Curiosity eliminates the exploration-exploitation dilemma

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

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

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

Introduction

Exploratory behavior can have two very different explanations depending on the expected result. If rewards are expected, then exploration is cast as a search for reward (1–6). This is how exploration is commonly framed (2). If instead there is no reason to expect rewards, exploration is cast as a search for information, novelty, or curiosity (7–13). For example, a rat placed in a novel maze will explore, extensively, even when no rewards could be expected (14).

An open problem in decision science is to unify exploration–of any kind–with exploitation, often defined by taking most rewarding action. When exploration optimizes for reward, which is the the classic view, this union leads to the famous exploration-exploitation dilemma (15–20), which we illustrate in Figure 1a. In this paper, we offer an alternative approach. We unify exploitation with curiosity instead, which we illustrate in Figure 1b.

It is easy to justify our use of curiosity, in this way. Curiosity is a primary drive (7, 21, 22). It is a drive that is as strong as the drive for reward, if not stronger (9, 11, 12, 21, 23–25). Curiosity is highly effective at solving difficult optimization problems, both in the long-term and the short-term (8, 26–33).

We will first develop the theory behind our union. This consists of mathematical study of curiosity alone, then with reward collection. The second half presents initial simulations where we compare to, and often exceed, more standard approaches. We conclude with some questions and answers about our approach.

Results

Theory. For our theoretical work, we first introduce reinforcement learning, and its basis in the Bellman equation. Next we develop a new mathematical view of information value, curiosity, and a new understanding of ideal curious search. We then use these results to show exploration-for-curiosity leads to optimal value solutions for all explore-exploit problems with deterministic dynamics.

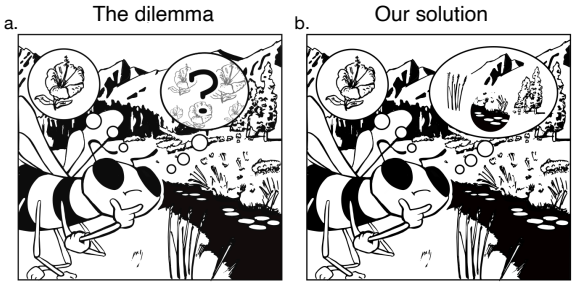

Fig. 1. Two views of exploration and exploitation. **a**. The dilemma: either exploit an action with a known reward (e.g., return to the previous flower) or explore other actions on the chance they will return a better outcome. The central challenge here is that the outcome of exploration is uncertain,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

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

When the environment is unknown, to maximize reward an animal must trade-off 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' (2).

Formality. A standard way to study reward collection is by assuming Markov decisions. Our Markov decision process (MDP) consists of states of real valued vectors $\mathbf{S}$ from a finite set $\mathcal{S}$ of size n. Actions $\mathbf{A}$ are 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, $\Lambda(\mathbf{S}, \mathbf{A})$. In our mathematics, we assume Λ is a deterministic function of its inputs. In our simulated experiments we assume it is a stochastic function.

In general, we use italic capital letters for sets. Upper-case letters are used for elements. Quantities that are real-valued vector spaces are written in bold. Parameters are denoted by greek letters. A full set of the symbols used in this paper is shown in Table 1.

Significance Statement

It's optimal to be curious.

The authors have no conflicts of interest to declare.

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

PREPRINT

Table 1. Symbols.

Notation	Description
$\mathcal{S}$	The state space of the environment.
$\mathcal{A}$	The action space of the agent.
$\mathbf{S}$	A state.
$\mathbf{S}'$	A next state in a sequence.
$\Lambda(\mathbf{S}, \mathbf{A})$	The transition function. Generates $\mathbf{S}'$.
$\mathbf{A}$	An action.
R	Reward value from the environment.
$\hat{R}$	Subjective reward value.
$\mathcal{X}$	The observation space.
$\mathbf{X}$	An observation.
$T(\mathbf{S}, \mathbf{A}, \mathbf{S}', R, \mathbf{M})$	The observation function. Generates $\mathbf{X}$.
$\mathbf{M}$	Memory (subjective).
$f(\mathbf{X}, \mathbf{M})$	A learning function. Generates $\mathbf{M}'$.
$\hat{E}$	Subjective information value.
$E(\mathbf{M}, \mathbf{M}')$	Information value function. Generates $\hat{E}$.
π	An action policy. Generates $\mathbf{A}$.
π_R	A policy to maximize reward (R, $\hat{R}$).
π_E	A policy to maximize information ($\hat{E}$).
$\pi(\mathbf{S})$	A deterministic action policy.
$\pi(\mathbf{S}\|\mathbf{A})$	A stochastic action policy.
Π	A set of possible policies. A meta-policy.
J	A cost function.
J_E	A cost function for information value ($\hat{E}$).
J_R	A cost function for reward value (R, $\hat{R}$).
$\|.\|$	Denotes the absolute value.
$\|\|.\|\|$	Denotes a vector norm.
δ	Denotes a difference, between two scalars or vectors.
∇	Denotes a gradient.
∇^2	Denotes the Laplacian.
α	Step-size (0,1].
β	Softness parameter for softmax sampling (0, ∞).
ϵ	Probability of random choice for e-greedy policy [0,1].
η	Boredom bias [0,∞).
ρ	Reward bias [0,∞).

Formality in place we can then write down a value definition for reward, V_R (2), and its Bellman equation (34). For simplicity we assume the time horizon is finite and bounded by T.

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

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

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

A central challenge in reinforcement learning in efficiently estimating $V_R^*(.)$ is finding good exploration methods (2), which is the problem we take up here. In our eventual union with curiosity, we assume the reward learning rule can be any reinforcement learning algorithm (**?**)Sutton2018). We use the TD(0) learning rule in our simulations. See, *Methods*.

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

The information collection problem. Imagine we wish to arrive at an equation to maximize curious behavior, in terms of information value. Assume, as a stand-in, $\hat{E}$ represents the information value of some observation. This hypothetical equation we want could be written to match Eq. 2. As,

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

We add a "hat" to this term E to indicate it is a subject value. In contrast R, the external reward defined by the environment, has no "hat".

So how can we arrive at Eq. 2? How can we ensure $\hat{E}$ is general, and therefore suitable for use across species? How can we do so when we choose to limit ourselves to non-Markovian kinds of learning and memory?

Observations. Animals at different stages in development, and in different environments, express curious behavior about the environment, their own actions, and their own mental processes. We assume a parsimonious way to handle this is be assuming curious animals often "bind" (35) the elements of environment, the Markov process $(\mathbf{S}, \mathbf{A}, \mathbf{S}', R)$, into single observations $\mathbf{X}$. For convenience, we assume observations are embedded in a finite real space, $\mathbf{X} \in \mathcal{X}$ of size l. This binding is imagined to happen via a communication channel, T, identical to our observation function, $T(\mathbf{S}, \mathbf{A}, \mathbf{S}', R, \mathbf{M})$. ($\mathbf{M}$ is the animal's memory, that we define below.)

PREPRINT

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

Memory. 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 into a vector space. In fact, most any memory system can be so defined.

We define memory as a vector of size p, embedded in a real-valued space, that is closed and bounded. A learning function is then any function f that maps observations $\mathbf{X}$ into memory $\mathbf{M}$. We do not assume memory $\mathbf{M}$ is Markovian. Using recursive notation, learning is denoted by $\mathbf{M} \leftarrow f(\mathbf{X}, \mathbf{M})$. Though we will sometimes use $\mathbf{M}'$ to denote the updated memory in place of exact recursion. Other times we will add subscripts, like $(\mathbf{M}_t, \mathbf{M}_{t+1}, \ldots)$, to express the path memory follows. We also define a forgetting function, $f^{-1}(\mathbf{X},\ \mathbf{M}') \rightarrow \mathbf{M}$.

We assume f and f^{-1} have been selected by evolution and experience to be learnable (36, 37), and have sufficient inductive bias to generalize that learning (38, 39).

Axioms for information value. To make our eventual union proper, we need to separate reward value from information value. We do this axiomatically. Our definition of information value is inspired by the definition of information on a communication channel (40), with one critical difference. Evenly distributed learning provides minimum value. In a sense, maximum entropy is the minimum information value.

We reason the value of any information animal depends entirely on what an animal learns by making that observation—literally how it changes that animal's memory. This definition lets us sidestep a problem with most working models of curiosity. They tend to conflate the learning rule, with curiosity itself (32, 41–49). So then if we have a memory $\mathbf{M}$ which has been learned by f over a history of observations, $(\mathbf{X_0}, \mathbf{X_1}, \ldots)$, can we measure how much value the next observation $\mathbf{X_t}$ should have?

We think this, our hypothetical measure $\hat{E}$, should have certain intuitive properties:

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

A corollary of Axiom 1 is that an observation that doesn't change memory, has no value. That is, if $\sum |\Delta\mathbf{M}| = 0$ then $\hat{E} = 0$. See Figure 2a for an example of $\hat{E}$ defined only in memory space.

Axiom 2 (Axiom of Induction)**.** *If all p elements $|\Delta m_i|$ are equal, $|\Delta m_i| = \frac{1}{p}\sum |\Delta\mathbf{M}|$, then $\hat{E}$ should be minimized.*

That is, information which is learned evenly in memory cannot induce any inductive bias, and is therefore the least useful learning possible (38, 50). It should therefore be the least valuable information for our system. In a sense then, information value can sometimes act as the reciprocal of entropy.

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

That is, information value is in principle positive even if its consequences are later found to be negative.

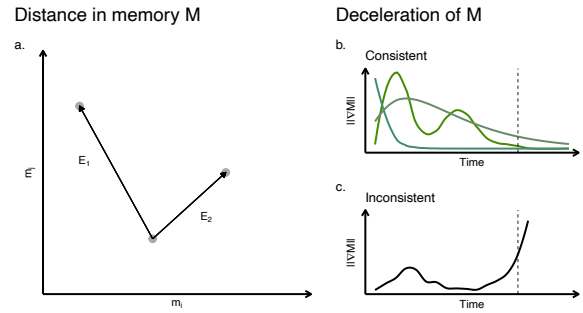

Fig. 2. Information value. **a**. This panel illustrates a two dimensional memory. The information value of two observations $\mathbf{X}_1$ and $\mathbf{X}_2$ depends on the distance in memory, $\mathbf{E_1}$ and $\mathbf{E_2}$ respectively, with learning. Here we show this distance as a euclidean norm, 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 that our account of information value and curiosity does not work when learning is inconsistent.

Axiom 4 (Axiom of Equilibrium)**.** *Given the same observation $\mathbf{X}_i$, $\hat{E}$ should approach 0 in finite time.*

That is, learning on $\mathbf{M}$ should eventually make progress toward equilibrium (i.e. self-consistency) for each observation. We use the term consistency to denote a memory approaching equilibrium. By this we mean the memory has become consistent in itself in time. We show examples of consistent and inconsistent learning in Figure 2b-c.

The practical example of information value we use throughout this paper is the Jenson-Shannon metric (*Methods*). This is a good choice because in our simulations we focus probability and classic bandit learning problems.

The information value axioms are equally well satisfied by taking the norm of the learning gradient directly. That is, $\hat{E} = ||\nabla\mathbf{M}||$. This geometric view may prove useful for expressing curiosity in differentiable learning systems, and for manifold-style analyses of learning. Both of which are of great interest to biologists, neuroscientists, and computer scientists alike.

A bellman solution for information value. The common way to arrive at the Bellman solution is to again assume a Markov decision space. This is a problem for our definition of memory, which we assume has a natural long-term path dependence. This breaks the Markov assumption, and its needed claim of optimal substructure. To get around this we proved that exact forgetting of the last observation is another way to achieve optimal substructure, and so to derive a Bellman solution.

In the theorem below, which shows the optimal substructure of $\mathbf{M}$, we implicitly assume that $\mathbf{X} = \mathbf{S}$, $\mathbf{A}$, f, and Λ are given, and that Λ is deterministic,

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

PREPRINT

Given Theorem 1 and an arbitrary starting value, $E_0 > 0$, the best solution to maximizing $\hat{E}$ is therefore given by Equation 4) (34).

$$V^*_{\hat{E}}(\mathbf{S}) = \underset{\mathbf{A}\in\mathbb{A}}{\text{argmax}} \left[\hat{E}_t + V^*_{\hat{E}}(\Lambda(\mathbf{S},\mathbf{A}))\right] \quad [4]$$

As long as learning about the different observations $\mathbf{X}$ is independent, there are many equal optimal solutions to the information collection problem above. Under these conditions we can simplify Eq. 4 giving Eq. 5.

$$V^*_{E}(\mathbf{X}) = \underset{\mathbf{A}}{\text{argmax}} \left[E_t + E_{t+1}\right] \quad [5]$$

A hypothesis about boredom. To limit curiosity we introduce a penalty term, η, which we treat as synonymous with boredom (8, 51, 52). We hypothesize boredom an adaptive trait, in other words. As in (53)we treat boredom as a tunable parameter, here defined by $\eta \geq 0$. We use boredom to ignore marginal information value by requiring curious exploration to stop at $E \leq \eta$.

Optimal exploration. Having a policy to maximize $\hat{E}$ does not ensure that exploration is of good quality, or is optimal. We think it is reasonable to call an exploration optimal if it meets the criteria below.

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 actions as possible to achieve 1 and 2.

In Theorems 2 and 3 we prove that our Bellman solution satisfies these, when $\eta > 0$ and when the observations all $\mathbb{X}$ contain complete state information, $\mathbb{S} \subset \mathbb{X}$. The complete proofs are provided in the Appendix.

Theorem 2 (Complete exploration)**.** *Given some arbitrary value E_0, an exploration policy governed by $\pi^*_{\hat{E}}$ will visit all states $\mathbf{S} \in \mathbb{S}$ in a finite number of steps T.*

Theorem 3 (Efficient exploration)**.** *An exploration policy governed by $\pi^*_{\hat{E}}$ will only revisit states for where information value is marginally useful, defined by $\hat{E} > \eta$.*

A hypothesis about equal importance. Must searching for information be less efficient than a direct search for reward? To answer this question, we have two hypotheses. If confirmed, they would let us answer the dilemma in what seems to be a new way.

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

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

If reward and information are equally important, and curiosity is sufficient, how can an animal optimally balance them? If you accept the conjectures, then answering this question is the same as solving the dilemma. To solve what is essentially a dual value learning problem–information and reward maximization—we need an algorithm that can maximize, $V_{\hat{E}R}$. The value of a mixed R and $\hat{E}$ sequence. For example,

$$V^*_{\hat{E}R}(\mathbf{S}) = \underset{\mathbf{A}\in\mathbb{A}}{\text{argmax}} \left[\sum_{k=0}^{T} \hat{E}_t + \hat{E}_{t+1} + R_{t+2} + \hat{E}_{t+3} + R_{t+4} + \dots\right] \quad [6]$$

A similar dual value problem was solved by Robbins (54) in his original derivation of the win-stay lose-switch rule. This WSLS approach has since proven useful in decision making and reinforcement learning (55, 56), Bayesian sampling (57), and especially in game theory (58).

We have come to think about our two policies π_R and $\pi_{\hat{E}}$ as two "players" in game playing for behavioral control of an animal's moment-by-moment actions (55). A simple solution to this game is a WSLS rule. We denote this rule as Π_π, and define it by the inequalities below.

$$\Pi_\pi = \begin{cases} \pi^*_{\hat{E}} & : \hat{E} - \eta > R + \rho \\ \pi_R & : \hat{E} - \eta < R + \rho \end{cases} \quad [7]$$

No matter which policy is in control, in Eq. 7, each policy can learn from the other policies actions and observations. That is, while the two policies do compete for control, they also learn from each other in a cooperative way.

In addition to η we introduce a reward bias term ρ, where $\rho > \eta$. This bias ensures no ties can occur, and that as $\hat{E}$ decreases to 0, as it must by Axiom 4, the default policy is reward collection. That is, by π_R. In this paper we assume ρ is set to be just a modest amount larger than η. Its role is then to break ties, and little else. Larger values of ρ could be used to bias reward exploitation in "irrational" ways.

To provide some intuition for Eq.7, imagine that in the previous action an animal observed, arbitrarily, 0.7 units of information value and no reward (i.e. $R = 0$). Using our rule on the next action the animal should then choose its exploration policy $\pi^*_{\hat{E}}$. If it was explored last time, this is a "stay" action. If not, it is a "switch" action. Let's then say that with this next observation, $\hat{E}$ decreases to 0.2 and a reward value of 1.0 is observed. Now the animals should switch its strategy for the next round, to exploit instead.

When we compare reward R and information value $\hat{E}$ as above we have not ensured they have the same "units", or a common scale. We sidestep this 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 to be positive and non-zero, $(E_0 - \eta) \geq 0$. These constraints are enough to ensure optimal exploration (defined above) which, in turn, ensures that optimal reward learning is possible.

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

The WSLS rule in Eq 7 when applied to exploration-exploitation problems can guarantee three things, shown below.

Theorem 4 (No regret - reward value)**.** *When π_R is in control under Π_π, all actions are zero regret in terms of V_R. That is, $\sum_{k=0}^{T} G = 0$.*

PREPRINT

Theorem 5 (No regret - information value)**.** *When $\pi_{\hat{E}}$ is in control under Π_π, all actions are zero regret in terms of $V_{\hat{E}}$. That is, $\sum_{k=0}^{T} G = 0$.*

Theorem 6 (No regret - mixed values)**.** *When either $\pi_{\hat{E}}$ or π_R is in control under Π_π, all actions are zero regret in terms of $V_{\hat{E}R}$. That is, $\sum_{k=0}^{T} G = 0$.*

In each of theses Theorems we assume our definitions for $\mathbf{X}$, $\mathbf{A}$, $\mathbf{M}$, f, Λ, and Π_π are implicitly given, and a finite horizon T. Complete proofs for each are presented in the *Appendix*.

Homeostatic adjustments. We have been studying cases where the environmental reward R is equal to its subjective value to the animal, $\hat{R}$. In natural behavior though the value of reward often declines. For example, as the animal reaches satiety (59–61). This phenomenon is called reward homeostasis, and it is commonly formulated using control theory.

The goal in a control setting is not to maximize R, but to minimize the difference between collected R and a set point, $\bar{R}$ (59–61). So, in a homeostatic setting reward value is subjective, and approximated by $\hat{R} \approx \bar{R} - R$.

Fortunately, a reward collection policy designed for one greedy reward collection, will work for homeostasis as well (59). However, to use a form similar to our Eq. 7 with homeostatic reward value, we need to make one modification. This is shown in Eq. 8.

$$\Pi_\pi = \begin{cases} \pi^*_{\hat{E}} & : \hat{E} - \eta > R - \rho \\ \pi_R & : \hat{E} - \eta < R - \rho \end{cases} \qquad [8]$$

Under homeostasis ties between values in Eq. 7 should be broken in favor of exploration. Pursuing reward exploitation instead, when homeostasis is satisfied and $\hat{R} = 0$, disturbs homeostasis. Exploitation, in the control theoretic sense of homeostasis, is therefore a suboptimal result.

Experiments. Does our theory work in practice? The remainder of this paper is devoted to studying performance in simulated bandit tasks. Each task was designed to either test exploration performance in a way that matches recent experimental studies, or to test the limits of curiosity.

Information collection. 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 let's us resolve the dilemma without regret. A deterministic view of exploration is at odds with how exploration is generally viewed, which involves using some randomness.

We confirmed and illustrated how a deterministic curious search is best using an information foraging task (Task 1; Figure 3). This variation of a bandit task (2) 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 Figure 6a. When the relative color probabilities are about even, 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

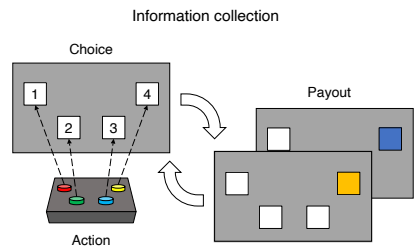

Fig. 3. A simple four choice information foraging task. The information in this task is a yellow or blue stimulus, which can change from trial to trial. A good learner in this task is one who tries to learn the probabilities of all the symbols in each of the four choices. The more random the stimuli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

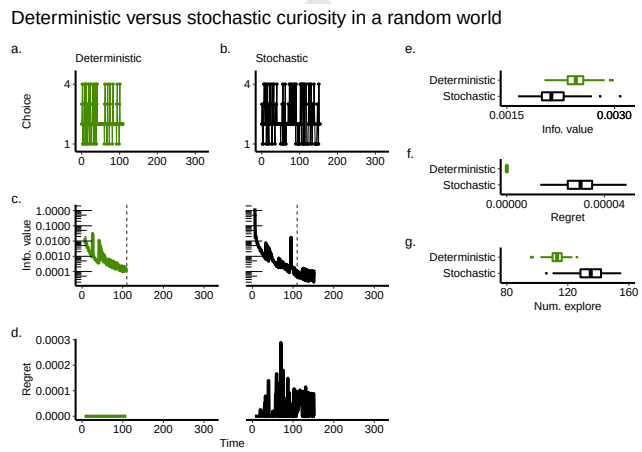

Fig. 4. Comparing deterministic versus stochastic variations of the same curiosity algorithm, in a simple information foraging task (Task 1). Deterministic results are shown in the left column, and stochastic results are shown in the right. The hyper-parameters for both models were found by random search, 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.

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. We cannot mathematically ensure a mixed value solution to explore-exploit problems will maximize reward value better than other well-established methods. To find out if this is so, we measured reward collected over 7 tasks and 10 agents. Each agent's parameters were independently optimized for each task. For a brief description of each agent, see Table 2. 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 intrinsic reward or bonus (62, 63).

The general form of the 7 tasks are depicted in Fig 5. In

PREPRINT

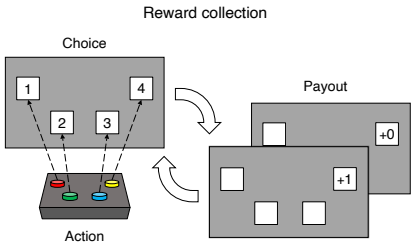

Fig. 5. A 4 choice reward collection task. The reward is numerical value, here a 1 or 0. A good learner in this task is one who collects the most rewards. The task depicted here matches that in Fig 6b. Every other reward collection task we study has the same basic form, only the number of choices increases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

Table 2. 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
Random	Random	Pure random exploration
Reward	Directed	Softmax sampling of reward value
Bayesian	Directed	Softmax sampling of reward value + information value
Novelty	Directed	Softmax sampling of reward value + novelty signal
Entropy	Directed	Softmax sampling of reward value + action entropy
Count (EB)	Directed	Softmax sampling of reward value + visit counts
Count (UCB)	Directed	Softmax sampling of reward value + visit counts

the main text we present a short summary of each task. For complete details, see the *Methods*. As with our first task (Figure 3) they are all variations of the classic multi-armed bandit (2). The payouts for each of the tasks are shown separately in Figure 6. 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 Figure 7. All agents, including ours, used the same exploitation policy based on the temporal difference learning rule (2) (Methods).

Task 1 is an information gathering task, discussed above. We present the design and results here to contrast with the reward collection tasks.

Task 2 examined reward collection, in a simple classic 4-armed probabilistic reward setting, very common in the decision making literature (64–70). Here we expected all the exploration strategies to succeed. Still, our deterministic curiosity was the top-performer in terms of median rewards collected, though by a small margin (Figure 7**a**).

Task 3 was designed with very sparse rewards (71, 72) and an expanded set of 10 choices. Both made this a more difficult task (Figure 6**c**). Sparse rewards are common in the natural world (73, 74). Most reinforcement learning algorithms struggle in this setting, because the thing they need to learn, that is rewards, are mostly absent. In this task we saw quite a bit more variation in performance. Our approach was a close second, with the novelty-bonus strategy taking the top slot. The only time it does so.

Task 4 was designed with deceptive rewards. By deceptive we mean that the best long-term option presents itself initially with a decreasing reward value (Figure 6**d**). It is well known that classic reinforcement learning rules struggle in this kind of task (2, 75). Here ours was the only strategy that reaches above chance performance. Median performance of all other strategies are similar to the random control (Fig 7**c**).

Task 5 was designed to fool curiosity, specifically our algorithm, by presenting information that was irrelevant to reward collection, but had very high entropy and so "interesting" to our algorithm. We anticipated this task would fool our algorithm, with all others performing well since they are agnostic to information entropy. However, despite being designed to fail, curiosity produced competitive performance (Fig 7**d**) with no regret (Fig **d**).

Tasks 6-7 were paired of tasks, both having 121 choices, and a complex payout structure. Problems of this size are at the limit of human performance (76). We first trained all algorithms on Task 6, measured performance, and then tested them in Task 7. Task 7 was identical to 6 except the best payout arm was reset to be the worst choice (Figure 6**e-f**). In other words, Tasks 6 and 7 measure learning in a changing, high dimensional, environments.

In Task 6 curiosity performed well, securing a second place finish in both reward maximization (Figure 6**e**) and regret (Figure **e**). The Bayesian agent outperformed ours. When switching tasks the top Bayesian model became the worst performer. Curiosity took the top spot. Note how the other less-biased-toward-reward-value exploration models–count-based, novelty, and entropy models–also saw gains on Task 7.

Robustness. It is common to focus on best case performance, as we have above. However worst case performance is just as important in judging the usefulness of an algorithm. We therefore examined the progression from best case hyperparameters, to the worst case.

Unlike idealized simulations, animals cannot know with perfect fidelity how their environment may change. So they cannot perfectly know the search parameters, in other words. To test the robustness of our algorithm, and all other exploration strategies we considered, we reexamined reward collection performance across all hyperparameters from our tuning set. Results for this are shown in Figure 9. We plotted performance ranked by parameters, from best to worst.

No matter the parameter choices, exploration-as-curiosity produced top-3 performance (Figure. 9a-bd). Most interesting to note is the performance on the worst 10 parameters. Here curiosity was markedly better, with substantially less variability.

All other exploration strategies use hyperparameters to tune the *degree* of exploration. For example, the temperature of noise in softmax. With our approach we tune when exploration should stop, by η. It seems that performance is more robust to "mistuning" of the stopping point, rather than the degree

PREPRINT

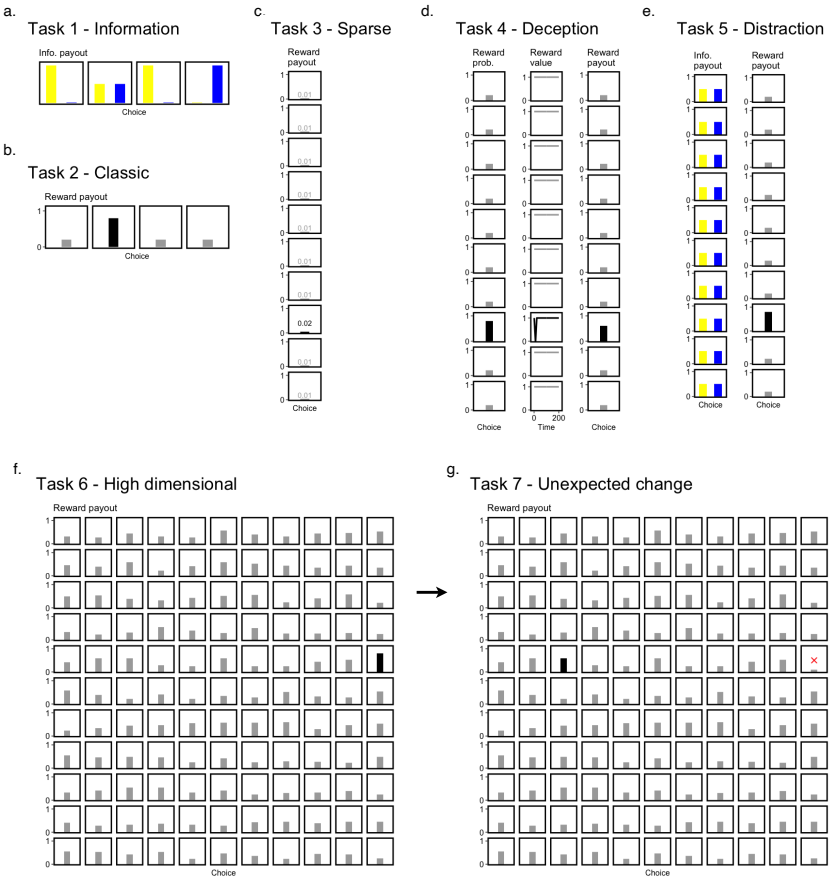

Fig. 6. Payouts for *Tasks 1 - 7*. Payouts can be information, reward, or both. For comments on general task design, see Fig 5. **a.** A classic four-choice design for information collection. A good learner should visit each arm, but quickly discover that only arm two is information bearing. **b.** A 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.

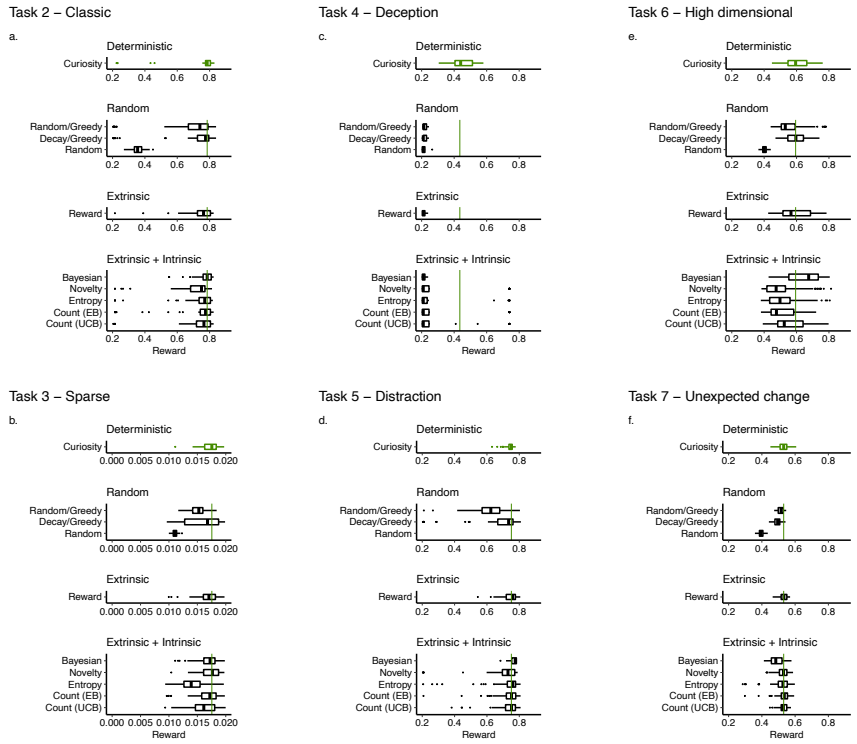

Fig. 7. Summary of reward collection (*Tasks 2-7*). The strategies in each panel are grouped according to the class of search they employed (Curiosity. Random, Extrinsic reward or Extrinsic + Intrinsic rewards). **a.** Results for Task 2, which has four choices and one clear best choice. **b.** Results 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.

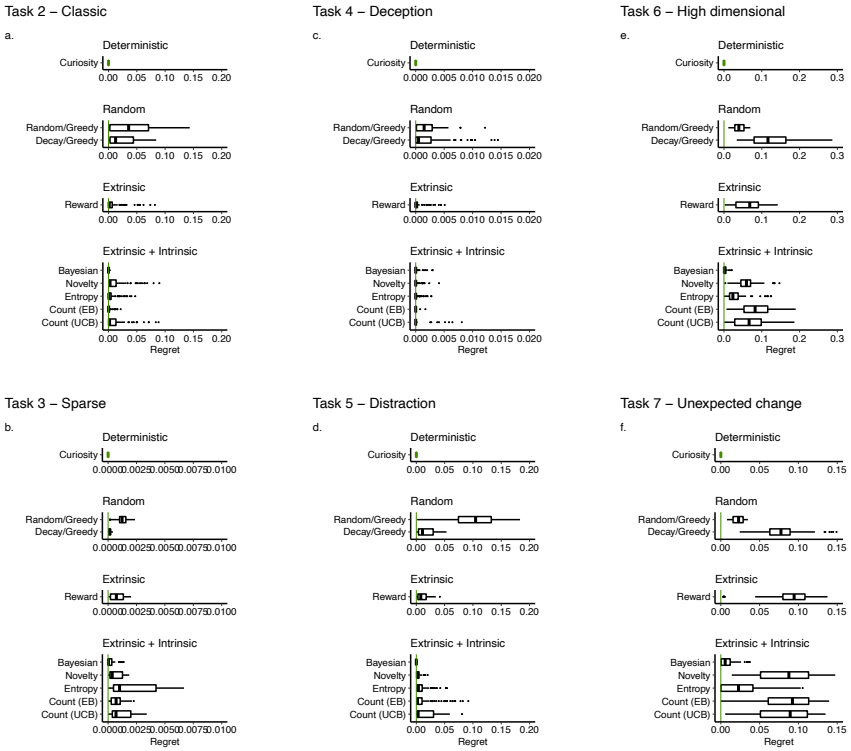

Fig. 8. Summary of total regret (*Tasks 2-7*). See the previous figure for details.

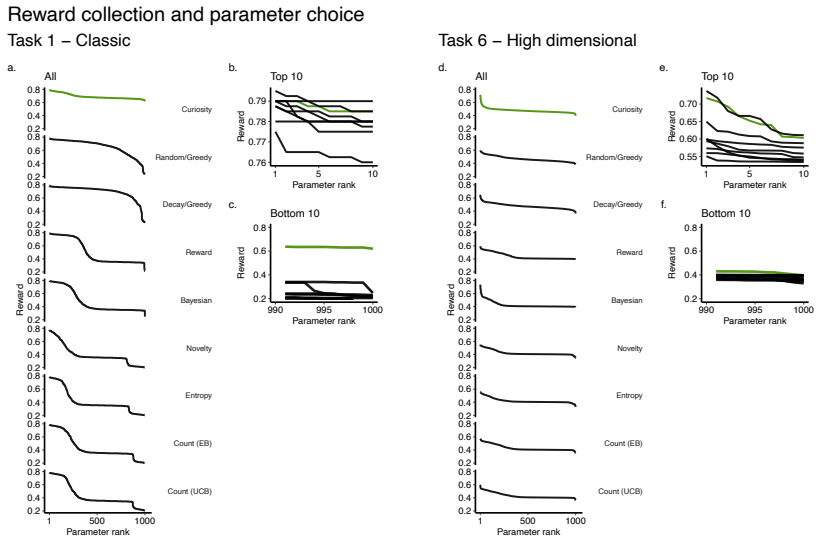

Fig. 9. Parameter sensitivity (*Task 2* and *6*). Here we evaluated how robust performance was for random hyperparameters, rather than those which were carefully chosen. A very robustness exploration algorithm would produce strong performance with both the best parameter choices, and the worst parameter choices. **a,d** Total reward for 1000 randomly chosen hyperparameters, after ranking. Curiosity (green) compared to all the others. **b,e** Performance with the top 10 hyperparameters. **c.f** Performance with the bottom 10 hyperparameters.

of exploration.

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. 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-switch strategy. When information is more valuable than rewards, be curious, otherwise seek rewards. We prove this rule has zero formal regret, when maximizing both reward and information value on equal terms.

Arguing the dilemma can be solved without regret is a controversial claim. So we will address some potential criticisms, as a series of questions. The final question explains how this approach can be tested.

Why curiosity? Our use of curiosity rests on three well established facts. 1. Curiosity is a primary drive in most, if not all, animal behavior (22). 2. It is as strong, if not sometimes stronger, than the drive for reward (9, 21, 23). 3. Curiosity as an algorithm is highly effective at solving difficult optimization problems (8, 26–33).

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

Weaver (77) in his classic introduction to the topic, describes information as a communication problem with three levels. a. The technical problem of transmitting accurately. b. The semantic problem of meaning. c. The effectiveness problem of 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.

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 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 Figure 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.

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

Questions about parsimony can be sometimes resolved by considering the benefits versus the costs to adding complexity. 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

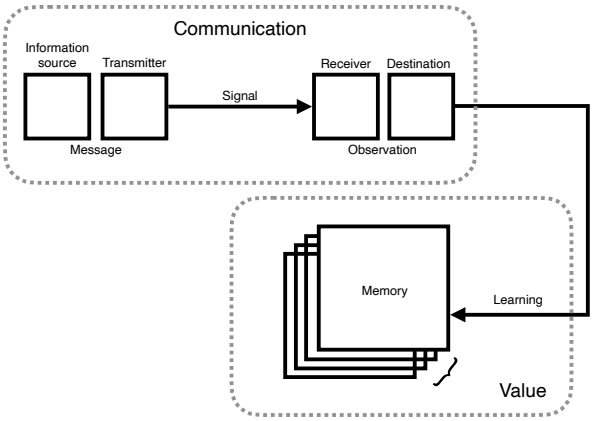

Fig. 10. The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel directly. Value is derived from learning about observations from the channel, which is in turn dependent on memory and its past history.

also be building a model of the environment, useful for later planning (78, 79), creativity, imagination (80), while also building, in some cases, diverse action strategies (29, 30, 75, 81).

The cost of curiosities is indirectness, which can make it slower to converge. Our numerical simulation suggests this cost is not fatal for finding practical successes.

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 (7, 9–12, 16, 23, 82? –84). If it predicts neural structures (20, 85), that would be some evidence for it (we discuss this more below). If this theory proves useful in machine learning, that would be some evidence for it (8, 28, 30, 31, 43, 45, 75, 84, 86, 87).

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

Does value as a technical problem even make sense? Having a technical definition of value, free from any meaning, might seem counterintuitive property for a value measure. We argue that it is not more or less counterintuitive than stripping information of meaning in the first place. The central question for our account of value is, will it prove useful?

What about mutual information, or truth? In other prototypical examples, information value is based on how well

PREPRINT

what is learned relates to the environment (88–90), that is the mutual information, the internal memory and the external environment. Colloquially, one might call this "truth seeking".

As a people we pursue information which is fictional (91), or based on analogy (92), or outright wrong (93). Conspiracy theories, disinformation, and our many more mundane errors, are far too commonplace for information value to be based on mutual information, as defined above. This does not mean that holding false beliefs cannot harm survival, but this is a second order question as far as learning goes.

So you suppose there is always a positive value for learning of all fictions, disinformation, and deceptions? We must. This is our most unexpected prediction. Yet, logically, it is consistent with our analysis and we believe qualitatively consistent with the behavior of humans, and animals alike. The fact that humans consistently seek to learn false things calls we as scientissts accept that learning alone is a motivation for behavior. In other words, it is more than a heuristic (42).

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, machine learning, game theory, psychology, neuroscience, biology, and 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? If reward is defined as more-or-less any quantity that motivates behavior, then our definition of information value is a reward. This last point does not mean, however, that information value and environmental rewards are interchangeable. Rewards from the environment are a conserved resource, information is not. For example, if a rat shares a potato chip with a cage-mate, it must break the chip up leaving it less food for itself. While if a student shares an idea with a classmate, that idea is not divided up.

In other words, it depends.

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 (9, 11). This is why we limit curiosity with boredom, and counter it with a drive for reward collecting (i.e., exploitation).

There is a useful analogy between science and engineering here. Science is an open-ended inquiry, and engineering often is a purpose driven enterprise. They each have their own pursuits, but they also learn from each other, often in alternating iterations (1). It is their different objectives though which makes them such good 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 (2, 94), that is unless significant prior knowledge can be brought to bear (2, 95) or the problem can be itself compressed (28, 96). 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 (97).

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 (7, 9, 46). Computer science tends to define it as needed, for the task at hand (29, 30, 75, 81, 98). Before creating a taxonomy of curiosity it is important to first define it, mathematically.

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 the 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 can animals tune boredom? There is tentative evidence for this. Recently, (53) showed this in a series of three experiments. They reported that altering the expectations of future learning in turn alters self-reports of boredom. Others have shown how boredom and self-control interact to drive exploration (51, 52, 99).

How can an animal tune boredom for a new setting? The short answer is that one would need to guess and check. Our work in Fig 9 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 (2, 7, 9). Indeed, we have highlighted how they are separate problems, with links to different basic needs: gathering resources to maintain physiological homeostasis (59, 60) and gathering information to decide what to learn and to plan for the future (2, 36). Here we suggest that though they are separate problems, they are problems that can, in large part, solve one another. This insight is the central idea to our view of the explore-exploit decisions.

Yet there are hints of this independent cooperation of curiosity and reinforcement learning out there. Cisek (2019) has traced the evolution of perception, cognition, and action circuits from the Metazoan to the modern age (85). The circuits for reward exploitation and observation-driven exploration appear to have evolved separately, and act competitively, exactly the model we suggest. In particular he notes that exploration circuits in early animals were closely tied to the primary sense organs (i.e. information) and had no input from the homeostatic circuits (59, 60, 85). This neural separation

PREPRINT

for independent circuits has been observed in "modern" high animals, including zebrafish (100) and monkeys (12, 101).

What about aversive values? We certainly do not believe this is a complete theory of rewarding decisions. For example, we do not account for any consequences of learning, or any other aversive events. Accounting for these is another important next step.

The presence of many values fits well within our notion of competition between simple (pure) policies, in a win-stay loose-shift game (55). To take other values into account we need a different notation, from the inequalities of Eq. 7. For example, if we added a fear value F to our homeostatic scheme we can move to argmax notation,

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

$$\underset{\Pi^{\pi}}{\operatorname{argmax}} \left[E-\eta,\ \hat{R}-\rho,\ F+\phi\right] \quad [10]$$

If we also assume the fear/aversive term should take precedence over all other options when there is a tie, then biases should be set so $\phi > \rho > \eta$.

What about other values? Another more mundane example would be to add a policy for a grooming drive, C.

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

$$\underset{\Pi^{\pi}}{\operatorname{argmax}} \left[E-\eta,\ \hat{R}-\rho,\ C+\chi\right] \quad [12]$$

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

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 (102), or probably-approximately correct approaches (36), or other bounded forms of reinforcement learning (103). Such methods can guarantee that you find the best reward value actions, eventually. These methods do not consider regret, as we have. But we argue that regret is critical to consider when resources and time are scarce, as they are in most natural settings.

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.

What predictions does this theory make? In what sense can your mathematical union be tested? Normally, to test a model or theory of learning one tests whether or not the learning algorithm appears to be at work in the animal of interest. However our axioms, and their learning rule independence, means nearly any rule will do. Testing must therefore be qualitative, for which we have four strong predictions to make.

1. **Invariant exploration strategy**. The presence of reward should not change the exploration strategy. This is tricky, because differing sense information is often associated with the presence of absence of reward. This sense of information would in many cases change learning and so change curious exploration. *Auxiliary assumptions.* Learning is possible because there is some novelty ($\hat{E} > \eta$), and some observations available to learn from. These criteria rule out for example situations where Levy walks would be optimal (104).

2. **Information preference following reward absence**. When reward is absent, an animal should immediately choose to do an information collection with it's next action. *Auxiliary assumptions.* Learning is possible because there is some novelty ($\hat{E} > \eta$), and some observations available to learn from.

3. **Deterministic search**. Environmental noise is the only source of noise in our model's exploration. It's search was purely deterministic. To evaluate this in real agents we must move from population-level distribution analysis, to moment-by-moment predation of animal behavior (105, 106). *Auxiliary assumptions.* Perfect memory, and no neural noise.

4. **Well defined pure-periods**. There should be a well defined period of exploration and exploitation. *Auxiliary assumptions.* There are no other behaviors present. For example–fear,avoidance or grooming.

Of course the presence of any four criteria does not indicate our theory is at work, mechanistically. Mechanistic claims for any biological or behavioral problems are difficult if not impossible to be certain of. Instead, we view meeting each of these criteria as some evidence for our account. We think the nice mathematical properties of our union makes it a good general candidate for solving explore-exploit problems, and ultimately we wish to test how true that is.

Summary. We have argued one can put aside any intuitions that suggest a curious search is too inefficient to be practical for reward collection. We have shown it can be made very practical, and mathematically optimal. We have shown there is a zero-regret way to solve the dilemma, a problem that has seemed unsolvable. Simply, bee curious.

Methods and Materials

Tasks. We studied seven tasks, each designed to test exploration in a common setting or test the limits of curiosity. Despite their differences each task had the same basic structure. On each trial there were a set of n choices, and the learner should try and learn the best one. Each choice action returns a "payout" according to a predetermined probability. Payouts are information, reward, or both (Figure 6). Note that information was symbolic, denoted by a color code, "yellow" and "blue" and as is custom reward was a positive real number.

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

PREPRINT

"yellow" of "blue". This was the maximum entropy choice, The ideal choice for information foraging. See Figure 6**a**.

Task 2. was designed to examine reward collection, in setting very common in the decision making literature. At no point does the task generate symbolic information. Rewards were 0 or 1. There were again four choices. The best choice had a payout of $p(R = 1) = 0.8$. This is a much higher average payout than the others ($p(R = 1) = 0.2$). See Figure 6**b**.

Task 3. was designed with very sparse rewards (71, 72). There were 10 choices. The best choice had a payout of $p(R = 1) = 0.02$. The other nine had, $p(R = 1) = 0.01$ for all others. See Figure 6**c**.

Task 4. was designed with deceptive rewards. By deceptive we mean that the best long-term option presents itself initially with lower value. The best choice had a payout of $p(R > 0) = 0.6$. The others had $p(R > 0) = 0.4$. Value for the best arm dips, then recovers. This is the "deception" It happens over the first 20 trials. Reward were real numbers, between 0-1. See Figure 6**d**.

Task 5. designed to distract our curiosity algorithm with irrelevant information. Like task 1, there is one best option ($p(R = 1) = 0.8$) and 9 others ($p(R = 1) = 0.2$). In addition to reward, the task generated information. Information was max entropy and uncorrelated with rewarding payouts. Every choice had even probability of generating either "yellow" of "blue". See Figure 6**e**.

Tasks 6-7. were designed with 121 choices, and a complex payout structure. Tasks of this size are at the limit of human performance (76). We first trained all learners on *Task 6*, whose payout can be seen in Figure6**e-f**. This task, like the others, had a single best payout $p(R = 1) = 0.8$. After training for this was complete, final scores were recorded as reported, and the agents were then challenged by *Task 7*. *Task 7* was identical except for the best option was changed to be the worst $p(R = 0) = 0.2$ (Figure 6**f**).

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

Task memory. The tasks fit a simple discrete probabilistic model, with a memory "slot" for each choice. The memory space we defined for this is a probability space Figure 2**a**.

To measure distances in this memory space we used the Jenson-Shannon metric (JS) (107), a variation of the Kullback–Leibler divergence. The Kullback–Leibler divergence (KL) is a widely used information theory, and Bayesian modelling. It measures the information gained or lost by replacing one distribution with another. It used in machine learning (108), Bayesian reasoning (109), visual neuroscience (109), experimental design (110), compression (111**?**) and information geometry (112).

We used JS in place of KL to measure information value because it is a metric, while KL is not. This is necessary for our axioms. The Bayesian exploration strategy in the results used KL.

Where P and Q are two distributions, KL and JS are given by Eq. 13 and 14. In our implementation Q corresponded to the updated memory $\mathbf{M}_{t_1}$ and P was a "copy" of the older memory $\mathbf{M}$.

$$KL(Q, P) = \sum_{s \in S} Q(s) \log \frac{Q(s)}{P(s)} \qquad [13]$$

$$JS(Q, P) = \frac{1}{2} \sqrt{KL(Q, P) + KL(P, Q)} \qquad [14]$$

Tie breaking in search. By definition a greedy policy like $\pi_{\hat{E}}$ does not distinguish between tied values; There is after all no mathematical way to rank equal values. Theorems 3 and 2 ensure any tie breaking strategy is valid, so in the long term the choice is arbitrary. In the short-term, tie-breaking is important as it will affect choices. We expect different tie-breaking schemes will better fit different animals, their bodies, and environments.

We studied several breaking strategies, "take the left/right option", "take the closest option", or "take the option with the lowest likelihood". For simplicity we choose to use "take the next option" in all our numerical studies. Most any tie breaking scheme is fine, but should it be deterministic.

Initializing Π_π. In these simulations we assume that at the start of learning an animal should have a uniform prior over the possible actions $\mathbb{A}$. Thus $p(\mathbf{A}_k) = 1/K$ for all $\mathbf{A}_k \in \mathbb{A}$. We transform this uniform prior into the appropriate units for our JS-based $\hat{E}$ using by, $\hat{E}_0 = \sum_K p(\mathbf{A}_k) \log p(\mathbf{A}_)$.

Regret calculations. In line with past efforts on studying value loss during optimization we quantified lost value in terms of it regret G, where G is defined as $G = V^* - V_t$ and V is the value at the current time point and V^* is the best possible value that has been experienced up until t.

Reward learning equations. Reinforcement learning was always done with the TD(0) learning rule (2). This is given as Eq. 15 below.

$$V(s) = V(s) + \alpha(R_t - V(s) \qquad [15]$$

Where $V(s)$ is the value for each state (choice), R_t is the *return* for the current trial, and α_R is the learning rate $(0 - 1]$. See the *Hyperparameter optimization* section for information on how α_R chosen for each agent and task.

Hyperparameter optimization. The hyperparameters for each learner were tuned independently for each task by random search (113). Generally we reported results for the top 10 values, sampled from between 1000 possibilities.

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

PREPRINT

by the Pennsylvania Department of Health Formula Award SAP4100062201, and National Science Foundation CAREER Award 1351748.

1. Gupta AA, Smith K, Shalley C (2006) The Interplay between Exploration and Exploitation. *The Academy of Management Journal* 49(4):693–706.
2. Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning Series. (The MIT Press, Cambridge, Massachusetts), Second edition edition.
3. Woodgate JL, Makinson JC, Lim KS, Reynolds AM, Chittka L (2017) Continuous Radar Tracking Illustrates the Development of Multi-destination Routes of Bumblebees. *Scientific Reports* 7(1).
4. Lee MD, Zhang S, Munro M, Steyvers M (2011) Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research* 12(2):164–174.
5. Schulz E, et al. (2018) Exploration in the wild, (Animal Behavior and Cognition), Preprint.
6. Calhoun AJ, Chalasani SH, Sharpee TO (2014) Maximally informative foraging by Caenorhabditis elegans. *eLife* 3:e04220.
7. Berlyne D (1950) Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology* 41(1):68–80.
8. Schmidhuber (1991) A possibility for implementing curiosity and boredom in model-building neural controllers. *Proc. of the international conference on simulation of adaptive behavior: From animals to animats* pp. 222–227.
9. Kidd C, Hayden BY (2015) The Psychology and Neuroscience of Curiosity. *Neuron* 88(3):449–460.
10. Jaegle A, Mehrpour V, Rust N (2019) Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Arxiv* 1901.02478:13.
11. Sumner ES, et al. (2019) The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss. *PsyArxiv* h437v:11.
12. Wang MZ, Hayden BY (2019) Monkeys are curious about counterfactual outcomes. *Cognition* 189:1–10.
13. Auersperg AM (2015) Exploration Technique and Technical Innovations in Corvids and Parrots in *Animal Creativity and Innovation.* (Elsevier), pp. 45–72.
14. Rosenberg M, Zhang T, Perona P, Meister M (2021) Mice in a labyrinth: Rapid learning, sudden insight, and efficient exploration. *bioRxiv* 426746:36.
15. Kelly JL (1956) A New Interpretation of Information Rate. *the bell system technical journal* p. 10.
16. Berger-Tal O, Nathan J, Meron E, Saltz D (2014) The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE* 9(4):e95693.
17. Dayan P, Sejnowski TJ (1996) Exploration bonuses and dual control. *Mach Learn* 25(1):5–22.
18. Thrun SB (1992) Eficient Exploration In Reinforcement Learning. *NIPS* p. 44.
19. Mehlhorn K, et al. (2015) Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision* 2(3):191–215.
20. Kobayashi K, Hsu M (2019) Common neural code for reward and information value. *Proc Natl Acad Sci USA* 116(26):13061–13066.
21. Loewenstein G (1994) The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116(1):75.
22. Inglis IR, Langton S, Forkman B, Lazarus J (2001) An information primacy model of exploratory and foraging behaviour. *Animal Behaviour* 62(3):543–557.
23. Gottlieb J, Oudeyer PY (2018) Towards a neuroscience of active sampling and curiosity. *Nat Rev Neurosci* 19(12):758–770.
24. Gopnik A (2020) Childhood as a solution to explore–exploit tensions. *Phil. Trans. R. Soc. B* 375(1803):20190502.
25. Song M, Bnaya Z, Ma WJ (2019) Sources of suboptimality in a minimalistic explore–exploit task. *Nature Human Behaviour* 3(4):361–368.
26. Pathak D, Agrawal P, Efros AA, Darrell T (2017) Curiosity-Driven Exploration by Self-Supervised Prediction in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. (IEEE, Honolulu, HI, USA), pp. 488–489.
27. Stanton C, Clune J (2018) Deep Curiosity Search: Intra-Life Exploration Can Improve Performance on Challenging Deep Reinforcement Learning Problems. *arXiv:1806.00553 [cs].*
28. Fister I, et al. (2019) Novelty search for global optimization. *Applied Mathematics and Computation* 347:865–881.
29. Mouret JB, Clune J (2015) Illuminating search spaces by mapping elites. *arXiv:1504.04909 [cs, q-bio].*
30. Colas C, Huizinga J, Madhavan V, Clune J (2020) Scaling MAP-Elites to Deep Neuroevolution. *Proceedings of the 2020 Genetic and Evolutionary Computation Conference* pp. 67–75.
31. Cully A, Clune J, Tarapore D, Mouret JB (2015) Robots that can adapt like animals. *Nature* 521(7553):503–507.
32. Schwartenbeck P, et al. (2019) Computational mechanisms of curiosity and goal-directed exploration. *eLife* (e41703):45.
33. Laversanne-Finot A, Péré A, Oudeyer PY (2018) Curiosity Driven Exploration of Learned Disentangled Goal Spaces. *arXiv:1807.01521 [cs, stat].*
34. Bellmann R (1954) The theory of dynamic programming. *Bull. Amer. Math. Soc* 60(6):503–515.
35. Robertson LC (2003) Binding, spatial attention and perceptual awareness. *Nat Rev Neurosci* 4(2):93–102.
36. Valiant LG (1984) A theory of the learnable. *Commun. ACM* 27(11):1134–1142.
37. Rich P, de Haan R, Wareham T, van Rooij I (2021) How hard is cognitive science?, (PsyArXiv), Preprint.
38. Mitchell TM (1980) The Need for Biases in Learning Generalizations. *New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.*. pp. 184–191.
39. Thrun S, Möller K (1992) Active Exploration in Dynamic Environments. *Advances in neural information processing systems* pp. 531–538.
40. Shannon CE (1948) A Mathematical Theory of Communication. *The Bell system technical journal* 27(3):55.
41. Schmidhuber J (1991) Curious model-building control systems in *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. (IEEE, Singapore), pp. 1458–1463 vol.2.
42. Oudeyer PY (2018) Computational Theories of Curiosity-Driven Learning. *arXiv:1802.10546 [cs].*
43. Burda Y, et al. (2018) Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat].*
44. Zhang S, Yu AJ (2013) Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. *NeurIPS* 26.
45. de Abril IM, Kanai R (2018) Curiosity-Driven Reinforcement Learning with Homeostatic Regulation in *2018 International Joint Conference on Neural Networks (IJCNN)*. (IEEE, Rio de Janeiro), pp. 1–6.
46. Zhou D, Lydon-Staley DM, Zurn P, Bassett DS (2020) The growth and form of knowledge networks by kinesthetic curiosity. *arXiv:2006.02949 [cs, q-bio].*
47. Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014) Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General* 143(6):2074–2081.
48. Lehman J, Stanley KO (2011) Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19(2):189–223.
49. Velez R, Clune J (2014) Novelty search creates robots with general skills for exploration in *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation - GECCO '14*. (ACM Press, Vancouver, BC, Canada), pp. 737–744.
50. Thrun S (1998) Lifelong Learning Algorithms in *Learning to Learn*, eds. Thrun S, Pratt L. (Springer US, Boston, MA), pp. 181–209.
51. Hill AB, Perkins RE (1985) Towards a model of boredom. *British Journal of Psychology* 76(2):235–240.
52. Bench S, Lench H (2013) On the Function of Boredom. *Behavioral Sciences* 3(3):459–472.
53. Geana A, Daw N (2016) Boredom, Information-Seeking and Exploration. *CogSci* p. 6.
54. Robbins H (1952) Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 58(5):527–535.
55. Estes W (1994) Toward a statistical theory of learning. *Psychological Review* 101:94–107.
56. Worthy DA, Todd Maddox W (2014) A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: A tribute to W.K. Estes. *Journal of Mathematical Psychology* 59:41–49.
57. Bonawitz E, Denison S, Gopnik A, Griffiths TL (2014) Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology* 74:35–65.
58. Nowak M, Sigmund K (1993) A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364(1):56–58.
59. Keramati M, Gutkin B (2014) Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife* 3:e04811.
60. Juechems K, Summerfield C (2019) Where does value come from?, (PsyArXiv), Preprint.
61. Münch D, Ezra-Nevo G, Francisco AP, Tastekin I, Ribeiro C (2020) Nutrient homeostasis — translating internal states to behavior. *Current Opinion in Neurobiology* 60:67–75.
62. Ng A, Harada D, Russell S (1999) Policy invariance under reward transformations: Theory and application to reward shaping. *In Proceedings of the Sixteenth International Conference on Machine Learning* pp. 278–287.
63. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning. (MIT Press, Cambridge, Mass).
64. Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience* 27(47):12860–12867.
65. Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306(5703):1940–1943.
66. Cavanagh JF, Masters SE, Bath K, Frank MJ (2014) Conflict acts as an implicit cost in reinforcement learning. *Nature communications* 5(1):1–10.
67. Jahfari S, et al. (2019) Cross-task contributions of frontobasal ganglia circuitry in response inhibition and conflict-induced slowing. *Cerebral Cortex* 29(5):1969–1983.
68. Collins AG, Frank MJ (2014) Opponent actor learning (opal): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review* 121(3):337.
69. Collins AG, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017) Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biological Psychiatry* 82(6):431–439.
70. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66(4):585–595.
71. Silver D, et al. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.
72. Silver D, et al. (2018) Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *Science* 362(6419):1140–1144.
73. Anderson O (1984) Optimal foraging by largemouth bass in structured environments. *Ecology* 65(3):851–861.
74. Westphal C, STEFFAN-DEWENTER I, Tscharntke T (2006) Foraging trip duration of bumblebees in relation to landscape-wide resource availability. *Ecological Entomology* 31(4):389–394.
75. Lehman J, Stanley KO (2011) Novelty Search and the Problem with Objectives in *Genetic Programming Theory and Practice IX*, eds. Riolo R, Vladislavleva E, Moore JH. (Springer New York, New York, NY), pp. 37–56.
76. Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B (2018) Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour* 2(12):915–924.

PREPRINT

77. Shannon C, Weaver W (1964) *The Mathematical Theory of Communication*. (The university of Illinois Press).
78. Ahilan S, et al. (2019) Learning to use past evidence in a sophisticated world model. *PLoS Comput Biol* 15(6):e1007093.
79. Poucet B (1993) Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psycholocial Review* 100(1):162–183.
80. Schmidhuber J (2010) Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2(3):230–247.
81. Lehman J, Stanley KO, Miikkulainen R (2013) Effective diversity maintenance in deceptive domains in *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference - GECCO '13*. (ACM Press, Amsterdam, The Netherlands), p. 215.
82. Colas C, et al. (2020) Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration. *arXiv:2002.09253 [cs]*.
83. Rahnev D, Denison RN (2018) Suboptimality in perceptual decision making. *Behavioral and Brain Sciences* 41:e223.
84. Wilson RC, Bonawitz E, Costa V, Ebitz B (2020) Balancing exploration and exploitation with information and randomization, (PsyArXiv), Preprint.
85. Cisek P (2019) Resynthesizing behavior through phylogenetic refinement. *Atten Percept Psychophys*.
86. Stanley KO, Miikkulainen R (2004) Evolving a Roving Eye for Go in *Genetic and Evolutionary Computation – GECCO 2004*, eds. Kanade T, et al. (Springer Berlin Heidelberg, Berlin, Heidelberg) Vol. 3103, pp. 1226–1238.
87. Pathak D, Gandhi D, Gupta A (2019) Self-Supervised Exploration via Disagreement. *Proceedings of the 36th International Conference on Machine Learning* p. 10.
88. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10(9):1214–1221.
89. Kolchinsky A, Wolpert DH (2018) Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8(6):20180041.
90. Tishby N, Pereira FC, Bialek W (2000) The Information Bottleneck Method. *Arxiv* 0004057:11.
91. Sternisko A, Cichocka A, Van Bavel JJ (2020) The dark side of social movements: Social identity, non-conformity, and the lure of conspiracy theories. *Current opinion in psychology* 35:1–6.
92. Gentner D, Holyoak KJ (1997) Reasoning and learning by analogy: Introduction. *American psychologist* 52(1):32.
93. Loftus EF, Hoffman HG (1989) Misinformation and memory: The creation of new memories. *Journal of experimental psychology: General* 118(1):100.
94. MacKay DJC (2003) *Information Theory, Inference, and Learning Algorithms*. Second edition.
95. Zhang M, Li H, Pan S, Liu T, Su S (2020) One-Shot Neural Architecture Search via Novelty Driven Sampling in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. (International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan), pp. 3188–3194.
96. Ha D, Schmidhuber J (2018) Recurrent World Models Facilitate Policy Evolution. *NeurIPS* p. 13.
97. Wolpert D, Macready W (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Computat.* 1(1):67–82.
98. Stanley KO, Miikkulainen R (2004) Evolving a Roving Eye for Go in *Genetic and Evolutionary Computation – GECCO 2004*, eds. Kanade T, et al. (Springer Berlin Heidelberg, Berlin, Heidelberg) Vol. 3103, pp. 1226–1238.
99. Wolff W, Martarelli CS (2020) Bored Into Depletion? Toward a Tentative Integration of Perceived Self-Control Exertion and Boredom as Guiding Signals for Goal-Directed Behavior. *Perspect Psychol Sci* 15(5):1272–1283.
100. Marques J, Meng L, Schaak D, Robson D, Li J (2019) Internal state dynamics shape brainwide activity and foraging behaviour. *Nature* p. 27.
101. White JK, et al. (2019) A neural network for information seeking, (Neuroscience), Preprint.
102. Gittins JC (1979) Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(2):148–177.
103. Brafman RI, Tennenholtz M (2002) R-max – A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. *Journal of Machine Learning Research* 2:19.
104. Viswanathan GM, Afanasyev V, Buldyrev SV, Havlin S, Stanley HE (2000) Levy flights in random searches. *Physica A* p. 12.
105. Peters O, Gell-Mann M (2016) Evaluating gambles using dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26(2):023103.
106. Mangalam M, Kelty-Stephen DG (2021) Point estimates, Simpson's paradox, and nonergodicity in biological sciences. *Neuroscience & Biobehavioral Reviews* 125:98–107.
107. Endres D, Schindelin J (2003) A new metric for probability distributions. *IEEE Trans. Inform. Theory* 49(7):1858–1860.
108. Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning*. (MIT Press).
109. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. *Vision Research* 49(10):1295–1306.
110. López-Fidalgo J, Tommasi C, Trandafir PC (2007) An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2):231–242.
111. Still S, Sivak DA, Bell AJ, Crooks GE (2012) The thermodynamics of prediction. *Physical Review Letters* 109(12).
112. Ay N (2015) Information Geometry on Complexity and Stochastic Interaction. *Entropy* 17(4):2432–2458.
113. Bergstra J, Bengio Y (2012) Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13:281–305.
114. Gershman SJ (2018) Deconstructing the human algorithms for exploration. *Cognition* 173:34–42.
115. Roughgarden T (2019) *Algorithms Illuminated (Part 3): Greedy Algorithms and Dynamic Programming*. Vol. 1.
116. Strogatz SH (1994) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, Studies in Nonlinearity. (Addison-Wesley Pub, Reading, Mass).
117. Hocker D, Park IM (2019) Myopic control of neural dynamics. *PLOS Computational Biology* 15(3):24.

PREPRINT

Mathematical Appendix.

Optimal substructure in memory. To find an optimal value solution for $\hat{E}$ using the Bellman equation 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 $\mathbf{M}$, and therefore maximization of $\hat{E}$ can be partitioned into a collection, or series of memories, each of which is itself an max value solution.

This opaque term of optimal substructure can be intuitivily understood by looking in turn at another theoretical construct, Markov spaces.

In Markov spaces there are a set of states $(\mathbf{S}_0, \mathbf{S}_1, \ldots)$, where the transition to the next $\mathbf{S}_t$ depends only on the previous state S_{t-1}. This limit means that if we were to optimize over these states, as in reinforcement learning, we know that we can treat each transition as its own "subproblem", and ignore the history's $(\mathbf{S}_0, \mathbf{S}_1, \ldots)$ effect on value, and therefore the overall situation has "optimal substructure".

The problem for our defintion of information value 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 the Bellman equation, we need to find another way to establish optimal substructure. This is the focus of Theorem 1. In this theorem we implicitly assume that $\mathbf{X} = \mathbf{S}$, $\mathbf{A}$, $\mathbf{M}$, f, and Λ are given, and that Λ is deterministic,

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

Proof. Given a known optimal value V^* given by $\pi_{\hat{E}}$ we assume for the sake of contradiction there also exists an alternative policy $\hat{\pi}_{\hat{E}} \neq \pi_{\hat{E}}$ that gives a memory $\hat{\mathbf{M}}_{t-1} \neq \mathbf{M}_{t-1}$ and for which $\hat{V}^*_{t-1} > V^*_{t-1}$.

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

This proof requires two things. First, we need to a mechanism of forgetting, of a very particular kind. We assume that the last learning step $f(\mathbf{X}, \mathbf{M}) \to \mathbf{M}'$ can be undone by a new vector valued function f^{-1}, such that $f(\mathbf{X}, \mathbf{M}') \to \mathbf{M}$. In other words we must assume what was last remembered, can always be exactly forgotten. Second, we assume the environmental dynamics, given by the transition function Λ, are deterministic.

Determinism is consistent with the natural world which does evolve in a deterministic way, at the scale of animal behavoir that we concerned with at least. This assumption is however at odds with much of reinforcement learning theory (2) and past experimental work. For example, (114). Both tend to study stochastic environments. We addressed this discrepancy in the main text, using numerical simulations.

Bellman optimal information collection. We aim to find a series of actions $(\mathbf{A}_1, \mathbf{A}_2, ..\mathbf{A}_T)$, drawn from a set $\mathbb{A}$, that maximize each payout $(\hat{E}_0, \hat{E}_1, \ldots, \hat{E}_T)$ so the total payout received is as large as possible. If there is a policy $\mathbf{A} = \pi(\mathbf{X})$ to take actions, based on a series of observations $(X_0, X_1, ..X_T)$,, given by Λ, then an optimal policy π^* will always find the maximum total value $V^* = \text{argmax}_{\mathbf{A} \in \mathbb{A}} \sum_T \hat{E}_t$. 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 this last problem simpler by breaking it down into a small set of subproblems 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. (34)

Mathematically Bellman's principle allows us to translate the full problem, $V^* = \text{argmax}_\pi \sum_T \hat{E}_1, \hat{E}_2, ..., \hat{E}_T$ to a recursive one. Having already proven the optimal substructure of $\mathbf{M}$, and given an arbitrary starting value E_0, the Bellman solution to curiosity optimization of $\hat{E}$ is therefore given by,

$$V^*_{\hat{E}}(\mathbf{S}) = \underset{\mathbf{A} \in \mathbb{A}}{\text{argmax}} \left[\hat{E}_t + V^*_{\hat{E}}(\Lambda(\mathbf{S}, \mathbf{A})) \right] \quad [16]$$

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

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.

If $\pi_{\hat{E}}$ to a deterministic policy this makes proving these three properties amounts to solving sorting problems on $\hat{E}$, and assuming $\mathbf{X} = \mathbf{S}$. That is, if a state is visited by our algorithm it must have the highest value, by definition. So if every state must be visited every state must, at one time or another, give the maximum value. This is certain to happend if we know that all values will begin contracting towards zero, in finite time.

Definitions. Let $\mathbb{Z}$ be the set of all visited states, where $\mathbb{Z}_0$ is the empty set $\{\}$ and $\mathbb{Z}$ is built iteratively over a path P, such that $\mathbb{Z} \to \{\mathbf{S} | \mathbf{S} \in \mathbb{S} \text{ and } x \notin \mathbf{Z}\}$.

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

$$a \leq b \Leftrightarrow \exists\ c;\ b = a + c \quad [17]$$

$$a > b \Leftrightarrow (a \neq b) \wedge (b \leq a) \quad [18]$$

Theorem 2 (Complete exploration)**.** *Given some arbitrary value $\hat{E}_0$, an exploration policy governed by $\pi_{\hat{E}}$ will visit all states $\mathbf{S} \in \mathbb{S}$ in finite number of steps T.*

Proof. Let $\mathbf{E} = (\hat{E}_1, \hat{E}_2, ...)$ be ranked series of $\hat{E}$ values for all states $\mathbf{S}$, such that $(\hat{E}_1 \geq \hat{E}_2, \geq ...)$. To swap any pair of values $(\hat{E}_i \geq \hat{E}_j)$ so $(\hat{E}_i \leq \hat{E}_j)$ by Eq. 17 $\hat{E}_i - c = \hat{E}_j$.

Therefore, again by Eq. 17, $\exists \int \delta \hat{E} \to -c$.

Recall: Axiom 4 - $\nabla^2 \mathbf{M} < 0$ after a finite time T^*.

However if we wished to instead swap $(\hat{E}_i \leq \hat{E}_j)$ so $(\hat{E}_i \geq \hat{E}_j)$ by definition $\nexists c; \hat{E}_i + c = \hat{E}_j$, as $\nexists \int \delta \hat{E} \to c$.

PREPRINT

To complete the proof, assume that some policy $\hat{\pi}_{\hat{E}} \neq \pi^*_E$. By definition policy $\hat{\pi}_{\hat{E}}$ can be any action but the maximum, leaving $k-1$ options. Eventually as $t \to T*$ 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 4 holds. Therefore, the policy $\hat{\pi}_{\hat{E}}$ will leave at least 1 option unexplored and $S \neq Z$. □

Theorem 3 ([Efficient exploration)**.** *An exploration policy governed by $\pi^*_{\hat{E}}$ will revisit all states in exact proption to their information value.*

Proof. Recall: Theorem 2. *Recall*: π^*_E is a maximum value deterministic algorithm. *Recall*: Axiom 2. Each time π^*_E visits a state $\mathbf{S}$, so $\mathbf{M} \to \mathbf{M}'$, and after T^* it is axiomatically true $\hat{E}' < \hat{E}$

By induction then, if $\pi^* E$ will visit all $\mathbf{S} \in \mathbb{S}$ in T^* trials, it will revisit them at most $2T^*$, therefore as $t \to \infty$, $E \to \eta$, where $eta > 0$. □

These exploration proofs come with some fine print, for practical work. 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.

Optimality of Π_π. Recall that in the main text we introduce the equation below as a candidate with zero regret solution to exploration-exploitation problems, set in terms the mixed value sequence $V_{\hat{E}R}$.

$$\Pi_\pi = \begin{cases} \pi^*_{\hat{E}} & : \hat{E} - \eta > R + \rho \\ \pi_R & : \hat{E} - \eta < R + \rho \end{cases} \quad [19]$$

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 7 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 (34, 115).

Theorem 4 (π_π is unbiased)**.** *Let any S, $\mathbf{A}$, $\mathbf{M}$, 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 $T \to \infty$, 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 (59, 60).

In classic scheduling problems the value of any job is known ahead of time (34, 115). 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. 7 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 (116). 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 (117). We therefore assume that last value is the best predictor of the next value, and use this assumption along with Theorem 4 to complete a trivial proof that Eq. 7 maximizes total value.

A win-stay, lose-switch solution. If we prove π_π has optimal substructure, then using the same replacement argument (115) as in Theorem 4, a greedy policy for π_π will maximize total value.

Theorem 5 (No regret - mixed values)**.** *When either $\pi_{\hat{E}}$ or π_R is in control under Π_π, all actions are zero regret in terms of $V_{\hat{E}R}$. That is, $\sum_{k=0}^{T} G = 0$.*

PREPRINT