

# Human-Aligned Skill Discovery: Balancing Behaviour Exploration and Alignment

Maxence Hussonnois

$A^2I^2$ , Deakin University  
Geelong, Australia

m.hussonnois@deakin.edu.au

Thommen George Karimpanal

School of IT, Deakin University  
Geelong, Australia

thommen.karimpanalgeorge@deakin.edu.au

Santu Rana

$A^2I^2$ , Deakin University  
Geelong, Australia

santu.rana@deakin.edu.au

## ABSTRACT

Unsupervised skill discovery in Reinforcement Learning aims to mimic humans' ability to autonomously discover diverse behaviours. However, existing methods are often unconstrained, making it difficult to find useful skills, especially in complex environments, where discovered skills are frequently unsafe or impractical. We address this issue by proposing *Human-aligned Skill Discovery - HaSD*<sup>1</sup>, a framework that incorporates human feedback to discover safer, more aligned skills. HaSD simultaneously optimises skill diversity and alignment with human values. This approach ensures that alignment is maintained throughout the skill discovery process, eliminating the inefficiencies associated with exploring unaligned skills. We demonstrate its effectiveness in both 2D navigation and SafetyGymnasium environments, showing that HaSD discovers diverse, human-aligned skills that are safe and useful for downstream tasks. Finally, we extend HaSD by learning a range of configurable skills with varying degrees of diversity-alignment trade-offs that could be useful in practical scenarios.

## KEYWORDS

Skill Diversity; Human Preferences; Reinforcement Learning

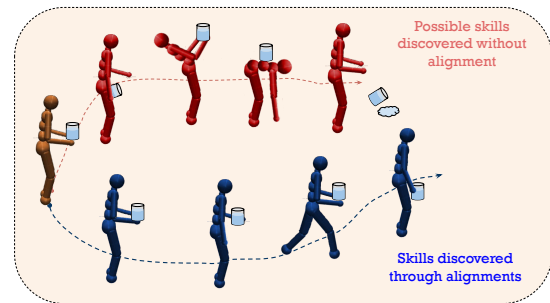
### ACM Reference Format:

Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. 2025. Human-Aligned Skill Discovery: Balancing Behaviour Exploration and Alignment. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, Detroit, Michigan, USA, May 19 – 23, 2025, IFAAMAS, 9 pages.

## 1 INTRODUCTION

Deep reinforcement learning [21] aims to solve sequential decision-making problems by maximising pre-specified rewards over time. Despite its proven success in a number of applications ranging from Atari games to robotics [19, 21], the framework is typically task-specific, resulting in agents with poor generalisation abilities.

Humans, on the other hand, can autonomously discover diverse and complex skills that can be combined later for better generalisation. In line with this objective, Unsupervised Skill Discovery (USD) methods [4, 7, 24, 28] aim to learn a library of policies (skills) driven by an intrinsic reward. In addition to locomotion and manipulation



**Figure 1: Without an alignment signal, discovering desirable skills in complex environments is like searching for a needle in a haystack, often leading to skills that achieve tasks in undesirable ways, such as carrying a glass of water awkwardly (red robots). Aligning skills during discovery ensures they meet human preferences (blue robots).**

tasks [25], these methods have also demonstrated promising results with pixel-based agents [26].

However, to discover dynamic behaviours these methods rely on correlating skills with changes in the environment regardless of the underlying safety or desirability of these changes. As illustrated in Figure 1, the same skill, that of delivering a glass of water, can have very different degrees of desirability. Without an alignment objective, although this task may be successfully completed, say, by the humanoid carrying the glass on its back, such skills are in general, not desirable, nor expected. This suggests a need to constrain skill discovery with alignment signals from human.

Recent works [15] have suggested relying on expert demonstrations to guide the agent. Such demonstrations are generally expensive and thus not available in large quantities, negatively impacting skill diversity. More generally, Kim et al. [14] introduces the problem of safety-aware skill discovery (SASD), which aims to discover diverse skills that satisfy user-predefined safety constraints. However, these methods are limited to a user's domain knowledge. Therefore, they cannot adapt to unforeseen unsafe scenarios that the agent may discover. To enable such adaptation, we advocate for the development of an online approach, where a human oversees training while providing necessary feedback.

In this regard, Human-in-the-loop reinforcement learning [34], in which humans are integrated with the agents' training to provide meaningful feedback, has been able to tackle problems such as safe reinforcement learning [27] and reward engineering [6] while being practical, scalable and efficient [18]. However, its incorporation

<sup>1</sup>See code here: [link to github](https://github.com)



This work is licensed under a Creative Commons Attribution International 4.0 License.

into unsupervised skill discovery methods remains mostly under-explored. In this context, Hussonnois et al. [10] framed the problem of controlling skill diversity as restraining goal-based skills to a certain region of the environment deemed more desirable by human preference. However, an agent may reach a desired goal in an unaligned manner as illustrated in Figure 1. This can be remedied only by aligning the entire behaviour. To address this issue, we formulate the problem of discovering diverse skills with aligned behaviour as a multi-objective problem, where the primary components are a skill discovery objective and an alignment objective. We call this framework Human-aligned Skill Discovery (HaSD). Additionally, we propose a method to optimise this objective that maximises the combination of a skill discovery reward and an alignment reward learned with the Preference-Based RL [6] framework. As human preferences evolve in response to the discovery of more diverse skills, it enables the discovery of increasingly more complex and aligned skills. Furthermore, we propose  $\alpha$ -HaSD, a more general framework that addresses the problem of balancing both rewards by conditioning the skills on the diversity-alignment trade-off variable  $\alpha$ . By doing so,  $\alpha$ -HaSD can produce a range of skills corresponding to varying degrees of diversity-alignment trade-offs. In summary, the main contributions of this work are:

- *Human-aligned Skill Discovery* (HaSD), a novel framework to discover diverse and aligned skills.
- *Configurable Human-Aligned Skills* ( $\alpha$ -HaSD), to discover diverse and aligned skills that are conditioned on the diversity-alignment trade-off.
- Qualitative and quantitative evaluation of the proposed methods, with suitable comparisons with existing baselines for learning diverse and aligned skills.

## 2 RELATED WORK

**Unsupervised Skill Discovery.** In unsupervised skill discovery approaches, the goal is to explore the space of learnable behaviours to acquire a set of useful and diversified temporally extended actions or skills [30] without a reward function. DIAYN [7], VIC [13], VALOR [1] and DADS [28] suggested to maximise mutual information between states and skills. By maximising skill distinguishability, they learned diverse skills in locomotion environments. On the other hand, these methods learn to maximise the mutual information between state and skills with only small state variations, resulting in mostly static skills. Thus Park et al. [24] proposed to maximise state variations with novel distance-maximising skill discovery objectives that learn more dynamic skills in locomotion tasks [24], manipulation tasks [25] and pixel-based environments [26]. However, skill discovery methods do not take into account the underlying context of the environment. This results in an inefficient discovery process that leads to unsafe and unusable skills. Our proposed method addresses this issue by integrating Unsupervised Skill Discovery methods with alignment techniques such as preference-based RL [6]. In recent work, Kim et al. [14] examined safety-aware skill discovery, which focused on finding inherently safe skills. Specifically, they proposed regularising skill discovery using a safety critic [29] that learns from any user-defined safety constraints. In contrast, our work focuses on discovering skills that align with human values learned during training.

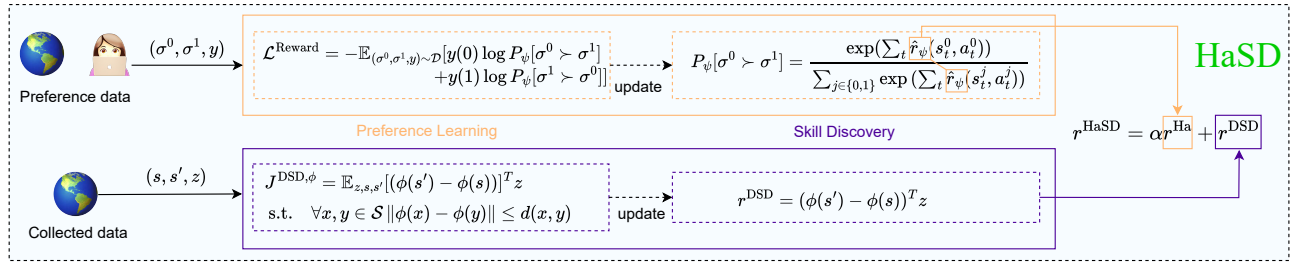
**Human-in-the-Loop and Preference Based RL.** Human in the loop reinforcement learning (HIL-RL) methods focus on learning through feedback from humans during training. In this regard, Preference-based RL [6] uses human preferences over agents' trajectories to infer a reward model and train an agent with it. Since human preferences are expensive, PEBBLE [18] aimed to mitigate this via improved sample and feedback efficiency by leveraging exploration methods and off-policy learning. SURF [23] and REED [20] further improved feedback efficiency by using supervised learning techniques and self-supervised representation learning. We follow this line of work to align skills with human preference. However, the settings differ in that the agent also optimises a skill discovery reward, which reduces the need for pre-training phases.

**Unsupervised Skills Discovery with Preferences.** Unsupervised Skill Discovery with Preferences aims to learn more desirable skills which is still an under-explored area of research. In this regard, Skill Preferences (SkiP) [31] used preferences over an offline dataset to extract relevant human-aligned skills. Alternatively, CDP [10] uses preferences in an online setting to first identify rewarding regions of the state space and then learn goal-based skills within those regions. However aligning goals of goal-based skills do not ensure that the entire trajectory of the skills is aligned. To overcome this limitation, we align the entire skill's trajectory by optimising both skill discovery and alignment (via preferences) objectives.

**Quality-Diversity Policy Optimisation.** Quality-Diversity Policy Optimisation [5, 16, 33] focuses on discovering multiple strategies for solving a given task. Generally, these methods are based on a sophisticated combination of diversity reward and task reward from the environment. In SMERL [16] the optimal policies are diversified by adding the DIAYN diversity reward to transitions along trajectories that produce a known optimal return. Alternatively, DGPO [5] and DOMINO [33] achieve better results by alternately constraining the diversity of the strategies while simultaneously constraining the extrinsic reward. Similar to our work, these works combine an extrinsic reward with a diversity reward. However, their motivation and approach differ fundamentally from ours, as they first seek to optimise performance over task reward, then diversify those performances, whereas we first diversify policies by discovering skills, then adjust them to fit desirability through human preference. Our approach is set up as such, as unlike the former methods, we do not assume direct access to the task rewards. We thus tackle a more challenging setting where an agent must simultaneously infer this knowledge from human feedback.

## 3 APPROACH

To discover diverse and complex skills that are more aligned with human values, agents must be able to adjust the skill discovery process based on human feedback. To this end, we consider the problem of discovering diverse skills that satisfy humans' preferences. We first detail the problem settings and the resulting Human-aligned Skill Discovery (HaSD) Objective. We then present the details of the optimisation in Section 3.5. Furthermore, in Section 3.6 we propose an extension of HaSD by learning configurable skills that can produce a range of diversity-alignment trade-offs.



**Figure 2: Illustration of the HaSD reward components. Skill Discovery rewards are computed using the Distance-Maximising Skill Discovery (DSD) objective and data collected from interaction with the environment. The reward encourages skills to be more dynamic and diverse. Then, we add a  $r_{Ha}$  human-aligned reward learned with preference learning through data collected from interaction with the environment and human preferences. This reward encourages skills to align with human preferences.**

### 3.1 Problem Settings

We consider a Markov Decision Process (MDP) without a reward function, defined as  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$ , where  $\mathcal{S}$  and  $\mathcal{A}$  respectively denote the state and action spaces, and  $\mathcal{P}$  is the transition function governing the agent-environment interaction dynamics. Consistent with prior work [4, 25], we also consider Skills as policies  $\pi(a|s, z)$  conditioned on latent vector  $z \in \mathcal{Z}$ . Finally, we assume that there exists a human in the loop who has an opinion on the desirability of the agent’s behaviour. The human communicates this opinion during training to the agent using a set of preferences  $\zeta$ . We aim to learn a set of diverse, useful behaviours  $\pi(a|s, z)$  that align with human preferences  $\zeta$ .

### 3.2 Human-aligned Skill Discovery Objective

First, for  $\pi(a|s, z)$  to learn diverse skills, we follow previous skill discovery approaches that rely on maximising the mutual information between  $\mathcal{S}$  and  $\mathcal{Z}$ , denoted as  $I(\mathcal{S}; \mathcal{Z})$ . Then, to align skills with human values, we also want to maximise the probability  $Pr$  of realising trajectories with the policies  $\pi(a|s, z)$  that satisfy the set of preferences  $\zeta$ . Thus we propose the following novel multi-objective problem:

$$\max_{\pi} (I(\mathcal{S}; \mathcal{Z}), Pr(\zeta)), \quad (1)$$

where the resulting policies will learn to discover diverse skills with a high chance of satisfying human preferences. In the following section, we describe how each objective will be addressed in our method.

### 3.3 Distance-Maximising Skill Discovery

To maximise  $I(\mathcal{S}; \mathcal{Z})$  with  $\pi(a|s, z)$ , we use the Distance-Maximising Skill Discovery (DSD) objectives proposed by Park et al. [24, 25, 26]. Although our approach is not limited to a specific skill discovery method, we chose the DSD objective since it achieves state-of-the-art performance in discovering dynamic and diverse skills. Specifically, instead of maximising  $I(\mathcal{S}; \mathcal{Z})$ , DSD maximizes  $I_w(\mathcal{S}; \mathcal{Z})$ , which measures the Wasserstein dependency between states and skills Ozair et al. [22]. DSD objectives are based on two elements, the state-representation function  $\phi : \mathcal{S} \rightarrow \mathcal{Z}$  as well as a non-negative, arbitrary distance function  $d$  that can be learned or specified (such as the Euclidean distance as in LSD [24]). The function  $\phi$  is trained

to represent the displacement in the state space under the distance function  $d$  while staying aligned with the skill variable  $z$ . Thus,  $\phi$ ’s objective is the following :

$$J^{DSD, \phi} := \mathbb{E}_{z, s, s'} [(\phi(s') - \phi(s))^T z] \quad \text{s.t.} \quad \forall x, y \in \mathcal{S} \quad \|\phi(x) - \phi(y)\| \leq d(x, y) \quad (2)$$

This constrained objective can be optimised with dual gradient descent [2] as described in Park et al. [25].

We then use  $\phi$  to train the skills  $\pi(a|s, z)$  in order to generate trajectories with high differences in the latent state space. The skill-discovery reward per step is then given by:

$$r^{DSD} = (\phi(s') - \phi(s))^T z. \quad (3)$$

This reward allows us to discover a continuous set of diverse and dynamic skills.

### 3.4 Reward Learning from Preferences

To maximise  $Pr(\zeta)$  with  $\pi(a|s, z)$ , we follow prior works in preference-based RL [6, 18, 32], where a human is presented with two trajectory segments (state-action sequences)  $\sigma^1$  and  $\sigma^2$  and is asked to indicate their preference  $\zeta_i$  for one over the other. For instance, a preference for the first segment over the second is denoted as  $\zeta_i = \sigma^1 \succ \sigma^2$  and would result in the label  $y = (1, 0)$  and be stored in a buffer  $\mathcal{D}$  as  $(\sigma^1, \sigma^2, y)$ . Then we model the human’s internal reward function  $\hat{r}_\psi$  responsible for the indicated preferences via the Bradley-Terry model [3] as follows:

$$P_\psi[\sigma^1 \succ \sigma^2] = \frac{\exp(\sum_t \hat{r}_\psi(s_t^1, a_t^1))}{\sum_{j \in \{1,2\}} \exp(\sum_t \hat{r}_\psi(s_t^j, a_t^j))}. \quad (4)$$

As in Lee et al. [18], we model the reward function as a neural network with parameters  $\psi$ , which is updated by minimising the following loss:

$$\mathcal{L}^{Reward} = -\mathbb{E}_{(\sigma^1, \sigma^2, y) \sim \mathcal{D}} [y(1) \log P_\psi[\sigma^1 \succ \sigma^2] + y(2) \log P_\psi[\sigma^2 \succ \sigma^1]]. \quad (5)$$

Thus a policy  $\pi(a|s, z)$  maximising the reward function  $\hat{r}_\psi$  would also maximise  $Pr(\zeta)$ .

### 3.5 Human-Aligned Skill Discovery (HaSD)

---

**Algorithm 1:** Human-aligned Skill Discovery ( $\alpha$ -HaSD)

---

```

Initialise  $\mathcal{B}$ ,  $\pi_z$ ,  $r_\psi$ ,  $\phi$  and  $A$ ;
for each epoch do
  // Collect data;
  for each episode do
    Sample  $z \sim p(z)$  and  $\alpha \sim A$ ;
    Sample trajectory  $\tau$  with  $\pi_\theta(a_t|s_t, z, \alpha)$ ;
    Store trajectory  $\tau$  in  $\mathcal{B}$ 
  end for
  if it's time to update the preference then
    for each 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 for
    Update  $\hat{r}_\psi$  with gradient descent on  $\mathcal{L}^{\text{Reward}}$  (5);
  end if
  Update  $\phi$  with gradient ascent on  $J^{\text{DSD}}$ ;
  Update  $\pi_\theta(a|s, z, \alpha)$  using SAC [9] and TQC [17] with
  reward  $r^{\text{HaSD}}$ ;
end for

```

---

In this section, we present our novel Human-aligned Skill Discovery (HaSD) method that learns latent-conditioned policies  $\pi(a|s, z)$  that maximise the Human-aligned Skill Discovery Objective in Equation 1. HaSD linearly combines the skill discovery reward and the preference reward as follows:

$$r^{\text{HaSD}} = \alpha r^{\text{Ha}} + r^{\text{DSD}} \quad (6)$$

where :

$$r^{\text{Ha}} = \hat{r}_\psi(s, a) \quad \text{and} \quad r^{\text{DSD}} = (\phi(s') - \phi(s))^T z. \quad (7)$$

$\alpha$  is a hyper-parameter that represents diversity-alignment trade-offs,  $\hat{r}_\psi$  is the reward learned from human feedback as described in Section 3.4 and  $\phi$  is the state-representation function from the DSD objective presented in Section 3.3.

*Intuition and selecting  $\alpha$ .* Intuitively, Equation 1 implies that agents should learn diverse and dynamic skills in the latent space while maximising some degree of alignment ( $\alpha$ ) with human values. In practice,  $\alpha$  must not be negative, as it would lead in discovering skills contrary to human values. Generally, a higher  $\alpha$  will result in more aligned skills but a less diverse skill set. A lower  $\alpha$  will have the opposite effect. In the following section, we address the issue of selecting a suitable  $\alpha$  by learning multiple diversity-alignment trade-offs.

*Pre-training phase.* At the start of training, state coverage or coherent behaviours are limited, resulting in non-informative queries [18]. To mimic the pre-training phase of PEBBLE [18], we let the policy be trained only on the skill discovery reward  $r^{\text{DSD}}$ , which is

equivalent to setting  $\alpha$  dynamically as:

$$\alpha = \begin{cases} 0 & \text{if } t \leq \tau \\ c(\text{constant}) & \text{otherwise} \end{cases} \quad (8)$$

where  $\tau$  is a hyper-parameter that indicates the time step from which when we start to elicit and learn from human feedback.

### 3.6 Configurable Human-Aligned Skills ( $\alpha$ -HaSD)

Despite the consideration of alignment and skill discovery rewards  $r^{\text{Ha}}$  and  $r^{\text{DSD}}$ , we may not know what the Diversity-Alignment trade-offs across objectives should be. Performing a hyper-parameter search over  $\alpha$  would require multiple trials, which would be impractical, considering the cost associated with collecting human feedback. Additionally, individual users may have different preferences and thus it may be useful to learn across the whole range of trade-offs and let an user choose a trade-off value at run-time. In this case, we can extend HaSD to learn a conditional policy on the trade-off value, thereby learning Configurable Human-aligned Skills ( $\alpha$ -HaSD)  $\pi(a|s, z, \alpha)$ . Then we could apply any search method over  $\alpha$  to the trained conditional policy without the need for additional human feedback. We train this policy by augmenting states with a variety of trade-offs, corresponding to a range of  $\alpha$  values and then optimising the  $\alpha$ -HaSD objective. Our overall method is described in Algorithm 1.

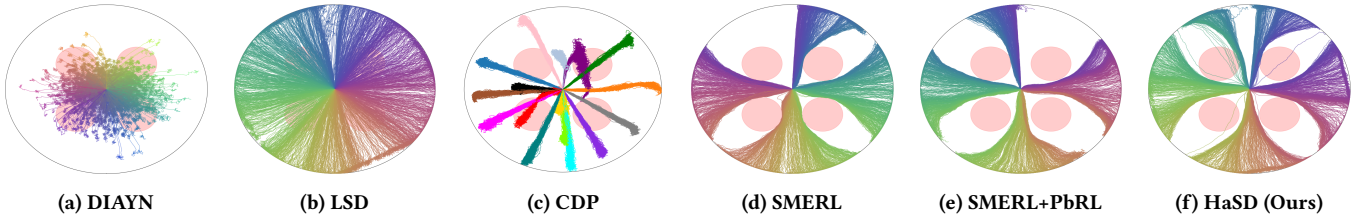
## 4 EXPERIMENTS

In this section, we demonstrate that HaSD discovers diverse skills that align with human preferences and that  $\alpha$ -HaSD learns a wide range of diversity-alignment trade-offs. To this end, we first show how our method works in simple 2D navigation with safety costs in Section 4.3. Then we demonstrate in Safety-Gymnasium [12] the scalability of our methods across a variety of robots and environments in higher dimensions. We consider environments with different types of alignment objectives in Sections 4.6 and 4.7.

In one set of experiments, the alignment objective is safety, and in another set of experiments, the alignment objective involves specific interaction with an object. We note that as such, there are no inherent prerequisites for an alignment objective—just that a human should be able to indicate their preference in terms of this objective, given a pair of trajectories. For each experiment, the alignment objective is described as a sentence that could be given to a human annotator. Following previous work on preference-based RL [18], we simulate human preferences with a ground truth reward function which is assumed to reflect human annotator preferences. Unless explicitly states, we did not limit human feedback budget to ensure alignment. However, we examine the sensitivity of our method to the number of feedback samples in Section 4.4, and later, the sensitivity of our method to real human feedback in Section 4.8. We include additional results and implementation details in full version [11].

### 4.1 Baselines

We compare HaSD with 5 baselines ranging from using no human information to using preference information, and finally to having direct access to human information. Wherever possible, we



**Figure 3: Skill sets learned from all baselines in the 2D navigation environment. LSD covers a larger region of the environment than DIAYN, while HaSD avoids hazardous areas while maintaining good coverage in safe places. SMERL methods are well aligned, but the coverage is not optimal.**

implement baselines over LSD[24], the base for unsupervised skill discovery for our method. We overview the baselines below:

- **Unaligned Skill discovery:** We train DIAYN[7] and LSD[24] without any information about human preferences, serving as baselines for unaligned skill discovery.
- **Human-aligned Skill discovery with Complete Information:** We train original SMERL[16] which first maximises the ground truth rewards, and then diversifies skills, serving as a baseline with complete information.
- **Human-aligned Skill discovery with Preferences:** We train CDP[10] and extend SMERL to incorporate human preferences (SMERL+PbRL), serving as a baseline using human preferences.

## 4.2 Environment Descriptions

**4.2.1 Nav2D.** The 2D navigation environment consists of a two-dimensional circular room enclosed by walls that confine the agent to the circular area. The environment contains 4 fixed hazardous areas (red regions in Figure 3) associated with a safety cost. The agent begins each episode in the center of the room, until episode termination, which occurs after 75 steps. The agent only accesses its horizontal and vertical coordinates (X, Y). It can deterministically change its direction and amplitude of steps to freely move in the environment. State and action spaces are continuous. The ground truth reward function<sup>2</sup> is designed to mimic the following preference: ‘Trajectories that travel as far as possible from the initial position while avoiding unsafe regions should be preferred’.

**4.2.2 Safety Gymnasium Environments.** In the Safety Gymnasium environment, we evaluated our method with five different agents (point, car, racecar, doggo and ant) across two environments. Each agent has its own state and action space, and learning the corresponding agent controller increases in complexity - i.e., point (least complex) to ant (most complex).

**4.2.3 Hazardous-Room.** The Hazardous-Room environment consist of a two-dimensional square room enclosed by walls that doesn’t restrain the agent but is associated with some safety (alignment) cost. The environment contains four fixed hazardous areas associated with some safety (alignment) cost. The agent begins each episode in the center of the room until episode termination, which occurs after 200 steps. The agent accesses internal state and lidar

information about the hazardous areas. The ground truth reward function<sup>2</sup> is designed to mimic the following preference ‘Trajectories that travel as far as possible from the initial position while avoiding unsafe regions and staying in the enclosed area should be preferred’.

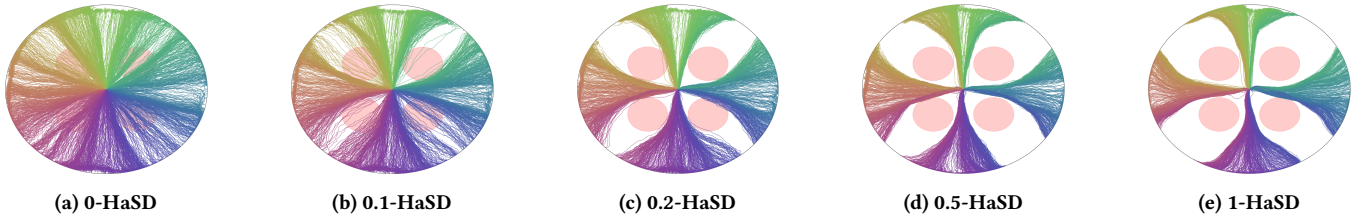
**4.2.4 Push-Room.** The Push-Room environment consists of a two-dimensional square room enclosed by walls that do not restrain the agent, but is associated with some safety (alignment) cost. The environment contains a box in the center of the room that can be pushed. The agent begins each episode in the northeast corner of the room until episode termination, which occurs after 400 steps. The agent accesses internal state and lidar information about the box. The ground truth reward function<sup>2</sup> is designed to mimic the following preference ‘Trajectories that make the box travel as far as possible from its initial position should be preferred’.

## 4.3 Qualitative and Quantitative Comparisons in Nav2D

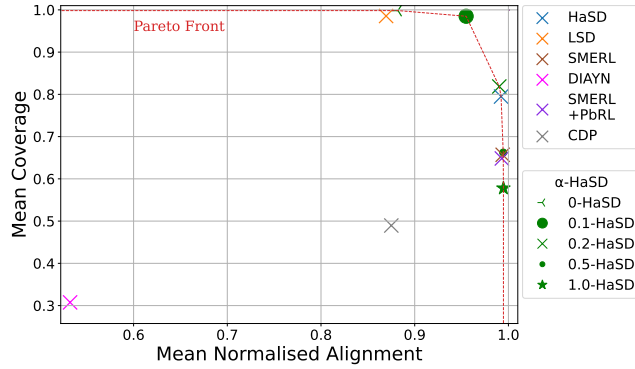
In this section, we demonstrate that HaSD and  $\alpha$ -HaSD can acquire a continuous set of skills that cover the environment while avoiding unsafe regions according to human preference. To this end, we sampled 2000 skills from policies trained across 5 seeds with DIAYN, LSD, SMERL, SMERL+PbRL, CDP HaSD and  $\alpha$ -HaSD, and compared them qualitatively and quantitatively.

Qualitatively, Figure 3b shows how LSD’s skills cover the entire room, but ignore the unsafe region. In contrast, HaSD’s skills as shown in Figure 3f adequately cover the environment while avoiding unsafe areas. On the other hand,  $\alpha$ -HaSD can learn to generate a range of diversity-alignment trade-offs. Figures 4a to 4e illustrates how  $\alpha$  can be adjusted to zero to produce a skills set similar to LSD as in Figure 4a, corresponding to a skill set that does not consider human preferences. By increasing  $\alpha$ , we gradually lose diversity in favour of alignment as seen in the figures 4c and 4e. Figures 3d and 3e illustrates that SMERL and SMERL+PbRL are also able to learn diverse skills that avoid unsafe areas however they will generally cover less than HaSD as they focus on optimising the alignment reward first. CDP successfully uses preferences to discover diverse safe goals however it learns to reach some of the desired goals in an unaligned manner as illustrated in Figure 3c. We report in Figure 5 each skill set (solutions) obtained with LSD or HaSD the normalised total coverage of a skill set (y-axis) and the normalised mean alignment value generated by skills in a skill set (x-axis). Coverage is measured by the number of  $0.1 \times 0.1$  square

<sup>2</sup> Details of the corresponding precise ground truth reward function used to simulate this objective can be found in the full version [11]



**Figure 4: Comparing visually skill set obtained with  $\alpha$ -HaSD by changing  $\alpha$  in the 2D navigation environment. When alpha is set to 0 the skill set is similar to LSD. The higher the  $\alpha$ , the lower the diversity and coverage is.**



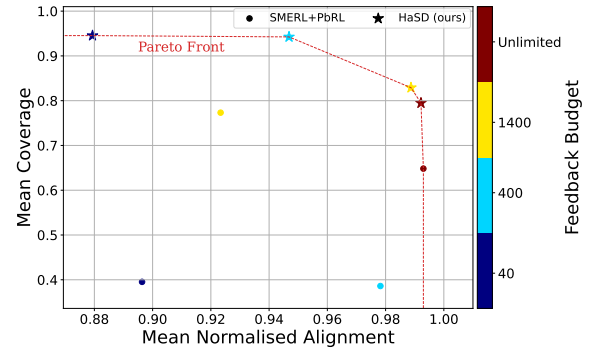
**Figure 5: The approximated Pareto front shows that in the 2D navigation environment, LSD solutions achieve high coverage with low alignment on the left side. HaSD solutions are more to the right, offering high alignment while maintaining coverage.  $\alpha$ -HaSD covers more areas with diverse diversity-alignment trade-offs. SMERL methods attain high alignment but have lower coverage compared to HaSD.**

bins occupied by the agent at least once, while we used the ground truth reward as the alignment value. We highlight the approximate Pareto frontier obtained. Quantitatively, we observe that LSD solutions achieve high coverage but low alignment. In contrast, HaSD solutions deliver high alignment while maintaining high coverage. Meanwhile, as expected SMERL and SMERL+PbRL have marginally better alignment than HaSD, but with significantly less coverage. Finally, DIAYN and CDP demonstrates a lower performance than LSD, HaSD and SMERL methods. Generally, HaSD offers a better balance between exploration and alignment than SMERL or LSD. Lastly, we can observe that  $\alpha$ -HaSD solutions lie on/close to the approximated Pareto frontier meaning that we learn qualitative solutions over both objectives.

#### 4.4 Sensitivity to Human Feedback Budget

**4.4.1 HaSD and SMERL+PbRL Sensitivity:** In this section we analyse the sensitivity of our methods to available budget feedback. The degree to which skills are aligned depends on how well the reward model captures human values. This naturally depends on the availability of feedback labels. As illustrated in Figure 6, we found that as we decreased the number of feedback received, HaSD’s performance over the alignment objective decreased. At the lowest

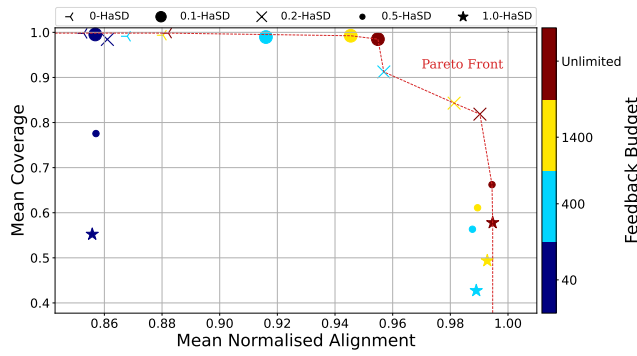
(40 feedback instances), its performance was comparable to that of LSD. Additionally, we found that HaSD was more robust to a reduction of human feedback instances than SMERL+PbRL. This is because SMERL’s Objective first seeks to optimise performance over task reward, which here is poorly approximated. This illustrates the limitations of SMERL’s reward in our situation, where both rewards and a skill discovery objective are required.



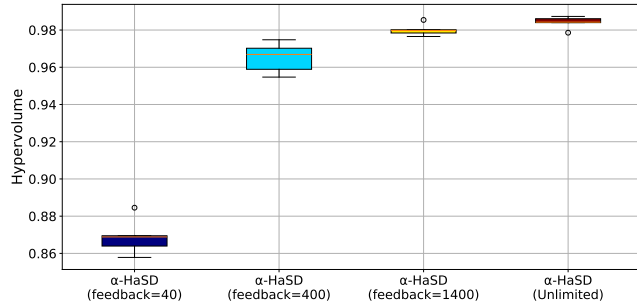
**Figure 6: Approximated Pareto front considering SMERL+PbRL and HaSD with different human feedback budget in Nav2d environment. HaSD is more robust to reduction feedback than SMERL+PbRL.**

**4.4.2  $\alpha$ -HaSD Sensitivity:** We also analysed the sensitivity of  $\alpha$ -HaSD to human feedback budgets. Figure 7 illustrates the approximate Pareto frontier obtained with all solutions.  $\alpha$ -HaSD exhibits similar behaviour as HaSD when the budget is restricted, resulting in less accurate alignment rewards. Nevertheless,  $\alpha$ -HaSD is still able to learn a range of alignment values even on very restrictive budgets.

**4.4.3 Hypervolume:** To more accurately quantify the impact of restricted human feedback on  $\alpha$ -HaSD, we showcase in Figure 8 the Hypervolume[35] computed for each set of solutions. Hypervolume is a popular metric in multi-objective problems[8] to measure how a set of solutions is diverse but simultaneously performant as well. The Hypervolume corresponds to the set of points “contained” within an n-dimensional space. As expected we can see that as the amount of feedback diminishes, the Hypervolume metrics also decline, indicating a reduction in the quality of the solution set. Additionally, the variability of the solution set as feedback decreases.



**Figure 7: Approximated Pareto front considering  $\alpha$ -HaSD with different human feedback budget in Nav2d environment.**



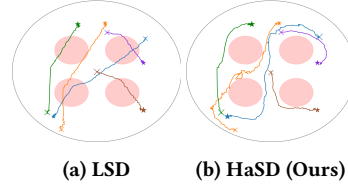
**Figure 8: Hypervolume computed for each set of solutions produced by  $\alpha$ -HaSD with different human feedback budget in Nav2d environment.**

#### 4.5 Nav2D: Downstream Task

In this section, we demonstrate how we can use skills learned with HaSD to achieve downstream tasks in more challenging settings. To this end we implemented a meta-controller in our 2D navigation environment to demonstrate the utility of the discovered skills. In this setup:

- The agent starts at a random location in the environment, excluding unsafe areas, and aims to reach a goal at another random location, also excluding unsafe areas.
- The environment provides sparse rewards. Specifically, the agent receives a reward signal only upon reaching the goal; otherwise, it receives no reward.
- The meta-controller selects a skill which is then executed for one environment-step, replacing the traditional action space with the learned skill space.
- Importantly, no explicit information about unsafe areas is provided during this downstream task.

We provide more details in the full version [11]. We evaluated the meta-controller over 1000 sample goals and reported (across 9 seeds) both the average goal achievement rate and the mean cost generated to achieve them in Table 1. In addition, we compute a ranking score for each method based on their separate ranking on scores and costs. Our results demonstrate that HaSD attains a goal achievement rate



**Figure 9: Goal-reaching trajectories produced with LSD skills and HaSD skills in the Nav2d environment.**

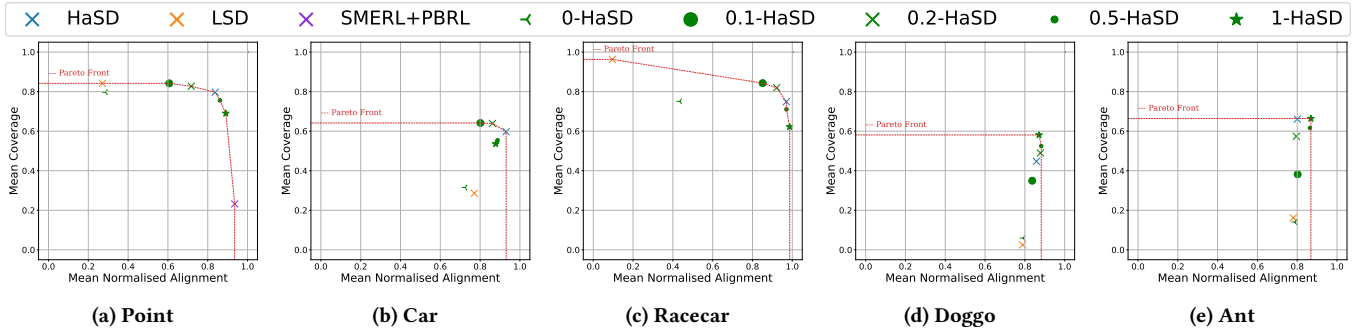
comparable to LSD while maintaining significantly lower costs. Compared to HaSD, SMERL and SMERL+PbRL achieve significantly lower goal achievement. This may be due to the lower coverage of these methods, which highlights the importance of our approach to first discovery skill and then aligning with human preferences. The qualitative results in Figure 9 demonstrate that the skill space acquired with HaSD retains alignment information, which allows our meta-controller to avoid unsafe areas more than with LSD, despite no information regarding safety being provided during the downstream task. This evaluation substantiates the fact that skills discovered through our method are not only safe but also useful in downstream tasks.

**Table 1: Comparison of baselines (DIAYN, LSD, HaSD, SMERL, and SMERL+PbRL) based on their performance scores and associated costs**

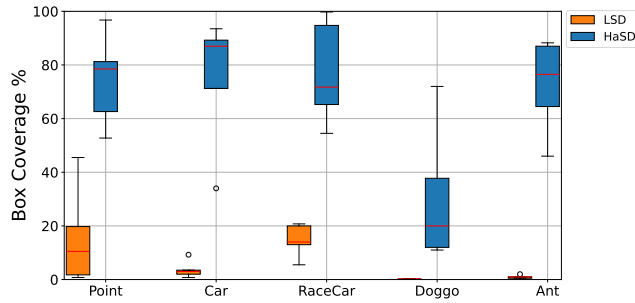
| Method            | Score            | Cost            | Rank           |
|-------------------|------------------|-----------------|----------------|
| <b>HaSD(Ours)</b> | $98.5\% \pm 1.0$ | $6.7 \pm 0.5$   | <b>4 (2,2)</b> |
| LSD               | $99.5\% \pm 0.6$ | $12.01 \pm 0.3$ | 5 (1,4)        |
| SMERL+PbRL        | $68.9\% \pm 4.1$ | $3.86 \pm 1.23$ | 5 (4,1)        |
| SMERL             | $75.6\% \pm 4.0$ | $9.49 \pm 2.0$  | 6 (3,3)        |
| DIAYN             | $23.9\% \pm 3.3$ | $23.9 \pm 7.2$  | 10 (5,5)       |

#### 4.6 Safety Gymnasium: Hazard-Room

In this section we present results obtain in the Hazard-Room environment described in Section 4.2.4. All results are presented after sampling 1000 skills from the policy and averaging across 5 seeds. We report a similar pareto front as in Section 4.3 and highlight the approximate Pareto frontier obtained. We trained SMERL+PbRL only on the point agent to minimise experimental workload. We observe similar trends as in nav2d, where SMERL+PbRL achieves high alignment but very low coverage as shown in Figure 10a. Although we have not run SMERL+PbRL on other agents, we expect similar characteristics (ie., high alignment and low coverage) for these agents as well. Quantitatively, In Figures 10a and 10c, Point and RaceCar display a similar pattern as Section 4.3 where LSD achieves high coverage but low alignment while HaSD provides a better trade-off between metrics and  $\alpha$ -HaSD generates diverse favourable trade-offs. In Figures 10b, 10d and 10e Car, Doggo, and Ant LSD fails to learn relevant skills, resulting in a low-coverage skill set that does not incur much cost. It is also important to note that the LSD’s objective also fails to drive exploration in our method, leading to HaSD methods relying mostly on the alignment objective. We report qualitative results in the full version [11].



**Figure 10: Approximated Pareto Front obtained with LSD, HaSD and  $\alpha$ -HaSD for each agent in the Hazardous-Room environment. In general, solutions are arranged from left to right, as alignment objectives are more influential.**



**Figure 11: Comparing box-pushing agent’s coverage of skill sets discovered by LSD and HaSD with the different robots in the Push-Room environment.**

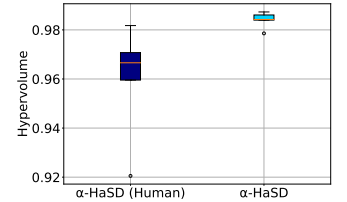
#### 4.7 Safety Gymnasium: Push-Room

In this section we present results obtained in the Push-Room environment described in Section 4.2.4. All results are presented after sampling 1000 skills from the policy across 5 seeds. We report results regarding the box’s coverage of the room in Figure 11. This coverage is measured by the number of  $0.1 \times 0.1$  square bins occupied by the box at least once. Having a high coverage by the box indicates that the agent interacts with the box in a wide range of ways. HaSD obtains better coverage with every agent than LSD. This means that the alignment objective helps indicate desired behaviour while allowing the diversity objective to discover diverse skills. We report qualitative results in the full version [11].

#### 4.8 Sensitivity to Human Feedback

In this section, we analysed the sensitivity of  $\alpha$ -HaSD to real human feedback. For practical reasons, we train offline a reward model from actual human feedback subjects (the authors) familiar with the task. This reward is then used to simulate human preferences during training. The reward model was trained with 1400 feedback which took 1h to collect. We provide more detail in the full version [11]. We show in Figure 12 the hypervolume obtained with  $\alpha$ -HaSD(Human) which used the real human process presented. The hypervolume indicates that  $\alpha$ -HaSD(Human) can obtain similar results with actual human than ground truth reward. However we can observe that  $\alpha$ -HaSD(Human). This is largely due to the noise

**Figure 12: Comparing hypervolume computed for each set of solutions produced by  $\alpha$ -HaSD with actual human feedback budget and ground truth reward in Nav2d environment.**



inherent in the process. This method is inherently noisier than scenarios where preferences are directly provided during training. The accumulation of approximation errors while learning the reward offline can contribute to this noise, and these errors may carry over to the use of this reward as simulated human input during  $\alpha$ -HaSD training. Other than that we are able to obtain similar results from actual human feedback.

## 5 DISCUSSION AND CONCLUSION

In this work, we introduced a novel approach to address under-constrained skill discovery. Our proposed approach, Human aligned skill discovery, is based on the idea of optimising both a skill discovery objective and an alignment objective. Using these objectives, we explicitly designed HaSD, an approach to learn diverse skills that align with human preference over their entire trajectory. With different alignment objectives, we empirically demonstrated that HaSD learned diverse skills that align with human preferences in various navigation and box-pushing environments. Additionally, we showed that we can condition skills on the diversity alignment trade-off variable to produce a range of skills relevant to different diversity-alignment trade-offs. One of the inherent limitations of this work is that the degree of alignment depends on the accuracy of the reward model. This can heavily depend on the number of feedback available. Although this is a common limitation of HIL-methods, it is exacerbated in USD settings due to their inherent long training time, sample inefficiency and the size of the behaviour space. In these circumstances, the questions of when to seek human feedback and what queries to select to maximise information gain require more sophisticated methods. In future work, we will investigate how to leverage the diversity and amount of interactions generated by the skill discovery objective to increase reward model accuracy with low feedback budgets in USD settings.

## REFERENCES

- [1] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. 2018. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299* (2018).
- [2] Stephen P Boyd and Lieven Vandenbergh. 2004. *Convex optimization*. Cambridge university press.
- [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] Wentse Chen, Shiyu Huang, Yuan Chiang, Tim Pearce, Wei-Wei Tu, Ting Chen, and Jun Zhu. 2024. DGPO: discovering multiple strategies with diversity-guided policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11390–11398.
- [6] 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>
- [7] Benjamin Eysenbach, Julian Ibarz, Abhishek Gupta, and Sergey Levine. 2019. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019*.
- [8] Florian Felten, Lucas Nunes Alegre, Ann Nowe, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno Castro da Silva. 2023. A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=jfwRLudQyJ>
- [9] 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>
- [10] Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. 2023. Controlled Diversity with Preference: Towards Learning a Diverse Set of Desired Skills. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1135–1143.
- [11] Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. 2025. Human-Aligned Skill Discovery: Balancing Behaviour Exploration and Alignment. *arXiv preprint arXiv:2501.17431* (2025).
- [12] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. 2023. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems* 36 (2023).
- [13] Danilo Rezende Karol Gregor and Daan Wierstra. 2016. Variational Intrinsic Control. *International Conference on Robotic Learning* 0, 0 (2016), 0.
- [14] Sunin Kim, Jaewoon Kwon, Taeyoon Lee, Younghyo Park, and Julien Perez. 2022. Safety-Aware Unsupervised Skill Discovery. In *2023 International Conference on Robotics and Automation (ICRA)*. IEEE.
- [15] Even Klemsdal, Sverre Herland, and Abdulmajid Murad. 2021. Learning Task Agnostic Skills with Data-driven Guidance. *arXiv preprint arXiv:2108.01869* (2021).
- [16] Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. 2020. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *Advances in Neural Information Processing Systems* 33 (2020), 8198–8210.
- [17] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. 2020. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*. PMLR, 5556–5566.
- [18] 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).
- [19] 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).
- [20] Katherine Metcalf, Miguel Sarabia, and Barry-John Theobald. 2022. Rewards Encoding Environment Dynamics Improves Preference-based Reinforcement Learning. *arXiv preprint arXiv:2211.06527* (2022).
- [21] 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, Dharmashan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.
- [22] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein Dependency Measure for Representation Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/f9209b7866c9f69823201c1732cc8645-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/f9209b7866c9f69823201c1732cc8645-Paper.pdf)
- [23] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2021. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. In *International Conference on Learning Representations*.
- [24] Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. 2021. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*.
- [25] Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. 2023. Controllability-Aware Unsupervised Skill Discovery. In *International Conference on Machine Learning*. PMLR, 27225–27245.
- [26] Seohong Park, Oleh Rybkin, and Sergey Levine. 2023. METRA: Scalable Unsupervised RL with Metric-Aware Abstraction. In *The Twelfth International Conference on Learning Representations*.
- [27] William Saunders, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. 2018. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2067–2069.
- [28] 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=HJgLR4KvH>
- [29] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. 2020. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603* (2020).
- [30] 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](https://doi.org/10.1016/S0004-3702(99)00052-1)
- [31] Xiaofei Wang, Kimin Lee, Kourosh Hakhmaneshi, 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.
- [32] Aaron Wilson, Alan Fern, and Prasad Tadepalli. 2012. A Bayesian Approach for Policy Learning from Trajectory Preference Queries. In *NIPS*.
- [33] Tom Zahavy, Yannick Schroecker, Feryal Behbahani, Kate Baumli, Sebastian Flennerhag, Shaobo Hou, and Satinder Singh. 2023. Discovering Policies with DOMINO: Diversity Optimization Maintaining Near Optimality. In *The Eleventh International Conference on Learning Representations*.
- [34] Ruohan Zhang, Faraz Torabi, Garrett Warnell, and Peter Stone. 2021. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021).
- [35] Eckart Zitzler. 1999. Evolutionary algorithms for multiobjective optimization: methods and applications. <https://api.semanticscholar.org/CorpusID:14157879>