A way around the exploration-exploitation dilemma

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

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

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

Abstract The exploration-exploitation dilemma is considered a fundamental but intractable problem in the life sciences. Here we challenge the basic formulation and define independent mathematical objectives for exploration and exploitation. Using theory and experiments we prove a competition between exploration-as-curiosity and exploitation-for-reward provides a tractable solution to the dilemma. This solution suceeds when rewards are sparse, deceptive, or non-stationary. That is, it suceeds when the regular approaches often fail.

The regular approach to the explore-exploit dilemma decisions are understood in terms of reward. For example, if a bee goes a familiar direction *to gather nectar* it is said to be exploiting past knowledge. If it goes in an unknown direction *to gather nectar*, it is said to be exploring. But because the world seems stochastic to the bee it cannot know with certainty if it should be choosing either exploitation or exploration *to gain more nectar*. Such an innate uncertainty about rewarding outcomes means the choice is a real dilemma, with no tractable mathematical solution ***Thrun and Möller*** (***1992***); ***Dayan and Sejnowski*** (***1996***); ***Ishii et al.*** (***2002***); ***Şimşek and Barto*** (***2006***); ***Gershman*** (***2018***). We illustrate this in Fig. 1a.

This dilemma is faced routinely by learners of all kinds, including by foraging bees, business orgainizations, humans, worms, monkeys, rodents, birds, children, and computer algorithms ***Gupta et al.*** (***2006***); ***Sutton and Barto*** (***2018***); ***Woodgate et al.*** (***2017***); ***Lee et al.*** (***2011***); ***Schulz et al.*** (***2018***); ***Calhoun et al.*** (***2014***); ***Wang and Hayden*** (***2019***); ***Sumner et al.*** (***2019***); **?**. But is reward fundamental to the problem?

In the natural world exploration finds two explanations. If there is no reason to expect a reward, exploration is described theoretically as a search for information, what we will call curiosity ***Berlyne*** (***1950***); ***Schmidhuber*** (***1991***); ***Kidd and Hayden*** (***2015***); ***de Abril and Kanai*** (***2018***); ***Jaegle et al.*** (***2019***); ***Friston et al.*** (***2016***). On the other hand, if reward is expected exploration gets re-interpreted as a search for reward which then leads to the dilemma ***Kelly*** (***1956***); ***Berger-Tal et al.*** (***2014***); ***Dayan and Sejnowski*** (***1996***); ***Thrun*** (***1992***); ***Mehlhorn et al.*** (***2015***); ***Kobayashi and Hsu*** (***2019***). We argue here that all exploration should be treated as a curiosity search. We illustrate this in Fig. 1b.

The first half of this paper is devoted to developing a new view of curiosity, that can span all fields. This is, it turns out, the same formal problem as learning to value information. We then develop proofs for what we can expect from any curious search. The second half will use the first results to prove curiosity can eliminate this dilemma, if two conjectures we make are true.

Results

To understand curiosity in animals we have become interested in curiosity in artificial life. This is because algorithmic curiosity can,

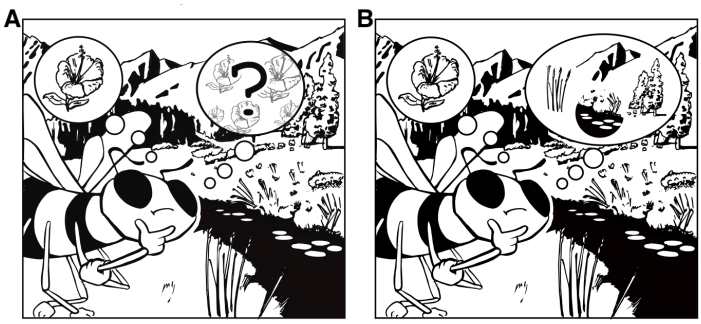

Figure 1. Two views of exploration and exploitation. **a**. The classic dilemma: either exploit an action with a known reward (e.g., return to the previous flower) or explore other actions on the chance they will return a better outcome. The central challenge here is that the outcome of exploration is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

- solve sparse reward problems **?**,
- solve deceptive reward problems **?**,
- solve multi-objective problems **?**,
- lead to diverse skills that are
- robust to instability and **?**
- robust to damage **?**.

This list is unusually broad for a learning algorithm. There is however no mathematics to let us understand this.

So what is curiosity? We define it as an open-ended search to learn any information ***Kidd and Hayden*** (***2015***). We assume this search should be efficient, and is motivated by value. What we would like is a way to describe these ideas mathematically, for kind of any learning algorithm. This is because curiosity depends on many different kinds of learning. In the broadest case, curiosity would then allow for any kind of learning. So this is the setting we have chosen.

Memory and learning

Any notion of information value seems to require that there must be a memory Θ, a way to learn f. For our general mathematical theory we will define memory as,

- A set of real valued numbers
- Embedded in a finite space, with N dimensions
- That is closed and bounded

Let us then define learning in a matching way then. We say a learning function *can be any function* which maps observations x into memory. Using the function-arrow notation this kind of learning is denoted by, $f(x);\Theta) \rightarrow \Theta'$. Likewise a forgetting function *can be any function* which inverts the memory, $f^{-1}(x;\ \Theta') \rightarrow \Theta$.

So what are observations? It is convenient to assume all observations are also embedded in a finite real space, $x \sim X \in \mathbb{R}^m$, as are actions $a \sim A \in \mathbb{R}^k$ and that rewards are positive real numbers, $R \in (R^+)$.

In calling them "observations" we do not mean to imply they must from the senses. Observations could be internally generated from thought, or externally generated from the senses, or both in combination. Observations may contain action information, or they may not. Observations may contain reward information, or they may not. We have no preference and wish to accommodate all the different kinds of cognition **?**.

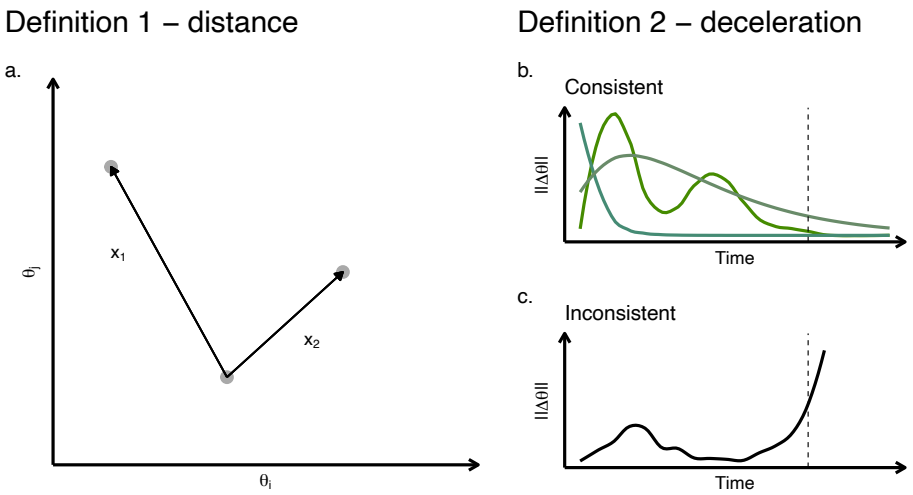

Figure 2. A diagram of our definitions. **a**. This panel illustrates a two dimensional memory. The information value of two observations x_1 and x_2 depends on the norm of memory with learning. Here we show this distance as a euclidean norm, denoted as a black arrow. **b-c** This panel illustrates learning dynamics with time (over a series of observations that are not shown). If information value becomes decelerating in finite time bound, then we say that learning is consistent with Def. 2. This is shown in panel b. The bound is depicted as a dotted line. If learning does not decelerate, then it is said to be inconsistent with Def. 2 (Panel c). *It is important to note:* our account of curiosity does not apply when learning is inconsistent.

Information value

Let's say we have a memory Θ which has been learned by f over a history of observations, $[f(x_0;\Theta_0), f(x_1;\Theta_1), f(x_2;\Theta_2), \ldots]$. Can we measure how much value the next observation $f(x_t)$ should have? If there is such a measure E, we think it reasonable to require it to have certain properties.

1. E should depend only on memory, and what can be immediately learned. (x, Θ, and f)
2. An observation that is known completely, or cannot be learned, should have a value of 0. (If $\Delta\Theta = 0$, then $E = 0$)
3. E should be non-negative because learning is never itself harmful. ($E \geq 0$)
4. E should increase monotonically with the change in memory. ($E \propto \Delta\Theta$)
5. E for the same observation should become decelerating in finite time. ($E \to 0$ as $t \to T^*$)

These properties are in part inspired by Shannon's axioms for information theory, and in part found in the theories of information maximization, and information foraging **?**. What we have done is make their aspects which were implicit in their learning algorithms become explicit and independent of any kind of learning. This is important because anything we can prove this way must then be true for most practical learning algorithms ***MacKay*** (***2003***).

We can satisfy all these by taking a geometric view of learning, as given in the definitions below. They are also shown graphically in Fig. 2 and described in detail in Boxes 1-2.

Definition 1 Let E be the norm of $||f_\Theta(x) - \Theta||$.

Definition 2 Let $||f_\Theta(x)$ be constrained such that $\nabla^2\Theta < 0$ for all $t \geq T^*$.

Exploration as a dynamic programming problem

Curiosity is a search to maximize information value, again in the boarded sense possible. We decided to seek a dynamic programming solution to maximizing curiosity. This is useful because it would guarantee value is maximized, which because of the way we have made our definitions also

Box 1. Norms.

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

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

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

Box 2. The Laplacian.

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

Box 3. Dynamic programming and Bellman optimality.

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

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

Mathematically this allows us to translate the full problem, $V^* = \text{argmax}_\pi \sum_T r_1, r_2, ..., r_T$ to a recursive one, which is given below. Using this new form we can find an optimal solution by moving though the sequence of action iteratively, and solving each step independently.

$$V^* = \underset{\pi}{\text{argmax}} \left[r_0 + V(x_1) \right] \tag{1}$$

The question which remains though is how can we ensure our problem can be broken down this way? This is addressed in Box 4.

---

guarantees that learning is complete. It is also convenient because dynamic programming is the basis for reinforcement learning ***Sutton and Barto*** (***2018***).

The normal route to find a dynamic programming solution is to first prove your problem has "optimal substructure", then to derive the final result using Bellman optimality, and this is the route we will follow. For a more complete introduction to optimal substructure, see Box 4. For more on dynamic programming and Bellman optimality, see3.

In Theorem 1 (mathematical appendix) we prove our definition of memory has optimal substructure. This proof relies on the forgetting function f^{-1} to succeed. Even though we have said little about it until now, this why we include forgetting in our early definitions.

The Bellman derivation leads to an especially practical and simple result (Eq. 9). Given an arbitrary starting value $E_0 > 0$, the dynamic programming solution to maximizing E is given by 2. A full derivation is provided in the mathematical appendix. We denote our policy that follows this solution as π_E.

$$V_E^\pi(x) = \underset{A}{\text{argmax}} \left[E_0 + V_E(x_1) \right] \tag{2}$$

However as long as the search for E continues until learning is at a steady-state, and as result information value is zero, then are a range of equally good solutions (Theorem **??**). These let us simplify Eq. 2 further, giving Eq. 3.

$$V_E^\pi(x) = \underset{A}{\text{argmax}} \left[E_0 + E(x_1) \right] \tag{3}$$

Optimal exploration, and the importance of boredom

In our opening we said curiosity is a search for information, but we also said it should be an efficient search. So what would it mean for a curiosity search to be efficient? Does our Bellman solution

Box 4. Optimal substructure

To use Bellman's principle of optimality (Box 3) the problem we want to solve needs to have **optimal substructure**. This opaque term can be understood by looking in turn at another theoretical construct, Markov spaces.

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

In reinforcement learning the problem of reward collection is defined by fiat as happening in a Markov space ***Sutton and Barto*** (***2018***). The problem for curiosity is it relies on memory and memory, being necessarily composed of many past observations in many orders, cannot be a Markov space. So if we wish to use dynamic programming to maximize E, we need to find another way to establish optimal substructure. This is the focus of Theorem 1.

ensure it?

We have discussed maximizing curiosity by means of E but to study efficiency must now discuss limiting it. This is necessary because exploration is limited in practice by the available energy, hunger, greed, thirst or other biological priorities. But more importantly search must be limited to avoid the central limit of curiosity: it can become distracted by minutia. A good example of minutia was given ***Pathak et al.*** (***2017***) who asks us to, "Imagine a scenario where the agent is observing the movement of tree leaves in a breeze. Since it is inherently hard to model breeze, it is even harder to predict the location of each leaf". This will imply, they note, that a curious agent will always remain curious about the leaves, to its own detriment.

The solutions to both problems are the same. To limit curiosity we will draw on its opposite, boredom ***Schmidhuber*** (***1991***). Others have considered boredom an adaptive trait as well, arguing it is a useful way to motivate aversive tasks ***Bench and Lench*** (***2013***). We take a slightly different view and consider boredom as a means to ignore the marginal, and as a tunable free parameter, $\eta \geq 0$. That is, we say curious exploration should cease once E becomes marginal as given by Eq. 4.

$$E \leq \eta : \pi_E \rightarrow \pi_{\emptyset} \tag{4}$$

Here we use $\pi_{\emptyset}$ to denote any other action policy that is not π_E. Informally what Eq. 4 this means that once E is below the boredom threshold, E is marginal, and the learner should change its behavior from exploration to some other policy. This change in behavior could be towards an aversive policy (as in ***Bench and Lench*** (***2013***)) or it could be towards reward seeking, as we will assume.

We think it is reasonable to call a curious search optimal if it meets the three criteria below. In Theorem **??** we prove that π_E satisfies these, when $\eta > 0$.

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

The importance of determinism in search

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 E. (It is also this determinism which will let us resolve the dilemma, later on.) But a deterministic view of exploration is at odds with how the problem is generally viewed today, which involves some amount of randomness. Whereas if our analysis is correct, randomness is not needed or desirable, as it

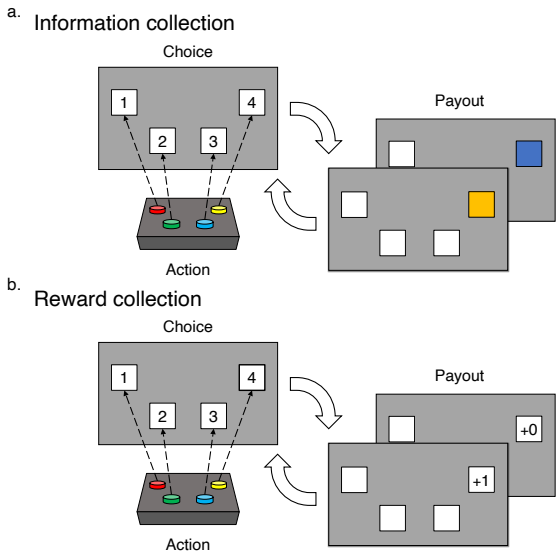

Figure 3. The two basic kinds of tasks we will study. **a.** Represents a very simple information foraging task. The information is a yellow or blue stimulus, which can change probabilistically from trial to trial. A good learner in this task is one who tries to learn the probabilities of all the symbols in each of the four choices. The more random the stimuli are in a choice, the more potential information/entropy there is to learn. **b.** Represents a very simple 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.

must lead to less value and a longer search. We confirm and illustrate this result using a simple information foraging task (Fig.**a**). The results of which are shown in Figure **??**.

Besides offering a concrete example, performance in this Task 1 is a first step in showing that even though π_E's optimality is proven for a deterministic environment in our proofs, it will perform well if not optimal in stochastic settings.

The dilemma

How can we justify using curiosity for the dilemma problem when the goal of exploitation is reward collection? Intuitive suggests that searching for in an open-ended way must be less efficient than a direct search. We have an answer in two conjectures:

- **Conjecture 1** Reward value and information value are equally important.
- **Conjecture 2** Curiosity is a sufficient solution to exploration (when learning is the goal)

If reward and information are equally important, how can an animal go about balancing them? We will argue that answering this question is the same as answering the exploration-exploitation dilemma. So can we solve the dilemma with independent policies?

We will answer this last question using the framework of regret minimization, which is the standard view. Regret is a numerical quantity that represents the difference between the most valuable choice, and the value of the choice that was made. To approximate regret G we use Eq. 5, where V is the value of the chosen action, and V^* is the maximum value.

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

An optimal value algorithm with zero regret always makes the most valuable choice. But this is impossible to achieve when exploring any variable environment, using stochastic search.

To find the algorithm that maximizes both information and reward value with zero regret, we imagined the policies for exploration and exploitation acting as two possible "jobs" competing for control of behavior, a fixed resource. We know (by definition) each of these jobs produces non-negative values which an optimal job scheduler could use: E for information or R for reward/reinforcement learning. We can also ensure (by definition) that each of the jobs takes a

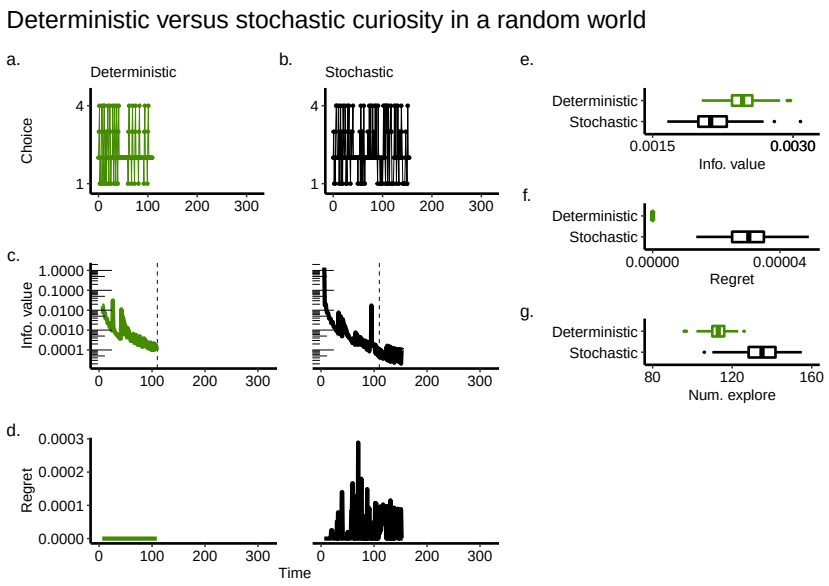

Figure 4. Comparing deterministic and stochastic curiosity in a simple information foraging task (Task 1). Deterministic results are shown in the left column, and stochastic are shown in the right. The hyperparameters for both models were found by random search, which is decribed in the Methods. **a-b**. 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.

constant amount of time and that each policy can only take one action at a time. These two properties, and one further assumption, are it turns out sufficient.

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

We arrived at our solution by taking the simplest possible path available. We just wrote down the simplest equation that could have zero regret answer, and began to study that. This is Eq. 19. We refer to it as a meta-greedy algorithm, as it decides in greedy fashion which of our independent policies to use.

$$\pi^{\pi} = \left[\pi_E,\ \pi_R\right] \tag{6}$$

$$\underset{\pi^{\pi}}{\operatorname{argmax}}\ \left[E_{t-1},\ R_{t-1}\right]_{\mathbf{(2,\,1)}} \tag{7}$$

We have in Eq. 19 modified traditional argmax notation. It now includes a vector $(2, 1)$ to denote a ranking system for which value/policy should be prefered in the event of a tie. (Smaller numbers are more preferred.)We add this notation because an exact rule for tie breaking is critical to our mathematical proofs (which follow below).

In Eq. 19 we also limit the probability of having zero reward to be itself non-zero. That is, $p(R_t = 0) > 0$. This is necessary to ensure exploration. For reasons of mathematical convenience we also limit E_t to be positive and non-zero when used with boredom, $(E_t - \eta) \geq 0$.

In Theorem 5 we prove Eq. 19 is Bellman optimal, and so it will always find a maximum value sequence of E and R values. Given that is, an initial value $E_0 > 0$, and that the environment is deterministic. In Theorem 5 we also prove that with a judicious choice in boredom η, the reward policy π_R will also converge to its optimal value, assuming it has one. For full details the mathematical appendix.

Box 5. Algorithmic run time.

It is common in computer science to study the runtime of an algorithm. So what is the runtime of our independent policies 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 $T_E + T_R$.

Box 6. Reward homeostasis.

We have been studying a kind of reinforcement learning where the value of a reward is fixed. We'll call this the economic style. It is the most common kind of model found in machine learning, psychology, neuroeconomics, and neuroscience ***Sutton and Barto*** (***2018***). In natural behavior though it is common that the motivational value of reward declines as the animal gets "full", or reaches satiety. We'll call this the homeostatic style ***Keramati and Gutkin*** (***2014***); ***Juechems and Summerfield*** (***2019***); ***Münch et al.*** (***2020***) Fortunately it has already been proven that a reward collection policy designed for one style will work for the other (at least this is so an infinite time-horizon ***Keramati and Gutkin*** (***2014***)). However to use Eq. 19 for reward homeostasis requires we must make one intuitive modification. Ties between values should be broken in favor of exploration, and information seeking. (In the economic reinforcement learning ties break in favor of exploiting rewards). This modification leads to Eq. 8.

$$\underset{\pi^\pi}{\operatorname{argmax}} \left[E_{t-1},\ R_{t-1} \right]_{(1,\,2)} \tag{8}$$

Table 1. Exploration strategies.

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

Simulations of reward collection

We will now focus on practical performance, believing it is a compelling way to make the case for curiosity as sufficient. The general form of the 7 tasks we will study is depicted in Fig 3. They are all variations of the classic multi-armed bandit ***Sutton and Barto*** (***2018***). The payouts for each are shown in Fig. **??**-**??**.

Each task was designed to either test exploration performance in a way that matches recent experimental study(s), or to test the limits of curiosity. On each trial was with a set of n choices. Each choice returns a "payout", according to a predetermined probability. Payouts are information, reward, or both (Fig. **??**). Note that information was symbolic, denoted by a color code, "yellow" and "blue" and as is custom reward was a positive real number. A summary for each task, and a visual depiction of its payouts, are shown in Fig 5.

We have chosen a range of other exploration strategies to compare against. We feel it is broad enough that we have made a fair representation of today's state of the art. They are drawn from both neuroscience, and machine learning. They have in common that their central design goal was to maximize reward value, often supplemented by some other goal. For a brief description of each, see Table 1. The results in full for all tasks and exploration strategies are shown in Fig. 6. Note that all learners, including ours, used the same exploitation policy, based on the TD(0) learning rule ***Sutton and Barto*** (***2018***); **?**.

It is not possible for any one learning algorithm to work best in all situations **?**. A good outcome for our algorithm seems to be falling in the top 2 or 3 or so for all tasks. This is what we observed (Fig 6). Given our aim to test our method of independent policies it is not interesting to detail the varieties of why performance was found to vary between strategies and tasks. The no free lunch theorem forbids any general no one strategy from prevailing on all problems **?**.

In terms of curiosity's performance though its worth noting that on *Task 4* ours is the only strategy that reaches above chance performance. This was then deception task, which was included to confirm that in a small amount deception derails exploration in the classic form and to confirm that our form of curiosity can overcome deception as we would expect it to

In terms of curiosity, *Task 5* is also worth a special note (Fig 6**c**). It was designed to fool curiosity by presenting information that was irrelevant. What we saw though was that with judicious use of boredom, curiosity still produced a very high level of performance. In the supplementary figure we also show that if boredom is mistuned, curiosities performance suffers, but not catastrophically (TODO - add this figure).

Tasks 6-7 are also worth noting (Fig 6**e-f**). *Task 6* was a 121 choice experiment that is at the limit of human performance **?**. In *Task 7* then we had learning continue from Task 6, and all was the

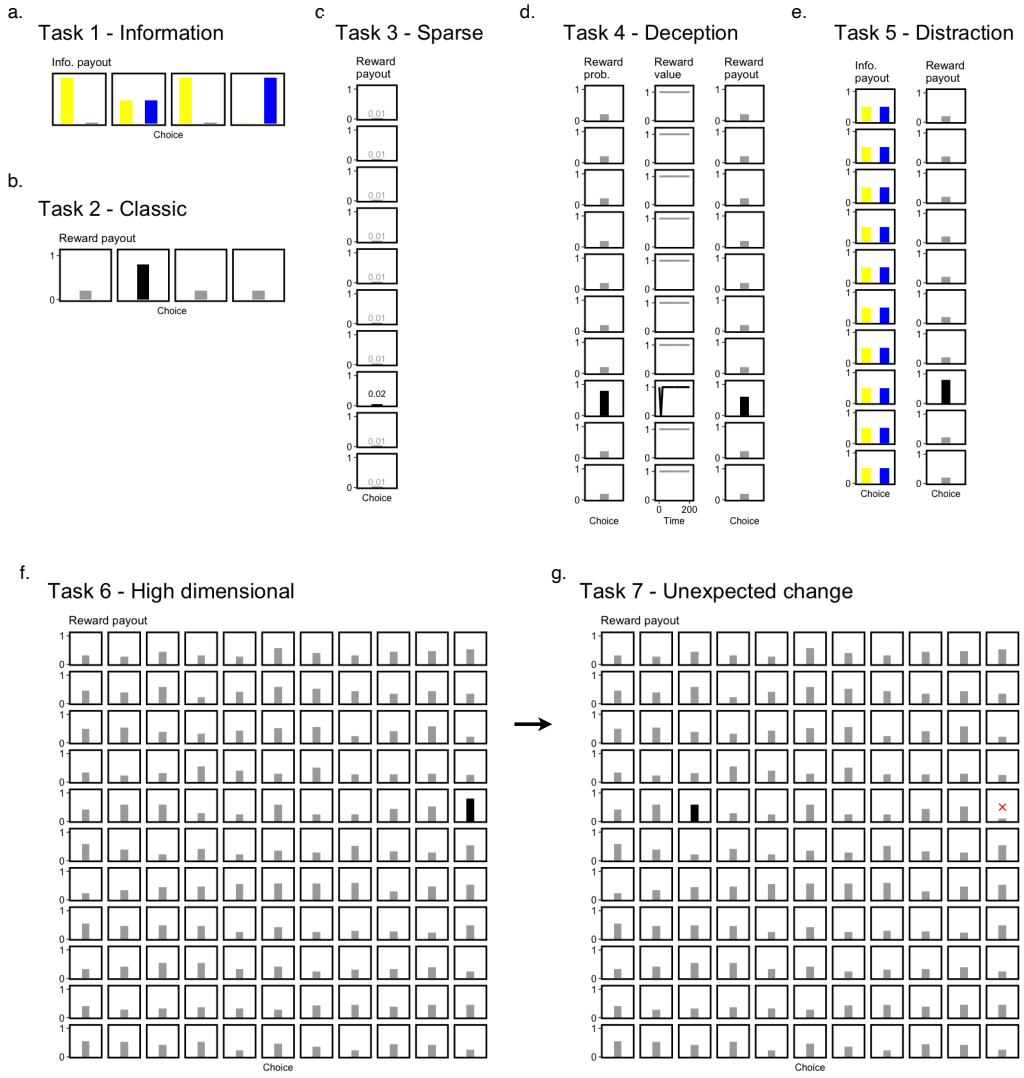

Figure 5. Payouts for *Tasks 1 - 5*. Payouts can be information, reward, or both. For comments on general task design, see Fig 3. **a.** A classic four-choice design for information collection. A good learner should visit each arm, but quickly discover that only arm two is information bearing. **b.** A classic four-choice design for testing exploration under reward collection. The learner is presented with four choices and it must discover which choice yields the highest average reward. In this task that is Choice 2. **c.** A ten choice sparse reward task. The learner is presented with four choices and it must discover which choice yields the highest average reward. In this task that is Choice 8 but the very low overall rate of rewards makes this difficult to discover. Solving this task with consistency means consistent exploration. **d.** A ten choice deceptive reward task. The learner is presented with 10 choices, but the choice which is the best on the long-term (>30 trials) has a lower value in the short term. This value first declines, then rises (see column 2). **e.** A ten choice information distraction task. The learner is presented with both information and rewards. A good learner though will realize the information does not predict reward collection, and so will ignore it. **f.** A 121 choice task with a complex payout structure. This task is thought to be at the limit of human performance. A good learner will eventually discover choice number 57 has the highest payout. **g.** This task is identical to *a.*, except for the high payout choice being changed to be the lowest possible payout. This task tests how well different exploration strategies adjust to simple but sudden change in the environment.

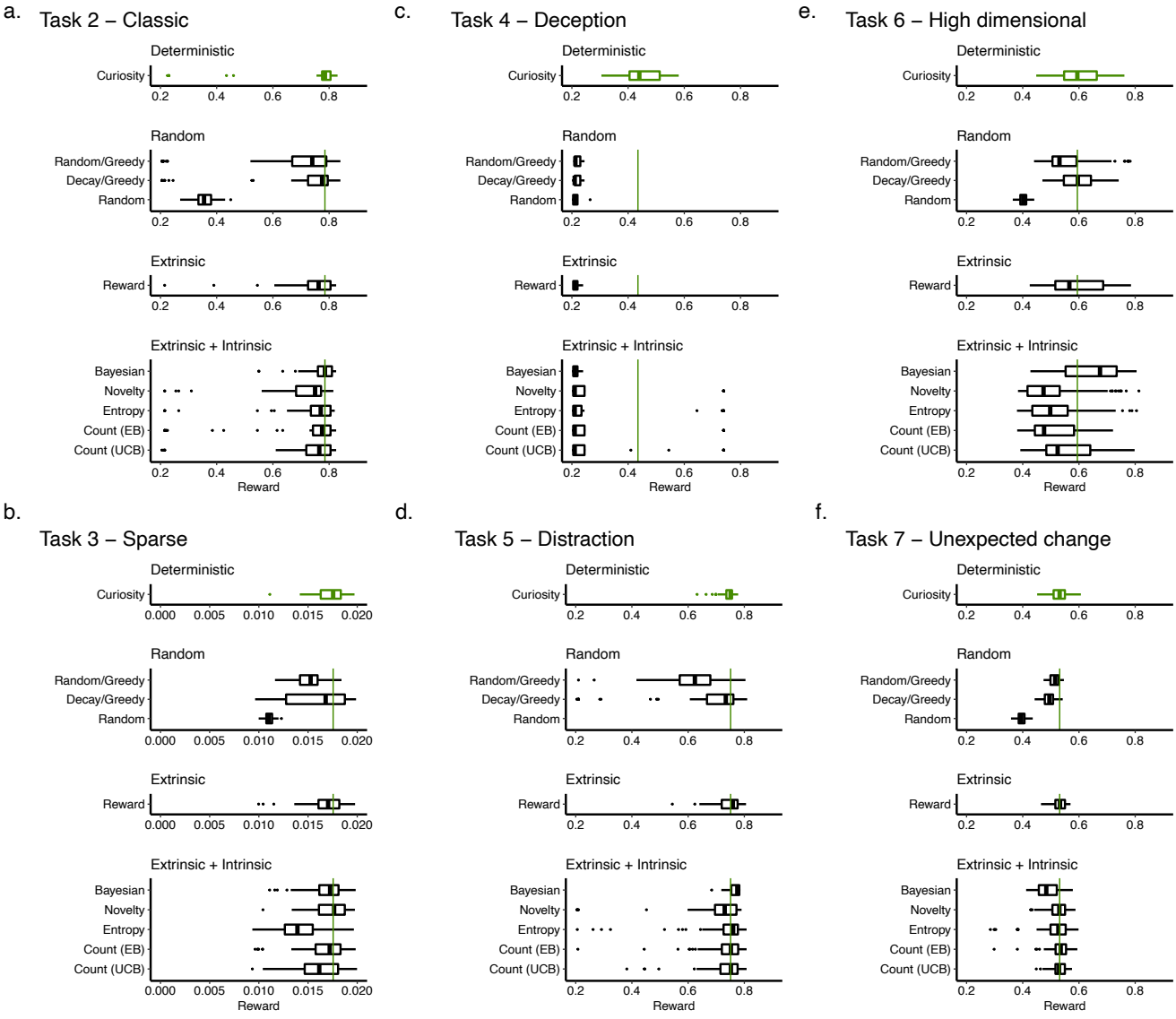

Figure 6. Summary of reward collection (*Tasks 2-7*). The strategies in each panel are grouped according to the class of search they employed (Curiosity. Random, Extrinsic reward or Extrinsic + Intrinsic rewards). **a.** Results for Task 2, which has four choices and one clear best choice. **b.** Results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

Figure 7. The importance of boredom during reward collection (Task 2). One hundred example experiments, each began with a different random seed. **a**. Behavioral choices with optimized boredom. **b**. Behavioral choices with unoptimized boredom. **c,d**. Average reward as a function of time for optimized (c) and unoptimized (d) boredom.

same except that what was the best choice in Task 6 was suddenly and unexpectedly was the worst choice in Task 7. Under this sudden nonstationarity the top strategy (Bayesian) became the worst, and curiosity took the top spot. (It was already second place in Task 6). This is exactly the kind of robustness we'd predict for any curiosity algorithm **?**. Curiosity drove an unbiased model of the value structure that allowed it to adapt quickly. As did it's worth noting all the other more unbiased exploitation models (the count, novelty, and entropy models; Fig 6**f**).

When using Eq 19 a good choice of boredom eta is critical, and was optimized carefully above. We show an example of this importance in Fig. 7 and in Fig 3b. On the left hand side we did a random search over 10000 possible parameters to find a value for η that produced excellent results. On the right hand side we choose a value for boredom we knew would lead to a relatively poor result. We then ran the 100 different randomly generated simulations which gives us the results seen in Fig. 7).

The optimized boredom converged far more quickly (Fig. 7a-b), generating more reward over time (Fig. 7c-d), and showed far more in consistency during exploration (Fig. 7e-f).

Fig. 7 is also useful in showing the wide variance in behavior which is possible with what is otherwise a deterministic search. That is, exploration in these figures is deterministic, as prescribed by Fig **??**. What we see experimentally however is a wide range of behaviors (top panels), with very different distributions for their actions choices (bottom panels). This is no because exploration is stochastic, but can instead be understood by variations in the environment changing what is learned, which then changes what the best learner should choose to learn next.

Unlike idealised simulations, animals cannot pre-optimize their search parameters for every task. We therefore explored reward collection as a function of 1000 randomly chosen search parameters, and reexamined performance on two tasks. These were chosen to represent an "easy" task, and a "hard one". The results are shown in Fig. 8.

As above exploitation by deterministic curiosity produced top-3 performance on both tasks (Fig. 8a-b). Most interesting is the performance for the worst parameters. Here deterministic curiosity was markedly better and the other still. We can explain this in the following way. All the other exploration strategies use parameters to tune the degree of exploration. With deterministic

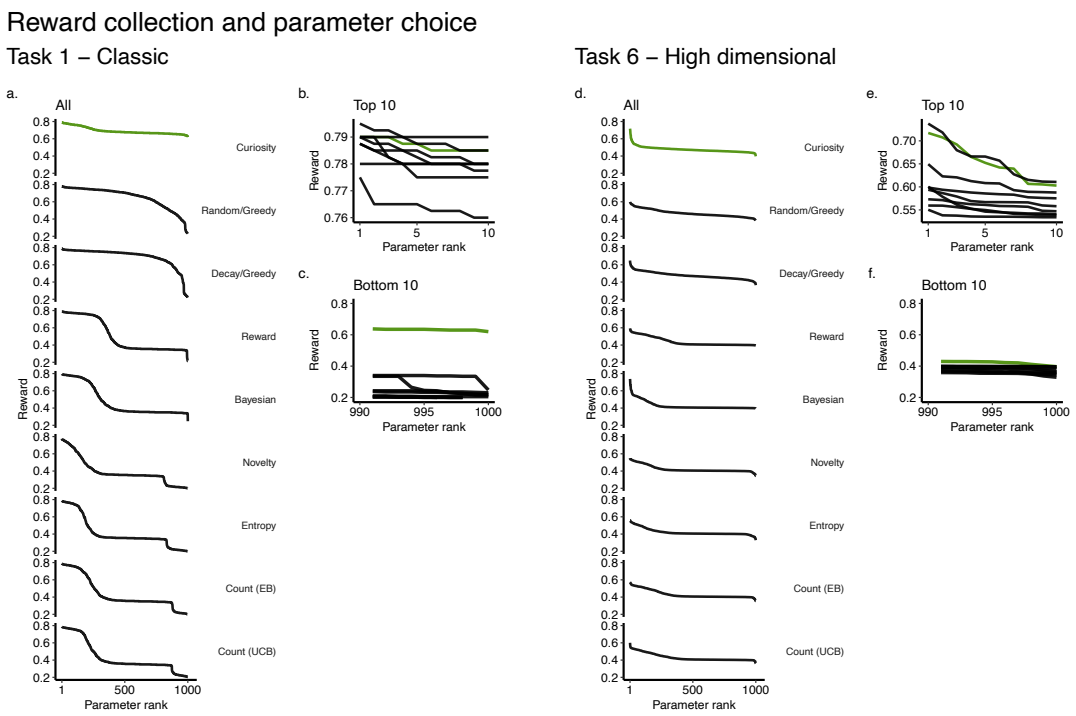

Figure 8. Parameter selection and search performance (*Task 2* and *6*). We wanted to evaluate how robust performance was to poor and random hyperparameters. A very robustness exploration algorithm would produce strong performance with both the best parameter choices, and the worst parameter choices. **a,d** Total reward for 1000 search hyperparameters, for each of strategies we considered. **b,e** Performance with the top 10 hyperparameters, curiosity (green) compared to all the others. **c.f** Performance with the bottom 10 hyperparameters. **TODO** - add error bars

curiosity though we tune only when exploration should stop. The degree of exploration–measure by how close it is to maximum entropy–is fixed by the model itself. And it seems that here at least it is far more robust to tune the stopping point, rather than the degree of exploration.

If our deterministic solution is working as the mathematics predict it should, then adding a random aspect to either exploration or exploitation should only act to degrade performance. We showed an example of this in Fig. 9. Here adding two levels of random noise to the choices only served to increase variability of convergence (left) and reduce the total rewards collected (right). When we then compare performance between curiosity and the other strategies as noise smoothly changes we see performance eventually decline to chance levels.

In studying noise in Fig 10 many of the strategies showed a peak in performance as noise increased, suggesting a minimum level is needed. There were two who did not–entropy and Bayesian. They showed some ability to operate well with low stochasticity. If this trend when removing stochasticity completely, then that would be half of the ingredients needed to find another set of zero regret solutions. When however we "forced" determinism on these and other strategies, performance declined to chance (Fig 10b). The exception to this is the Bayesian strategy, whose median performance approaches the ideal (0.8). Hope for this exception is limited however by its having substantially more variance than any others. In other words, it appeared to be unreliable.

Figure 9. Behavior during reward collection, using either deterministic or stochastic curiosity (*Task 2*). **a-c** Examples of exploration choices using Eq 19. In the left panel is the deterministic version. The other two panels show the same algorithm with two levels of noise added to both curiosity and reward collection policies. **d.** The effect of noise on rewards collected, over 100 independent experiments.

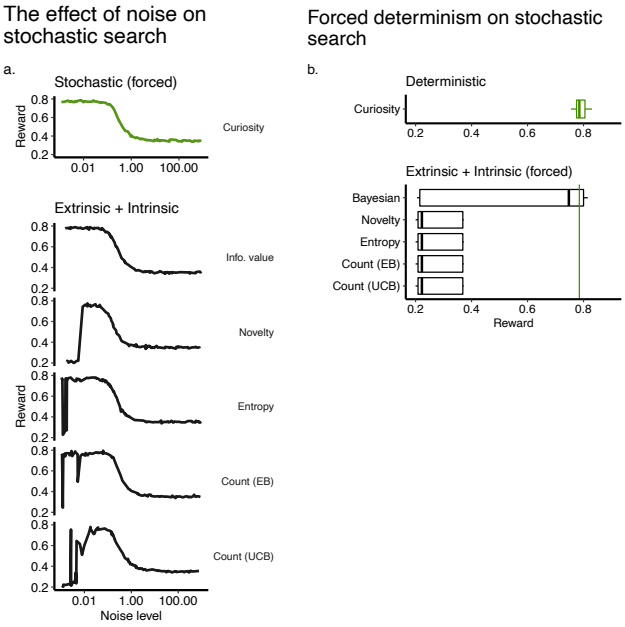

Figure 10. Adding noise, or forcing determinism (*Task 2*). **a.** In this panel we forced our deterministic approach to behavior stochastically. We compared ours to all the others as the level of noise increased. **TODO** - add error bars **b.** In this panel we forced several of the search strategies which rely in part on stochastic search to operate in a deterministic way. We plot total reward for these agents, compared to our deterministic approach (green).

Discussion

We have offered a way around the dilemma where curiosity is how animals explore, in all circumstances. This is a controversial claim. So we'll address some of its criticisms, as a series of questions.

Is this too complex?

Perhaps turning a single objective into two, as we have, is an unneeded complication. If this is true this it would mean ours is not a parsimonious solution to the dilemma, and so we could or should reject it on that alone.

Questions about parsimony are resolved by considering the benefits versus the costs. The benefits to curiosity-based search is that it leads to no regret solutions to exploration, to exploration-exploitation, and that it so far seems to do very well in practice, at reward collection. At the sametime curiosity-as-exploration can also be building a model of the environment, useful for later planning ***Ahilan et al.*** (***2019***); ***Poucet*** (***1993***), creativity, imagination ***Schmidhuber*** (***2010***), while also building in some cases diverse action strategies ***Lehman and Stanley*** (***2011***); ***Lehman et al.*** (***2013***); ***Mouret and Clune*** (***2015***); ***Colas et al.*** (***2020***). In other words, we argue it is a needed complication.

Is this too simple?

Solutions to the regular dilemma are complex and require sophisticated mathematical efforts and complex computations when compared to our solution. Or put another way, our solution may seem too simple. So have we "cheated" by changing the problem as we have?

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

The answer to this question is empirical. If our account does well in describing and predicting animal behavior, that would be some evidence for it. If it predicts neural structures ***Cisek*** (***2019***), that would be some evidence for it. If this theory proves useful in machine learning and artificial intelligence research, that would be some evidence for it.

Is this a slight of hand, theoretically?

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

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 ***Sumner et al.*** (***2019***); ***Kidd and Hayden*** (***2015***). This is why we limit curiosity with boredom, and counter it with a competing drive for reward collecting / exploitation.

We believe there is a useful analogy between science and engineering. Science can be considered an open-ended inquiry, and engineering a purpose driven enterprise. They each have their own pursuits but they also learn from each other, often in alternating iterations ***Gupta et al.*** (***2006***). It is their different objectives though which make them such good long-term collaborators.

But what about curiosity on big problems?

A significant drawback to curiosity, as an algorithm, is that it will struggle to deliver timely results when the problem is very large, or highly detailed. First, it's worth noting that most learning algorithms struggle with this setting ***MacKay*** (***2003***); ***Sutton and Barto*** (***2018***), unless that is unless significant prior knowledge can be brought to bear ***Zhang et al.*** (***2020***); ***Sutton and Barto*** (***2018***) or

the problem can be itself compressed ***Ha and Schmidhuber*** (***2018***); ***Fister et al.*** (***2019***). Because however "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 afterall ***Wolpert and Macready*** (***1997***).

But what about...?

We did not intend to dismiss the niceties of curiosity, which 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 ***Kidd and Hayden*** (***2015***); ***Berlyne*** (***1950***); **?**. Computer science tends to define it as needed, for the task at hand ***Stanley and Miikkulainen*** (***2004***); ***Friston et al.*** (***2016***); ***Lehman and Stanley*** (***2011***); ***Lehman et al.*** (***2013***); ***Mouret and Clune*** (***2015***); ***Colas et al.*** (***2020***). Before dividing it 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 a next problem we will work on and it is very important. The central challenging comes in deriving boredom while keeping the abstract approach which we have used, and which seems critical.

Does this mean you are hypothesizing that animals not only become bored but that a part of their physiology has optimized this?

We are predicting exactly that.

But I don't feel like I am tuning my boredom?

Perhaps it happens unconsciously. The relationship between learning, function, consciousness and unconscious is not well enough understood for how something does or does not feel to be an argument in and of itself.

How can an animal tune boredom in a new setting?

The short answer is that it would need to guess and check. Our work in Fig 8 suggests this can be somewhat robust.

What about information theory?

Weaver ***Shannon*** (***1948***) 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.

Unfortunately valuing information is not at its core a communications problem. Consider for example Bob and Alive 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 if we argue that he values 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 the personal history defines what is valuable on a channel, and this is independent of the channel and the messages sent. We have summarized this diagrammatically, in Fig. 11.

There is however an analogous set of levels for information value as to those Weaver describes. a. There is the technical problem of judging how much was learned. b. There is the semantic

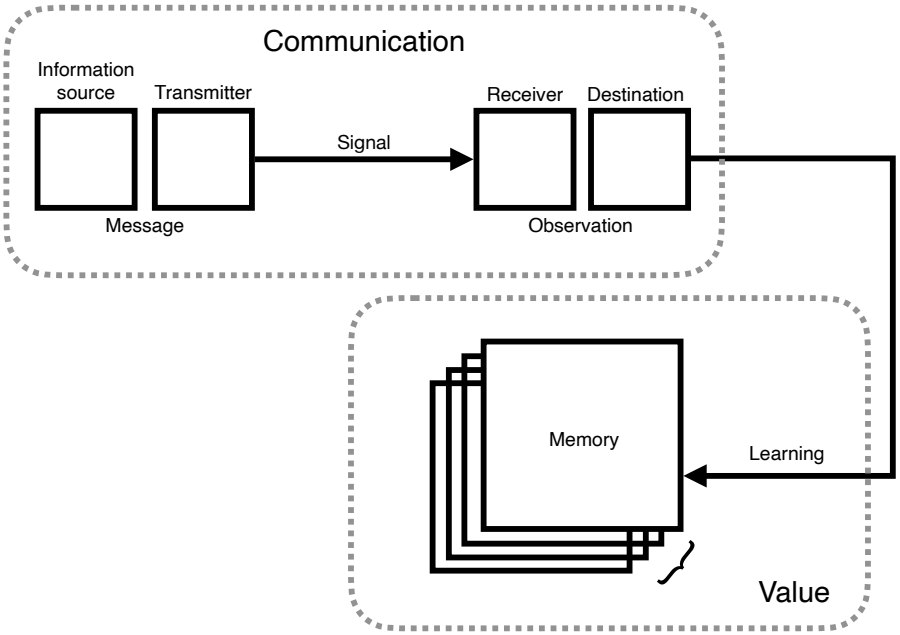

Figure 11. The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel. Value is derived from learning about observations, which is in turn dependent on memory and its past history.

problem of both what this learning "means", and also what its consequences are. c. There is an effectiveness problem of using what was learned to some other effect. In many ways b and c are similar to those in the information theory. It is which is most different.

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

(

Does value as a technical problem even make sense?) Information value has been taken up as a problem in meaning, or usefulness **?**. These are based on what we will call b-type questions (a notion we defined in the section above). It is not that b questions are not important or are not valuable. It is that they are second order questions. The first order questions are the 'technical' or a-type questions (also defined above).

Having a technical definition for value, free from any meaning, might seem counterintuitive for a value measure. But we argue that it is not more or less counterintuitive than stripping information of meaning in the first place. (This is how Shannon got his information theory). The central question for our account of value is, is it useful? It is enough to ensure a complete but perfectly efficient search for information/learning in any finite space. It is enough to play a critical part in solving/replacing the dilemma, which has for many years looked intractable.

What about mutual information, or truth?

In other prototypical examples, information value is based on how well what is learned relates to the environment **?**. Their mutual information that is. (Colloquially, one might call this truth.) The larger the mutual information between the environment and the memory, the more value it is said

that we should then expect. This view seems to us at odds with actual learning behavior, especially in humans.

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

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

This is our most unexpected prediction: We must.

What about information foraging?

Our work is heavily inspired by information foraging. (We are also aware of its limits **?**). Our motivation in developing was that information value should not depend on only a probabilistic reasoning, as it does in information foraging. This is for the simple reason that many of the problems animals need to learn are not probabilistic problems *from their point of view*. (Most any learning problem can be described as probabilistic from an outsider's perspective; the one scientists' rely on). We also felt the algorithmic curiosity looked so important that it needed as general a way to study it, which is why we developed an axiomatic approach that ended up addressing the technical problem of information value.

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

No. It was an idealistic choice, which worked out. The field of curiosity studies has shown 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 the ideas will span fields, addressing the problem as it appears in computer science, psychology, neuroscience, biology, and perhaps economics.

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

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

Is information a reward?

It depends. If reward is defined as more-or-less any quantity that motivates behavior, then our definition of information value is a reward. This does not mean however that information value and environmental rewards are interchangeable. In particular, if you add reward value and information value to motivate search you can drive exploration in practical and useful ways. But this doesn't change the intractability of the dilemma proper ***Thrun and Möller*** (***1992***); ***Dayan and Sejnowski*** (***1996***); ***Findling et al.*** (***2018***); ***Gershman*** (***2018***). So exploration will still generate substantial regret (as we show in Fig 6). But perhaps more important is that these kinds of intrinsic rewards ***Schmidhuber*** (***1991***); ***Berger-Tal et al.*** (***2014***); ***Itti and Baldi*** (***2009***); ***Friston et al.*** (***2016***); ***Kobayashi and Hsu*** (***2019***) introduce a bias that will drive behavior away from what, could otherwise be, ideal skill learning ***Ng et al.*** (***1999***); ***Şimşek and Barto*** (***2006***).

Animal behavior and neural circuits

In psychology and neuroscience, curiosity and reinforcement learning have developed as separate disciplines ***Berlyne*** (***1950***); ***Kidd and Hayden*** (***2015***); ***Sutton and Barto*** (***2018***). And er hghilighted how they are separate problems, with links to different basic needs: gathering resources to maintain physiological homeostasis ***Keramati and Gutkin*** (***2014***); ***Juechems and Summerfield*** (***2019***) and gathering information to decide what to learn and to plan for the future ***Valiant*** (***1984***); ***Sutton and Barto*** (***2018***). 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. We'll now review evidence that evolution has respected this distinction.

Cisek (2019) has traced the evolution of perception, cognition, and action circuits from the Metazoan to the modern age ***Cisek*** (***2019***). 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, historically anyway, had no input from the homeostatic circuits needed for reward valuation ***Keramati and Gutkin*** (***2014***); ***Cisek*** (***2019***); ***Juechems and Summerfield*** (***2019***). This neural separation for independent circuits has been observed in "modern" high animals, including zebrafish **?**, mouse **?**, humans **?**, and monkey ***White et al.*** (***2019***); ***Wang and Hayden*** (***2019***). It is difficult to know the exact objectives of these circuits, but their independence aof activity and independence during behavior suggests they are, by degrees, independent policies.

What about aversive values?

We do not believe this is complete theory. For example, we do not account for any consequences of learning, or any other aversive events. The need to add

Fitting deterministic models?

The theoretical description of exploration in scientific settings is probabilistic ***Calhoun et al.*** (***2014***); ***Song et al.*** (***2019***); ***Gershman*** (***2018***); ***Schulz et al.*** (***2018***). By definition probabilistic models can't make exact predictions of behavior, only statistical ones. Our approach is deterministic, and so does make exact predictions. Our theory predicts that it should be possible to guide exploration in real-time using, for example, optogenetic methods in neuroscience, or well timed stimulus manipulations in economics or other behavioral sciences.

In species ranging from human to *Caenorhabditis elegans*, there are hundreds perhaps thousands of exploration-exploitation experiments. Analysis of their behavior has generally been limited to distributions. A deterministic theory can, in principle, open up entirely new avenues for reanalysis by using our model to make exact predictions.

TODO - discuss practical fits:

1. Burn in
2. Population E0, and refinement, replicators?
3. Data fusion (bayes methods for deterministic equations are legit)

Does regret matter?

Another way around the dilemma is to ignore regret. This is the domain of pure exploration methods, like the Gitten's Index **?**, or PAC approaches ***Valiant*** (***1984***), or other bounded forms of reinforcement learning **?**. Pure random search is another useful example. Such methods can guarantee you find the best value, often in an unbiased way. They can in some cases do so with a bound on the number of actions needed. These methods do not consider regret or missed value encountered during search. Regret might not be important in some settings, but we expect that when time and resources are scarce minimizing regret must be a critical skill for survival.

Summary

There will be certainly cases in an animal's life when their exploration must focus on reward. But even here the search for reward information alone will make the search more robust to deception, and more reliable in the future. Think of this as curiosity about reward. But there are many more cases where curiosity can include observations of the environment, and the self. This paper has argued that one should put aside intuitions which suggest open-ended curious search is too inefficient to be practical. We have shown how it can be made very practical. We have argued one should put aside intuitions for what exploration behavior represents, and used this to offer a zero-regret way around the dilemma – a problem which has seemed unsolvable.

Mathematical Appendix.

Information value as a dynamic programming problem

To find a dynamic programming solution based on the Bellman equation (Eq **??**) we must prove our memory Θ 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 Θ can be partitioned into a collection, or series of memories, each of which is itself a solution. That is, it lets us find a kind of recursive structure in memory learning, which is what we need to apply the Bellman. Proving optimal substructure also allows for simple proof by induction ***Roughgarden*** (***2019***), a property we will take advantage of.

Theorem 1 (Optimal substructure)**.** *Let X, A, Θ, f_Θ, (Def. 1), π_E and δ be given. Assuming transition function δ is deterministic, if $V^*_{\pi_E}$ is the optimal information value given by policy π_E, a memory Θ_{t+1} has optimal substructure if the the last observation x_t can be removed from Θ_t, by $\Theta_t = f^{-1}(\Theta_{t+1}, x_t)$ such that the resulting value $V^*_{t-1} = V^*_t - E_t$ is also optimal.*

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

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

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

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

Bellman solution

Knowing the optimal substructure of Θ and given an arbitrary starting value E_0, the Bellman solution to curiosity optimization of E is given by,

$$V_E(x) = E_0 + \underset{a}{\operatorname{argmax}} \left[E_t \right] \tag{9}$$

To explain this derivation we'll begin simply with the definition of a value function and work from there. A value function $V(x)$ is defined as the best possible value of the objective for x. Finding a Bellman solution is discovering what actions a to take to arrive at $V(x)$. Bellman's insight was to break the problem down into a series of smaller problems, leading a recursive solution. Problems

that can be broken down this way have optimal substructure. The value function for a finite time horizon T and some payout function F is given by Eq 10. It's Bellman form is given by Eq. 11

$$V(x) = \underset{a}{\operatorname{argmax}} \left[\sum_{T} F(x, a) \right] \tag{10}$$

$$V(x) = \underset{a}{\operatorname{argmax}} \left[F(x_0, a_0) + V(x_1) \right] \tag{11}$$

In reinforcement learning the value function is based on the sum of future rewards giving Eq. 12-13,

$$V_R(x) = \underset{a}{\operatorname{argmax}} \left[\sum_{T} R_t \right] \tag{12}$$

$$V_R(x) = \underset{a}{\operatorname{argmax}} \left[R_0 + V(x_1) \right] \tag{13}$$

In our objective the best possible value is not found by temporal summation as it is during reinforcement learning. E is necessarily a temporally unstable function. We expect it to vary substantially before contracting to approach 0. It's best possible value is well approximated then by only looking at the maximum value for the last time step, making our optimization myopic, that is based on only last values encountered ***Hocker and Park*** (***2019***).

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

$$\begin{aligned} V_E(x) &= \underset{a}{\operatorname{argmax}} \left[E_0 + V(x_1) \right] \\ &= E_0 + \underset{a}{\operatorname{argmax}} \left[E_t \right] \end{aligned} \tag{15}$$

Good exploration by curiosity optimization

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

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

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

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

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

$$a \leq b \Leftrightarrow \exists\ c;\ b = a + c \tag{16}$$

$$a > b \Leftrightarrow (a \neq b) \wedge (b \leq a) \tag{17}$$

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

Proof. Let $\mathbf{E} = (E_1, E_2, ...)$ be ranked series of E values for all states X, such that $(E_1 \geq E_2, \geq ...)$. To swap any pair of values $(E_i \geq E_j)$ so $(E_i \leq E_j)$ by Eq. 16 $E_i - c = E_j$.

Therefore, again by Eq. 16, $\exists \int \delta E(s) \rightarrow -c$.

Recall: $\nabla^2 f_\Theta(x) < 0$ after a finite time T.

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

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

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

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

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

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

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

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

Optimality of π_π

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

$$\pi^\pi = \left[\pi_E,\ \pi_R\right] \tag{18}$$

$$\underset{\pi^\pi}{\operatorname{argmax}} \left[E_{t-1},\ R_{t-1}\right]_{(2,\,1)} \tag{19}$$

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 18 is certain to maximize total value. The total value of R and 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 **?*Roughgarden* (*2019*)**.

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

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

In proving this optimality of π_π we limit the probability of a positive reward to less than one, denoted by $p(R_t = 1) < 1$. Without this constraint the reward policy π_R would always dominate π_π when rewards are certain. While this might be useful in some circumstances, from the point of view π_E it is extremely suboptimal. The model would never explore. Limiting $p(R_t = 1) < 1$ is a reasonable constraint, as rewards in the real world are rarely certain. A more naturalistic way to handle this edge case is to introduce reward satiety, or a model physiological homeostasis ***Keramati and Gutkin*** (***2014***); ***Juechems and Summerfield*** (***2019***).

Optimal scheduling for dual value learning problems

In classic scheduling problems the value of any job is known ahead of time ***Bellmann*** (***1954***); ***Roughgarden*** (***2019***). 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. 18 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 ***Strogatz*** (***1994***). 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 ***Hocker and Park*** (***2019***). We therefore assume that last value is the best predictor of the next value, and use this assumption along with Theorem 5 to complete a trivial proof that Eq. 18 maximizes total value.

Optimal total value

If we prove π_π has optimal substructure, then using the same replacement argument ***Roughgarden*** (***2019***) as in Theorem 5, a greedy policy for π_π will maximize total value.

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

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

Recall: The memory Θ has optimal substructure (Theorem 1.

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

Recall: The controller π_π is deterministic and greedy.

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

Acknowledgments

We wish to thank Jack Burgess, Matt Clapp, Kyle "Donovank" Dunovan, Richard Gao, Roberta Klatzky, Jayanth Koushik, Alp Muyesser, Jonathan Rubin, and Rachel Storer for their comments on earlier drafts. We also wishes to thank Richard Grant for his illustration work in Figure 1, and Jack Burgess for his assitance with Figure 3.

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

References

Ahilan S, Solomon RB, Breton YA, Conover K, Niyogi RK, Shizgal P, Dayan P. Learning to Use Past Evidence in a Sophisticated World Model. PLoS Comput Biol. 2019 Jun; 15(6):e1007093. doi: 10.1371/journal.pcbi.1007093.

Bellmann R. The Theory of Dynamic Programming. Bull Amer Math Soc. 1954; 60(6):503–515.

Bench S, Lench H. On the Function of Boredom. Behavioral Sciences. 2013 Aug; 3(3):459–472. doi: 10.3390/bs3030459.

Berger-Tal O, Nathan J, Meron E, Saltz D. The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. PLoS ONE. 2014 Apr; 9(4):e95693. doi: 10.1371/journal.pone.0095693.

Berlyne D. Novelty and Curiosity as Determinants of Exploratory Behaviour. British Journal of Psychology. 1950; 41(1):68–80.

Calhoun AJ, Chalasani SH, Sharpee TO. Maximally Informative Foraging by Caenorhabditis Elegans. eLife. 2014 Dec; 3:e04220. doi: 10.7554/eLife.04220.

Cisek P. Resynthesizing Behavior through Phylogenetic Refinement. Atten Percept Psychophys. 2019 Jun; doi: 10.3758/s13414-019-01760-1.

Colas C, Huizinga J, Madhavan V, Clune J. Scaling MAP-Elites to Deep Neuroevolution. Proceedings of the 2020 Genetic and Evolutionary Computation Conference. 2020 Jun; p. 67–75. doi: 10.1145/3377930.3390217.

Dayan P, Sejnowski TJ. Exploration Bonuses and Dual Control. Mach Learn. 1996 Oct; 25(1):5–22. doi: 10.1007/BF00115298.

de Abril IM, Kanai R. Curiosity-Driven Reinforcement Learning with Homeostatic Regulation. In: *2018 International Joint Conference on Neural Networks (IJCNN)* Rio de Janeiro: IEEE; 2018. p. 1–6. doi: 10.1109/IJCNN.2018.8489075.

Findling C, Skvortsova V, Dromnelle R, Palminteri S, Wyart V. Computational Noise in Reward-Guided Learning Drives Behavioral Variability in Volatile Environments. Neuroscience; 2018.

Fister I, Iglesias A, Galvez A, Del Ser J, Osaba E, Fister I, Perc M, Slavinec M. Novelty Search for Global Optimization. Applied Mathematics and Computation. 2019 Apr; 347:865–881. doi: 10.1016/j.amc.2018.11.052.

Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, O'Doherty J, Pezzulo G. Active Inference and Learning. Neuroscience & Biobehavioral Reviews. 2016 Sep; 68:862–879. doi: 10.1016/j.neubiorev.2016.06.022.

Gershman SJ. Deconstructing the Human Algorithms for Exploration. Cognition. 2018 Apr; 173:34–42. doi: 10.1016/j.cognition.2017.12.014.

Gupta AA, Smith K, Shalley C. The Interplay between Exploration and Exploitation. The Academy of Management Journal. 2006; 49(4):693–706.

Ha D, Schmidhuber J. Recurrent World Models Facilitate Policy Evolution. NeurIPS. 2018; p. 13.

Hocker D, Park IM. Myopic Control of Neural Dynamics. PLOS Computational Biology. 2019; 15(3):24.

Ishii S, Yoshida W, Yoshimoto J. Control of Exploitation–Exploration Meta-Parameter in Reinforcement Learning. Neural Networks. 2002 Jun; 15(4-6):665–687. doi: 10.1016/S0893-6080(02)00056-4.

Itti L, Baldi P. Bayesian Surprise Attracts Human Attention. Vision Research. 2009 Jun; 49(10):1295–1306. doi: 10.1016/j.visres.2008.09.007.

Jaegle A, Mehrpour V, Rust N. Visual Novelty, Curiosity, and Intrinsic Reward in Machine Learning and the Brain. Arxiv. 2019; 1901.02478:13.

Juechems K, Summerfield C. Where Does Value Come From? PsyArXiv; 2019.

Kelly JL. A New Interpretation of Information Rate. the bell system technical journal. 1956; p. 10.

Keramati M, Gutkin B. Homeostatic Reinforcement Learning for Integrating Reward Collection and Physiological Stability. eLife. 2014 Dec; 3:e04811. doi: 10.7554/eLife.04811.

Kidd C, Hayden BY. The Psychology and Neuroscience of Curiosity. Neuron. 2015 Nov; 88(3):449–460. doi: 10.1016/j.neuron.2015.09.010.

Kobayashi K, Hsu M. Common Neural Code for Reward and Information Value. Proc Natl Acad Sci USA. 2019 Jun; 116(26):13061–13066. doi: 10.1073/pnas.1820145116.

Lee MD, Zhang S, Munro M, Steyvers M. Psychological Models of Human and Optimal Performance in Bandit Problems. Cognitive Systems Research. 2011 Jun; 12(2):164–174. doi: 10.1016/j.cogsys.2010.07.007.

Lehman J, Stanley KO. Novelty Search and the Problem with Objectives. In: Riolo R, Vladislavleva E, Moore JH, editors. *Genetic Programming Theory and Practice IX* New York, NY: Springer New York; 2011.p. 37–56. doi: 10.1007/978-1-4614-1770-5_3.

Lehman J, Stanley KO, Miikkulainen R. Effective Diversity Maintenance in Deceptive Domains. In: *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference - GECCO '13* Amsterdam, The Netherlands: ACM Press; 2013. p. 215. doi: 10.1145/2463372.2463393.

MacKay DJC. Information Theory, Inference, and Learning Algorithms. Second ed.; 2003.

Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, Braithwaite VA, Hausmann D, Fiedler K, Gonzalez C. Unpacking the Exploration–Exploitation Tradeoff: A Synthesis of Human and Animal Literatures. Decision. 2015 Jul; 2(3):191–215. doi: 10.1037/dec0000033.

Mouret JB, Clune J. Illuminating Search Spaces by Mapping Elites. arXiv:150404909 [cs, q-bio]. 2015 Apr; .

Münch D, Ezra-Nevo G, Francisco AP, Tastekin I, Ribeiro C. Nutrient Homeostasis — Translating Internal States to Behavior. Current Opinion in Neurobiology. 2020 Feb; 60:67–75. doi: 10.1016/j.conb.2019.10.004.

Ng A, Harada D, Russell S. Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping. In Proceedings of the Sixteenth International Conference on Machine Learning. 1999; p. 278–287.

Pathak D, Agrawal P, Efros AA, Darrell T. Curiosity-Driven Exploration by Self-Supervised Prediction. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* Honolulu, HI, USA: IEEE; 2017. p. 488–489. doi: 10.1109/CVPRW.2017.70.

Poucet B. Spatial Cognitive Maps in Animals: New Hypotheses on Their Structure and Neural Mechanisms. Psycholocial Review. 1993; 100(1):162–183.

Roughgarden T. Algorithms Illuminated (Part 3): Greedy Algorithms and Dynamic Programming, vol. 1; 2019.

Schmidhuber. 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. 1991; p. 222–227.

Schmidhuber J. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). IEEE Transactions on Autonomous Mental Development. 2010 Sep; 2(3):230–247. doi: 10.1109/TAMD.2010.2056368.

Schulz E, Bhui R, Love BC, Brier B, Todd MT, Gershman SJ. Exploration in the Wild. Animal Behavior and Cognition; 2018.

Shannon C. A Mathematical Theory of Communication. The Bell System Technical Journal. 1948; 27:379–423, 623–656,.

Şimşek Ö, Barto AG. An Intrinsic Reward Mechanism for Efficient Exploration. In: *Proceedings of the 23rd International Conference on Machine Learning - ICML '06* Pittsburgh, Pennsylvania: ACM Press; 2006. p. 833–840. doi: 10.1145/1143844.1143949.

Song M, Bnaya Z, Ma WJ. Sources of Suboptimality in a Minimalistic Explore–Exploit Task. Nature Human Behaviour. 2019 Apr; 3(4):361–368. doi: 10.1038/s41562-018-0526-x.

Stanley KO, Miikkulainen R. Evolving a Roving Eye for Go. In: Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Deb K, editors. *Genetic and Evolutionary Computation – GECCO 2004*, vol. 3103 Berlin, Heidelberg: Springer Berlin Heidelberg; 2004.p. 1226–1238. doi: 10.1007/978-3-540-24855-2_130.

Strogatz SH. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Studies in Nonlinearity, Reading, Mass: Addison-Wesley Pub; 1994.

Sumner ES, Li AX, Perfors A, Hayes BK, Navarro DJ, Sarnecka BW. The Exploration Advantage: Children's Instinct to Explore Allows Them to Find Information That Adults Miss. PsyArxiv. 2019; h437v:11.

Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Second edition ed. Adaptive Computation and Machine Learning Series, Cambridge, Massachusetts: The MIT Press; 2018.

Thrun S, Möller K. Active Exploration in Dynamic Environments. Advances in neural information processing systems. 1992; p. 531–538.

Thrun SB. Eficient Exploration In Reinforcement Learning. NIPS. 1992; p. 44.

Valiant L. A Theory of the Learnable. Communications of the ACM. 1984; 27(11):1134–1142.

Wang MZ, Hayden BY. Monkeys Are Curious about Counterfactual Outcomes. Cognition. 2019 Aug; 189:1–10. doi: 10.1016/j.cognition.2019.03.009.

White JK, Bromberg-Martin ES, Heilbronner SR, Zhang K, Pai J, Haber SN, Monosov IE. A Neural Network for Information Seeking. Neuroscience; 2019.

Wolpert DH, Macready WG. No Free Lunch Theorems for Optimization. IEEE Trans Evol Computat. 1997 Apr; 1(1):67–82. doi: 10.1109/4235.585893.

Woodgate JL, Makinson JC, Lim KS, Reynolds AM, Chittka L. Continuous Radar Tracking Illustrates the Development of Multi-Destination Routes of Bumblebees. Scientific Reports. 2017 Dec; 7(1). doi: 10.1038/s41598-017-17553-1.

Zhang M, Li H, Pan S, Liu T, Su S. One-Shot Neural Architecture Search via Novelty Driven Sampling. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization; 2020. p. 3188–3194. doi: 10.24963/ijcai.2020/441.