs proper.

183 **Theorem 1** (State search: breadth). *Given some arbitrary value $E_0$, an exploration policy governed by*
184 $\pi_{\hat{E}}^*$ *will visit all states $\mathbf{S} \in \mathbb{S}$ in finite number of steps $T$.*

185 **Theorem 2** (State search: depth). *An exploration policy governed by $\pi_{\hat{E}}^*$ will only revisit states for where*
186 *information value is marginally useful, defined by $\hat{E} > \eta$.*

## Two key conjectures

188 Searching in an open-ended way for information must be less efficient than a direct search for
189 reward? In answer to this question, we make two conjectures.

190 **Conjecture 1** (A hypothesis of equal important). *Reward value and information value are equally*
191 *important for survival, in changing environments*

192    It's worth reiterating here, that curiosity is a primary drive in most, if not all, animals **?**. It is as
193 strong, if not sometimes stronger, than the drive for reward **???**.
194    If both reward and information are equally important, then time is not wasted in a curious
195 search, and so the process is not *necessarily* inefficient or even indirect. Instead, we offer a new
196 question. Is curiosity practical enough to serve as a useful "trick" to solve reward collection search
197 problems? We answer this with a second conjecture,

198 **Conjecture 2** (The curiosity trick). *When learning is possible, curiosity is a sufficient solution for all*
199 *exploration problems.*

200    It's worth noting here that curiosity, as an algorithm in machine learning, is highly effective at
201 solving broad kinds of optimization problems **??????????**.

## A win-stay, lose-switch solution

203 If reward and information are equally important, and curiosity is sufficient, how can an animal go
204 about balancing them? If you accept our conjectures, then answering this question optimally is the
205 same as solving the dilemma. To solve this dual value learning problem we've posed—information
206 and reward maximization—we need an algorithm that can maximize the value of a mixed sequence,
207 $V_{\hat{E}R}$. For example,

$$V_{\hat{E}R}^*(\mathbf{S}) = \underset{\mathbf{A} \in \mathbb{A}}{\mathrm{argmax}} \left[ \sum_{k=0}^{T} \hat{E}_t + \hat{E}_{t+1} + R_{t+2} + \hat{E}_{t+3} + R_{t+4} + \dots \right] \tag{8}$$

208    A similar dual value optimization problem was posed, solved, with its regret bounded, by Robbins
209 **?** in his derivation of the win-stay lose-swtich rule. This WLSL approach has since proven uesful in

210 decision making and reinforcement learning **??**, Bayesian sampling **?**, and especially in game theory
211 **?**. We also put it to use here.

212 We have come to think about our two optimal value policies $\pi_R$ and $\pi_{\hat{E}}$ as two "players", who are
213 competing for behavioral control of an animals moment-by-moment actions **?**. We a express simple
214 solution to this game as a set of inequalities, that implement a WSLS rule, and which we term $\Pi_\pi$.

$$\Pi_\pi = \begin{cases} \pi_{\hat{E}}^* & : \hat{E} - \eta > R + \rho \\ \pi_R & : \hat{E} - \eta < R + \rho \end{cases} \tag{9}$$

215 Here *eta* is the boredom tern we defined earlier. We also introduce the reward bias term $\rho$,
216 where $\rho > \eta$. This bias ensures no ties can occur, and that as $\hat{E}$ decreases to 0, as it must, the default
217 policy is reward collection. That is, by $\pi_R$.

218 In Eq. 25 we compare reward $R$ and information value $\hat{E}$ but have not made sure they have
219 same "units", or are otherwise comparible. We sidestep this isssue by limiting the probability of
220 having zero reward to be, itself, non-zero. That is, $p(R_t = 0) > 0$. We also limit $E_0$, the first set
221 of values, to be positive and non-zero when used with boredom, $(E_0 - \eta) > 0$. These constraints
222 ensure complete "good" exploration (defined above),which in turn ensures optimal reward learning
223 is possible irrespitive of the reltative scales of $R$ and $\hat{E}$.

224 To provide some intuition for Eq.25, imagine that in the previous action an animal observed,
225 arbitrarily, 0.7 units of information value, $\hat{E}$, and no reward (i.e. $R = 0$). Using our rule on the next
226 action the animal should then choose its exploration policy $\pi_{\hat{E}}^*$. If it explored last time, this is a "stay"
227 action. If not, it is a "swtich". Let's then say that with this next observation, $\hat{E}$ decreases to 0.2 and a
228 reward value of 1.0 is observed. Now the animals should swtich its strategy for the next round, to
229 exploit instead.

230 In general, if there was more reward value then information value last round ($R > \hat{E}$), our rule
231 dictates an animal should choose the reward policy. If information value dominated, $R < \hat{E}$, then it
232 should choose the information gathering policy.

233 The WSLS rule in Eq 25 when applied to exploration-exploitation problems can gauentee three
234 things, shown below. In each Theoreum we assume our a finite horizon $T$ and that definitions for **X**,
235 **A**, **M**, $f$, $\Lambda$, and $\Pi_\pi$ are implicitly given.

236 **Theorem 3.** *When $\pi_R$ is in control under $\Pi_\pi$, all actions are zero regret in terms of $V_R$. That is, $\sum_{k=0}^{T} G = 0$.*

237 **Theorem 4.** *When $\pi_{\hat{E}}$ is in control under $\Pi_\pi$, all actions are zero regret in terms of $V_{\hat{E}}$. That is, $\sum_{k=0}^{T} G = 0$.*

238 **Theorem 5.** *When either $\pi_{\hat{E}}$ or $\pi_R$ is in control under $\Pi_\pi$, all actions are zero regret in terms of $V_{\hat{E}R}$. That*
239 *is, $\sum_{k=0}^{T} G = 0$.*

240 Proofs for Theorems 4 and 4 follow from both reinforcement learning and our definition of
241 information optimization having a Bellman optimality. The proof for the optimality of $V_{\hat{E}R}$ (Th 5) in
242 provided in the Appendix. The intuition for this proof is simple that a greedy algorithm made over
243 two Bellman solutions is itself a Bellman solution.

244 In Eq **??** we compare reward $R$ and information value $E$ but have not ensured they have same
245 "units", or are comparible valuables. We sidestep this isssue by limiting the probability of having
246 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
247 be positive and non-zero when used with boredom, $(E_0 - \eta) \geq 0$. These constraints ensure complete
248 "good" exploration (defined above) which in turn ensures optimal reward learning is possible.

249 ## A note about subjective reward value

250 We have been reward value is fixed. That is, where environment reward value $R$ is always equal to
251 it subjective value to the animal $\hat{R}$. In natural behavior though the motivational value of reward
252 often declines as the animal reaches satiety (**???**). Fortunately, it has already been shown a reward
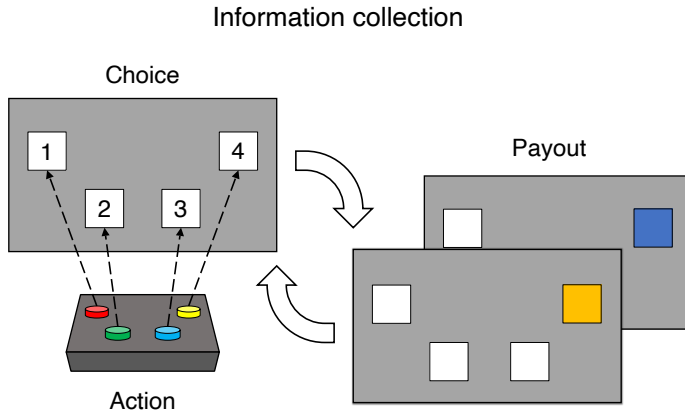
Information collection

Choice



**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 ).

253 collection policy designed for one style will work for the other (**?**). However to use a form similiar
254 to Eq. 25 for subjective or homeostatic reward value we need to make one modification. Under
255 homeostasis ties between values in Eq. 25 should be broken in favor of exploration because if this
256 is not so, and exploitation remains the defult, homeostasis is not possible. This modification leads
257 to Eq. 10.

$$\Pi_\pi = \begin{cases} \pi^*_{\hat{E}} & : \hat{E} - \eta > R - \rho \\ \pi_R & : \hat{E} - \eta < R - \rho \end{cases} \tag{10}$$

### Information collection in practice

259 The work so far has built up the idea that the most valuable, and most efficient, curious search will
260 come from a deterministic algorithm. That is, every step strictly maximizes $\hat{E}$. It is this determinism
261 which will let us resolve the dilemma, later on. A deterministic view of exploration seems at odds
262 with how the problem is generally viewed today, which involves searching with some amount of
263 randomness. Whereas if our analysis is correct, randomness is not needed or desirable, as it must
264 lead to less value and a longer search.
265 We confirm and illustrate this result using an example of a simple information foraging task
266 (Task 1; Fig.). This variation of the bandit task (**?**) replaces rewards with information, in this case
267 colors. On each selection, the agent sees one of two colors according to a specific probability shown
268 in Fig. 6a. When the relative color probabilities are more similar that arm has more entropy and so
269 more to learn and more information value. Arms that are certain to present only one color lose
270 their informative value quickly.
271 The results of this simple information seeking task are shown in Figure 4. Deterministic curiosity
272 in this task generated more information value, in less time, and with zero regret when compared
273 against a stochastic agent using more directed random search. As our work here predicts, noise
274 only made the search happen slower and lead to regret. Besides offering a concrete example,
275 performance in Task 1 is a first step in showing that even though $\pi_E$'s optimality is proven for a
276 deterministic environment, it can perform well in other settings.

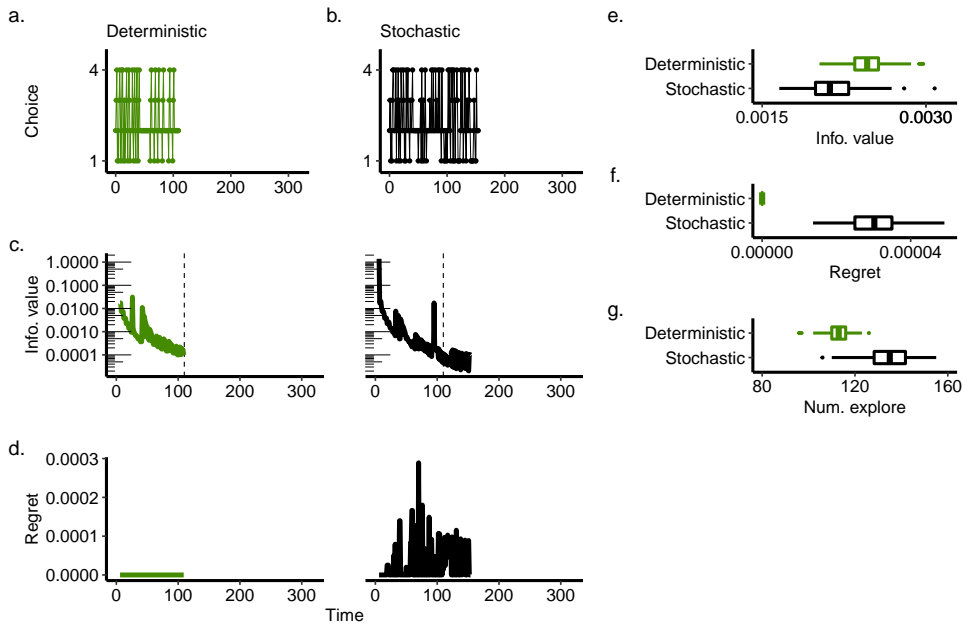## Deterministic versus stochastic curiosity in a random world



**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.
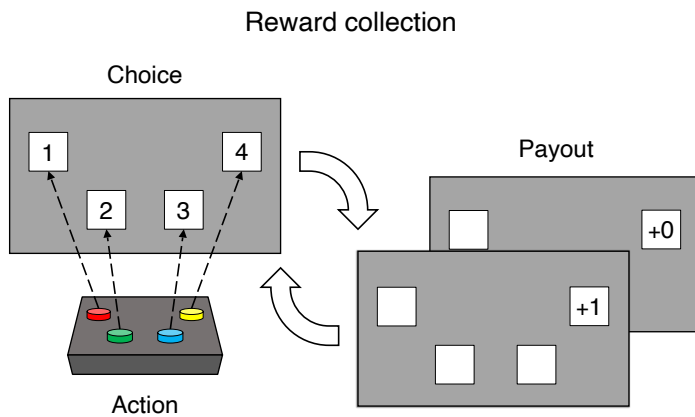
**Figure 5.** A 4 choice reward collection task. The reward is numerical value, here a 1 or 0. A good learner in this task is one who collects the most rewards. The task depicted here matches that in Fig 6b. Every other reward collection task we study has the same basic form, only the number of choices 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 $\epsilon$. |
| Decay/Greedy | Random | The $\epsilon$ parameter decays with a half-life $\tau$ |
| 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 |

## Reward collection in practice

Our approach does not mathematically ensure a dual value solution to the dilemma maximizes reward value comparerd with other well-established methods. To findout we meaasured total reward collected over 7 tasks and 10 agents, including ours. Each agent's parameters were independently optimized for each task. For a brief description of each agent, see Table 1. They all have in common that their central goal is to maximize total reward value, though this is often supplemented by some other goal, an intrinic reward or bonus **??**.

The general form of the 7 tasks are depicted in Fig 5. As with our first task (Fig.  ) they are all variations of the classic multi-armed bandit (**?**). The payouts for each of the tasks are shown separately in Fig. 6. Each task was designed to either test exploration performance in a way that matches recent experimental studies, 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.

The results in full for all tasks and exploration strategies are shown in Fig. **??**. All agents, including ours, used the same exploitation policy based on the temporal difference learning rule (**?**) (Methods).
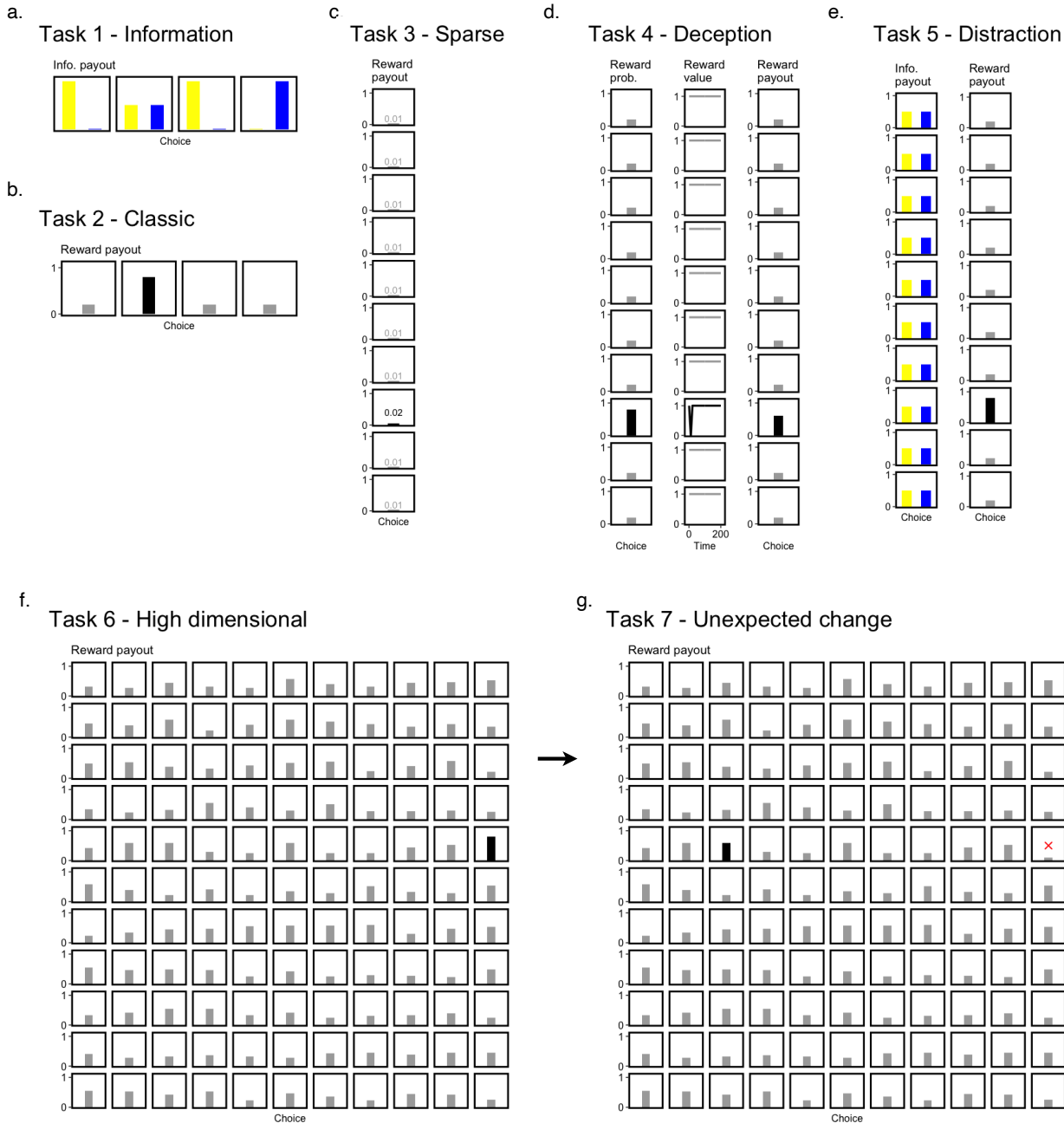
**Figure 6.** Payouts for *Tasks 1 - 7*. Payouts can be information, reward, or both. For comments on general task design, see Fig 5. **a.** A classic four-choice design for information collection. A good learner should visit each arm, but quickly discover that only arm two is information bearing. **b.** 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.
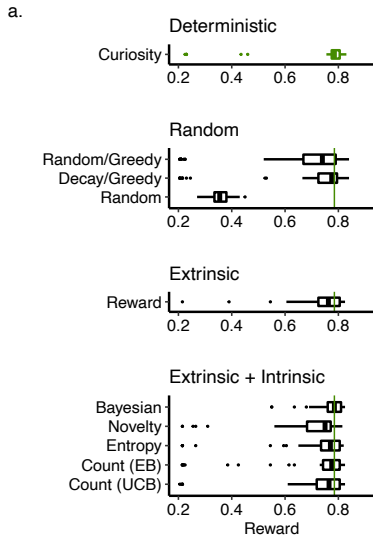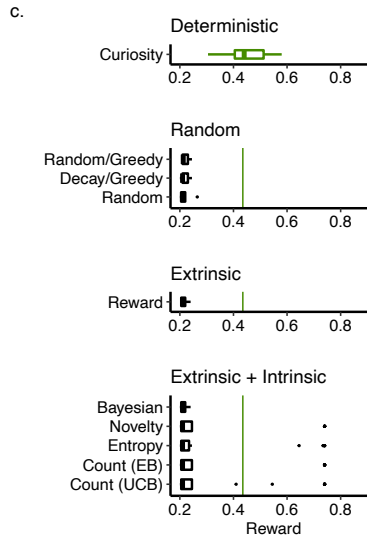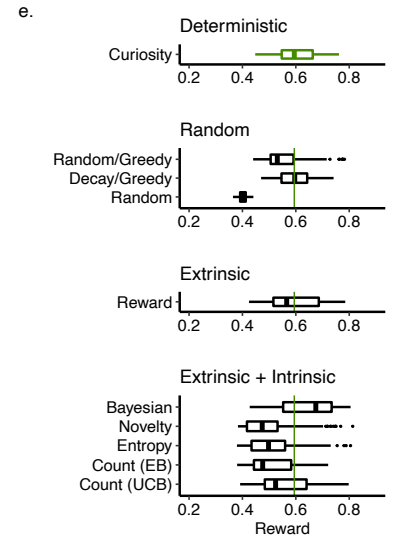
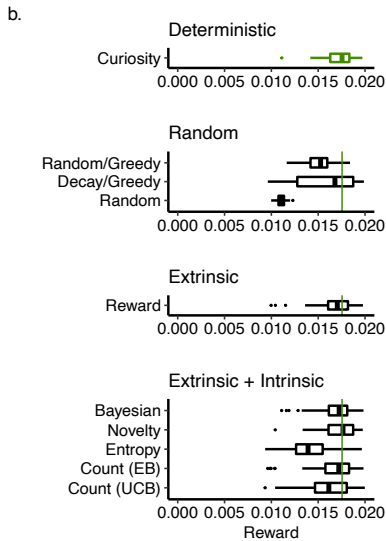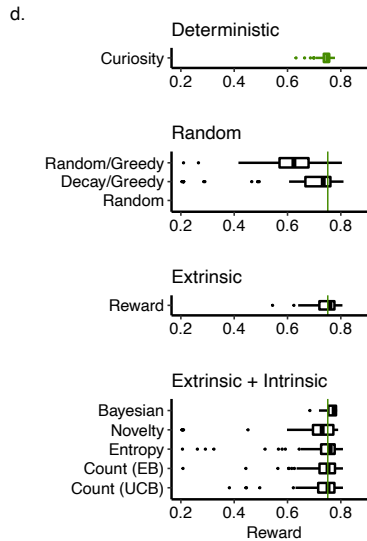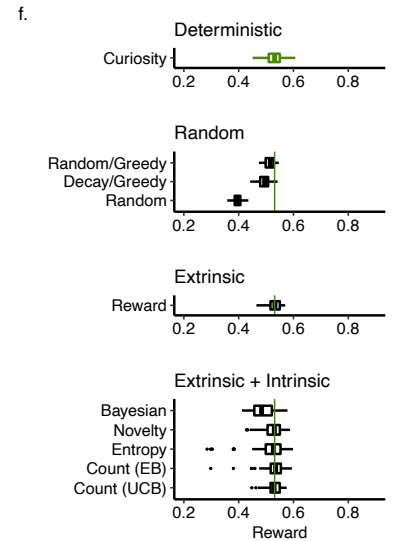**Figure 7.** Summary of reward collection (*Tasks 2-7*). The strategies in each panel are grouped according to the class of search they employed (Curiosity. Random, Extrinsic reward or Extrinsic + Intrinsic rewards). **a.** Results for Task 2, which has four choices and one clear best choice. **b.** 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.

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

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

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

309　　　*Task 4.* was designed with deceptive rewards. By deceptive we mean that the best long-term
310　　option presents itself initially with a decreasing reward value (Fig. **??d**). Such small deceptions
311　　abound in many natural contexts where one must often make short-term sacrifices (**?**). It is
312　　well known that classic reinforcement learning will often struggle in this kind of task (**??**). Here
313　　our deterministic curiosity is the only strategy that reaches above chance performance. Median
314　　performance of all other strategies are similar to the random control (Fig **??c**).

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

320　　　*Tasks 6-7* were designed as a pair, with both having 121 choices, and a complex payout structure.
321　　Tasks of this size are at the limit of human performance (**?**). We first trained all learners on *Task*
322　　*6*, then tested them in Task 7 which identical to 6, except the best payout arm is reset to be worst
323　　(Fig. **??e**-**f**). In other words Tasks 6 and 7 were joined to measure learning in a high dimensional, but
324　　shifting, environments.

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

### Robustness

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

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

342　　　These experiments in setting $\eta$ are also useful in demonstrating the wide variance in behavior
343　　that is possible with a deterministic search. Exploration in these figures is deterministic, as pre-

Optimizing boredom for reward collection
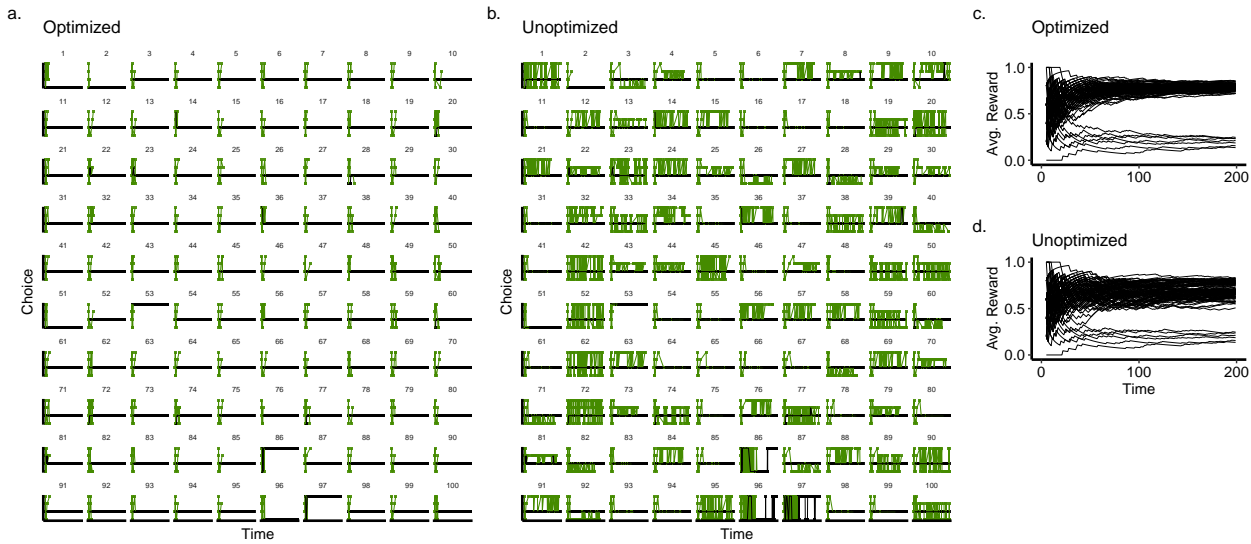
100 simulated experiments



**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.

344    scribed by Eq. **??**. The environment was stochastic however and so the variations come from the
345    environment changing what is learned, which then changes what the best learner should choose to
346    learn next.

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

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

### Discussion

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

364    This is a controversial claim. So we will address some of its criticisms as a series of questions.
365 The final question explains how this approach can be tested, with behavioral experiments.

### Why curiosity?

367 Curiosity is an ecologically rational behavoir **?**. This arguemnt rests on what we feel are three well
368 established facts. Curiosity is just as important for survival as reward collection **?**. Curiosity is a

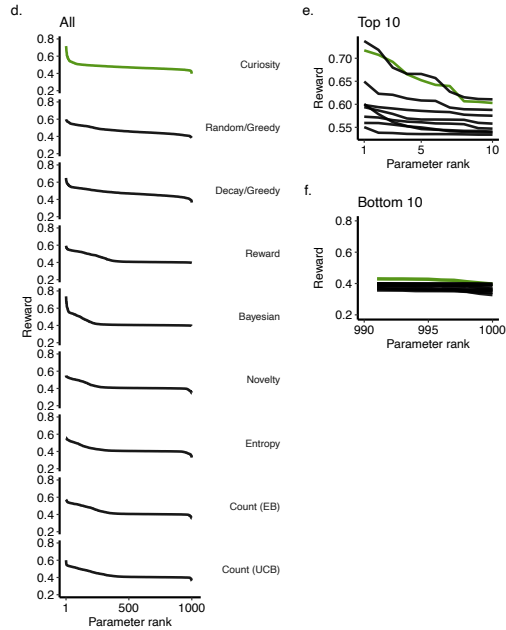## Reward collection and parameter choice
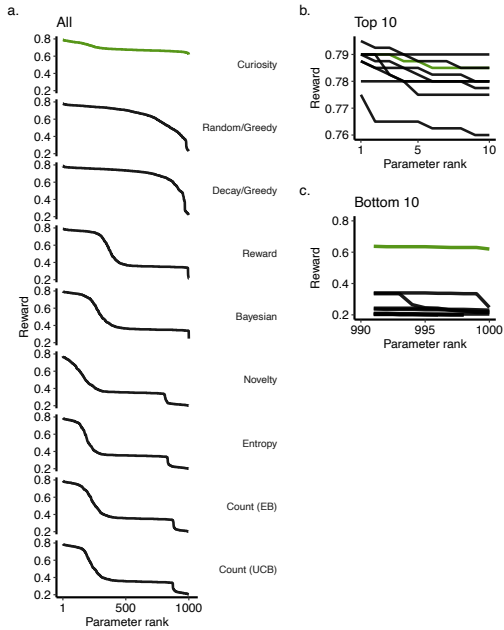


**Figure 9.** Parameter selection and search performance (*Task 2* and *6*). We wanted to evaluate how robust performance was to poor and random hyperparameters. A very robustness exploration algorithm would produce strong performance with both the best parameter choices, and the worst parameter choices. **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).
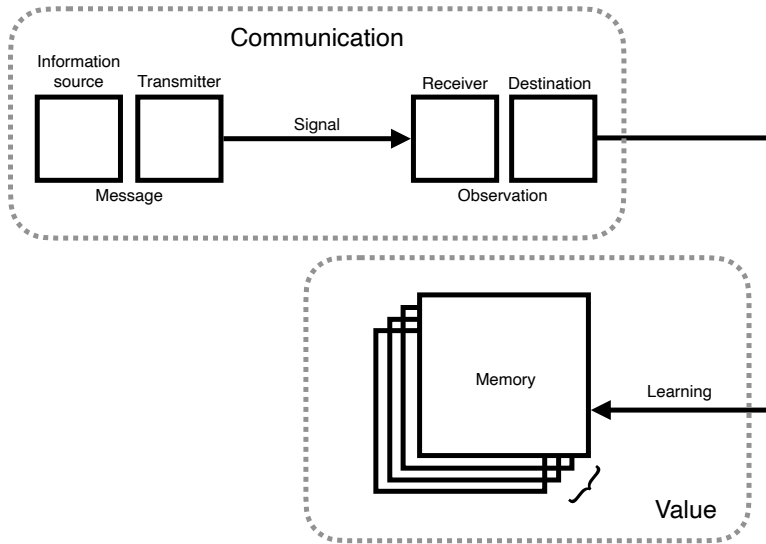
**Figure 10.** The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel directly. Value is derived from learning about observations from the channel, which is in turn dependent on memory and its past history.

369 primary drive in most, if not all, animal behavoir **?**. It is as strong, if not sometimes stronger, than
370 the drive for reward **???**. Curiosity as an algorithm is highly effective at solving difficult optimization
371 problems **???????????**.

372 In other words, curiosity is not irrational , nor a paradox , nor a heuristic , nor even quitesssen-
373 tially cognitive **?**. It is a basic homeostatic drive **?**.

374 **What about information theory?**

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

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

382 Unfortunately valuing information is not, at its core, a communications problem. Consider the
383 very simple example Bob and Alice, having a phone conversation and then Tom and Bob having the
384 exact same conversation. The personal history of Bob and Alice will determine what Bob learns in
385 their phone call, and so determines what he values in that phone call. These might be completely
386 different from what he learns when talking to Tom, even if the conversations are identical. What we
387 mean by this example is that the personal history defines what is valuable on a channel, and this is
388 independent of the channel and the messages sent. We have summarized this diagrammatically, in
389 Fig. 10.

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

395 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

529 needs: gathering resources to maintain physiological homeostasis (**??**) and gathering information to
530 decide what to learn and to plan for the future (**??**). Here we suggest that though they are separate
531 problems, they are problems that can, in large part, solve one another. This insight is the central
532 idea to our view of the explore-exploit decisions.

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

### What about aversive values?

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

545     The prescence of mnay values fits well within our notion of competition between simple (pure)
546 policies, in a win-stay loose-shift game **?**. To take other values into account we need a different
547 notation, from the inequalities of Eq. 25. For example, if we added a fear value $F$ to our homeostatic
548 scheme we can move to argmax notation,

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

$$\underset{\Pi^{\pi}}{\mathrm{argmax}} \left[ E - \eta, \ \hat{R} - \rho, \ F + \phi \right] \tag{12}$$

549     Adding bias term to each value provides a direct way to break ties, and set default behavoirs.
550 Additional to our standarrd assumptions, if we also assume the fear/aversive term should take
551 precedence over all other option when there is a tie, then $\phi > \rho > \eta$.

### What about other values?

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

$$\Pi^{\pi} = \left[ \pi_E, \ \pi_R, \ \pi_C \right] \tag{13}$$

$$\underset{\Pi^{\pi}}{\mathrm{argmax}} \left[ E - \eta, \ \hat{R} - \rho, \ C + \chi \right] \tag{14}$$

554     Where we assume the grooming term should take precedence over all other options, when
555 there is a tie. That is, $\chi > \rho > \eta$.

### What about fitting deterministic models?

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

589 **Box 1. Norms.**

590 A **norm** is a mathematical way to measure the overall size or extent of an "object" in a
591 space. The object we are concerned with is a mathematical object with a form given by,
592 $\Delta\mathbf{M} = f_{\mathbf{M}}(x) - \mathbf{M}$. But to define the norm let us as the first step consider a generic vector $v$.
593 Keeping with convention will have its norm denoted by $||v||$, and defined by three properties:

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

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

---

### Does regret matter?

563
564 Another way around the dilemma is to ignore regret altogether. This is the domain of pure
565 exploration methods, like the Gitten's Index (**?**), or probably-approximately correct approaches (**?**),
566 or other bounded forms of reinforcement learning (**?**). Such methods can guarantee that you find
567 the best reward value actions, eventually. These methods do not consider regret, as we have. But
568 we argue that regret is critical to consider when resources and time are scarce, as they are in most
569 natural settings.

### Is the algorithmic run time practical?

570
571 It is common in computer science to study the run time of an algorithm as a measure of its efficiency.
572 So what is the run time of our win stay, loose switch solution? There is unfortunately no fixed
573 answer for the average or best case. The worst case algorithmic run time is linear and additive in
574 the independent policies. If it takes $T_E$ steps for $\pi_E$ to converge, and $T_R$ steps for $\pi_R$, then the worst
575 case run time for $\pi_\pi$ is $T_E + T_R$. Of course this is the worst case run time and would still be within
576 reasonable ranges for learning in most naturalistic contexts.

### Summary

577
578 There will certainly be cases in an animal's life when their exploration must focus on tangible physical
579 rewards. Even here, searching for information about physical rewards, rather than searching for
580 their value, will make the search more robust to deception, and more reliable in the future. We think
581 of this as curiosity about reward. There are many more cases though where curiosity can include
582 observations of the environment and the self. We have argued one should put aside intuitions
583 that suggest an open-ended curious search is too inefficient to be practical. We have shown it
584 can be made very practical. We have argued one should put aside traditional intuitions for what
585 exploratory behavior in animals represents. We have shown how there is a zero-regret way to solve
586 the dilemma, a problem that has seemed unsolvable. Simply be curious.

### Mathematical Appendix.

587

### Information value as a dynamic programming problem

610
611 To find a dynamic programming solution based on the Bellman equation (Eq **??**) we must prove
612 our memory $\mathbf{M}$ has optimal substructure. This is because the normal route, which assumes the
613 problem rests in a Markov Space, is closed to us. By optimal substructure we mean that the process
614 of learning in $\mathbf{M}$ can be partitioned into a collection, or series of memories, each of which is itself a

## Box 2. The Laplacian.

Here we use $\nabla^2$ 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.

## 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.

---

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), $\pi_E$ and $\delta$ be given. Assuming transition function $\delta$ is deterministic, if $V^*_{\pi_E}$ is the optimal information value given by policy $\pi_E$, 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 $\pi_E$ 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 $\hat{V}^*_{t-1} > V^*_{t-1}$.

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 $\hat{\pi}_E$. $\square$

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 $\delta$ 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.

**652 Box 4. Dynamic programming and Bellman optimality.**

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

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

666 Mathematically this allows us to translate the full problem, $V^* = \text{argmax}_\pi \sum_T r_1, r_2, ..., r_T$ to a
668
667 recursive one, which is given below.
669

670
$$V^* = \underset{\pi}{\text{argmax}} \left[ r_0 + V(x_1) \right] \tag{15}$$
671

672 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
673 can we ensure our problem can be broken down this way? This is addressed in Box 3.
674

## Bellman solution

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

$$V_E(x) = E_0 + \underset{a}{\text{argmax}} \left[ E_t \right] \tag{16}$$

678 To explain this derivation we'll begin simply with the definition of a value function and work
679 from there. A value function $V(x)$ is defined as the best possible value of the objective for $x$. Finding
680 a Bellman solution is discovering what actions $a$ to take to arrive at $V(x)$. Bellman's insight was to
681 break the problem down into a series of smaller problems, leading a recursive solution. Problems
682 that can be broken down this way have optimal substructure. The value function for a finite time
683 horizon $T$ and some payout function $F$ is given by Eq 17. It's Bellman form is given by Eq. 18

$$V(x) = \underset{a}{\text{argmax}} \left[ \sum_T F(x, a) \right] \tag{17}$$

$$V(x) = \underset{a}{\text{argmax}} \left[ F(x_0, a_0) + V(x_1) \right] \tag{18}$$

684 In reinforcement learning the value function is based on the sum of future rewards giving
685 Eq. 19-20,

$$V_R(x) = \underset{a}{\text{argmax}} \left[ \sum_T R_t \right] \tag{19}$$

$$V_R(x) = \underset{a}{\text{argmax}} \left[ R_0 + V(x_1) \right] \tag{20}$$

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 (**?**).

$$V_E(x) = \underset{a}{\text{argmax}} \left[ E_t \right] \tag{21}$$

$$V_E(x) = \underset{a}{\text{argmax}} \left[ E_0 + V(x_1) \right]$$
$$= E_0 + \underset{a}{\text{argmax}} \left[ E_t \right] \tag{22}$$

## 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 $\pi_E$ 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 $(a, b, c)$,

$$a \leq b \Leftrightarrow \exists\, c;\; b = a + c \tag{23}$$
$$a > b \Leftrightarrow (a \neq b) \wedge (b \leq a) \tag{24}$$

**Theorem 2** (State search: breadth). *Given some arbitrary value $E_0$, an exploration policy governed by $\pi_E^*$ 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. 23 $E_i - c = E_j$.

Therefore, again by Eq. 23, $\exists \int \delta E(s) \rightarrow -c$.

*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 $S \neq Z$. $\qquad\square$

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

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

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

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

723 $T \to \infty$, $E \to \eta$. □

724 We assume in the above the action policy $\pi_E$ can visit all possible states in $X$. If for some reason

725 $\pi_E$ can only visit a subset of $X$ then proofs above apply only to that subset.

726 These proofs come with some fine print. $E_0$ can be any positive and finite real number, $E_0 > 0$.

727 Different choices for $E_0$ will not change the proofs, especially their convergence. So in that sense

728 one can chosen it in an arbitrary way. Different choices for $E_0$ can however change individual

729 choices, and their order. This can be quite important in practice, especially when trying to describe

730 some real data. This choice will not change the algorithm's long-term behavior *but* different choices

731 for $E_0$ will strongly change the algorithm's short-term behavior. This can be quite important in

732 practice.

## Optimality of $\pi_\pi$

734 Recall that in the main text we introduce the equation below as a candidate with zero regret solution

735 to the exploration-exploitation dilemma.

$$\pi^\pi = \begin{bmatrix} \pi_E, & \pi_R \end{bmatrix} \tag{25}$$

$$\underset{\pi^\pi}{\text{argmax}} \begin{bmatrix} E_{t-1} - \eta, & R_{t-1} \end{bmatrix}_{(2, 1)} \tag{26}$$

736 In the following section we prove two things about the optimality of $\pi_\pi$. First, if $\pi_R$ had any

737 optimal asymptotic property for value learning before their inclusion into our scheduler, they retain

738 that optimal property under $\pi_\pi$ when $\eta = 0$, or is otherwise sufficiently small. Second, show that

739 if both $\pi_R$ and $\pi_E$ are greedy, and $\pi_\pi$ is greedy in its definition, then Eq 25 is certain to maximize

740 total value. The total value of $R$ and $\hat{E}$ is the exact quantity to maximize if information seeking

741 and reward seeking are equally important, overall. This is, as the reader may recall, one of our key

742 assumptions. Proving this optimality is analogous to the classic activity selection problem from the

743 job scheduling literature (**??**).

744 **Theorem 4** ($\pi_\pi$ is unbiased)**.** *Let any $S$, $A$, $\mathbf{M}$, $\pi_R$, $\pi_E$, and $\delta$ be given. Assuming an infinite time horizon,*

745 *if $\pi_E$ is optimal and $\pi_R$ is optimal, then $\pi_\pi$ is also optimal in the same sense as $\pi_E$ and $\pi_R$.*

746 *Proof.* The optimality of $\pi_\pi$ can be seen by direct inspection. If $p(R = 0) > 0$ we are given an

747 infinite horizon, then $\pi_E$ will have a unbounded number of trials meaning the optimally of $P^*$ holds.

748 Likewise, $\sum E < \eta$ as $T \to \infty$, ensuring $pi_R$ will dominate $\pi_\pi$ therefore $\pi_R$ will asymptotically converge

749 to optimal behavior. □

750 In proving this optimality of $\pi_\pi$ we limit the probability of a positive reward to less than one,

751 denoted by $p(R_t = 1) < 1$. Without this constraint the reward policy $\pi_R$ would always dominate $\pi_\pi$

752 when rewards are certain. While this might be useful in some circumstances, from the point of view

753 $\pi_E$ it is extremely suboptimal. The model would never explore. Limiting $p(R_t = 1) < 1$ is a reasonable

754 constraint, as rewards in the real world are rarely certain. A more naturalistic way to handle this

755 edge case is to introduce reward satiety, or a model physiological homeostasis (**??**).

### Optimal scheduling for dual value learning problems

757 In classic scheduling problems the value of any job is known ahead of time (**??**). In our setting,

758 this is not true. Reward value is generated by the environment, *after* taking an action. In a similar

759 vein, information value can only be calculated *after* observing a new state. Yet Eq. 25 must make

760 decisions *before* taking an action. If we had a perfect model of the environment, then we could

761 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. 25 maximizes total value.

## Optimal total value

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

**Theorem 5** (Total value maximization of $\pi_\pi$). *Let any $S$, $A$, $\mathbf{M}$, $\pi_R$, and $\delta$ be given. If $\pi_R$ is defined on a Markov Decisions, then $\pi_\pi$ 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 $\pi_R$ and $\pi_E$ are independent under $\pi_\pi$ (Theorem 5

*Recall*: The controller $\pi_\pi$ is deterministic and greedy.

If both $\pi_R$ and $\pi_E$ have optimal substructure and are independent, then $\pi_\pi$ must also have optimal substructure. If $\pi_\pi$ has optimal substructure, and is greedy-deterministic then it is Bellman optimal. □

## Acknowledgments

Task 2 – Classic

a.



792

Task 3 – Sparse

b.



Task 4 – Deception

c.



Task 5 – Distraction

d.



Task 6 – High dimensional

e.



Task 7 – Unexpected change

f.



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

**Task 2 – Classic**

a.



**Task 4 – Deception**

c.



**Task 6 – High dimensional**

e.



793

**Task 3 – Sparse**

b.



**Task 5 – Distraction**

d.



**Task 7 – Unexpected change**

f.



**Figure 7–Figure supplement 2.** Summary of information value (*Tasks 2-7*)
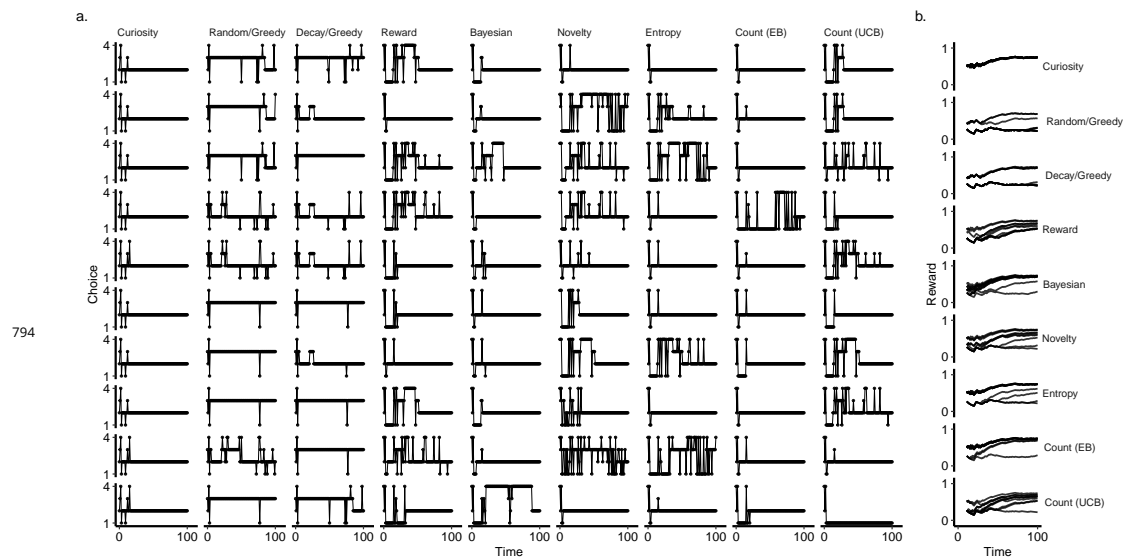
Exploration strategies
Ten 'animals'



**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.

Reward collection with independent policies:
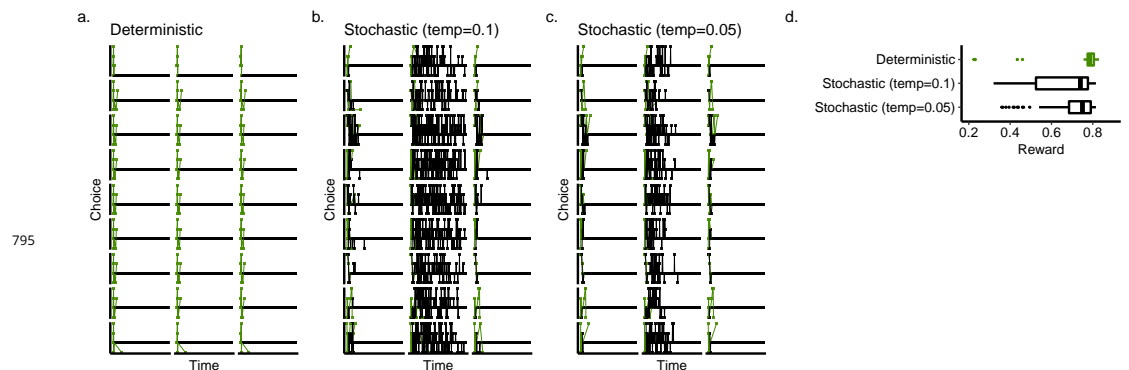deterministic versus stochastic



**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*.
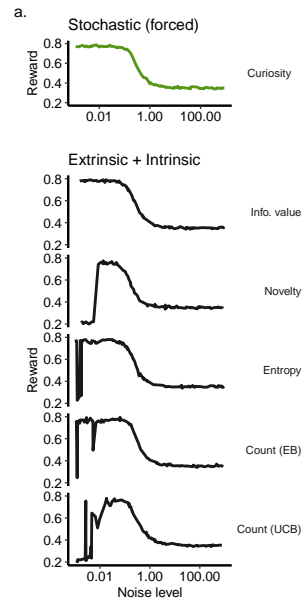
The effect of noise on
stochastic search

a.



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