Explore and Control with Adversarial Surprise

Anonymous Authors[1]

Abstract

Reinforcement learning (RL) provides a framework for learning goal-directed policies given user-specified rewards. However, since rewards can be sparse and task-specific, we are interested in the problem of learning without rewards, where agents must discover useful behaviors in the absence of domain-specific incentives. Intrinsic motivation is a family of unsupervised RL techniques which develop general objectives for an RL agent to optimize that lead to better exploration or the discovery of skills. In this paper, we propose a new unsupervised RL technique based on an adversarial game which pits two policies against each other to compete over the amount of surprise an RL agent experiences. The policies each take turns controlling the agent. The Explore policy maximizes entropy, putting the agent into surprising or unfamiliar situations. Then, the Control policy takes over and seeks to recover from those situations by minimizing entropy. The game harnesses the power of multi-agent competition to drive the agent to seek out increasingly surprising parts of the environment while learning to gain mastery over them, leading to better exploration and the emergence of complex skills. Theoretically, we show that under certain assumptions, this game pushes the agent to fully explore the latent state space of stochastic, partially-observed environments, whereas prior techniques will not. Empirically, we demonstrate that even with no external rewards, Adversarial Surprise learns more complex behaviors, and explores more effectively than competitive baselines, outperforming intrinsic motivation methods based on active inference, novelty-seeking (Random Network Distillation (RND)), and multi-agent unsupervised RL (Asymmetric Self-Play (ASP)).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Despite promising results across a number of domains (e.g., (Berner et al., 2019; Kober et al., 2013; Levine et al., 2016; Vinyals et al., 2019)), a major challenge in reinforcement learning (RL) is that the effectiveness of current methods depends heavily on task-specific rewards (Riedmiller et al., 2018; Vecerik et al., 2017). For many tasks, rewards are sparse, making learning slow and difficult (Riedmiller et al., 2018; Vecerik et al., 2017). Yet humans and animals easily learn from their own experience without being constantly told what to do. Therefore, significant prior research has focused on developing a class of unsupervised RL methods that optimize for *intrinsic motivation* (IM) – alternative objectives (such as seeking novelty) that incentivize the agent to autonomously explore and learn about the world without being entirely dependent on explicit goals and task incentives.

Many prior works have approached the problem of unsupervised RL as one of pure novelty-seeking, training agents to optimize for surprise or prediction error in the hopes of facilitating better exploration, which can be helpful for learning downstream tasks. While better exploration is a laudable goal, such objectives may be ill-suited to stochastic environments, because they can become distracted by inherently high-entropy elements; this is known as the "noisy TV" problem (Schmidhuber, 2010). In contrast, active inference methods (Friston, 2009; Friston et al., 2009; Berseth et al., 2019) deal with highly entropic environments by attempting to *minimize* surprise. However, these techniques suffer from the "dark room" problem (Friston et al., 2012); that is, they may get stuck because they prefer to stay in highly predictable, unsurprising states. This is especially problematic in partially-observed environments. Yet humans maintain a balance between optimizing for both novelty and familiarity. Thus, we propose that the role of unsupervised RL is more nuanced: beyond just exploring to find *surprising* states, the goal should be to also achieve meaningful *control* over a diverse set of states. For example, a child in a play room does not just try to toss their toys on the floor in every possible pattern, but instead tries to stack them together, find new uses for parts, or combine them in various ways. In the same way, an effective unsupervised RL agent should aim to explore the state space while still learning increasingly complex and interesting behaviors.

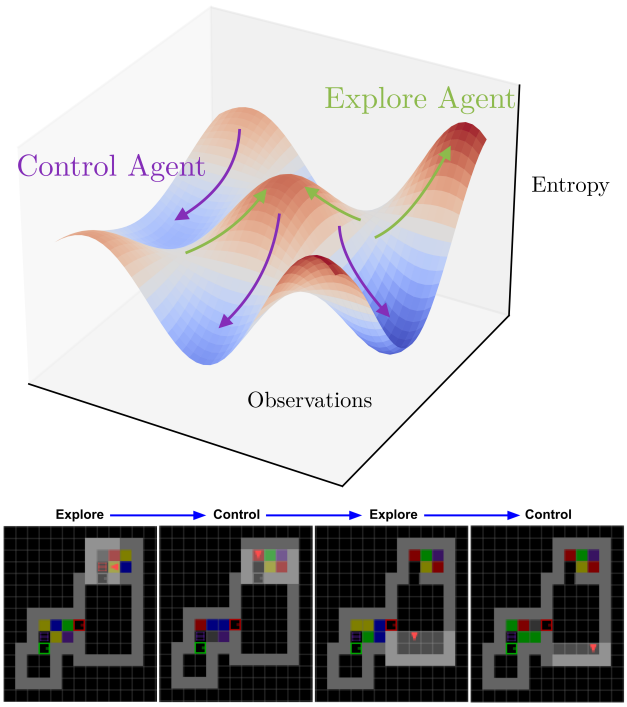

Figure 1: Adversarial Surprise is a multi-agent competition in which two policies take turns controlling an RL agent. The Explore policy acts first, and tries to put the agent into surprising, high entropy states. On its turn, the Control policy tries to minimize surprise by finding familiar, low-entropy, predictable states. Figure (b) shows an example rollout of a game. In the Explore policy's first turn, it attempts to maximize entropy by remaining in the first room with flashing lights (a *noisy TV* state). When the Control policy takes over, it is able to take an action to control the environment, by flipping the switch to stop the lights from flashing. Therefore, when the Explore policy acts again (c) it moves into the next room; this is predicted by our theoretical results, which show that the Explore policy must maximize entropy over the true state space. During the Control policy's next turn it remains in the dark room. This process continues until all of the rooms have been visited.

In this paper, we frame the quest for both exploration and control as an adversarial game between two policies, which each take turns sequentially acting for the same RL agent. The *Explore* policy is novelty-seeking, and attempts to maximize surprise (or observation prediction error) over the course of the episode, putting the agent into a diverse range of novel states. In turn, the *Control* policy must minimize surprise, by learning to manipulate its environment in order to return to safe and predictable states. When combined, the two adversaries engage in an arms race, repeatedly putting the agent into challenging new situations, then attempting to gain control of those situations. Figure 1 shows an illustration of the method, including a sample interaction. Rather than simply adding noise to the environment, the Explore policy learns to adapt to the Control policy, and search for increasingly challenging situations from which the Control policy must recover. Thus our method, *Adversarial Surprise* (AS), leverages the power of multi-agent training to generate a curriculum of increasingly challenging exploration and control problems. Both our theoretical and empirical results demonstrate that while other IM and unsupervised RL techniques can fail in stochastic, partially-observed environments, AS learns to both fully explore the state space, and control it.

Our main contribution is the Adversarial Surprise (AS) algorithm, a method that frames the unsupervised RL problem as an adversarial game between an Explore and Control policy. We derive our method, and provide a theoretical analysis in the recently introduced Block MDP setting (Du et al., 2019) which shows that our method maximizes state coverage in partially-observed, stochastic environments. We present a practical instantiation based on deep reinforcement learning, and provide an empirical evaluation that compares our approach to recent intrinsic motivation and unsupervised RL techniques. Our results demonstrate that our method learns complex behaviors in the absence of any external rewards. In tasks involving both distracting stochastic elements and complex environment dynamics, Adversarial Surprise outperforms both surprise-minimizing techniques like SMiRL (Berseth et al., 2019), and novelty-seeking methods like Random Network Distillation (RND) (Burda et al., 2018). We also show that AS learns more quickly than a recently-proposed unsupervised RL method based on multi-agent competition, Asymmetric Self Play (ASP) (Sukhbaatar et al., 2017; OpenAI et al., 2021). AS learns to explore more effectively than these prior works, takes more actions to control its environment, and learns complex behavior purely through optimizing its intrinsic motivation.

2. Related Work

Novelty-seeking intrinsic motivation: One of the most frequently studied forms of intrinsic motivation is novelty-seeking, or curiosity (e.g., (Bellemare et al., 2016; Pathak et al., 2017; Schmidhuber, 1991; Still and Precup, 2012)). A novelty-seeking agent is motivated to explore its environment, through incentives to maximize model uncertainty (Still and Precup, 2012; Shyam et al., 2019; Pathak et al., 2019), entropy (Eysenbach et al., 2018; Hazan et al., 2019; Lee et al., 2019; Pong et al., 2019), prediction error (Pathak et al., 2017; Yamamoto and Ishikawa, 2010), or surprise (Achiam and Sastry, 2017; Schmidhuber, 1991; Sun et al., 2011). However, naïvely maximizing uncertainty, predic-

tion error, or entropy leads to methods that are vulnerable to the *noisy TV* problem, where the agent becomes distracted by inherently high-entropy, unpredictable elements in the environment, such as white noise (e.g. static on a TV) (Schmidhuber, 2010). In this case, a curious agent will not be able to learn meaningful behaviours. We will show that AS overcomes this problem, and fully explores the state space even when there are highly stochastic elements.

Surprise minimization and the free energy principle: Rather than maximizing surprise, the free energy principle (Faraji et al., 2018; Friston, 2009; Friston et al., 2009; 2016; Ueltzhöffer, 2018) proposes that biological agents actually *minimize* surprise (or state prediction error) by seeking familiar, stable states, and controlling their environment to make it more predictable. Inspired by this idea, Berseth et al. (2019) built a surprise-minimizing RL agent, SMiRL, which outperforms curiosity-maximizing methods like ICM (Pathak et al., 2017) in high-entropy environments. However, in partially observed or low-entropy environments, surprise minimization is vulnerable to the *dark room* problem: if the agent can simply stay in a highly-predictable part of the environment where nothing happens, it will (Friston et al., 2012). In this scenario, a surprise-minimizing agent cannot learn meaningful behaviors either. AS is designed to avoid this problem, because the Explore policy will not allow the Control policy to remain in a dark room.

Mutual information and empowerment: Alternative approaches to intrinsic motivation may be based on mutual information (e.g. (Houthooft et al., 2016)). For example, the goal of *empowerment* (Klyubin et al., 2005; Salge et al., 2014b) is to maximize the mutual information between the agent's actions and its future states. Empowerment encourages agents to "keep their options open" by exploring many states, while still maintaining a high degree of control in those states. Maximizing measures of empowerment between agents has also been explored (Guckelsberger et al., 2016; 2018; Jaques et al., 2019). However, calculating empowerment in high dimensional environments is intractable, leading to various methods for approximating it (e.g., (Karl et al., 2015; Salge et al., 2014a; Zhao et al., 2020; de Abril and Kanai, 2018; Zhang et al., 2020)) for example, using variational inference (Mohamed and Rezende, 2015). Unfortunately, these methods can also be difficult to get working with high dimensional function approximation (Gregor et al., 2016). In contrast, we show that Adversarial Surprise works with deep neural networks applied to pixel inputs in Atari environments.

Multi-agent curriculum generation: Multi-agent competition can provide a mechanism for driving RL agents to automatically learn increasingly complex behavior (Leibo et al., 2019). As each agent adapts, it makes the learning problem for the other agent increasingly difficult, leading to the emergence of an automatic curriculum of challenging learning tasks (Baker et al., 2019; Dennis et al., 2020; Xu et al., 2020). For example, Schmidhuber (1997) proposed having two classifiers compete by repeatedly selecting examples which they can classify but which the other cannot. Asymmetric Self Play (ASP) (Sukhbaatar et al., 2017; OpenAI et al., 2021) rewards one agent, Alice, for executing the shortest trajectory that a second agent, Bob, cannot copy (or reverse). ASP is limited to training goal-conditioned policies, and is thus less applicable to standard RL tasks where the reward-maximizing states are not known a priori. Adversarial Surprise does not have this limitation, and instead serves as a general algorithm that can speed learning across a variety of contexts. More importantly, we show in this work that ASP can fail to explore stochastic environments, because Alice can easily produce a random goal state which Bob is not able to reproduce. In contrast, AS works well in stochastic environments, outperforming ASP.

3. Background

Partially Observed Markov Decision Process: A POMDP is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{O}, r, \gamma)$, where $s \in \mathcal{S}$ are states, $a \in \mathcal{A}$ are actions, $r(a, s)$ is the reward function, and $\gamma \in [0, 1)$ is a discount factor. The environment is partially observed, so the agent cannot observe the true state s, but rather observes $o \sim p(O|s)$. At each timestep t, the agent selects an action a_t according to its policy $\pi(a_t|o_t)$, receives reward $r(a_t, s_t)$, and the environment transitions to the next state according to $\mathcal{T}(s_{t+1}|s_t, a_t)$. We are interested in stochastic environments, in which the emission distribution distribution is inherently entropic for some states, i.e. $\exists s : H(p(O|s)) > 0$.

Block Markov Decision Process: A BMDP (Du et al., 2019) is a POMDP with an additional disjointness assumption: for any $s, s' \in S$, $s \neq s' \Rightarrow \text{supp}(p(O|s)) \cap \text{supp}(p(O|s)) = \emptyset$. Our theoretical results show that the AS algorithm fully covers the latent state space of a large family of BMDPs.

Intrinsic motivation: IM can either be used in combination with a task objective, in which case intrinsic motivation serves to facilitate exploration, or by itself, in which case the agent receives no external task rewards, and aims only to maximise its intrinsic objective, leading it to learn skills that may potentially be useful for downstream tasks. In this paper, we study how an agent can learn skilled behaviour without rewards. Therefore, we consider agents that seek to optimize cumulative intrinsic reward over the episode: $R = \sum_{t=0}^{T} \gamma^t r^i(a_t, s_t)$.

Surprise minimizing agents (SMiRL): The free energy principle (Faraji et al., 2018; Friston, 2009; Friston et al., 2009; 2016; Ueltzhöffer, 2018) suggests that biological

agents may minimize surprise, or state entropy, in order to remain in safe and stable states. Minimising surprise for an RL agent requires keeping track of the state history of the agent via a density model, $p_\theta(s)$, which bounds the state marginal density of the policy, d^π, given states seen so far in an episode, $\tau_t = \{s_0, ...s_t\}$ (Berseth et al., 2019). Berseth et al. (2019) shows that we can use the intrinsic reward $r^i(s_t) = \log p_\theta(s)$ to minimize the entropy of the state marginal distribution $H(d^\pi(s))$. A SMiRL agent trained with this objective learns emergent behaviors to reduce entropy in stochastic environments—such as stable walking robots or playing Tetris—even in the absence of any external reward. However, when applied to partially observed environments, the agent is susceptible to the dark room problem; rather than learning to control the environment, it can simply control its observations by remaining in unsurprising parts of the environment. Or, simply turning to look at a wall. In our method, we build on surprise minimization, incorporating it into a two player game that alleviates this shortcoming.

4. Adversarial Surprise

The goal of Adversarial Surprise (AS) is to facilitate exploring and learning to control stochastic, partially-observed environments. To this end, AS pits two policies against each other in a two-player competition over the amount of surprise an RL agent experiences. Specifically, we learn an Explore policy, π^E, and a Control policy, π^C. The goal of the Control policy is to minimize its own surprise, or observation entropy, using a learned model $p_\theta(o)$. However, the Explore policy's goal is to maximize the surprise that the Control policy experiences.

The policies take turns taking actions for the agent, switching back and forth throughout the episode. The policy controlling the RL agent change every k steps, such that:

$$a_t \sim \begin{cases} \pi^E(a_t|o_t) & \text{if } \exists n, t \in [2nk, (2n+1)k[\\ \pi^C(a_t|o_t) & \text{otherwise} \end{cases} \quad (1)$$

Each policy is given several steps to act because it enables it to reach states that will be challenging for the other policy to recover from, thus facilitating learning more complex and long-term exploration and control behaviors (see Figure 1).

To estimate surprise, the policies share a learned density model, which estimates the likelihood of experiencing observation o, $p_\theta(o)$. Because the Control policy is surprise-minimizing, its reward is $r^{iC}(s_t) = \log p_\theta(o_t)$, which resembles SMiRL (Berseth et al., 2019), except using the observation in place of the state. The goal of the Explore policy is to maximize the observation surprise of the RL agent when the Control policy is in control. This creates an adversarial game, in which the Explore policy attempts to find surprising situations with which to expose the Control policy, and the Control policy's objective is to recover from them. Therefore, the Explore policy's reward is based on the surprise for the observations of the Control policy. Assume that the Control policy's turn begins at timestep t^C, and it receives a total reward of $R^i = \sum_{t=t^C}^{t^C+k} \gamma^k r^i(a_t, s_t)$ for that turn. Then, the Explore policy's reward is $-R^i$, and is applied to the last timestep of the Explore policy's turn (i.e. timestep $t^C - 1$). The full training procedure for Adversarial Surprise is given in Algorithm 1.

Algorithm 1 Adversarial Surprise

Randomly initialize ϕ^E and ϕ^C
for episode = $0, ..., M$ **do**
  Initialize $\theta, R^i = 0$, $\beta \leftarrow \{\}$, explore_turn = True, $t^C = k$, $s_0 \sim p(s_0), o_0 \sim p(O_0|s_0)$
  for $t = 0, ..., T$ **do**
    if explore_turn **then**
      $a_t \sim \pi^E(o_t, h_t^E)$
    else
      $a_t \sim \pi^C(o_t, h_t^C)$
    end if
    $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t)$, $o_{t+1} \sim p(O_{t+1}|s_{t+1})$
    $r_t^i = \log p_\theta(o_{t+1})$
    if not explore_turn and $t - t^C > k/2$ **then**
      $R^i = R^i + r^i$
    end if
    $\beta = \beta \cup \{o_t, a_t, o_{t+1}\}$
    if $t == t^C$ **then**
      explore_turn = not explore_turn
      $t^C = t^C + 2k$
    end if
    θ_{t+1} =MLE_update(β, θ_t)
  end for
  ϕ^E =RL_update$(\beta, -R^i)$
  ϕ^C =RL_update(β, R^i)
end for

Thus, Adversarial Surprise defines the following adversarial game between the two agents:

$$\max_{\pi^E} \min_{\pi^C} -E\left[\sum_{t=t^C}^{t^C+k} \log p_\theta(o_t)\right], \quad (2)$$

where the Explore policy can only effect $p(o_t)$ through the final state that it produces at the end of its turn, which is the initial state for the Control policy. We show in the Appendix that the objective of the Explore policy approaches maximizing the entropy of the Control policy's observations:

$$\mathcal{J}^E = -E\left[\sum_{t=t^C}^{t^C+k} \log p_\theta(o_t)\right] \approx H(d^{\pi_C}(o)), \quad (3)$$

Analogously, the Control policy's goal is to minimize entropy:

$$\mathcal{J}^C = E\left[\sum_{t=t^C}^{t^C+k} \log p_\theta(o_t)\right] \approx -H(d^{\pi_C}(o)) \quad (4)$$

Implementation details: We parameterize the policies for the Explore and Control policy using deep neural networks (NN) with parameters ϕ^E and ϕ^C, respectively. We policy is based on a convolutional NN which conditions on a stack the last 4 observation frames. The networks are trained using Proximal Policy Optimization (PPO) (Schulman et al., 2017); further details and hyperparameters are given in the Appendix. Following (Berseth et al., 2019), the shared density model $p_\theta(o)$ is re-initialized each episode and trained using maximum likelihood estimation (MLE) to fit the observations of both agents within a single episode, which are stored in a buffer β. The density model is either represented using a Gaussian distribution as in (Berseth et al., 2019), or using independent categorical distributions. We have found it helpful to only compute the surprise reward using observations from the second half of the Control policy's turn; this gives the agent greater ability to take actions that may lead to initial surprise, but reduce entropy over the long term. We also experiment with when to reset the buffer β; we find that resetting the buffer after each *round* (after the Explore policy and Control policy each take one turn) can sometimes improve performance. Finally, we allow the Explore and Control policys to act for a different number of timesteps, tuning the emphasis on exploration or control depending on the environment.

5. Adversarial Surprise maximizes state coverage in Block MDPs

In this section, we show that our method covers the state space under some restrictions on the structure of the POMDP, and the density of states with low observation entropy. Full proofs are in the Appendix.

We define the marginal observation distribution as: $d^\pi(o) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(o_t = o)$. We are interested in maximizing the state marginal entropy $d^\pi(s)$ by using a density model of the observation, $p_\theta(o)$. To that end, we prove the following lemm (see Appendix):

Lemma 1. *The cumulative surprise measured by the observation density model $p_\theta(o)$ forms an upper bound of the observation marginal entropy $H(d^\pi(o))$, which becomes tight when the observation density model fits the observation marginal $d^\pi(o)$: $-E_\pi \sum_{t=0}^{\infty} \log p_\theta(o_t) \geq H(d^\pi(o))$*

We start with an assumption on the structure of the POMDP:

Assumption 1 (Block MDP)**.** *We suppose that every two different states have disjoint emission supports: for any $s, s' \in S$, $s \neq s' \Rightarrow \text{supp}(p(O|s)) \cap \text{supp}(p(O|s)) = \emptyset$*

Under this assumption, we show a useful relation between observation marginal entropy and state marginal entropy:

Lemma 2. *In a block MDP (BMDP) (Misra et al., 2020), we can decompose the observation marginal entropy:*

$$H(d^\pi(o)) = E_{d^\pi(s)} H(p(O|S=s)) + H(d^\pi(s)) \quad (5)$$

See Appendix for the proof. Equation 5 shows that maximizing entropy in the marginal observation distribution $d^\pi(o)$ amounts to maximizing two terms: the emission entropy $H(p(O|S))$, and the state marginal entropy $H(d^\pi(s))$.

Suppose that we have a small number of latent states with rich observations, i.e. where the entropy of the emission distribution is very high: $\exists s : H(p(O|S=s)) \gg \log|\mathcal{S}|$. We can think of these states as "*noisy TVs*". In this case, if we are trying to maximize the marginal observation entropy, we have:

$$\max_{d^\pi(s)} H(d^\pi(o)) \approx \max_{d^\pi(s)} E_{d^\pi(s)} H(p(O|S=s)) \quad (6)$$

The RHS is maximized by taking:

$$d^\pi(s) = \mathbb{1}(s = \arg\max_{s\in\mathcal{S}} H(p(O|S=s))) \quad (7)$$

This shows that conventional methods that focus on maximizing entropy over observations can trivially maximize $d^\pi(o)$ by remaining in state s. We will show that Adversarial Surprise is not subject to this problem.

To this end, we define a (semiquasi)metric on the latent state space: $\tilde{d}(s,s') = \min\{k : \exists\pi, P_k^\pi(s'|s) = 1\}$, where by convention $P_0^\pi(s|s) = 1$ for all s. In other words, $\tilde{d}(s,s') = k$ if there is a policy that reaches s' from s in k steps with probability 1. We symmetrize this metric by defining the following semimetric: $d(s,s') = \max\{\tilde{d}(s,s'), \tilde{d}(s',s)\}$

We now give a formal definition of dark rooms. We say that a state s is a dark room if it has minimal emission entropy: $H(p(O|S=s)) = \min_{s\in S} H(p(O|S=s))$. We will use the following assumption about the density of dark rooms in the latent state space:

Assumption 2. *We suppose that for every state s, there is a dark room such that $d(s,s') \leq \frac{T}{2}$. That is, the set of dark rooms is a $\frac{T}{2}$-cover of the state space with respect to d.*

We can now state our main result:

Theorem 1. *Under Assumptions 1 and 2 the Markov chain induced by the following AS game:*

$$\max_{d^{\pi_E}(s_0)} \min_{d^{\pi_C}_{1:\frac{T}{2}}(s|s_0)} H\left(d^{\pi_C}_{\frac{T}{2}}(o)\right) \quad (8)$$

$\frac{T}{2}$-covers the state space, i.e., for all states s, there is a state s' such that $d^\pi(s') > 0$ and $d(s, s') \leq \frac{T}{2}$, where d^π is the marginal induced by the game between the Explore (π_E) and Control (π_C) agents.

Assumption 2 guarantees that for any states that the Explore policy reaches, the Control policy can find a low-emission-entropy state within its turn, such that $H(p(O|s))$ is minimized. Thus, from the perspective of the Explore policy, the first term in its objective in Eq. 5 is minimized, and it must focus on maximizing the second term, $H(d^\pi(s))$. In order to maximize entropy over the state marginal distribution $d^\pi(s)$, the Explore policy must fully explore the state space.

6. Experimental results

In this section we present experimental results designed to answer the following three questions: **Q1. Exploration and state coverage:** how well does AS explore the underlying state space in a stochastic, partially-observed world, as compared to alternative methods? Given our theoretical results in Section 5, we hypothesize that methods based on novelty-seeking will become distracted by noisy elements, while AS will fully explore the environment. We will use the number of rooms visited in a navigation task as a measurement of state coverage. **Q2. Control:** will AS learn to take actions to control its environment, and recover from surprising situations? We measure control as the number of actions taken that cause changes to elements in the environment, such as flipping a switch to stop flashing lights. **Q3. No-reward learning:** will AS enable the agent to learn meaningful behaviors in the absence of any external reward? To assess this, we train the IM methods using only intrinsic reward, then assess the amount of task reward they obtain in the standard Atari benchmark (Bellemare et al., 2013). While there is no reason to expect AS to always correlate with the objectives in arbitrary MDPs, we expect that the twin goals of maximizing coverage while achieving high control should correlate well with objectives in many reasonable MDPs, particularly video games of the sort present in Atari. Many of these games have a notion of progress, which roughly corresponds to coverage, but at the same time have many dangerous states that could result in 'death', which leads to an unexpected jump back to the starting state. Therefore, we hypothesize that AS should, without even being aware of the task reward, perform well in these environments. Comparing to prior methods in these domains is interesting, because prior work has variously argued that both novelty-seeking exploration methods (Burda et al., 2018) and surprise-minimization methods (Berseth et al., 2019) should be expected to achieve high scores in these domains. Note that when we assess all three metrics we look at the performance of the agent as a whole, which is jointly controlled by both the Explore and Control policy.

Baselines: We compare AS to three competitive unsupervised RL baselines: Asymmetric Self-Play (ASP) (Sukhbaatar et al., 2017) (a state-of-the-art multi-agent curriculum method), Random Network Distillation (RND) (Burda et al., 2018) (a state-of-the-art exploration method), and SMiRL (Berseth et al., 2019) (a recently proposed method based on surprise minimization). These three methods are all related to AS: like RND, the Explore policy aims to maximize surprise, though not instantaneously but rather over the Control policy's episode; like SMiRL, the Control policy aims to minimize surprise, but starting from the initial states that the Explore policy puts it in, and in a partially observed setting; like asymmetric self-play (ASP), AS consists of a two-player game, though the AS two-player game is symmetric and zero-sum, and based on observational entropy rather than goal-reaching. All methods use the same underlying PPO implementation, with hyperparameters given in the Appendix.

Environments: To evaluate **Q1** and **Q2**, we need partially-observed environments that present an exploration challenge, and which include stochastic phenomenaSince standard benchmarks do not consistently exhibit these properties, we constructed a custom family of procedurally generated **navigation** tasks based on Minigrid (Chevalier-Boisvert et al., 2018). These environments contain rooms that are either empty (dark), or contain stochastic elements such as flashing lights that randomly change color. They also contain elements such as doors that can be opened, and switches that, when flipped, stop the stochastic elements from changing. An example is shown in Figure 1. As in MiniGrid, the environments are partially observed; the agent only sees a 5x5 window of the true state. While these environments allow us to carefully study the effects of partial observability and stochasticity, we would also like to compare to prior work on a standardized benchmark. For this purpose, we use the **Atari** Arcade Learning Environment (ALE) (Bellemare et al., 2013) (see Figure 4), which was used by both SMiRL (Berseth et al., 2019) and RND (Burda et al., 2018) to establish their effectiveness. Due to limited computational resources, we do not conduct experiments in all possible Atari games (which is consistent with prior work (Berseth et al., 2019; Burda et al., 2018)), but we show results for each of the games that we test, both in Section 6 and the Appendix.

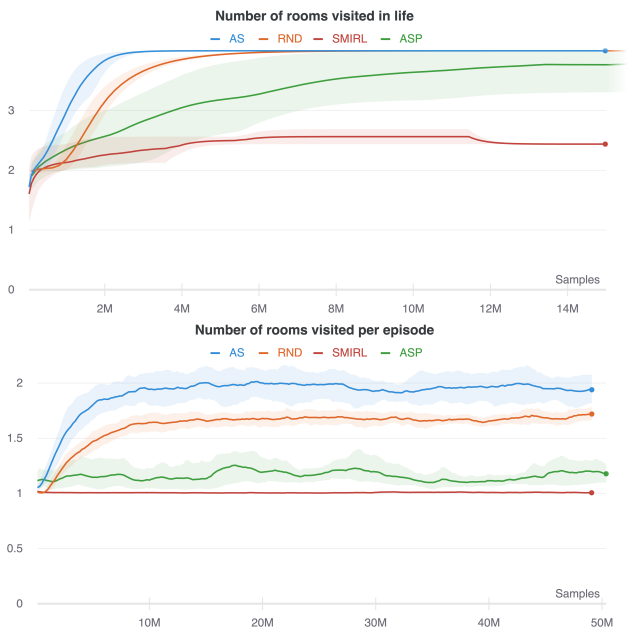

Figure 2: **Q1. Exploration and state coverage:** the number of rooms explored in partially-observed environments containing both stochastic elements and dark rooms, both (a) cumulatively over training, and (b) within an episode. Part (a) shows how well each method works as an exploration bonus to encourage collecting experience from all states, and part (b) provides a measure of how well the asymptotic policy learned by each method continues to explore the state space at convergence. AS outperforms surprise-minimizing SMiRL, which minimizes observation surprise by turning to face the wall, or remains in dark rooms. AS also explores better than the state-of-the-art exploration methods RND and ASP, which both become distracted by noisy elements.

Q1. Exploration and state coverage: Figure 2 shows the results of training each of the agents in the procedurally-generated navigation environments in terms of the number of rooms the agents learn to visit (our measure of state coverage). We measure the number of rooms cumulatively, over the course of training (Figure 2), to assess whether each method will lead the agent to collect experience from all possible states. This measure is relevant to whether the technique can be used as an effective exploration bonus to aid learning a downstream task. We also measure the number of rooms explored within each episode (Figure 2). This allows us to assess whether the asymptotic policy learned by the algorithms continues to explore once it has converged. As predicted by our theoretical analysis, we see that AS learns to more fully explore the environments, visiting significantly more rooms over a lifetime (Figure 2) and per episode (Figure 2) than competing methods. It learns more quickly and explores more thoroughly than RND, which becomes distracted by the inherently random elements in the environment which lead to high prediction error. The stochastic elements also hinder learning for ASP, since Alice can easily produce random goals that are difficult for Bob to replicate. Finally, we see that SMiRL, which is designed for fully observed environments, does not explore effectively because it suffers from the dark room problem – it prefers to stay within the empty rooms, and not venture into rooms with high-entropy, stochastic elements.

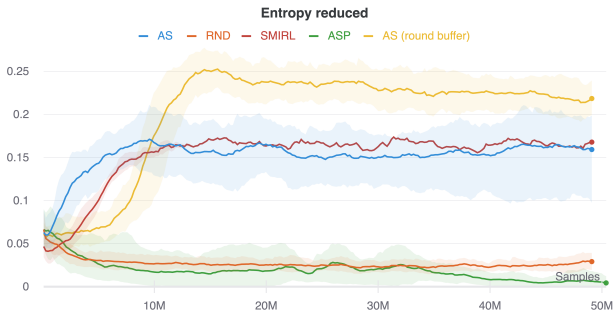

Figure 3: **Q2. Control:** the average number of times agents flip a switch to stop lights from flashing. ASP and RND do not learn to press the switch, while SMiRL and AS both press the switch a similar number of times. Resetting the AS buffer more frequently enables it to exceed even SMiRL in taking actions to control the environment.

Q2. Control: To measure whether agents learn to control their environment, we investigate how many times agents press switches in the navigation environment that stop the stochastic elements in the environment from changing color. The results are shown in Figure 3. Since RND has no incentive to learn to control the environment, it never learns to press the switch. A similar result is observed for ASP, since reducing the entropy would make it easier for the Bob agent to replicate the Alice agent's final state. Thus, ASP will not always lead agents to learn all possible behaviors relevant to controlling the environment. Both SMiRL and AS learn to take actions to reduce entropy. However, when we train AS by resetting the buffer β used to fit the density model $p_\theta(o)$ after each *round* (that is, after both the Explore and Control policy have taken one turn), rather than after each episode, we see that AS increases the number of actions it takes to reduce entropy even over SMiRL. This is likely because resetting the buffer removes any incentive to return to states that the agent has previously seen within its lifetime, and instead gives a stronger incentive to reduce entropy immediately.

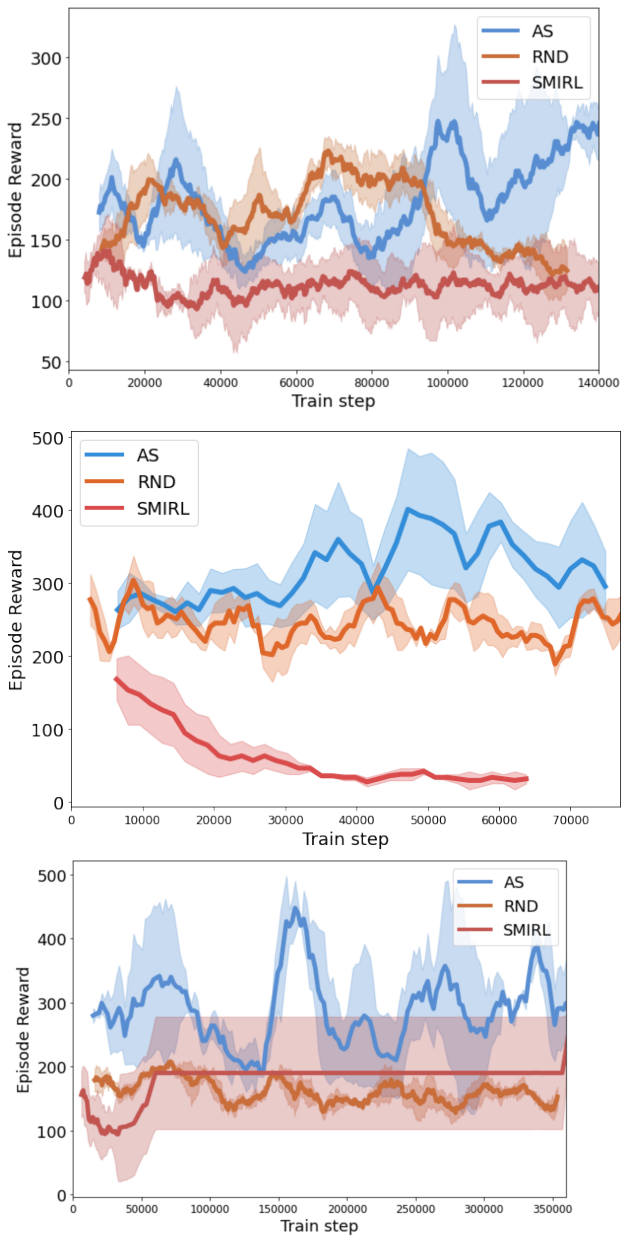

Figure 4: **Q3. No-reward learning:** Each IM method is trained in Atari using only intrinsic reward. The plot show how much of the true game reward the agent obtains, with error bars showing 95% Confidence Interval (CI) of three seeds. Since the games reward behavior such as staying alive and learning to shoot enemies, obtaining higher reward indicates the agent has learned meaningful behaviors. AS outperforms both RND and SMiRL, showing that AS provides a general way to learn useful behaviors across multiple environments, in the absence of external reward.

Q3. No-reward learning: Figure 4 shows how well each of the intrinsic motivation (IM) methods (SMiRL, RND, and AS) can be used to learn interesting behaviors in the absence of external reward. To assess whether they can learn useful skills, we measure the game reward obtained in several Atari environments. Because the games reward complex behaviors like shooting enemies, a high game reward indicates the agent has learned interesting skills, purely from optimizing the instrinsic objective. Across all three environments, AS performs better than both RND and SMiRL. The performance of RND decreases over time due the bonus from the prediction error shrinking as more states become familiar. SMiRL performs poorly in these environments, because it avoids entropy by hiding from enemies. In contrast, AS obtains higher returns, indicating that optimizing for both exploration and control provides a useful inductive bias for learning interesting behaviors in the absence of external reward. We include videos of the agents learning to play Atari games with no reward, as well as performing the navigation tasks, at https://sites.google.com/view/adversarial-surprise/home.

7. Discussion

We proposed Adversarial Surprise as a general approach for unsupervised reinforcement learning. Adversarial Surprise corresponds to a two-player adversarial game, in which two policies compete over the amount of surprise, or observation entropy, that an agent experiences. Reminiscent of Dr. Jekyll and Mr. Hyde, the Explore policy acts to expose the Control policy to highly entropic states from which it must recover by learning to manipulate the environment. AS provides a novel intrinsic motivation approach that leads to more effective learning in stochastic, partially observed environments. In such environments, prior methods either become distracted by noisy elements, or suffer from the "dark room" problem, in which observation entropy is minimized by simply hiding in a low-entropy part of the state space. We show both theoretically and empirically that AS is robust against these issues, and learns to explore the environment more thoroughly, and control it more effectively, than state-of-the-art prior works like RND, ASP, and SMiRL.

References

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael

Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. Smirl: Surprise minimizing reinforcement learning in dynamic environments. *arXiv preprint arXiv:1912.05510*, 2019.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv:1806.06505*, 2018.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *arXiv preprint arXiv:2012.02096*, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Mohammadjavad Faraji, Kerstin Preuschoff, and Wulfram Gerstner. Balancing new against old information: the role of puzzlement surprise in learning. *Neural computation*, 30(1):34–83, 2018.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

Karl Friston, Christopher Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130, 2012.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68: 862–879, 2016.

Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7): e6421, 2009.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

Christian Guckelsberger, Christoph Salge, and Julian Togelius. New and surprising ways to be mean. adversarial npcs with coupled empowerment minimisation. *arXiv preprint arXiv:1806.01387*, 2018.

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

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *arXiv preprint arXiv:1605.09674*, 2016.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR, 2019.

Maximilian Karl, Justin Bayer, and Patrick van der Smagt. Efficient empowerment. *arXiv preprint arXiv:1509.08455*, 2015.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer, 2005.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *arXiv preprint arXiv:1509.08731*, 2015.

OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique Ponde de Oliveira Pinto, et al. Asymmetric self-play for automatic goal discovery in robotic manipulation. *arXiv preprint arXiv:2101.04882*, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pages 5062–5071. PMLR, 2019.

Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, pages 4344–4353. PMLR, 2018.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819, 2014a.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014b.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

Jürgen Schmidhuber. What's interesting? 1997.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788. PMLR, 2019.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*, pages 41–51. Springer, 2011.

Kai Ueltzhöffer. Deep active inference. *Biological cybernetics*, 112(6):547–573, 2018.

Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

Kelvin Xu, Siddharth Verma, Chelsea Finn, and Sergey Levine. Continual learning of control primitives: Skill discovery via reset-games. *arXiv preprint arXiv:2011.05286*, 2020.

Naoyuki Yamamoto and Masumi Ishikawa. Curiosity and boredom based on prediction error as novel internal rewards. In *Brain-Inspired Information Technology*, pages 51–55. Springer, 2010.

Jin Zhang, Jianhao Wang, Hao Hu, Tong Chen, Yingfeng Chen, Changjie Fan, and Chongjie Zhang. Metacure: Meta reinforcement learning with empowerment-driven exploration. *arXiv preprint arXiv:2006.08170*, 2020.

Ruihan Zhao, Pieter Abbeel, and Stas Tiomkin. Efficient online estimation of empowerment for reinforcement learning. *arXiv preprint arXiv:2007.07356*, 2020.

References

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*, 2019.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Marc G Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *arXiv preprint arXiv:1606.01868*, 2016.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. Smirl: Surprise minimizing reinforcement learning in dynamic environments. *arXiv preprint arXiv:1912.05510*, 2019.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Ildefons Magrans de Abril and Ryota Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv:1806.06505*, 2018.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *arXiv preprint arXiv:2012.02096*, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Mohammadjavad Faraji, Kerstin Preuschoff, and Wulfram Gerstner. Balancing new against old information: the role of puzzlement surprise in learning. *Neural computation*, 30(1):34–83, 2018.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.

Karl Friston, Christopher Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130, 2012.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68: 862–879, 2016.

Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7): e6421, 2009.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Christian Guckelsberger, Christoph Salge, and Simon Colton. Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

Christian Guckelsberger, Christoph Salge, and Julian Togelius. New and surprising ways to be mean. adversarial npcs with coupled empowerment minimisation. *arXiv preprint arXiv:1806.01387*, 2018.

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

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *arXiv preprint arXiv:1605.09674*, 2016.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR, 2019.

Maximilian Karl, Justin Bayer, and Patrick van der Smagt. Efficient empowerment. *arXiv preprint arXiv:1509.08455*, 2015.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer, 2005.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *arXiv preprint arXiv:1509.08731*, 2015.

OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique Ponde de Oliveira Pinto, et al. Asymmetric self-play for automatic goal discovery in robotic manipulation. *arXiv preprint arXiv:2101.04882*, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pages 5062–5071. PMLR, 2019.

Vitchyr H Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. *arXiv preprint arXiv:1903.03698*, 2019.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International Conference on Machine Learning*, pages 4344–4353. PMLR, 2018.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819, 2014a.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014b.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

Jürgen Schmidhuber. What's interesting? 1997.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788. PMLR, 2019.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407*, 2017.

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal bayesian exploration in dynamic environments. In *International Conference on Artificial General Intelligence*, pages 41–51. Springer, 2011.

Kai Ueltzhöffer. Deep active inference. *Biological cybernetics*, 112(6):547–573, 2018.

Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

Kelvin Xu, Siddharth Verma, Chelsea Finn, and Sergey Levine. Continual learning of control primitives: Skill discovery via reset-games. *arXiv preprint arXiv:2011.05286*, 2020.

Naoyuki Yamamoto and Masumi Ishikawa. Curiosity and boredom based on prediction error as novel internal rewards. In *Brain-Inspired Information Technology*, pages 51–55. Springer, 2010.

Jin Zhang, Jianhao Wang, Hao Hu, Tong Chen, Yingfeng Chen, Changjie Fan, and Chongjie Zhang. Metacure: Meta reinforcement learning with empowerment-driven exploration. *arXiv preprint arXiv:2006.08170*, 2020.

Ruihan Zhao, Pieter Abbeel, and Stas Tiomkin. Efficient online estimation of empowerment for reinforcement learning. *arXiv preprint arXiv:2007.07356*, 2020.