Controlled Diversity with Preference : Towards Learning a Diverse Set of Desired Skills

Maxence Hussonnois
A^2I^2, Deakin University
Geelong, Australia
m.hussonnois@deakin.edu.au

Thommen George Karimpanal
A^2I^2, Deakin University
Geelong, Australia
thommen.karimpanalgeorge@deakin.edu.au

Santu Rana
A^2I^2, Deakin University
Geelong, Australia
santu.rana@deakin.edu.au

ABSTRACT

Autonomously learning diverse behaviors without an extrinsic reward signal has been a problem of interest in reinforcement learning. However, the nature of learning in such mechanisms is unconstrained, often resulting in the accumulation of several unusable, unsafe or misaligned skills. In order to avoid such issues and ensure the discovery of safe and human-aligned skills, it is necessary to incorporate humans into the unsupervised training process, which remains a largely unexplored research area. In this work, we propose *Controlled diversity with Preference* (CDP)[1] , a novel, collaborative human-guided mechanism for an agent to learn a set of skills that is diverse as well as desirable. The key principle is to restrict the discovery of skills to those regions that are deemed to be desirable as per a preference model trained using human preference labels on trajectory pairs. We evaluate our approach on 2D navigation and Mujoco environments and demonstrate the ability to discover diverse, yet desirable skills.

KEYWORDS

Skill Diversity; Human Preferences; Reinforcement Learning

ACM Reference Format:
Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. 2023. Controlled Diversity with Preference : Towards Learning a Diverse Set of Desired Skills. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

Deep Reinforcement learning [17] is a powerful computational approach for solving sequential decision making tasks by maximizing prespecified rewards over time. Despite its proven success in a number of applications ranging from Atari games to robotics [15, 17], the framework is typically task-specific, and the effectiveness of the learned policy is contingent on a carefully designed extrinsic reward function.

However, in the real world, an agent is likely to come across complex and unstructured tasks, for which it may need to learn several sub-behaviors or skills, possibly, without access to any extrinsic rewards. In order to autonomously discover and learn these skills, prior works have proposed information theory-based diversity objectives as an intrinsic reward to explore and learn diverse task-agnostic skills without a reward function [4, 6, 21].

[1]See code here: (https://github.com/HussonnoisMaxence/CDP)

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

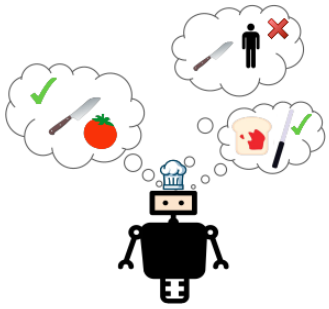


Figure 1: With unconstrained skill discovery, a cooking robot may discover undesirable skills (such as harming humans) using a kitchen knife.

While such unsupervised methods of skill discovery can produce promising results, their unconstrained nature may lead to the acquisition of useless, dangerous, or misaligned skills. For example, as depicted in Figure 1, a robot tasked with learning diverse skills with a kitchen knife may learn undesirable skills such as harming a human. This type of behavior can occur because the agent lacks context about the real world. Without context, the agent views all aspects of the environment as equally relevant, and learns to correlate its skills with any part of the environment regardless of its importance or safety.

In order to address this issue, Eysenbach et al. [6] attempted to limit the diversity of skills by manually selecting features for the agent to be diverse in. However, the effectiveness of this approach is limited, as it is still possible for agents to learn undesirable skills while being diverse about a specific feature. Recent works [11] have suggested relying on expert demonstrations to guide the agent towards expert-visited regions. Such demonstrations are generally expensive and thus would not be available in large quantities, thereby negatively impacting skill diversity. As such, designing online approaches for learning simultaneously diverse and desirable behaviors remains an important and challenging open problem.

In contrast to the approaches mentioned above, in this work, we contend that the agent can learn more desirable skills through guidance provided by humans in the loop during the learning of skills. Using human feedback to infer context allows much greater flexibility, adaptability and less engineering than relying on predefined extrinsic rewards. The key idea behind our approach is that we frame the problem of *controlling skill diversity* as finding regions of the environment where skill discovery will more likely produce desirable skills. Due to the difficulty of identifying such regions without human-provided context, we propose leveraging recent work in learning from human preferences [5, 12, 26] to infer

preferred regions in the environment. Intuitively, these are regions of the environment which are generally associated with favorable agent behaviors. We posit that such regions also correspond to suitable regions for learning a diverse set of skills. Once such regions are identified, we adapt recent exploration methods to direct the agent's exploration towards those preferred regions. Furthermore, by learning a representation of the state space from human preferences, we show that our approach scales to higher dimensional problems and learns skills that are discernibly diverse to human eyes. Thus, by restricting the diversity in skill discovery to human-preferred regions of the environment, we are capable of learning skills that are both diverse and desirable.

In summary, the main contributions of this work are:

- *Controlled diversity with Preference* (CDP), a novel method to control diversity in skill discovery using human preferences.
- Demonstration that our proposed approach guides the agent's exploration towards preferred state regions.
- Learning a representation of the state space for skill discovery, that contains features relevant to human preferences.
- Qualitative and quantitative evaluation of the proposed framework, with suitable comparisons with existing baselines for learning diverse skills.

2 RELATED WORK

Human in the loop and Preference based RL: Human in the loop reinforcement learning (HIL-RL) aims to improve reinforcement learning (RL) agents by using human knowledge. In contrast to imitation learning and inverse RL, HIL-RL uses human knowledge during the training process rather than prior to it.

To enable the use of human feedback for more complex and challenging tasks, Christiano et al. [5] learned a reward model from human preference labels over trajectories. Such preference-based frameworks offer the advantage of relatively easy/intuitive supervision, while being sample efficient enough to quickly learn a reward function.

PEBBLE [12] was an approach that further developed this framework to design a more sample- and feedback-efficient preference-based RL algorithm without any additional supervision. This was achieved by leveraging off-policy learning and utilizing unsupervised pre-training to collect data to substantially improve efficiency. Although we follow a PEBBLE-like approach to learn a reward function from preference labels, our approach differs from PEBBLE in that in addition to learning reward functions from preferences, we use this learned reward function to determine a distribution of states for guiding the agent's exploration. We essentially utilize the learned preference based rewards as a means for determining the dynamic space that humans prefer.

In the context of our proposed approach, Skill Preferences (SkiP) [25] is a related approach that combines skill learning and human preferences. SkiP was shown to learn a reward model over skills with human preferences and used that model to extract human-aligned skills from offline data. In contrast to SkiP, our approach targets the online learning setting, where skills are still under development when we obtain preferences.

Unsupervised Reinforcement learning and Skill discovery: Unsupervised RL is an approach for autonomously learning relevant behavior in any environment based on task-agnostic intrinsic rewards. Intrinsic rewards form the basis for many agent concepts such as curiosity [19] , novelty [16], and empowerment [20]. In contrast to curiosity [19], which guides exploration towards regions where predictive models perform poorly, novelty [16] guides exploration toward areas that are less frequently visited. In order to maximize the agent's future potential, empowerment approaches [20] direct the agent to explore regions that offer it more possible states to visit.

DIAYN [6], VIC [9], and VALOR [1] suggested an empowerment objective based on mutual information. This objective was shown to enable the discovery and acquisition of a variety of skills relevant to complex locomotion. To add predictability to the set of diverse skills, DADS [21] formulated a variation of an objective based on mutual information. EDL [4] showed that such skill discovery methods suffer from poor exploration, and proposed to split the process into three independent phases: exploration, discovery, and learning. In our work, we use the EDL framework, thereby separating the discovery and learning of skills. However, we integrate the discovery of skills within the exploration process by using them to gather data more from the preferred regions.

Despite various advances in the area of autonomous skill discovery, it remains challenging to learn and discover meaningful skills in high-dimensional state spaces due to the curse of dimensionality. Many works have mitigated this problem by learning representations of the state space, to distinguish skills based on more relevant features. Nieto et al. [18] leverages self-supervised learning of state representation techniques such as contrastive techniques to learn a compact latent representation of the states. IBOL [10] proposes a linearization of environments that promotes more diverse and distant state transitions. Unlike these works, we do not learn or change the representation of the state. Instead, we redirect diversity to specific regions of the state space likely associated with meaningful, desirable skills. We note that the aforementioned methods for dealing with high dimensionality remain orthogonal to our work, and could possibly be combined with our proposed framework to realise more scalable solutions.

As far as controlling diversity using human data is concerned, the work by Klemsdal et al. [11] is most closely related to ours. By leveraging prior expert data, they obtain a state projection that makes expert-visited states recognizable and, consequently, encourages skills to visit them. However, in contrast to this approach, our proposed framework does not require access to expert trajectories. It instead only assumes a finite number of human-generated preference labels based on the agent's trajectories. We contend that this type of feedback is relatively easier to collect, with minimal cognitive load on the human collaborator.

Restraining behavior: Several works have aimed at controlling the behavior of agents. For example, Giacomo et al. [7] introduced restraining bolts to restrain agents' behavior by offering additional rewards when logical specifications of desired actions are satisfied. In another direction, Alizadeh Alamdari et al. [2] also augments the reward of the agent to consider the future wellbeing of others and thus restraining its behavior to reduce negative side effects. Our

work differs from these in that, through interaction using human preferences, we learn how to regulate the diversity of skills.

3 PRELIMINARIES

In this paper, we consider the problem of controlling diverse skill discovery by combining the EDL framework with human guidance in the form of preference-based RL. Here, we briefly present related concepts, before describing our method in detail in Section 4.

3.1 Skill Discovery

Consistent with prior work [4] the skill discovery problem is formalized as a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$ without external rewards, $\mathcal{S}$ and $\mathcal{A}$ respectively denote the state and action spaces, and $\mathcal{P}$ is the transition function. Skills introduced by Sutton et al. [22] are temporally extended actions (sub-behavior), which consist of a sequence of primitive actions. We define skills as policies $\pi(a|s, z)$ conditioned on a fixed latent variable $z \in Z$.

Skill discovery methods aim to learn these latent-conditioned policies by maximising the mutual information between $\mathcal{S}$ and Z. Due to symmetry, the corresponding mutual information can be expressed in two forms:

$$I(\mathcal{S}; Z) = \underbrace{H(Z) - H(Z|\mathcal{S})}_{\text{reverse}} = \underbrace{H(\mathcal{S}) - H(\mathcal{S}|Z)}_{\text{forward}} \tag{1}$$

where $I(\cdot;\cdot)$ and $H(\cdot)$ are respectively the mutual information and the Shannon entropy. Following prior work, we refer to these as the reverse and forward forms. By using either of the two forms of the objective, prior works [4, 6, 21] have demonstrated the learning of latent-conditioned policies that execute diverse skills. Our method uses the forward form in Equation (1) to learn the latent-conditioned policies $\pi(a|s, z)$.

3.2 EDL Framework

EDL optimizes the same information theoretic objective in Equation (1), but separates the skill discovery into three distinct stages - exploration, discovery and learning of the skill.

3.2.1 Exploration stage. The Exploration stage aims to collect environment transitions; it can be achieved via any exploration method [14, 16, 19].

3.2.2 Skill Discovery stage. Given a distribution over states $p(s)$, the Skill Discovery stage trains a Vector-Quantized VAE (VQ-VAE) to model the posterior $p(z|s)$ as an encoder $p_\phi(z|s)$, and $p(s|z)$ as the decoder $q_\phi(s|z)$. The VQ-VAE has the advantage of having a discrete bottleneck, which in our case is the categorical distribution of $p(z)$. Typically, VQ-VAEs are trained to optimize for the objective:

$$\begin{aligned}\mathcal{L}^{\text{VQ-VAE}} = \mathbb{E}_{s\sim p(s)}[&\log(q_\phi(s|p_\phi(z|s))) \\ &+\|\text{sg}[z_e(s)] - e\| + \beta\|z_e(s) - \text{sg}[e]\|]\end{aligned} \tag{2}$$

where $z_e(s)$ and e are respectively the codebook vector and the codebook index, and $sg[.]$ is the operation 'stop gradient'. For more details, we refer the reader to van den Oord et al. [24].

3.2.3 Skill Learning stage. Finally, the Skill Learning stage consists of training the latent-conditioned policies $\pi_\theta(a|s, z)$ that maximize the forward form of the mutual information (Equation (1)) between states and latent variables. The corresponding reward function is then defined by:

$$r(s, z) = \log q_\psi(s|z), z \sim p(z) \tag{3}$$

where $q_\psi(s|z)$ is given by the decoder of the VQ-VAE at the discovery stage. This reward function reinforces the policy to visit states that the decoder generates for each latent variable z. Our proposed method builds upon the EDL Framework, although we enhance it via two novel contributions: (1) an exploration phase guided by preferences, which integrates with the skill discovery phase to improve coverage relevance, and (2) a way to transform $p(s)$ into a more suitable distribution for discovering desirable skills.

3.3 Reward Learning from Preferences

In this work, we use preference-based RL to identify preferred regions, which are then used to constrain the diversity of learned skills. In preference-based RL, a human is presented with two trajectory segments (state-action sequences) σ^i and σ^j, and is asked to indicate their preference y for one over the other. For instance, the label $y = (1, 0)$ would imply that the first segment is preferred over the second.

We follow the same framework as prior works in preference-based RL [5, 12, 26], where the aim is to model the human's internal reward function responsible for the indicated preferences. This is usually done via the Bradley-Terry model [3], which models a preference predictor using the reward function $\hat{r}_\psi$ as follows:

$$P_\psi[\sigma^i \succ \sigma^j] = \frac{\exp(\sum_t \hat{r}_\psi(s_t^i, a_t^i))}{\sum_{j\in\{0,1\}} \exp\left(\sum_t \hat{r}_\psi(s_t^j, a_t^j)\right)} \tag{4}$$

where $\sigma^i \succ \sigma^j$ denotes the event that the segment σ^i is preferable to the segment σ^j. As in Lee et al. [12], we model the reward function as a neural network with parameters ψ, which is updated by minimizing the following loss:

$$\begin{aligned}\mathcal{L}^{\text{Reward}} = -\mathbb{E}_{(\sigma^0,\sigma^1,y)\sim\mathcal{D}}[&y(0)\log P_\psi[\sigma^0 \succ \sigma^1] + \\ &y(1)\log P_\psi[\sigma^1 \succ \sigma^0]]\end{aligned} \tag{5}$$

In the current work, the above framework is used to infer context regarding the importance of each region of the environment by estimating the human's reward function $\hat{r}_\psi$ from preferences labels y. Specifically, we use these rewards to identify regions associated with favorable agent behaviors. For simplicity, we choose to work with the trajectory represented as state sequences rather than state-action sequences as introduced.

4 METHODS

In this section, we present **CDP** (**C**ontrolled **D**iversity with **P**reference), a skill discovery method that utilizes preference-based RL methods to control diversity and discover more preferred skills based on human feedback. Our main idea is that with the reward learned from human preference feedback, we can estimate a region of the state space where it is more likely for the agent to discover desirable skills and subsequently learn them. To this end, we introduce the concept of controlled diversity and preferred regions. Then,

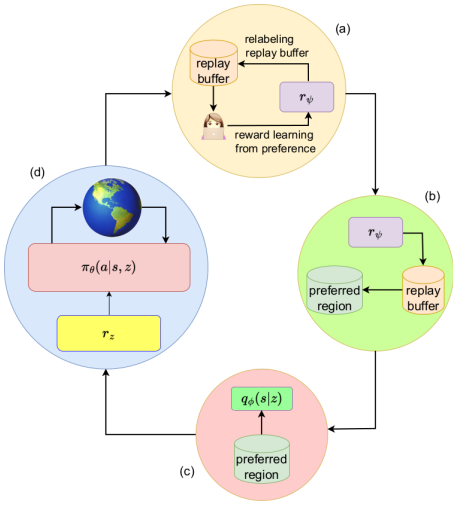


Figure 2: Illustration of the guided exploration process. The agent iterates through four steps to explore. First it learns a reward from human preferences (b) so that it can update its belief over the preferred region from the existing data in the buffer (c) then it discovers skills in this region (d) and finally it collects experience regarding the beliefs of the preferred region.

we present how to integrate them with EDL for a more efficient exploration of the preferred region.

4.1 Influencing Skill Discovery with Human Feedback

4.1.1 How to control diversity? We define controlled diversity as limiting diversity to a certain region of the state space. It differs from the standard setting for skill-discovery problems, where diversity is applied to the entire state space in an unconstrained manner. To achieve this, we follow the EDL framework, where we encourage skill discovery towards targeted behaviors by modifying priors through a distribution over a target region $p^*(s)$ of the state space. Performing skill discovery on $p^*(s)$ assigns latent variable z to regions within the target region. In other words, a carefully designed target region containing only desirable skills will enable us to correlate Z only with desirable skills.

It is however, difficult to design such a region of the state space or to gain direct access to it. Thus, we formulate our problem of 'controlled diversity' as finding an approximation of this target region. In this work, we identify such regions through their high preference rewards, learned from human preferences.

4.1.2 Preferred regions. We define a preferred region as those regions associated with high estimated preference rewards $\hat{r}_\psi$, where $\hat{r}_\psi$ is learnt using the preference based RL framework, as described in Section 3.3. Concretely, a preferred state region $\hat{S} \subseteq S$ is a region of the state space S where $\hat{r}_\psi(s) \geq \beta$, where $\beta \in [0, 1]$ is a preference reward threshold, and $\hat{r}_\psi(s)$ is normalised to be within the range $[0, 1]$. That is,

$$\hat{S} = \{\forall s \in S | \hat{r}_\psi(s) \geq \beta\} \tag{6}$$

Ideally, the preferred region would be aligned with the intended skills if the state space was fully explored. However, the assumption of full state coverage may not be realistic. Alternatively, we iteratively build a more accurate preference model by first using the current estimate of the preference model to sample more trajectories from the highly-preferred regions, and then updating the preference model with human labels on those trajectories. This directed sampling makes our method more query-efficient.

4.2 Exploration Towards a Preferred Region

In this section, we adapt the exploration phase of EDL to explore preferred regions more effectively. To this end, we add three components to the exploration phase - We learn a reward from human preferences, we estimate the potential preferred regions and we discover skills based on the potential preferred regions. Formally, we consider latent-conditioned policies $\pi(a|s, z)$, a reward function $\hat{r}_\psi$, a preferred state region $\hat{S}$ and a discriminator $q_\psi(s|z)$, which are updated by the following processes, as illustrated in Figure (2):

- Step (a): Reward Learning - We query human for preference over trajectories and update a reward function $\hat{r}_\psi$ from the preferences.
- Step (b): Preferred regions estimations - We update our belief about the preferred subset $\hat{S}$ as described in Section 4.1.2.
- Step (c): Discovery - We train the discriminator $q_\psi(s|z)$, following VQ-VAE training, based on the most recent belief about the preferred subset.
- Step (d): Exploration - We train latent-conditioned policies $\pi(a|s, z)$ using guided intrinsic motivation to explore and collect diverse experiences.

In the following sections, we explain how these components can be integrated into existing exploration methods to guide exploration towards preferred regions.

4.2.1 State Marginal Matching. We base the exploration phase of our work on SMM (State Marginal Matching [13]), although our approach is not limited to this method. SMM aims to learn a state marginal distribution $\log \rho_{\pi_z}(s)$ to match a given target distribution $p^*(s)$ by minimising their Kullback-Leibler (KL) divergence. Additionally, to explore more efficiently, Lee et al. [13] proposed to learn latent-conditioned policies $\pi(a|s, z)$ by adding the diversity objective from Eysenbach et al. [6]. Thus, the reward function is defined as:

$$r_z(s) = r_z^{\text{exploration}}(s) + r_z^{\text{diversity}}(s) \tag{7}$$

where :

$$r_z^{\text{exploration}}(s) = \underbrace{\log p^*(s)}_{(a)} - \underbrace{\log \rho_{\pi_z}(s)}_{(b)} \tag{8}$$

$$r_z^{\text{diversity}}(s) = \underbrace{\log q_\phi(z|s)}_{(c)} - \underbrace{\log(p(z))}_{(d)} \tag{9}$$

Intuitively, according to Lee et al. [13], the above equations imply that the agent should go to states (a) with high probability under the target state distribution, (b) where this agent has not been before, and (c) where this skill is clearly distinguishable from other skills.

The last term (d) encourages exploration in the space of mixture components z.

4.2.2 Adding reward from preferences. In order to direct the exploration towards preferred regions, we use $\hat{r}_\psi$ as the target distribution $p^*(s)$ in Equation (8) to motivate the agent to explore regions with high preference-based rewards. Therefore Equation (8) can be rewritten as:

$$r_z^{\text{exploration}}(s,a) = \hat{r}_\psi(s) - \log \rho_{\pi_z}(s) \tag{10}$$

4.2.3 Adding preferred regions and skills discovery. Following the definition of preferred region and skill discovery described in Section 3.2.2, we define a potential preferred region $\hat{S}$ with $\hat{r}_\psi$ according to (Equation (6)) on states collected online, and use it to train a discriminator q_ϕ presented in Section 3.2.2. The discriminator q_ϕ, encourages each skill to explore distinct regions related to the potential preferred region. In other words, we incentivize the agent to learn diverse skills within the preferred region. Therefore, we define the diversity reward as:

$$r_z^{\text{diversity}} = \log q_\phi(\hat{s}|z), \text{ with } \hat{s} \in \hat{S} \tag{11}$$

4.2.4 Overall objective. By combining each of the different reward components mentioned, the overall reward function to enable exploration towards preferred regions is given by:

$$r_z(s,a) = \underbrace{\hat{r}_\psi(s)}_{(a)} \underbrace{- \log \rho_{\pi_z}(s)}_{(b)} + \underbrace{\log q_\phi(\hat{s}|z)}_{(c)} \tag{12}$$

Intuitively, Equation (12) implies that the agent should go to (a) states with high preference rewards (b) states where the agent has not been before, and (c) to distinct regions within potential preferred regions. Our overall guided exploration method is described in Algorithm 1.

4.2.5 Learning skills. By following the previous objective in Equation (12) we explore the preferred region and train a discriminator q_ϕ to assign diverse regions of the preferred region to skills. We then use the discriminator q_ϕ to learn skills in the skill learning phase, as described in Section 3.2.3.

4.3 Preferred Latent Representation

Despite being able to restrict diversity to specific regions of the environment, skills discovered in the state space might not appear diverse from a human point of view. The state space in MuJoCo [23] environments, for example, is a concatenation of joint positions and velocities. Discovering skills in this space often results in static positions. Even though easily distinguishable by the discriminator, they may seem similar to the human eye. Hence, as recommended by Eysenbach et al. [6], we examine using prior knowledge to identify discernably diverse skills.

This prior can be represented as any function of the state space and used as a prior to condition the discriminator. In this case, the discriminator is defined as $q_\phi(f(s)|z)$ with $f(s)$ being the prior.

Although it can be useful to encourage the learning of specific types of skills by specifying a prior, relying on specifically designed priors may be limiting. Thus, we present an alternative to manually specifying this prior to learn skills that are more discernably diverse to human eyes. Specifically, we simply use the representation in the last hidden layer of the reward model $\hat{r}_\psi$ learnt from human preferences as the prior. The intuition is that the last hidden layer of the neural network that models the internal reward function of a human, should learn a latent state representation that captures features that matter for human preferences. We refer to this as the preferred latent representation.

Formally we can write $\hat{r}_\psi(s)$ as:

$$\hat{r}_\psi(s) = h_\psi(f_\psi(s)) \tag{13}$$

where f_ψ, represents all layers of the reward model all layers except the output layer and h_ψ is the output layer of the neural network. Hence, we define the discriminator in Equation (11) as $q_\phi(f_\psi(\hat{s})|z)$.

As depicted later in the experiments, this general approach for specifying priors achieves discernably diverse behaviors, while obviating the need for any additional training.

Algorithm 1: Guided exploration with preferences

Initialize $\mathcal{B}$, π_z, r_ψ, q_ϕ ;
foreach *timestep* **do**
  Sample $z \sim p(z)$;
  // Collect data ;
  for *each timestep t* **do**
    Sample action $a_t \sim \pi_\theta(a_t|s_t, z)$;
    Step environment $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$;
    Set reward $r_z(s)$ as in (12);
    Update policy (θ) to maximise r_z with SAC [8];
    Store transitions $\mathcal{B} \leftarrow \mathcal{B} \cup \{(s_t, a_t, s_{t+1}, r_z)\}$;
  end for
  if *it's time to update the preference* **then**
    // Query instructor ;
    foreach *query to instructor* **do**
      Sample $(\sigma^0, \sigma^1) \sim \mathcal{B}$;
      Collect preference from instructor $y = \sigma^0 \succ \sigma^1$;
      Store transitions $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\sigma^0, \sigma^1, y)\}$
    end foreach
    // Update reward model ;
    foreach *each gradient step* **do**
      Sample minibatch $(\sigma^0, \sigma^1, y)_{j=1}^{\mathcal{D}} \sim \mathcal{D}$;
      Optimize $\mathcal{L}^{\text{reward}}$ in (5) with respect to ψ
    end foreach
    // Estimate the preferred region ;
    $\hat{S} = \{\forall s \in \mathcal{B} | \hat{r}_\psi(s) \geq \beta\}$;
    // Skill Discovery phase ;
    foreach *query to instructor* **do**
      Sample minibatch $(s)_{j=1}^{\hat{S}} \sim \hat{S}$;
      Optimize $\mathcal{L}^{\text{VQ-VAE}}$ in (2) with respect to ϕ
    end foreach
  end if
end foreach

5 EXPERIMENTS

In this section, we examine our proposed method to control diversity with preference and to guide the agent's exploration towards preferred regions. We first demonstrate our approach on a 2D navigation environment, following which we also show the performance of our method in higher dimensional environments such as MuJoCo in Section 5.3 and 5.4. The 2D environment consists of a two-dimensional room enclosed by walls that restrain the agent. The agent begins each episode in the middle of the room, until episode termination, which occurs after 100 steps. The agent has only access to its horizontal and vertical coordinates (X,Y). It can deterministically change its direction and amplitude of steps to freely move in the environment. Both state space and action space are continuous. Following previous work on preference-based RL, we simulate human preference with an oracle 'true' reward function. The true reward function is designed to be a gaussian distribution, centered around a goal position, and the reward is computed as the negative distance to the goal.

5.1 Results in 2D navigation

In the 2D navigation environment, we intend to demonstrate that a preferred region can be used to define a relevant area of interest. In the interest of studying the effectiveness of preferred regions for discovering skills, in this section, we assume an ideal state coverage and an oracle reward function. We show both EDL and CDP results to demonstrate the full impact of the preferred region.

By applying the definition of the preferred region described in Section 4.1.2 to the assumed state coverage in Figure 3a, we identify the preferred region in the top right corner, as indicated in Figure 3b. We then discover and learn skills in those proposed regions. As illustrated in Figure 3c and 3e, EDL discovers and learns skills uniformly across the environment. In our case (CDP), the discriminator concentrates all skills' assigned regions in the top right corner, as shown in Figure 3d. Additionally, in Figure 3d, centroids (the most likely state under the discriminator for each skill) are located in the top corner, resulting in skills moving to the top right corner as illustrated in Figure 3f.

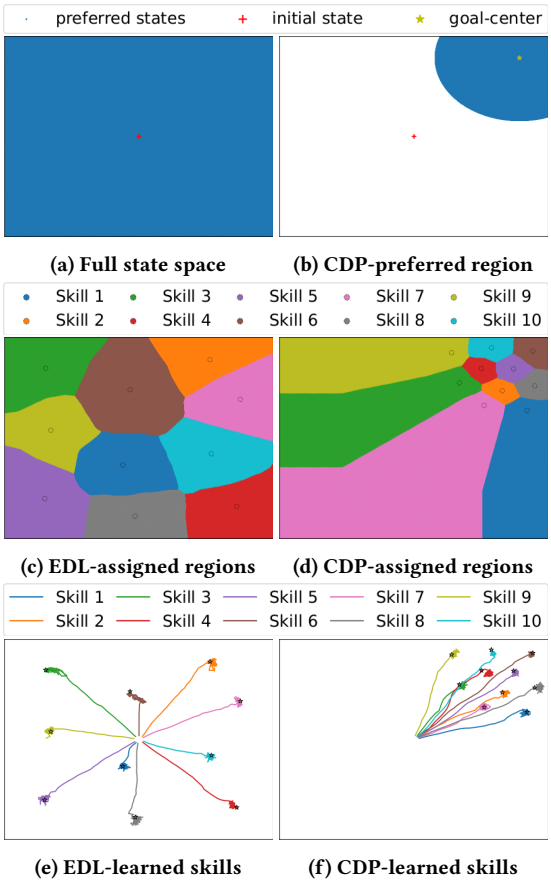


Figure 3: (a), (c) and (e) are respectively the full state space, the regions assigned to each skill by the discriminator trained on the full state space and the skills learned with this discriminator. (b), (d) and (f) are respectively the preferred regions of the state space obtained by our method, the regions assigned to each skill by the discriminator trained on the preferred region and the skills learned with this discriminator.

5.2 Exploration of the Preferred Region

This section aims to demonstrate that the modifications we made to the SMM method in Section 4.2 have significant advantages with regards to exploring the preferred region. We place ourselves in more realistic settings where we don't have full state coverage, or have access to the oracle reward. We compare our method with SMM as described in EDL and SMM+prior that uses the same prior as us. The prior is a reward function learned from preference, and used as described in Section 4.2.2.

As illustrated in Figure 4, we compared each method in terms of their average returns as per the target reward function. Intuitively, exploring more of the preferred region should result in a higher return. Results in Figure 4 suggest that our proposed method visits more states with higher rewards than the other methods, which implies that it explores the preferred region more efficiently.

From a qualitative perspective, Figure 5 depicts the states visited by each method and shows that our method visits more states in the top right corner (preferred region). Further, the presence of

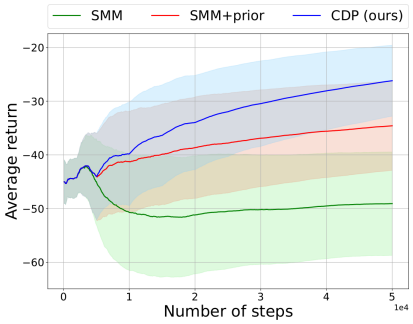
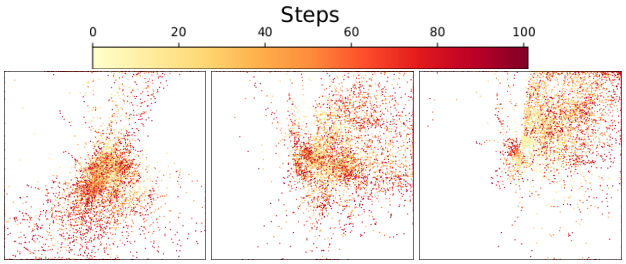


Figure 4: Average return achieved by each method.

darker shades (indicating the later stages of interaction) in the top right corner indicates that the skills from our method tend to end near or in the preferred region. This can be explained by the

discriminator incentivizing the agent to learn diverse skills within the preferred region. This is in contrast to the other methods in which the discriminator only encourages agents to acquire diverse skills, as indicated by the darker points in Figure 5a and 5b being relatively more evenly distributed in different state regions, and not particularly within the preferred region.

The comparisons with the first method are unfair since they do not have access to any information about the preferred region. In spite of this, we still feel that the comparison is relevant to emphasize the choice to use human preference to control diversity.



(a) SMM **(b) SMM+prior** **(c) CDP (ours)**

Figure 5: States visited by each of the methods, SMM (a), SMM+prior(b), ours(c).

5.3 Results using Preferred Latent Representations

In this section, we demonstrate that preferred latent representations facilitate the acquisition of appropriate skills in a general manner, that are capable of scaling to larger state and action spaces. To this end, we performed experiments on a MuJoCo-based modified Half Cheetah agent, in which moving backwards was preferred. It was specified by using a version of the original half-cheetah reward function modified (by multiplying the original reward by -1) to encourage the agent to move backwards as far as possible along the horizontal axis. In other words, we aim to achieve diverse velocities corresponding to the desired behavior of moving backwards.

As shown in Figure 7a, without additional prior knowledge about the state space, the agent does not learn any relevant skills. However, when using the preferred latent representation (Figure 7b), the agent is able to learn diverse skills that go backward at varying speeds similar to skills learned while using a manually-specified prior (the agent's velocity) over velocity (Figure 7c).

Additionally, we repeat the 2D navigation experiments from Section 5.1, but using preferred latent representations to learn diverse and desirable skills. As seen in Figure 6 the agent learns skills comparable to those in Section 5.1. We note that the trajectories in Figure 6 are relatively more noisy when compared to those in Figure 3f, probably due to the inherent noise associated with learning the preferred latent representations. However, the fact that preferred latent representations also enable the agent to learn the intended skills implies that they do indeed capture relevant features of the state space, be it in the navigation task, or the more complex backwards Half Cheetah environment.

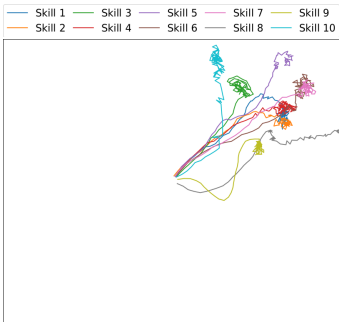
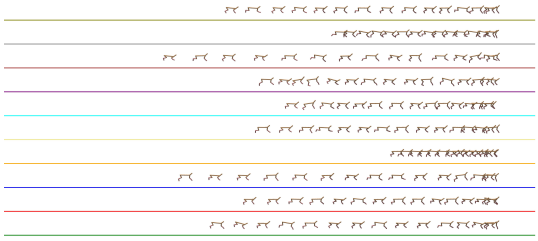
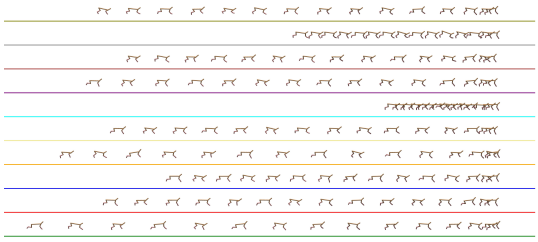


Figure 6: Skill learned using the preferred latent representation as a prior to discover skills.

(a) Modified Half cheetah's skill learned using the state space to discover skills



(b) Modified Half cheetah's skill using the preferred latent representation to discover skills



(c) Modified Half cheetah's skill using a manually specified prior over velocity to discover skills

Direction of motion

Figure 7: Modified Half cheetah's skill learned using different representations of the state space to discover skills

5.4 Effect of β

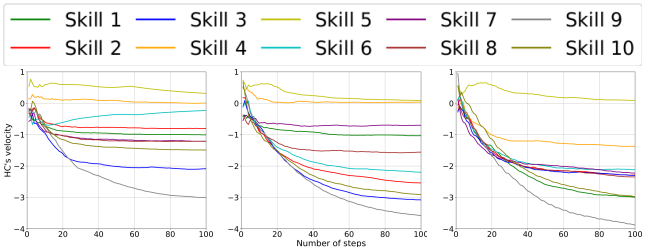
Here, we examine the effect of varying β (used in Equation (6)) on the resulting skills obtained. We show results for both the MuJoCo-based modified Half Cheetah and 2D navigation experiments. β

determines how much emphasis is placed on skill discovery centered around high rewards. This can be viewed as a parameter that controls how much we exploit the reward function to constrain skill discovery. In Figure 8, we use the setting described in Section 5.1 to show that a low β will produce skills that may be far away from the goal, while a high β will learn skills around the goal.

(a) β=0.1 **(b) β=0.5** **(c) β=0.9**

Figure 8: Regions assigned to each skill by the discriminator trained on preferred regions set by different values of β.

Similarly, in Figure 11, a low β results in skills that only cover shorter distances, as these are easier to learn. On the other hand, an agent that exploits the reward (high β) learns skills that cover larger distances. However, high β values may cause the agent to be overly exploitative, leading to a lower diversity of learned skills. This phenomenon is illustrated in Figures 9 and 10 which show that the variance of velocity across skills is relatively low for both high and low values of β, while it is the highest for the intermediate value of $\beta = 0.5$. Hence, a user favoring a more uniform distribution of skills might choose a more balanced β of 0.5, while one favoring skills more relevant to the task should select a relatively high β.



(a) β=0.3 **(b) β=0.5** **(c) β=0.9**

Figure 9: Average velocity over time for each skill.

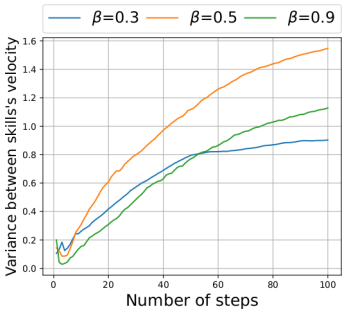
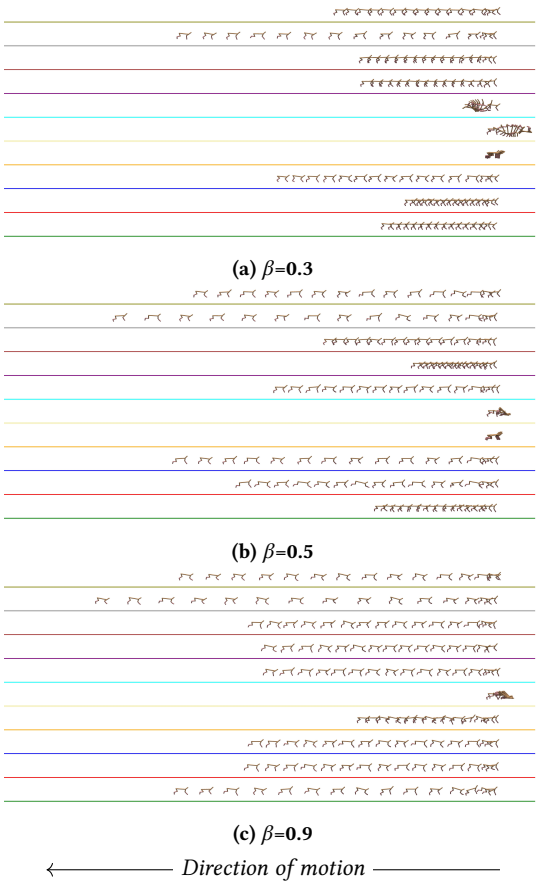


Figure 10: Comparaison of the variance between skill's velocity over time for each of the β values.



(a) β=0.3

(b) β=0.5

(c) β=0.9

Figure 11: Modified Half cheetah skills learned with different β values.

6 CONCLUSION

We introduced a novel approach for addressing the issue of under-constrained skill discovery. Our proposed approach, Controlled diversity with preference (CDP) was designed to leverage human feedback to identify human-preferred regions, following which we discovered diverse skills within those regions, thereby ensuring the learning of diverse and desirable skills. In addition, we show that our method can be used to guide exploration towards possible preferred regions. We validated our proposed approach in 2D navigation and Mujoco environments. Empirically, our agents demonstrated the ability to favor the exploration of the preferred regions and to learn diverse skills in these regions. We also empirically studied the effect of the user-controlled hyperparameter β to demonstrate its effects on the diversity of learned skills. As such, we believe that our approach presents a way to control the autonomous discovery of skills, while still obtaining safe, aligned and desirable skills.

REFERENCES

[1] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. 2018. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299* (2018).

[2] Parand Alizadeh Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2022. Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 18–26.

[3] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39 (1952), 324.

[4] Víctor Campos, Alexander Trott, Caiming Xiong, Richard Socher, Xavier Giro i Nieto, and Jordi Torres. 2020. Explore, Discover and Learn: Unsupervised Discovery of State-Covering Skills. In *ICML*.

[5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

[6] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is All You Need: Learning Diverse Skills without a Reward Function. (2018).

[7] Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, and Fabio Patrizi. 2018. Foundations for Restraining Bolts: Reinforcement Learning with LTLf/LDLf Restraining Specifications. In *International Conference on Automated Planning and Scheduling*.

[8] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1861–1870. https://proceedings.mlr.press/v80/haarnoja18b.html

[9] Danilo Rezende Karol Gregor and Daan Wierstra. 2016. Variational Intrinsic Control. *International Conference on Robotic Learning* 0, 0 (2016), 0.

[10] Jaekyeom Kim, Seohong Park, and Gunhee Kim. 2021. Unsupervised Skill Discovery with Bottleneck Option Learning. In *ICML*.

[11] Even Klemsdal, Sverre Herland, and Abdulmajid Murad. 2021. Learning Task Agnostic Skills with Data-driven Guidance. *arXiv preprint arXiv:2108.01869* (2021).

[12] Kimin Lee, Laura Smith, and Pieter Abbeel. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. *International Conference on Machine Learning* (2021).

[13] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. 2019. Efficient Exploration via State Marginal Matching. (2019).

[14] Youngwoon Lee, Jingyun Yang, and Joseph J. Lim. 2020. Learning to Coordinate Manipulation Skills via Skill Behavior Diversification. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ryxB2lBtvH

[15] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *CoRR* abs/1509.02971 (2016).

[16] Hao Liu and Pieter Abbeel. 2021. Behavior From the Void: Unsupervised Active Pre-Training. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 18459–18473. https://proceedings.neurips.cc/paper/2021/file/99bf3d153d4bf67d640051a1af322505-Paper.pdf

[17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.

[18] Juan José Nieto, Roger Creus, and Xavier Giro-i Nieto. 2021. Unsupervised Skill-Discovery and Skill-Learning in Minecraft. *arXiv preprint arXiv:2107.08398* (2021).

[19] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-Driven Exploration by Self-Supervised Prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), 488–489.

[20] Christoph Salge, Cornelius Glackin, and Daniel Polani. 2013. Empowerment - an Introduction. *ArXiv* abs/1310.1863 (2013).

[21] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. 2020. Dynamics-Aware Unsupervised Discovery of Skills. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HJgLZR4KvH

[22] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1 (1999), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

[23] Emanuel Todorov, Tom Erez, and Yuval Tassa. 2012. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 5026–5033. https://doi.org/10.1109/IROS.2012.6386109

[24] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

[25] Xiaofei Wang, Kimin Lee, Kourosh Hakhamaneshi, Pieter Abbeel, and Michael Laskin. 2022. Skill preferences: Learning to extract and execute robotic skills from human feedback. In *Conference on Robot Learning*. PMLR, 1259–1268.

[26] Aaron Wilson, Alan Fern, and Prasad Tadepalli. 2012. A Bayesian Approach for Policy Learning from Trajectory Preference Queries. In *NIPS*.