Proof

Making a game out of exploration-exploitation

Anonymous authors
Paper under double-blind review

Abstract

What is the best way for a agent to balance exploration with exploitation? In this paper we suggest an answer to this question that treats exploration and exploitation as independent players competing to maximize a joint objective. Through theory and simulations we show how a "game" played between two deterministic policies, one maximizing intrinsic curiosity and one maximizing extrinsic environmental rewards, yields a simple maximum value solution over both policies. The key assumption that allows for this is our conjecture that curiosity and reward seeking are equally valuable on evolutionary terms. We start by developing an axiomatic approach to defining information value that generalizes past approaches, while simplifying our ability to estimate such value in both artificial and biological memory systems. We then show how our deterministic solution performs at least as well as standard stochastic explore-exploit algorithms, but has the added benefit of being far more resilient to deceptive rewards (i.e., local minima), more efficient in high-dimensional action contexts, and robust to hyperparameter choices. Thus, the solution to our version of the gamified exploration-exploitation problem can be summarized by a simple heuristic: when the expected value of information is more than the expected value of rewards, be curious, otherwise seek rewards.

1 Introduction

One of the most fundamental questions in reinforcement learning is determining when to explore and when to exploit prior knowledge. As stated in Sutton & Barto (2018), "The agent has to exploit what it has already experienced in order to obtain reward, but it also has to explore in order to make better action selections in the future. The dilemma is that neither exploration nor exploitation can be pursued exclusively without failing at the task". In finding answers to this question it is often assumed that actions are taken based on a set of learned values (Sutton & Barto, 2018; Roughgarden, 2019), and exploitation is defined as choosing the most valuable action. Exploration gets defined in several different ways depending on the task, but a common theme is the idea that exploration is an essential if risky means to accomplish reward maximization (Dayan & Sejnowski, 1996; Bellemare et al., 2016; Schmidhuber, 1991; Wilson et al., 2021; Gershman, 2018; Berger-Tal et al., 2014; Asmuth et al., 2009; Wilson et al., 2014; Dam & Körding, 2009; Reddy et al., 2016; Dayan & Sejnowski, 1996; Berger-Tal et al., 2014). The risk inherent in exploration is that reward value may be lost in pursuing new options, as illustrated in Figure 1a, thus framing the problem as a dilemma or trade=off.

Here we wonder if the worry about the risk of lost rewards during exploration is necessarily justified. Or, to put it another way, should exploration in reinforcement learning consider reward maximization at all? We use theory and simulations to show how the character of the explore-exploit question changes if we treat environmental rewards and information value as equally important (Inglis et al., 2001). We show that if they are equally important, then agents can simply choose the larger expected value between them and know with certainty they have gotten the "best" return available for their immediate explore-exploit question. In other words, we put strategies for reward collection and curiosity in competition with one another, as illustrated in Figure 1b.

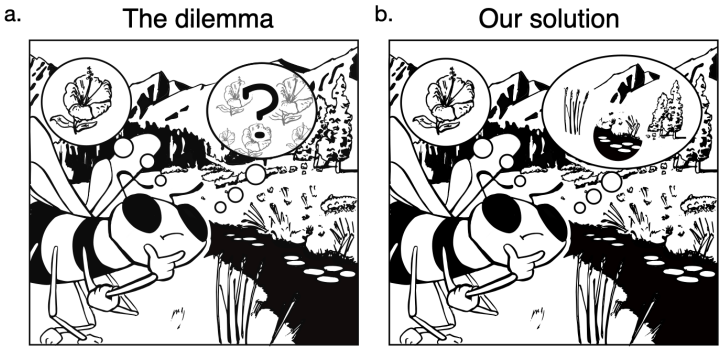

Figure 1: Two views of the exploration-exploration question. **a**. The classic trade-off: either exploit an action with a known reward by returning to the best flower or explore other flowers on the chance they will return a better outcome. The central challenge here is that the exploration of other flowers is innately uncertain in terms of the pollen collected, the extrinsic reward. **b**. Our alternative: an agent can have two goals: either exploit the pollen (reward) from the best flower (action), *or* explore to maximize pure learning with a curious search of the environment. To solve this second view, we set information and reward value on equal terms and propose a greedy rule to choose between them. *Artist credit*: Richard Grant.

We find it fruitful to view this competition as a maximum value "game", played out in an agent's "mind". The two policies compete for control of the agent's decisions, but share a common objective of maximizing the joint value of external rewards and intrinsic information value. To arrive at our game, and prove that it can be a general solution to explore-exploit questions, we need to show that information value, and by proxy curiosity, might be as important as collecting tangible rewards from the environment. We must also establish a general way to value information so as to derive a metric that is practical to measure directly from experimental observations of artificial memory systems, and perhaps biological ones.

One way to establish the usefullness of our game is to point out that animals in the natural world often strongly prefer to satisfy curiosity over receiving tangible rewards from their environment (food, water, etc). They do so even when that information is costly in terms of time, energy, or both (Song et al., 2019; Wang & Hayden, 2019; Taylor, 1975; Singh, 1970). A reader accustomed to thinking of animals as externally motivated by rewards might wonder why animals should show such curious preferences?

Prior work has shown that curiosity leads to the building of intuitive physics (Laversanne-Finot et al., 2018), and may be key to understanding causality (Sontakke et al., 2020). It can help ensure that an agent has generalizable world models (Kim et al., 2020; Wang & Hayden, 2020). It can drive imagination, play, and creativity (Schmidhuber, 2008; Auersperg, 2015; Haber et al., 2018; Friston et al., 2017). It can lead to the discovery of, and creation of, knowledge (Berlyne, 1954; Loewenstein, 1994; Zhou et al., 2020). It can ensure local minima, or other deceptive outcomes, are overcome when learning (Fister et al., 2019; Pathak et al., 2019) (an especially important point that we revisit in our experiments later on). Curiosity leads to the discovery of changes in the environment (Sumner et al., 2019), and ensures there are robust action policies to respond to these changes (Dubey & Griffiths, 2020). Curiosity can help in the recovery from injury (Cully et al., 2015). It can drive language learning (Colas et al., 2020). It is critical to evolution and the process of cognitive development (Oudeyer & Smith, 2016; Gopnik, 2020). It is the diverse benefits of curiosity that makes it so useful to agents, both biological and artificial. In fact, one could conjecture that curiosity is sufficient for all exploration in reinforcement learning (Groth et al., 2021; Fister et al., 2019; Inglis et al., 2001). Building on this, we further hypothesize the dual objectives of maximizing external rewards and information are on equal terms evolutionarily speaking: the latter aids survival in changing environments, while the former aids in more immediate ways by supplying food, water, and other biological requirements.

The diverse of uses of curiosity make it difficult, however, to describe curiosity in a general mathmatical way suitable for incorporation into canonical reinforcement learning. Should we therefore treat curiosity on a case-by-case or task-by-task basis? Like Spinka et al. (2001), who asked similar questions about animal play,

we believe that a case-by-case approach should be rejected for two reasons. First, despite the diversity of potential objectives, curious behavior appears to have a common root in the idea of "learning for learning's sake" (Berlyne, 1954; Loewenstein, 1994). Second, it is more parsimonious to assume that there is a unity to curiosity and set about discovering it. In our view, though, any unity must respect the diverse roles, objectives, and learning mechanisms that underlie curiosity in practice in the natural world. This is because we wish to develop a mathematical account equally well suited for biological and artificial agents.

1.1 Contributions.

The exploration-exploitation question has received substantial amounts of attention, but there has been little attention paid to the idea of exploration as being on fundamentally equal but independent terms to exploitation.

In this paper we develop:

- An axiomatic account of information value based on the recursive dynamics of memory. This metric is equally well suited for gradient-based artificial learning systems and biological nervous systems.
- We establish a Bellman optimality for this account of information value. This optimality is based on a novel idea of targeted forgetting and so removes the typical requirement of keeping the entire learning history intact in order to establish optimal substructure.
- We prove there exists a "meta-Bellman" rule to choose between pure (deterministic) policies for exploration and exploitation.
- We simplify this meta-policy to implement the win-stay lose-switch (WSLS) rule (Nowak & Sigmund, 1993; Bonawitz et al., 2014; Worthy & Todd Maddox, 2014).
- We study this WSLS rule in a number of computer simulations. We show it performs equivalently as state of the art methods for resolving explore-exploit in classic bandit environments, but also show that it is highly robust to deceptive rewards (i.e., local minima in the reward value function) and to hyperparameter selection. In other words, we show it performs as well as reward-motivated exploration and provides additional benefits for practical robustness.
- We also establish that the "white noise" problem found in artificial curiosity algorithms is naturally eliminated in our explore-exploit game without the need to constrain the objective of curiosity itself, which is a standard solution (Pathak et al., 2017; Kim et al., 2020).

1.2 Related work

We are not the first to propose a unity for curiosity and information value. For example, Friston et al. (2017) asserted that agents *must* minimize the entropy of their sensations, and so proposed to unify curiosity as a means to minimize uncertainty via active sampling and Bayesian reasoning. In other words, they prescribe an objective, representation, and mechanism for curiosity. This is just one of many prior approaches to curiosity and exploration (Berlyne, 1954; Loewenstein, 1994; Friston et al., 2017; Lydon-Staley et al., 2021; Oudeyer, 2018; Schmidhuber, 2019; Kim et al., 2020; Schmidhuber, 2008; Wilson et al., 2021). Our aim is not to debate the merits of any one of these. Instead we aim to develop a mathematical account that generalizes across *all* of them. Specifically, we develop a perspective of curiosity that does not require knowing the objective of curiosity itself (only the value of information), does not assume there is only one objective between rewards and information, does not assume a specific kind of representation, and does not prescribe mechanisms of learning beyond normative reinforcement learning. By not requiring these details we can achieve a strong generality, one that we hope is suited to artificial and biological agents alike.

Our approach to information value is built on ideas taken from the *learning progress* literature (Kaplan & Oudeyer, 2007; Lopes et al., 2012; Baranes & Oudeyer, 2013; Ten et al., 2021). The key point of difference between our approach and this prior work is that we do not include in our value metric the learning objective or loss function itself, as all the learning progress metrics that we are aware of do, either directly or indirectly

(Oudeyer, 2007; Graves et al., 2017; Oudeyer, 2018; Haber et al., 2018). Inclusion of the learning objective or loss function undercuts the desire to develop a general formalism for curiosity. As it turns out, removing these details may also be convenient in practice. We can measure information value directly, given only observations, for most any kind of memory system.

We will go on to formalize environmental reward collection and then contrast it to a formal definition of information value collection and curiosity. This might lead the reader to wonder if our information value is itself a reward or, to be consistent with past work should we just denote information value as an "intrinsic reward" and contrast it with "extrinsic rewards"? This is after all a fairly common distinction. So long as we are consistent in our definitions and use of terms our meaning should be clear. But we have refrained from these distinctions for two reasons. One, we mean something specific by information value and do not wish to conflate it with the more general intrinsic reward label. Two, the standard theoretical frameworks for reinforcement learning makes it seem as if extrinsic (i.e., from the environment) and intrinsic (i.e., from the agent) value are quite independent. However, the circuits for environmental reward valuation in real animals depend in part on satiety, reward history, and other preferences intrinsic to the animal. This suggests to us the split between intrinsic and extrinsic is less clear in practice than it is in theory.

As far as we are aware, we are the first to factor exploration-exploitation into a game between curiosity and reward. Many others have however separated out explore-exploit into independent parts. In one recent example of explore-exploit separation, (Colas et al.) suggested a fixed-length curious search ought to proceed with reinforcement learning in order to build up useful task-agnostic representations. Groth et al (Groth et al., 2021) suggested saving task-agnostic curiosity policies as a basis for learning task-specific reinforcement learning models. Zhang et al (Zhang et al., 2019) derived a randomized approach which alternated between exploration, using a specific representation, and exploitative reinforcement learning. In an effort to scale another learning algorithm, Clune et al experimented with a mixture of two policies for novelty-seeking and task performance (Colas et al., 2020). However the closest prior analog to our work seems to be found in an article from organizational psychology. In their work, Gupta et al (Gupta et al., 2006) argue it may be fruitful for business organizations to switch between periods of pure curious exploration for product ideas, and pure periods of exploitation of existing products.

2 Theory

2.1 Reward collection

We begin by restating the standard reinforcement learning problem of maximizing rewards (Sutton & Barto, 2018). To do this effectively, however, we need to clarify the following nomenclature. Here we will refer to to the intrinsic value of any external stimulus as simply "information value", though noting that it is just as capable of reinforcing learned behavior as any extrinsic reward might be. We will then use "reward" as a specific shorthand for value that is established and given by the external environment. We denote information value with a capital E, and reward with a capital R.

To develop our theory we first consider an agent that interacts with an environment over a sequence of states, actions and rewards. The agent's goal is to select actions in a way that maximizes the total reward, R, collected. How should an agent do this? The standard way of approaching exploration in reinforcement learning is to assume, "The agent must try a variety of actions and progressively favor those that appear to be the most rewarding" (Sutton & Barto, 2018).

To describe the environment we use a discrete time Markov decision process, $\mathcal{X}_t = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$. States are real valued vectors S from a finite set $\mathcal{S}$ of size n. Actions A are from the finite real set $\mathcal{A}$ of size k. We consider a set of policies that consist of a sequence of functions, $\pi = \{\pi_t, \pi_{t+1}, \ldots, \pi_{T-1}\}$. For any step t, actions are generated from states by policies, either deterministic $A = \pi(S)$ or random $A \sim \pi(S|A)$. In our presentation we drop the indexing notation on our policies, using simply π to refer to the sequence as a whole. Rewards are single valued real numbers, R, generated by a reward function $R \sim \mathcal{R}(S|A, t)$. Transitions from state S to new state $S\mathfrak{t}$ are caused by a stochastic transition function, $S' \sim \mathcal{T}(S|A, t)$. We leave open the possibility both $\mathcal{T}$ and $\mathcal{R}$ may be time-varying.

In general we use the $=$ to denote assignment, in contrast to $\sim$ which we use to denote taking random samples from some distribution. The standalone $|$ operator is used to denote conditional probability. An asterisk is used to denote optimal value policies, π^*.

So given a policy π_R, the value function in standard reinforcement learning is given by Eq. 1. This is the term we use to maximize reward collection.

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

But this equation is hard to use in practice because it requires that we integrate over all time, T. The Bellman equation (Eq. 2) is a desirable simplification because it reduces the entire action sequence in Eq. 1 into two (recursive) steps, t and $t+1$. This rule can then be recursively applied. The practical problem we are left with is finding a reliable estimate of $V_{\pi_R^*}(S_{t+1})$ (Eq. 2). This is a problem we return to further on.

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

Here we have focused on finite time horizons and we do not discount the rewards over time as is typically done in classic reinforcement learning. Both are done simply for simplicity. The basic approach should generalize to continuous spaces and discounted but infinite time horizons (Bertsekas, 2017)

2.2 Information collection

We next consider an agent that interacts with the environment over the same space as in reward collection, but whose goal is now to select actions in a way that maximizes information value (Schmidhuber, 1991; Oudeyer, 2018; Burda et al., 2018; Zhang & Yu, 2013; de Abril & Kanai, 2018; Zhou et al., 2020; Schwartenbeck et al., 2019; Wilson et al., 2014; Lehman & Stanley, 2011a; Velez & Clune, 2014). How should an agent do this? To find this answer we face two more questions: 1) how should information be valued?, 2) and, like for reward collection, is there a Bellman solution for information value?

If E represents information value, then the Bellman solution we want for E is given by Eq. 3.

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

We assume that maximizing E also maximizes curiosity. We refer to this greedy and deterministic approach to curiosity as E-exploration, or E-explore for short.

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

We define memory $\mathbf{M}$ as a vector of size p embedded in a real-valued space that is closed and bounded. This idea maps to a range of physical examples of memory, including firing rates of neurons in a population (Wang, 2021), strength differences between synapses (Stokes, 2015), or calcium dynamics (Higgins et al., 2014), as well as to the weight vectors common in deep reinforcement learning (Arulkumaran et al., 2017).

A learning function is then any function f that maps observations $\mathcal{X}$ into memory $\mathbf{M}$. This idea can be expressed using recursive notation, denoted by $\mathbf{M} \leftarrow f(\mathcal{X}, \mathbf{M})$. Though we will sometimes use $\mathbf{M}'$ to denote the updated memory instead. We also need to define a forgetting function, $f^{-1}(\mathcal{X},\ \mathbf{M}') \rightarrow \mathbf{M}$ which is important later on when establishing our novel Bellman result. As an auxiliary assumption we assume f has been selected by evolution to be formally learnable (Valiant, 1984).

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

such a function will be in operation "behind the scenes". In this paper we are not concerned with the objective y and the absolute error of the loss function, but only with the memory dynamics themselves and their equilibrium points. This approach lets us provide a *technical definition* for information value (Shannon, 1948).

2.2.1 Curiosity axioms

We next define a new metric of *learning progress* (Oudeyer, 2007) that is general, practical, and simple to measure for a memory system that aligns with our definition above. That is, we do not assume the learning algorithm, is known, singular, or fixed. We only assume that we have a memory with learning dynamics we can measure. We are, in other words, concerned only with how much was learned, not how it was learned. This is because we reason that the value of any piece of information depends entirely on the magnitude that the observation introduces a change in memory. We can formalize this idea with a formal set of axioms.

We start by thinking about how can we value the change in memory $\Delta\mathbf{M}$ that occurs from observing $\mathbf{X_t}$. We assume that there is such a value, which we have already denoted as E. There should be two basic properties of E up front.

Axiom 1 (Axiom of Memory)**.** *E is continuous with continuous changes in memory, $\delta\mathbf{M}$.*

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

In other words, information that is learned evenly in memory cannot induce any specific inductive bias, and is therefore safely regarded as the least valuable learning possible (see also(Mitchell, 1980)).

Axiom 3 (Axiom of Scholarship)**.** *Learning has an innately positive value. Therefore, $E \geq 0$.*

Thus, information value restricted to be positive by defined by by some absolute change in $\mathbf{M}$ (formalized more concretely below). We believe that this axiom holds in all cases even if for some particular learning instance information has negative consequences for the agent. That is, we choose regard learning as "good" even in the face of "bad" consequences.

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

This axiom means that the value of information is only relative to what is already known. This is largely consistent with Shannon's definition of information as well, which is based on the surprisal of an individual observation (Shannon, 1948): the more expected an observation becomes, the less value it will have since it fundamentally conveys less information.

Putting Axioms 1-4 together, we can now present a geometric perspective of information value. Figure 2a shows an example of E defined in a memory space where $\mathbf{M}_i$ and $\mathbf{M}_j$ reflect two entries, or dimensions, of this memory space. For example, consider a simple two unit system, say two reciprocally connected neurons labeled A and B (inset in Figure 2a). In this example, $\mathbf{M}_i$ reflects the weight of $A \rightarrow B$ and $\mathbf{M}_j$ would reflect the weight of $B \rightarrow A$. At time t the weights $\mathbf{M}_i$ and $\mathbf{M}_j$ have specific values (what those values are is not important for this example). Now, let's consider a potential observation, or experience, that the system can have at time $t + \tau$. Due to plasticity, we would see a change the weights between units A and B. The distance between the values of $\mathbf{M}_i$ and $\mathbf{M}_j$ from time t to time $t + \tau$ thus reflects the information value, E, of that learning experience. The larger the distance the larger the E, and thus more value to learning in $\mathbf{M}$.

We can next show examples of the dynamics of E over time. Changes consistent with our axioms are shown in Figure 2b. Even though these different trajectories have dramatically different patterns, their overall dynamics are consistent with changes in E by our definition. In contrast, Figure 2c shows a pattern of learning-related changes in E that would be inconsistent with our definition. These notions of consistency follow from fact we only ensure memory dynamics converge, and so they become self-consistent. However, we do and cannot ensure that errors are minimized and therefore learning and memory is "correct". It turns out that self-consistency is sufficient for a useful curious search, at least in out simulations (see below).

We should note that a corollary of Axiom 1 is that an observation that does not change memory has no value. So if $\sum_{i=1}^{p} |\Delta M_i| = 0$ then $E = 0$. It also follows from Axiom 3 and 4 that curiosity-driven exploration will

visit every state in $\mathcal{S}$ at least once in a finite time T, assuming $E_0 > 0$. In addition, as we stop exploration when $E < \eta$, only those states in which there is more to be learned will be revisited. This is a maximum entropy exploration policy Hazan et al. (2019). This, along with the deterministic nature of our policy, ensures perfectly efficient exploration in terms of learning progress.

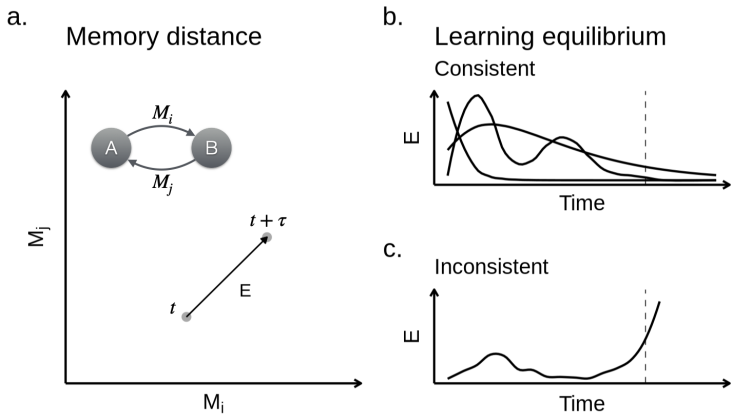

Figure 2: Information value. **a**. An example of information value in a two dimensional memory system. For example the directed weights from unit A to unit B and vice versa (inset model). The information value of an observation depends on the distance memory "moves" during learning, from timepoint t to $t+\tau$, denoted here by E. We depict memory distance as euclidean space, but this is one of many possible ways to realize E. **b** Examples of learning dynamics in time, made over a series of observations that are not shown. If information value becomes decelerating in finite time bound, then we say learning is consistent with Axioms 1-4. **c** If learning does not decelerate, then it is said to be inconsistent. The finite bound is represented here by a dotted line.

Our axioms for E are useful in three unique ways. First, they provide a measure of information value *without* needing to know the exact learning algorithm(s) being used. This leads to the second useful property of our axioms: if we want to measure any agent's representation of information value, we need only record differences in dynamics of their memory circuits. There is little need to decode, or interpret. Finally, and perhaps most importantly, these axioms properly generalize prior attempts to formalize curiosity. In other words, prior definitions of information value for curious exploration all fall under the definition we provide here.

Because of the breadth of our universal geometric defintion of information value, there are range of mathematical formulations that satisfy our axioms for the semi-metric $E(\Delta\mathbf{M})$ or more generally $E = d(\mathbf{M}, \mathbf{M}')$ for some measuring function d. Without loss of generality here we will make a *working definition* suitable for our multi-armed bandit simulations. Namely we use the Kullback-Leibler divergence (KL), a common and classic measure of information gain MacKay (2003) where we restrict $\mathbf{M}$ to be a memory for the probability of task events, $\mathbf{X}$.

$$E := d(\mathbf{M}, \mathbf{M}') = \mathrm{KL}(\mathbf{M}|\mathbf{M}') = \sum_{x \in \mathbf{X}} \mathbf{M}(x) \log \left(\frac{\mathbf{M}(x)}{\Delta\mathbf{M}'(x)} \right) \tag{4}$$

It is important to note however that the above definition is one of several options that can be freely chosen for E depending on the details of $\mathbf{M}$ and f. For example, the lp-norm of the difference between memories of vectors of real numbers, $\|\Delta\mathbf{M}\|_p$, satisfies our axioms. As does the norm of gradient, $\|\nabla\mathbf{M}\|_p$, a fact that is especially useful for studying curiosity and information value in differential programs and deep learning. Thus our definition of E allows it to take on many differnt realizations that are equally compatible.

2.2.2 The importance of boredom

Of course, curiosity has its limits. To avoid the white noise problem (Kim et al., 2020), other useless minutia (Pathak et al., 2017), and to stop exploration efficiently we rely on the threshold term, $\eta \geq 0$ (Ax. 4). We treat η as synonymous with boredom as defined elsewhere (Hill & Perkins, 1985; Schmidhuber, 1991; Bench & Lench, 2013). In other words, we hypothesize that boredom is an adaptive parameter tuned to fit the environment (Geana & Daw, 2016). Specifically, we use boredom to ignore arbitrarily small amounts of information value, by requiring exploration of some observation $\mathcal{X}$ to cease once $E \leq \eta$ for that $\mathcal{X}$.

2.3 The learning problem

Having defined both reward and information value, we now turn to how they are used during learning and decision making. Since maximizing both R and E are fundamentally a reinforcement learning problem, we first need to show that our definition of E satisfies the Bellman solution Bellmann (1954), which has already been used to prove the optimal learning policy for R.

2.3.1 A Bellman solution for information

A common way to arrive at the Bellman solution is to rely on a Markov decision space. This is a problem for our definition of memory as it has path dependence, which violates the Markov property assumption. To get around this, we prove that exact forgetting of the last observation is another way to find a Bellman solution.

To find a dynamic programming solution based on the Bellman equation (Eq 3) we must prove our memory $\mathbf{M}$ has optimal substructure. This is because the normal route, which assumes the problem rests in a Markov Space, is closed to us. By optimal substructure we mean that the process of learning in $\mathbf{M}$ can be partitioned into a collection, or series of memories, each of which is itself a solution. That is, it lets us find a kind of recursive structure in memory learning, which is what we need to apply the Bellman. Proving optimal substructure also allows for simple proof by induction (Roughgarden, 2019), a property we will take advantage of.

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

This can be validated as follows:

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 $\hat{\pi}_E$.*

Our proof requires 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 responses in stochastic environments. We will address this discrepancy later on using numerical simulations.

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

$$V_E(x) = E_0 + \arg\max_a \Big[E_t\Big] \tag{5}$$

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 6. It's Bellman form is given by Eq. 7

$$V(x) = \arg\max_a \Big[\sum_T F(x, a) \Big] \tag{6}$$

$$V(x) = \arg\max_a \Big[F(x_0, a_0) + V(x_1) \Big] \tag{7}$$

In reinforcement learning the value function is based on the sum of future rewards giving Eq. 8-9,

$$V_R(x) = \arg\max_a \Big[\sum_T R_t \Big] \tag{8}$$

$$V_R(x) = \arg\max_a \Big[R_0 + V(x_1) \Big] \tag{9}$$

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 & Park, 2019).

$$V_E(x) = \arg\max_a \ \Big[E_t \Big] \tag{10}$$

$$\begin{aligned} V_E(x) &= \arg\max_a \Big[E_0 + V(x_1) \Big] \\ &= E_0 + \arg\max_a \ \Big[E_t \Big] \end{aligned} \tag{11}$$

Having now proven the optimal substructure of $\mathbf{M}$, and assuming some arbitrary starting value $E_0 > 0$, it follows that the Bellman solution to information optimization is in fact given by Equation 3.

2.3.2 Reward and information collection

Finally, our union of independent curiosity with independent reward collection. We now model an agent who interacts with an environment who wishes to maximize both information *and* reward value, as shown in Eq 12. How can an agent do this? To answer this, let's make two conjectures:

Conjecture 1 (The hypothesis of equal importance)**.** *Reward and information collection are equally important in reinforcement learning.*

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

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

$$V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\left[\sum_{t=0}^{T} \max[E_t, R_t] \;\middle|\; S_t = S\right] \tag{12}$$

Having already established we have a Bellman optimal policy for E , and knowing reinforcement learning provides many solutions for R (Sutton & Barto, 2018; Bertsekas, 2017), we can write out the Bellman optimal solution for $V_{\pi_{ER}}$. This is,

$$\begin{aligned} V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\Big[&\max[E_t, R_t] + V_{\pi_{ER}}(S_{t+1}) \\ &\Big| \; S_t = S, A_t = A\Big]. \end{aligned} \tag{13}$$

From here we can substitute in the respective value functions for reward and information value. This gives us Eq.14, which translates to the Bellman optimal decision policy shown in Eq.15.

$$\begin{aligned} V_{\pi_{ER}}(S) = \max_{\pi_{ER}} \mathbb{E}\Big[&\max[V_{\pi_E}(S), V_{\pi_R}(S)] + V_{\pi_{ER}}(S_{t+1}) \\ &\Big| \; S_t = S, A_t = A\Big] \end{aligned} \tag{14}$$

$$\pi_{ER}(S) = \begin{cases} \pi_R(S) & \text{if } V_{\pi_R}(S) \geq V_{\pi_E}(S) \\ \pi_E(S) & \text{otherwise} \end{cases} \tag{15}$$

2.3.3 Simplifying with win-stay lose-switch

In our analysis of decision making under dynamic programming we have assumed that the value for the next state, $V(S_{t+1})$, is readily available with no uncertainty. In practice for most environments this is not the case. This presents a key distinction between the field of dynamic programming and reinforcement learning. If we no longer assume V_{π_E} and V_{π_R} are available, but must be estimated from the environment, then we are left with something of a paradox for information value.

In order to resolve this paradox, we wish to use value functions to decide between information and reward value but are simultaneously estimating those values. There are a range of methods to handle this (Sutton & Barto, 2018; Bertsekas, 2017), but we opted to further simplify π_{ER} in three ways. We first shift from using the full value functions to using only the last payout, E_{t-1} and R_{t-1}. Second, we remove all state-dependence leaving only time dependence. Third, we also include $\eta > 0$ to control the duration of exploration. These simplifications give Eq. 16, a win-stay lose-switch (WSLS) rule for exploration-exploitation.

$$\tilde{\pi}_{ER}(S) = \begin{cases} \pi_R(S) & \text{if } R_{t-1} \geq E_{t-1} - \eta \\ \pi_E(S) & \text{otherwise} \end{cases} \tag{16}$$

In Eq. 16, we think of π_R and π_E as two "players" in a game played for behavioral control (Estes, 1994). We feel the approach has several advantages. It is myopic and therefore can optimally handle nonlinear changes in either the reward function or environmental dynamics (Hocker & Park, 2019). In stationary settings, its regret is bounded (Robbins, 1952). It can approximate Bayesian inference (Bonawitz et al., 2014). It leads to cooperative behavior (Nowak & Sigmund, 1993). Further, WSLS has a long history in psychology, where it predicts and describes human and animal behavior (Worthy & Todd Maddox, 2014). But most of all, our version of WSLS is simple to implement and seems robust in practice (shown below).

3 Experiments

3.1 Tasks

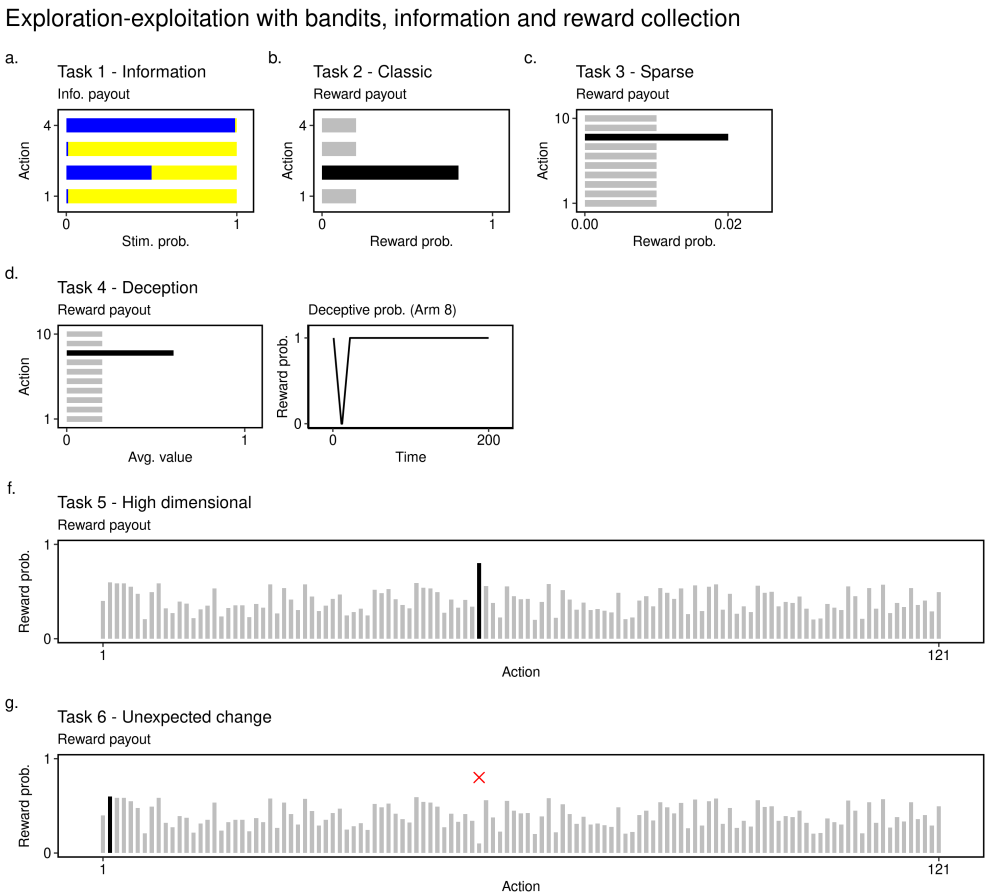

Figure 3: Multi-armed bandits – illustration and payouts. On each trial the agent must take on n actions. Each action generates a payout. Payouts can be information, reward, or both. For comments on general task design. **a.** A 4 choice bandit for information collection. In this task the payout is information, a yellow or blue "stimulus". A good agent should visit each arm, but quickly discover that only arm two is information bearing. **b.** A 4 choice design for reward collection. The agent is presented with four actions and it must discover which choice yields the highest average reward. In this task that is Choice 2. **c.** A 10 choice sparse reward task. Note the very low overall rate of rewards. Solving this task with consistency means consistent exploration. **d.** A 10 choice deceptive reward task. The agent is presented with 10 choices but the action which is the best in the long-term (>30 trials) has a lower value in the short term. This value first declines, then rises (see column 2). **e.** A 121 choice task with a complex payout structure. This task is thought to be at the limit of human performance. A good agent will eventually discover choice number 57 has the highest payout. **f.** 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.

We evaluated our WSLS exploration policy in seven variants of the bandit task popular in reinforcement learning Sutton & Barto (2018). The general structure of each task was similar. On each trial there were a set of n choices, and the goal of the agent to try and learn the best one, often determined by choosing the target that returns the most rewards, but not always (see below). Each choice action returns a "payout" according to a predetermined probability. Payouts are information, reward, or both (Figure 3).

Task 1 was designed to examine information foraging. There were no rewards. There were four choices. Three of these generated either a "yellow" or "blue" symbol, with a set probability. See Figure 3**a**.

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

Task 3 was designed with very sparse rewards (Silver et al., 2016; 2018). There were 10 choices. The best choice had a payout of $p(R = 1) = 0.02$. The other nine had, $p(R = 1) = 0.01$ for all others. See Figure 3c.

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

Tasks 5-6 were designed with 121 choices, and a complex payout structure. Tasks of this size are at the limit of human performance (Wu et al., 2018). We first trained all agents on *Task 6*, whose payout can be seen in Figure3f-g. This task, like the others, had a single best payout $p(R = 1) = 0.8$. After training for this was complete, final scores were recorded as reported, and the agents were then challenged by *Task 7*. *Task 7* was identical except that the best option was changed to be the worst $p(R = 0) = 0.2$ (Figure 3f-g).

3.2 Agents

We considered two kinds of specialized agents to play our set of tasks. Those suited to solve bandit tasks and those suited to solve foraging tasks.

Bandit: E-explore - This agent uses our algorithm defined above. It pursued either pure exploration based on E maximization or pure exploitation based on pure R maximization. All its actions are deterministic and greedy. Both E and R maximization was implemented as an actor-critic architecture (Sutton & Barto, 2018) where value updates were made in crtic according to Eq. 25. Actor action selection was governed by,

$$A_t = \arg\max A \in \mathbb{A} = Q(., A_t) \tag{17}$$

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

This agent has two parameters, the learning rate α (Eq. 25) and the boredom threshold η (Eq. 16). Its payout function was simply, $G = R_t$ (Eq. 25).

Bandit: Reward. An algorithm whose objective was to estimate the reward value of each action, and stochastically select the most valuable action in an effort to maximize total reward collected. It's payout function was simply, $G = R_t$ (Eq. 25). Similar to the E-explore agent, this agent used actor-critic, but its actions were sampled from a softmax / Boltzman distribution,

$$p(A_t) = \frac{e^{\gamma Q_R(.,A_t)}}{\sum_{A \in \mathbb{A}} e^{\gamma Q_R(.,A)}} \tag{18}$$

Where γ is the "temperature" parameter. Large γ generate more random actions. This agent has two parameters, the learning rate α (Eq. 25) and the temperature $\gamma > 0$ (Eq. 18).

Bandit: Reward+Info. This algorithm was based on the Reward algorithm defined previously but it's reward value was augmented or mixed with E. Specifically, the information gain formulation of E described below. Its payout is, $G_t = R_t + \beta E_t$ (Eq. 25). Critic values were updated using this G, and actions were selected according to Eq. 18. This agent has three parameters, the learning rate α (Eq. 25) the temperature γ (Eq. 18) and the exploitation weight $\beta > 0$. Larger values of β will tend to favor exploration.

Bandit: Reward+Novelty. This algorithm was based on the Reward algorithm defined previously but its reward value wass augmented or mixed with a novelty/exploration bonus (Eq. 19).

$$B_{A_t} = \begin{cases} B & \text{if } A_t \text{ not it } \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

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

This agent's payout is, $G_t = R_t + B_t$ (Eq. 25). Critic values were updated using this G and actions were selected according to Eq. 18. This agent has three parameters, the learning rate α (Eq. 25), the temperature γ (Eq. 18), and the bonus size $B > 0$.

Bandit: Reward+Entropy. This algorithm was based on the Reward algorithm defined previously but its reward value was augmented or mixed with an entropy bonus (Eq. 20). This bonus was inspired by the "softactor" method common to current agents in the related field of deep reinforcement learning (Haarnoja et al., 2018).

$$H(A_t) = \sum_{A \in \mathbb{A}} p(A) \log p(A) \tag{20}$$

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

This agent's payout is, $G_t = R_t + \beta H_t$ (Eq. 25). Critic values were updated using this G, and actions were selected according to Eq. 18. It has three free parameters, the learning rate α (Eq. 25), the temperature γ (Eq. 18), and the exploitation weight $\beta > 0$.

Bandit: Reward+EB. This algorithm was based on the Reward algorithm defined previously but it's reward value was augmented or mixed with an evidence bound (EB) statistic (Bellemare et al., 2016) (Eq 21).

$$\text{EB}(A) = \sqrt{C(A)} \tag{21}$$

Where $C(A)$ is a running count of each action, $A \in \mathbb{A}$. This agent's payout is, $G_t = R_t + \beta$ EB (Eq. 25). Critic values were updated using this G, and actions were selected according to Eq. 22.

$$A_t = \arg\max A \in \mathbb{A} = Q(., A_t) + \beta \text{ EB}(A) \tag{22}$$

It has three two parameters, the learning rate α (Eq. 25) and the exploitation weight $\beta > 0$.

Bandit: Reward+UCB. This algorithm was based on the Reward algorithm defined previously, but it's reward value was augmented or mixed with an upper confidence bound (UCB) statistic (Bellemare et al., 2016) (Eq 23).

$$\text{UCB}(A) = \frac{2 \log(N_A + 1)}{\sqrt{C(A)}} \tag{23}$$

Where $C(A)$ is a running count of each action, $A \in \mathbb{A}$ and N_A is the total count of all actions taken. This agent's payout is, $G_t = R_t + \beta$ UCB (Eq. 25). Critic values were updated using this G, and actions were selected according to Eq. 24.

$$A_t = \arg\max A \in \mathbb{A} = Q(., A_t) + \beta \text{ UCB}(A) \tag{24}$$

It has three two parameters, the learning rate α (Eq. 25) and the exploitation weight $\beta > 0$.

In the majority of simulations we have used a Bayesian or Information Gain (IG) formulation for curiosity and E. Each task's observations fit a simple discrete probabilistic model, with a memory "slot" for each

Table 1: Exploration strategies for bandit agents.

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

Table 2: Hyperparameter tuning - parameters and ranges.

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

action for each state. Specifically, probabilities were tabulated on state reward tuples, (S, R). To measure distances in this memory space we used the Kullback–Leibler divergence (Goodfellow et al., 2016; Itti & Baldi, 2009; López-Fidalgo et al., 2007; Schmidhuber, 2008; Ganguli & Sompolinsky, 2010; Ay, 2015).

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

$$V(S) = V(S) + \alpha[G_t - V(S_t)] \tag{25}$$

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

Value learning updates for all relevant agents in the foraging task were made using the TD(0) learning rule (Sutton & Barto, 2018),

$$V(S_t) = V(S_t) + \alpha[G_t - V(S_{t+1}) - V(S_t)] \tag{26}$$

We assume an agent will have a uniform prior over the possible actions $\mathbb{A}$ and set accordingly, $E_0 = \sum_K p(A_k) \log p(A_k)$.

The hyperparameters for each agent were tuned independently for each task by random search (Bergstra & Bengio, 2012). Generally, we reported results for the top 10 values, sampled from 1000 possibilities. Top 10 parameters and search ranges for all agents are shown in Supplemental Table 2. Parameter ranges were held fixed for each agent across all the bandit tasks.

3.3 Information collection simulations

The optimal policy to maximize information value is not a sampling policy. It is a deterministic greedy maximization (Eq 3). In other words, to minimize uncertainty during exploration an agent should not introduce additional uncertainty by taking random actions (Sehnke et al., 2010). To confirm our deterministic method is best we examined a multi-armed information task in Task 1 (Figure 3a). This variation of a bandit task replaced rewards with state information that must be intrinsically valued. In this case, state is simple colors represented internally as integers. On each selection, the agent saw one of two colors (integers) returned according to specific probabilities (Figure 3a).

Figure 4 shows the results of this head-to-head comparison of information-maximizing agents. None of the reward collection agents were run on this task because no rewards are provided, just feedback state, making optimization impossible for these agents. As expected, the deterministic information-maximizing agent was able to converge to its best estimate of the more informative target faster than the stochastic agent (Figure 4a-b), even though both agents used an otherwise identical curiosity-directed search. In addition, the deterministic agent was able to accrue a higher estimate of E than the stochastic agent (Figure 4c), with fewer exploratory steps (Figure 4d). This confirms the utility of our deterministic search policy.

3.4 Reward collection simulations

Now that we know that our agent can maximize information, we can look at how our *E-explore* agent compares with more traditional reward maximization agents in typical bandit tasks. Our overall goal is to answer the question: can curious search solve reward collection problems better than other approaches which use reward value? To find out we measured the total reward collected in several bandit tasks (Figure 3b-g). We considered eight agents, including ours (Table 1), described in Section 3.2.

Figure 5 presents the overall performance in some standard bandit tasks: simple (Figure 5b), sparse (Figure 5c), high dimensional (Figure 5e) and non-stationary (Figure 5f). Overall, our approach matches or sometimes exceeds all the other exploration strategies (Figure 5a)). Despite our E-exploration agent never optimizing for reward value, our method of pure exploration (Bubeck et al., 2010) matched, or in some cases outperformed standard approaches that rely on reward value. Indeed, with the exception of Task 4,

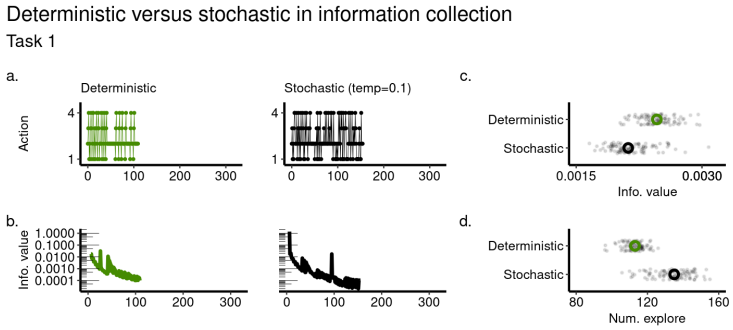

Figure 4: Comparing deterministic versus stochastic variations of the same curiosity algorithm in Task 1. Deterministic results are shown in the left column and stochastic results are shown in the right. **a**. Examples of target selection, across trials, for deterministic (green) and stochastic (black) information-seeking agents. The block ended when E converged to a minimum value. **b**. Information value, E, plotted across time. Matches the behavior shown in a. **c**. Average information value for 100 independent simulations. er of steps it took to reach the stopping criterion in c., e.g. the boredom threshold *eta* (described below). Smaller numbers of steps imply a faster search. **d**. Number of exploration steps until convergence for each agent.

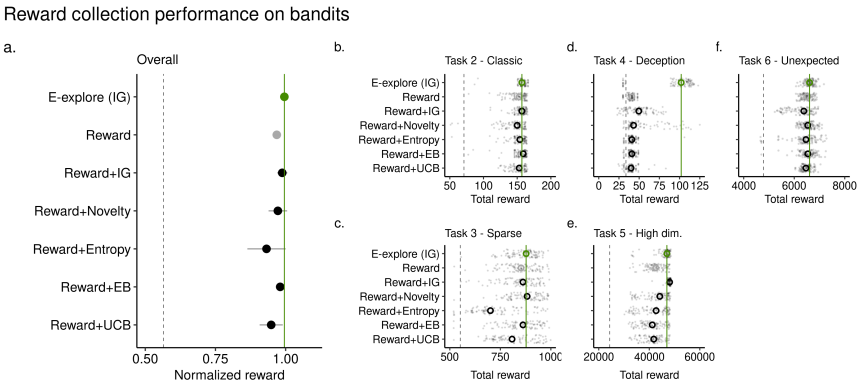

Figure 5: Reward collection performance. **b.** Overall results. Total reward collected for each of the four tasks was normalized. Dot represents the median value, error bars represent the median absolute deviation between tasks (MAD). **b.** Results for Task 2, which has four choices and one clear best choice. **c.** Results for Task 3, which has 10 choices and very sparse positive returns. **d.** Results for Task 4, whose best choice is initially "deceptive" in that it returns suboptimal reward value over the first 20 trials. **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 the worst. Agents were pre-trained on Task 6, then tested on 7. In panels b-e, grey dots represent total reward collected for individual experiments while the large circles represent the experimental median.

performance improvements, when present, were small but meaningful. This modest improvement in most tasks might not look noteworthy. We argue it is because our exploration strategy never optimizes for reward value explicitly. Yet, we meet or exceed exploration strategies which do. In other words, when intrinsic curiosity is used in an optimal trade-off with extrinsic reward, curiosity does seem sufficient in practice.

However, our information maximizing strategy was highly effective in one particular context that is typically challenging for reward maximizing agents: deceptive feedback. Task 4, the deception task (Figure 3d), involved an initial, misleading, 20 step decline in reward value for the optimal target. Comparing the E-exploration agent against the reward maximizing agents (Figure 5d), we see that the other exploration strategies produced little better than chance performance. When deception is present, externally motivated

exploration is a liability. Yet E-explore was highly effective in this context because the change in feedback structure over time creates a boost in E, pushing the agent to seek information on the changing target and becoming robust to this early deception paradigm.

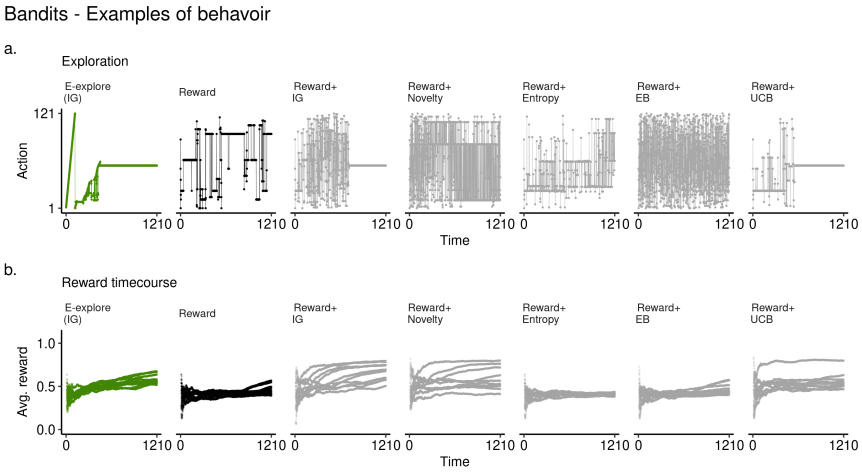

Figure 6: Behavior on a complex high-dimensional action space (Task 5). **a**. Examples of strategies for all agents (1 example). **b**. Average reward time courses for all agents (10 examples).

Finally, another place where we see a clear boost in performance with E-explore is in high-dimensional action spaces (Task 5). This task, modeled against the upper limit of human action selection (Wu et al., 2018), forced agents to explore across 121 potential discrete actions to find the optimal choice (Figure 3f). While we see that our E-explore agent performed as well as a standard reward maximization agent with a novelty boost (Figure 5e), our agent that maximizes E does so with greater efficiency (Figure 6a), converging on the optimal target faster than any other agent, and settling on a stable reward return by the end of each simulation run (Figure 6b). Of course, once the optimal target in this high-dimensional space becomes the *least* optimal target (Task 7), the E-exploration agent converges on the new optimal target, with a slightly higher reward return than the other agents (Figure 5f).

Of course, one could argue that the performance of our E-explore agent reflects the specific hyperparameter choices used to initialize the models. Thus we tested the robustness of all the agents to environmental mistunings (Figure 7). We re-examined total rewards collected across all model tunings or hyperparameters. Here the E-explore agent was markedly better than the other agents, with both substantially higher total reward when we integrate over all parameters (Figure 7a) and when we consider the tasks individually (Figure 7b-f).

4 Discussion

Here we presented a new way to think about solutions to the classic exploration-explioitation problem. Using theory and simulations we argue that reinforcement learning practitioners can safely set aside any intuitions that suggest a purely curious search is too inefficient to be generally practical when the goal is reward collection. To begin with, we derived a measure of information value that is possible to directly measure in most any memory system, especially with gradient based learning. This measure also properly generalizes many prior efforts to formalize curiosity in computer science, biology, and psychology. Using this measure of information value, E, we then showed how it can be maximized using standard reinforcement learning algorithms. This dynamic competition between pure curiosity and pure reward collection performs well on a variety of explore-exploit problems. Curiosity-driven exploration, tempered by boredom, matches or exceeds reward-based strategies when rewards are dense, sparse, high-dimensional, or non-stationary. In addition, curiosity driven exploration uniquely overcomes deceptive feedback schedules that can trip

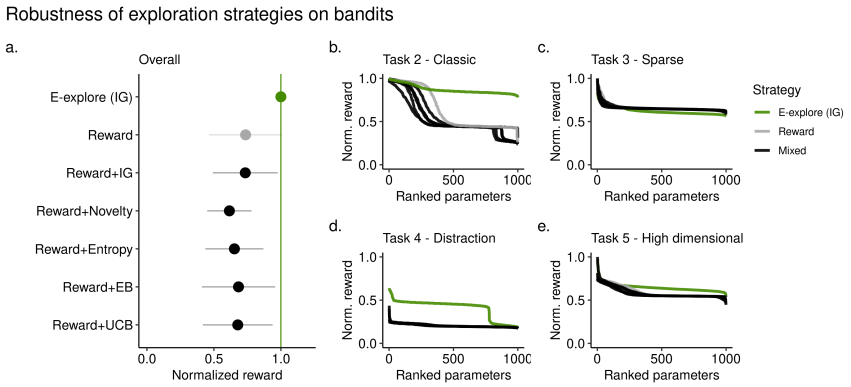

Figure 7: Exploration hyperparameter sensitivity. **a**. Integrated total reward (normalized) across 1000 randomly chosen exploration hyperparameters. Dots represent the median. Error bars are the median absolute deviation. **b-f**. Normalized total reward for exploration 1000 randomly chosen parameters, ranked according to performance. Each subpanel shows performance on a different task. Lines are colored according to overall exploration strategy - E-explore, reward only, or a mixed value approach blending reward and an exploration bonus).

up reward maximizing agents. Our game between curiosity and rewards also seems far more robust than standard algorithms to hyperparameter choices.

While the theoretical and empirical results presented here make a strong case for thinking of maximizing information versus maximizing reward as a scheduling problem, depending on relative immediate value, a lot of questions remain unanswered. Here we address some of the more immediate questions about our approach.

4.1 Questions and answers

Why curiosity? Not all situations involve immediately maximizing tangible rewards. Many natural (and artificial) behaviors involve exploring just to know. Curiosity as an algorithm is highly effective at solving difficult optimization problems (Schmidhuber, 1991; Pathak et al., 2017; Stanton & Clune, 2018; Fister et al., 2019; Mouret & Clune, 2015; Colas et al., 2020; Cully et al., 2015; Pathak et al., 2017; Schwartenbeck et al., 2019; Laversanne-Finot et al., 2018).

Is this explore-exploit game a slight of hand theoretically? Yes. We have taken one problem, the explore-exploit trade-off, that is formally considered intractable and replaced it with another problem that can be solved. In this replacement we swap one behavior, exploration with reward seeking in mind, for another, curiosity (i.e., exploration to gather information).

Is this too complex? Perhaps turning a single objective into two, as we have, is too complex an answer for reinforcement learning. If this is true, then it would mean our strategy is not parsimonious. Should we reject it on that alone?

Questions about parsimony can be sometimes resolved by considering the benefits versus the costs of added complexity. The benefits are an optimal value solution to the exploration-exploitation trade-off. A solution that seems especially robust to model-environment mismatch. At the same time, curiosity-as-exploration can build a model of the environment (Ha & Schmidhuber, 2018), useful for later planning (Ahilan et al., 2019; Poucet, 1993), creativity, imagination (Schmidhuber, 2010), while also building diverse action strategies (Lehman & Stanley, 2011b; Lehman et al., 2013; Mouret & Clune, 2015; Colas et al., 2020).

Is this too simple? Our game between curiosity and reward seems to be simple to solve. This is suspicious given how much work has been done on explore-exploit problems in the past. So have we "cheated" by changing the problem in the way we have?

The truth is that we might have cheated in one sense: the trade-off might really have to be as hard as it has seemed in the past. But the general case for curiosity is clear and backed up by the brute fact of its widespread presence in the animal kingdom Byrne (2013). The real questions are: 1) is curiosity so useful and so robust that it is sufficient for all exploration with learning (Fister et al., 2019), and 2) is curiosity really equally important as tangible rewards?

The answer to these questions is, we believe, empirical. If our account does well in describing and predicting animal behavior, that would be some evidence for it (see (Sumner et al., 2019; Wang & Hayden, 2019; Jaegle et al., 2019; Gottlieb & Oudeyer, 2018; Kidd & Hayden, 2015; Berlyne, 1950; Colas et al., 2020; Rahnev & Denison, 2018; Wilson et al., 2020; Berger-Tal et al., 2014)). If it predicts neural structures related to exploratory behaviors (see (Cisek, 2019; Kobayashi & Hsu, 2019)), then that would be other evidence for our theory. If our theory proves useful in machine learning applications (see (Burda et al., 2018; Schmidhuber, 1991; de Abril & Kanai, 2018; Fister et al., 2019; Lehman & Stanley, 2011b; Stanley & Miikkulainen, 2004; Colas et al., 2020; Cully et al., 2015; Wilson et al., 2020; Pathak et al., 2019)) then that would be be even more evidence for it. In other words how simple, complex, or parsimonious a theoretical idea is comes down to its usefulness. That is for follow up work to decide.

What about truth? In other prototypical approaches, information value comes from prediction errors or is otherwise measured by how well learning corresponds to the environment (Behrens et al., 2007; Kolchinsky & Wolpert, 2018; Tishby et al., 2000) or how useful the information might be in the future (Dubey & Griffiths, 2020). Colloquially, one might call this "truth seeking". As a people we pursue information that is fictional (**?**), or based on analogy (**?**), or outright wrong (**?**). It is our view that conspiracy theories, disinformation, and more mundane errors, are far too commonplace for information value to rest on mutual information or error alone. This does not mean that holding false beliefs cannot harm survival, but this harm might be a second order question as far as information value goes.

What about information theory? The short answer is that the problems of communicating information and the value of that information, are wholly different issues that require different theories. Shannon and Weaver (Shannon & Weaver, 1964) in their classic introduction to the topic, describe information as a communication problem with three levels. **A.** The technical problem of transmitting accurately. **B.** The semantic problem of meaning. **C.** The effectiveness problem of using information to change behavior. They then go on to describe how communication theory addresses only problem **A.**, which has lead to information theory finding a broad application over the last 75 years.

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

We have summarized this diagrammatically, in Figure 8.

Finally, we argue there is an analogous set of levels for information value as those that Shannon and Weaver describe for information. There is the technical problem of judging how much was learned. This is the one we address. There is the semantic problem of what this learning "means" (see Dretske (1981)) and also what its consequences are. There is also the effectiveness problem of using what was learned to some other effect. These are applications of the theoretical approach that we propose here.

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

So you suppose there is always a positive value for learning of all fictions, disinformation, and deceptions? According to our framework, we must. This is a most unexpected prediction. Yet it seems consistent with the behavior of humans, and animals alike. Humans do consistently seek to learn falsehoods.

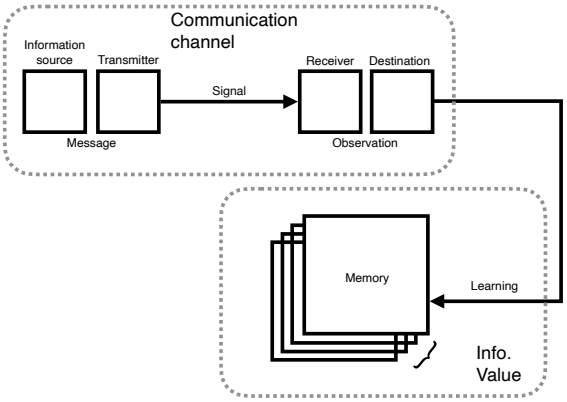

Figure 8: The relationship between the technical problem of communication with the technical problem of value. Note how value is not derived from the channel directly. Value comes from learning about observations taken from the channel, which in turn depends on memory.

Our axioms suggest why this is rational: it is good learning progress from the perspective of changes in memory.

Was it necessary to build a general theory for information value to describe curiosity? We made an idealistic choice that worked out. The field of curiosity studies has shown that there are many kinds of curiosity (Kidd & Hayden, 2015; Gottlieb & Oudeyer, 2018; Berlyne, 1950; Lehman & Stanley, 2011b; Pathak et al., 2017; Schmidhuber, 1991; 2008; Schwartenbeck et al., 2019; de Abril & Kanai, 2018; Stanton & Clune, 2018; Zhou et al., 2020; Loewenstein, 1994; Kashdan et al., 2019; Keller et al., 1994; Wang & Hayden, 2020). At the extreme limit of this diversity is a notion of curiosity defined for any kind of observation, and any kind of learning. This is what we offer. At this limit we can be sure to span fields, addressing the problem of information value as it appears in computer science, machine learning, game theory, psychology, neuroscience, biology, economics, among others.

What about other models of curiosity? Curiosity has found many specific definitions over the years (Berlyne, 1950; Oudeyer, 2018; Gottlieb & Oudeyer, 2018). Curiosity has been described as a prediction error that drives learning progress (Schmidhuber, 1991; Kaplan & Oudeyer, 2007). Schmidhuber (Schmidhuber, 1991) noted the advantage of looking at the derivative of errors, rather than errors directly, which falls within our definition of information value. Itti (Itti & Baldi, 2009) and others (Schmidhuber, 2008; Friston et al., 2017; Reddy et al., 2016; Calhoun et al., 2015) have taken a statistical and Bayesian view often using the KL divergence to estimate information gain or Bayesian surprise. This also falls under our general definition of information value. Other approaches have been based on adversarial learning, model disagreement (Pathak et al., 2017), or model compression (Schmidhuber, 2008). Some measures focused on past experience as a driver to seek information (Kim et al., 2020; Schmidhuber, 1991; Ha & Schmidhuber, 2018) while others focused on future planning and imagination (Savinov et al., 2019; Colas et al., 2020; Mendonca et al., 2021). These are, again, consistent within our general defition of information value. Finally, graphical models of memory have been useful in explaining how curiosity drives understanding (Colas et al., 2020), which are similar in farming to our geometric model.

What distinguishes our approach to defining information value is that we focus on memory dynamics, and base value on some general axioms that braodly cover these prior definitions. We try to embody the idea of curiosity "as learning for learning's sake", not done for the sake of some particular goal (Gabaix et al., 2006), or for future utility (Dubey & Griffiths, 2020). Our axioms were however designed with all these other definitions in mind. In other words, we do not aim to add a new metric. We aim to generalize others.

Is information a reward? If reward is any quantity that motivates behavior, then our definition of information value is a reward, an intrinsic reward. This last point does not mean that information value and environmental rewards are interchangeable however. 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 the Pythagorean theorem with a classmate, that student does not have less knowledge of the Pythagorean theorem as a result.

But isn't curiosity impractical for many problems?

Let's consider the answer to this by discussing colloquially how science and engineering often interact. Science is seen as an open-ended inquiry, whose goal is knowledge but whose practice is driven by learning progress. This is a fairly impractical aim in terms of finding solutions to problems. Engineering, on the other hand, is often seen as finding specific solutions to specific problems. This is both fairly practical, but also limited by the existing base of scientific knowledge. They each have their own pursuits, in other words, but they also learn from each other often in alternating iterations. Their different objectives is what makes them such good long-term collaborators.

In a related view Gupta et al (Gupta et al., 2006) encouraged managers in business organizations to strive for a balance in the exploitation of existing markets and ideas with exploration. They suggested managers should pursue periods of "punctuated equilibrium", where employees work towards either pure market exploitation or pure curiosity-driven exploration to drive future innovation.

Is the algorithmic run time for curiosity practical? According to our model, there is a lower limit in algorithmic run time for curiosity. The worst case algorithmic run time of our meta-policy method is linear and additive in its independent policies. If it takes T_E steps for π_E to converge, and T_R steps for π_R, then the worst case run time for π_{ER} is $T_E + T_R$. This puts our model at the same run time efficiency as standard reinforcement learning models.

Does this mean you are hypothesizing that boredom is actively tuned? Yes we are predicting that. In fact, Geana and Daw (Geana & Daw, 2016) showed preciselyl this in a series of experiments. They reported that altering the expectations of future learning in turn alters self-reports of boredom. Others have shown how boredom and self-control interact to drive exploration (Hill & Perkins, 1985; Bench & Lench, 2013; Wolff & Martarelli, 2020). So boredom does seem to be a pliable, and likely actively tuned, parameter.

Do you have any evidence of your theory in animal behavior or 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 (Berlyne, 1950; Kidd & Hayden, 2015; Sutton & Barto, 2018). Indeed, we have highlighted how they are separate problems, with links to different basic needs: gathering resources to maintain physiological homeostasis (Keramati & Gutkin, 2014; Juechems & Summerfield, 2019) and gathering information to decide what to learn and to plan for the future (Valiant, 1984; Sutton & Barto, 2018). Here we suggest that, though they are separate problems, they are problems that can generally 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 (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 had no input from the homeostatic circuits (Keramati & Gutkin, 2014; Cisek, 2019; Juechems & Summerfield, 2019). This neural separation for independent circuits has been observed in some animals, including zebrafish (Marques et al., 2019) and monkeys (White et al., 2019; Wang & Hayden, 2019).

Acknowledgments

References

Sanjeevan Ahilan, Rebecca B. Solomon, Yannick-André Breton, Kent Conover, Ritwik K. Niyogi, Peter Shizgal, and Peter Dayan. Learning to use past evidence in a sophisticated world model. *PLOS Computational Biology*, 15(6):e1007093, June 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007093.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

John Asmuth, Lihong Li, Michael L Littman, Ali Nouri, and David Wingate. A Bayesian Sampling Approach to Exploration in Reinforcement Learning. pp. 8, 2009.

Alice M.I. Auersperg. Exploration Technique and Technical Innovations in Corvids and Parrots. In *Animal Creativity and Innovation*, pp. 45–72. Elsevier, 2015. ISBN 978-0-12-800648-1. doi: 10.1016/B978-0-12-800648-1.00003-6.

Nihat Ay. Information Geometry on Complexity and Stochastic Interaction. *Entropy*, 17(4):2432–2458, April 2015. ISSN 1099-4300. doi: 10.3390/e17042432.

Adrien Baranes and Pierre-Yves Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, January 2013. ISSN 09218890. doi: 10.1016/j.robot.2012.05.008.

Timothy E J Behrens, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, September 2007. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn1954.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying Count-Based Exploration and Intrinsic Motivation. *arXiv:1606.01868 [cs, stat]*, June 2016.

Richard Bellmann. The theory of dynamic programming. *Bull. Amer. Math. Soc*, 60(6):503–515, 1954.

Shane Bench and Heather Lench. On the Function of Boredom. *Behavioral Sciences*, 3(3):459–472, August 2013. ISSN 2076-328X. doi: 10.3390/bs3030459.

Oded Berger-Tal, Jonathan Nathan, Ehud Meron, and David Saltz. The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. *PLoS ONE*, 9(4):e95693, April 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0095693.

James Bergstra and Yoshua Bengio. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

D. E. Berlyne. A theory of human curiosity. *British Journal of Psychology. General Section*, 45(3):180–191, 1954.

DE Berlyne. Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology*, 41(1):68–80, 1950.

Dimitri Bertsekas. *Dynamic Programming and Optimal Control, Vol. I.* Athena Scientific, fourth edition, 2017.

Elizabeth Bonawitz, Stephanie Denison, Alison Gopnik, and Thomas L. Griffiths. Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, 74:35–65, November 2014. ISSN 00100285. doi: 10.1016/j.cogpsych.2014.06.003.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure Exploration for Multi-Armed Bandit Problems. *arXiv:0802.2655 [cs, math, stat]*, June 2010.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355 [cs, stat]*, August 2018.

Richard W Byrne. Animal curiosity. *Current Biology*, 23(11):R469–R470, 2013.

Adam J. Calhoun, Ada Tong, Navin Pokala, James A.J. Fitzpatrick, Tatyana O. Sharpee, and Sreekanth H. Chalasani. Neural Mechanisms for Evaluating Environmental Variability in Caenorhabditis elegans. *Neuron*, 86(2):428–441, April 2015. ISSN 08966273. doi: 10.1016/j.neuron.2015.03.026.

Paul Cisek. Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, June 2019. ISSN 1943-3921, 1943-393X. doi: 10.3758/s13414-019-01760-1.

Cedric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms. pp. 10.

Cédric Colas, Joost Huizinga, Vashisht Madhavan, and Jeff Clune. Scaling MAP-Elites to Deep Neuroevolution. *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pp. 67–75, June 2020. doi: 10.1145/3377930.3390217.

Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14422.

Gregory Dam and Konrad Körding. Exploration and Exploitation During Sequential Search. *Cognitive Science*, 33(3):530–541, May 2009. ISSN 03640213, 15516709. doi: 10.1111/j.1551-6709.2009.01021.x.

Peter Dayan and Terrence J. Sejnowski. Exploration bonuses and dual control. *Machine Learning*, 25(1): 5–22, October 1996. ISSN 0885-6125, 1573-0565. doi: 10.1007/BF00115298.

Ildefons Magrans de Abril and Ryota Kanai. Curiosity-Driven Reinforcement Learning with Homeostatic Regulation. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, Rio de Janeiro, July 2018. IEEE. ISBN 978-1-5090-6014-6. doi: 10.1109/IJCNN.2018.8489075.

Fred I Dretske. Knowledge and the flow of information. 1981.

Rachit Dubey and Thomas L. Griffiths. A rational analysis of curiosity. *arXiv:1705.04351 [cs]*, August 2020.

W. Estes. Toward a statistical theory of learning. *Psychological Review*, 101:94–107, 1994.

Iztok Fister, Andres Iglesias, Akemi Galvez, Javier Del Ser, Eneko Osaba, Iztok Fister, Matjaž Perc, and Mitja Slavinec. Novelty search for global optimization. *Applied Mathematics and Computation*, 347: 865–881, April 2019. ISSN 00963003. doi: 10.1016/j.amc.2018.11.052.

Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference, Curiosity and Insight. *Neural Computation*, 29(10):2633–2683, October 2017. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_00999.

Xavier Gabaix, David Laibson, Guillermo Moloche, and Stephen Weinberg. Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *THE AMERICAN ECONOMIC REVIEW*, 96 (4):26, 2006.

Surya Ganguli and Haim Sompolinsky. Short-term memory in neuronal networks through dynamical compressed sensing. pp. 9, 2010.

Andra Geana and Nathaniel Daw. Boredom, Information-Seeking and Exploration. *CogSci*, pp. 6, 2016.

Samuel J. Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, April 2018. ISSN 00100277. doi: 10.1016/j.cognition.2017.12.014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Alison Gopnik. Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1803):20190502, July 2020. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2019.0502.

Jacqueline Gottlieb and Pierre-Yves Oudeyer. Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 19(12):758–770, December 2018. ISSN 1471-003X, 1471-0048. doi: 10.1038/s41583-018-0078-0.

Alex Graves, Marc G. Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated Curriculum Learning for Neural Networks. *arXiv:1704.03003 [cs]*, April 2017.

Oliver Groth, Markus Wulfmeier, Giulia Vezzani, Vibhavari Dasagi, Tim Hertweck, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Is Curiosity All You Need? on the Utility of Emergent Behaviours from Curious Exploration. *arXiv:2109.08603 [cs]*, September 2021.

Anil A. Gupta, Ken Smith, and Christina Shalley. The Interplay between Exploration and Exploitation. *The Academy of Management Journal*, 49(4):693–706, 2006.

David Ha and Jürgen Schmidhuber. World Models. *arXiv:1803.10122 [cs, stat]*, March 2018. doi: 10.5281/zenodo.1207631.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic Algorithms and Applications. *arXiv:1812.05905 [cs, stat]*, December 2018.

Nick Haber, Damian Mrowca, Li Fei-Fei, and Daniel L. K. Yamins. Learning to Play with Intrinsically-Motivated Self-Aware Agents. *arXiv:1802.07442 [cs, stat]*, October 2018.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019.

David Higgins, Michael Graupner, and Nicolas Brunel. Memory Maintenance in Synapses with Calcium-Based Plasticity in the Presence of Background Activity. *PLoS Computational Biology*, 10(10):e1003834, October 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003834.

A. B. Hill and R. E. Perkins. Towards a model of boredom. *British Journal of Psychology*, 76(2):235–240, May 1985. ISSN 00071269. doi: 10.1111/j.2044-8295.1985.tb01947.x.

David Hocker and Il Memming Park. Myopic control of neural dynamics. *PLOS Computational Biology*, 15 (3):24, 2019.

Ian R. Inglis, Steve Langton, Björn Forkman, and John Lazarus. An information primacy model of exploratory and foraging behaviour. *Animal Behaviour*, 62(3):543–557, September 2001. ISSN 00033472. doi: 10.1006/anbe.2001.1780.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, June 2009. ISSN 00426989. doi: 10.1016/j.visres.2008.09.007.

Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *Arxiv*, 1901.02478:13, 2019.

Keno Juechems and Christopher Summerfield. Where does value come from? Preprint, PsyArXiv, April 2019.

Frédéric Kaplan and Pierre-Yves Oudeyer. The progress drive hypothesis: An interpretation of early imitation. In Chrystopher L. Nehaniv and Kerstin Dautenhahn (eds.), *Imitation and Social Learning in Robots, Humans and Animals*, pp. 361–378. Cambridge University Press, Cambridge, 2007. ISBN 978-0-511-48980-8. doi: 10.1017/CBO9780511489808.024.

Todd Barrett Kashdan, David Disabato, Fallon Rachael Goodman, and Patrick McKnight. The Five-Dimensional Curiosity Scale Revised (5DCR): Briefer subscales while separating general overt and covert social curiosity. Preprint, Open Science Framework, October 2019.

Heidi Keller, Klaus Schneider, and Bruce Henderson (eds.). *Curiosity and Exploration*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. ISBN 978-3-540-54867-6 978-3-642-77132-3. doi: 10.1007/978-3-642-77132-3.

Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811, December 2014. ISSN 2050-084X. doi: 10.7554/eLife.04811.

Celeste Kidd and Benjamin Y. Hayden. The Psychology and Neuroscience of Curiosity. *Neuron*, 88(3): 449–460, November 2015. ISSN 08966273. doi: 10.1016/j.neuron.2015.09.010.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active World Model Learning with Progress Curiosity. *arXiv:2007.07853 [cs, stat]*, July 2020.

Kenji Kobayashi and Ming Hsu. Common neural code for reward and information value. *Proceedings of the National Academy of Sciences*, 116(26):13061–13066, June 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1820145116.

Artemy Kolchinsky and David H. Wolpert. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6):20180041, December 2018. ISSN 2042-8898, 2042-8901. doi: 10.1098/rsfs.2018.0041.

Adrien Laversanne-Finot, Alexandre Péré, and Pierre-Yves Oudeyer. Curiosity Driven Exploration of Learned Disentangled Goal Spaces. *arXiv:1807.01521 [cs, stat]*, November 2018.

Joel Lehman and Kenneth O. Stanley. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation*, 19(2):189–223, June 2011a. ISSN 1063-6560, 1530-9304. doi: 10.1162/EVCO_a_00025.

Joel Lehman and Kenneth O. Stanley. Novelty Search and the Problem with Objectives. In Rick Riolo, Ekaterina Vladislavleva, and Jason H. Moore (eds.), *Genetic Programming Theory and Practice IX*, pp. 37–56. Springer New York, New York, NY, 2011b. ISBN 978-1-4614-1769-9 978-1-4614-1770-5. doi: 10.1007/978-1-4614-1770-5_3.

Joel Lehman, Kenneth O. Stanley, and Risto Miikkulainen. Effective diversity maintenance in deceptive domains. In *Proceeding of the Fifteenth Annual Conference on Genetic and Evolutionary Computation Conference - GECCO '13*, pp. 215, Amsterdam, The Netherlands, 2013. ACM Press. ISBN 978-1-4503-1963-8. doi: 10.1145/2463372.2463393.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116 (1):75, 1994.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-yves Oudeyer. Exploration in Model-based Reinforcement Learning by Empirically Estimating Learning Progress. *NIPS*, 25:1–9, 2012.

J. López-Fidalgo, C. Tommasi, and P. C. Trandafir. 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, April 2007. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2007.00586.x.

David M. Lydon-Staley, Dale Zhou, Ann Sizemore Blevins, Perry Zurn, and Danielle S. Bassett. Hunters, busybodies and the knowledge network building associated with deprivation curiosity. *Nature Human Behaviour*, 5(3):327–336, March 2021. ISSN 2397-3374. doi: 10.1038/s41562-020-00985-7.

David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. Second edition, 2003.

Joao Marques, Li Meng, Diane Schaak, Drew Robson, and Jennifer Li. Internal state dynamics shape brainwide activity and foraging behaviour. *Nature*, pp. 27, 2019.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and Achieving Goals via World Models. *arXiv:2110.09514 [cs, stat]*, October 2021.

Tom M Mitchell. The Need for Biases in Learning Generalizations. *New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.*., pp. 184–191, 1980.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv:1504.04909 [cs, q-bio]*, April 2015.

Martin Nowak and Karl Sigmund. A strategy of win-stay lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364(1):56–58, 1993.

Pierre-Yves Oudeyer. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1, 2007. ISSN 1662-5218. doi: 10.3389/neuro.12.006.2007.

Pierre-Yves Oudeyer. Computational Theories of Curiosity-Driven Learning. *arXiv:1802.10546 [cs]*, June 2018.

Pierre-Yves Oudeyer and Linda B. Smith. How Evolution May Work Through Curiosity-Driven Developmental Process. *Topics in Cognitive Science*, 8(2):492–502, April 2016. ISSN 17568757. doi: 10.1111/tops.12196.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-Driven Exploration by Self-Supervised Prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 488–489, Honolulu, HI, USA, July 2017. IEEE. ISBN 978-1-5386-0733-6. doi: 10.1109/CVPRW.2017.70.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-Supervised Exploration via Disagreement. *Proceedings of the 36th International Conference on Machine Learning*, pp. 10, 2019.

B Poucet. Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psycholocial Review*, 100(1):162–183, 1993.

Dobromir Rahnev and Rachel N. Denison. Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, 41:e223, 2018. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X18000936.

Gautam Reddy, Antonio Celani, and Massimo Vergassola. Infomax strategies for an optimal balance between exploration and exploitation. *Journal of Statistical Physics*, 163(6):1454–1476, June 2016. ISSN 0022-4715, 1572-9613. doi: 10.1007/s10955-016-1521-0.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Tim Roughgarden. *Algorithms Illuminated (Part 3): Greedy Algorithms and Dynamic Programming*, volume 1. 2019.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic Curiosity through Reachability. *arXiv:1810.02274 [cs, stat]*, August 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*, pp. 222–227, 1991.

Juergen Schmidhuber. Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes. *arXiv:0812.4360 [cs]*, December 2008.

Juergen Schmidhuber. Unsupervised Minimax: Adversarial Curiosity, Generative Adversarial Networks, and Predictability Minimization. *arXiv:1906.04493 [cs]*, June 2019.

Jürgen Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, September 2010. ISSN 1943-0604, 1943-0612. doi: 10.1109/TAMD.2010.2056368.

Philipp Schwartenbeck, Johannes Passecker, Tobias U Hauser, Thomas HB FitzGerald, Martin Kronbichler, and Karl J Friston. Computational mechanisms of curiosity and goal-directed exploration. *eLife*, (e41703): 45, 2019.

Frank Sehnke, Christian Osendorfer, Thomas Rückstieß, Alex Graves, Jan Peters, and Jürgen Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, May 2010. ISSN 08936080. doi: 10.1016/j.neunet.2009.12.004.

C E Shannon. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):55, 1948.

Claude Shannon and Warren Weaver. *The Mathematical Theory of Communication*. The university of Illinois Press, 1964.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature16961.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *Science*, 362(6419):1140–1144, 2018.

Devendra Singh. Preference for bar pressing to obtain reward over freeloading in rats and children. *Journal of Comparative and Physiological Psychology*, 73(2):320–327, 1970. ISSN 0021-9940. doi: 10.1037/h0030222.

Mingyu Song, Zahy Bnaya, and Wei Ji Ma. Sources of suboptimality in a minimalistic explore–exploit task. *Nature Human Behaviour*, 3(4):361–368, April 2019. ISSN 2397-3374. doi: 10.1038/s41562-018-0526-x.

Sumedh A. Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal Representation Learning. *arXiv:2010.03110 [cs]*, October 2020.

Marek Spinka, Ruth C Newberry, and Marc Bekoff. Mammalian Play: Training for the Unexpected. *THE QUARTERLY REVIEW OF BIOLOGY*, 76:28, 2001.

Kenneth O. Stanley and Risto Miikkulainen. Evolving a Roving Eye for Go. In Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, and Kalyanmoy Deb (eds.), *Genetic and Evolutionary Computation – GECCO 2004*, volume 3103, pp. 1226–1238. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-22343-6 978-3-540-24855-2. doi: 10.1007/978-3-540-24855-2_130.

Christopher Stanton and Jeff Clune. Deep Curiosity Search: Intra-Life Exploration Can Improve Performance on Challenging Deep Reinforcement Learning Problems. *arXiv:1806.00553 [cs]*, November 2018.

Mark G. Stokes. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, 19(7):394–405, July 2015. ISSN 13646613. doi: 10.1016/j.tics.2015.05.004.

Emily S Sumner, Amy X Li, Amy Perfors, Brett K Hayes, Danielle J Navarro, and Barbara W Sarnecka. The Exploration Advantage: Children's instinct to explore allows them to find information that adults miss. *PsyArxiv*, h437v:11, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

George T. Taylor. Discriminability and the contrafreeloading phenomenon. *Journal of Comparative and Physiological Psychology*, 88(1):104–109, 1975. ISSN 0021-9940. doi: 10.1037/h0076222.

Alexandr Ten, Pramod Kaushik, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. Humans monitor learning progress in curiosity-driven exploration. *Nature Communications*, 12(1):5972, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26196-w.

Naftali Tishby, Fernando C Pereira, and William Bialek. The Information Bottleneck Method. *Arxiv*, 0004057:11, 2000.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782, 1557-7317. doi: 10.1145/1968.1972.

Roby Velez and Jeff Clune. Novelty search creates robots with general skills for exploration. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation - GECCO '14*, pp. 737–744, Vancouver, BC, Canada, 2014. ACM Press. ISBN 978-1-4503-2662-9. doi: 10.1145/2576768.2598225.

Maya Zhe Wang and Benjamin Y. Hayden. Monkeys are curious about counterfactual outcomes. *Cognition*, 189:1–10, August 2019. ISSN 00100277. doi: 10.1016/j.cognition.2019.03.009.

Maya Zhe Wang and Benjamin Y. Hayden. Curiosity, latent learning, and cognitive maps. Preprint, Neuroscience, June 2020.

Xiao-Jing Wang. 50 years of mnemonic persistent activity: Quo vadis? *Trends in Neurosciences*, pp. S0166223621001685, October 2021. ISSN 01662236. doi: 10.1016/j.tins.2021.09.001.

J Kael White, Ethan S Bromberg-Martin, Sarah R Heilbronner, Kaining Zhang, Julia Pai, Suzanne N Haber, and Ilya E Monosov. A neural network for information seeking. Preprint, Neuroscience, August 2019.

Robert C. Wilson, Andra Geana, John M. White, Elliot A. Ludvig, and Jonathan D. Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074–2081, 2014. ISSN 1939-2222, 0096-3445. doi: 10.1037/a0038199.

Robert C Wilson, Elizabeth Bonawitz, Vincent Costa, and Becket Ebitz. Balancing exploration and exploitation with information and randomization. Preprint, PsyArXiv, July 2020.

Robert C Wilson, Elizabeth Bonawitz, Vincent D Costa, and R Becket Ebitz. Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences*, 38:49–56, April 2021. ISSN 23521546. doi: 10.1016/j.cobeha.2020.10.001.

Wanja Wolff and Corinna S. Martarelli. Bored Into Depletion? toward a Tentative Integration of Perceived Self-Control Exertion and Boredom as Guiding Signals for Goal-Directed Behavior. *Perspectives on Psychological Science*, 15(5):1272–1283, September 2020. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691620921394.

Darrell A. Worthy and W. Todd Maddox. 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, April 2014. ISSN 00222496. doi: 10.1016/j.jmp.2013.10.001.

Charley M. Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D. Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915–924, December 2018. ISSN 2397-3374. doi: 10.1038/s41562-018-0467-4.

Jingwei Zhang, Niklas Wetzel, Nicolai Dorka, Joschka Boedecker, and Wolfram Burgard. Scheduled Intrinsic Drive: A Hierarchical Take on Intrinsically Motivated Exploration. *arXiv:1903.07400 [cs, stat]*, June 2019.

Shunan Zhang and Angela J Yu. Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. *NeurIPS*, 26, 2013.

Dale Zhou, David M. Lydon-Staley, Perry Zurn, and Danielle S. Bassett. The growth and form of knowledge networks by kinesthetic curiosity. *arXiv:2006.02949 [cs, q-bio]*, June 2020.