Learning to Assist Humans without Inferring Rewards

Vivek Myers[1] **Evan Ellis**[1]
Benjamin Eysenbach[2] **Sergey Levine**[1] **Anca Dragan**[1]
[1]UC Berkeley [2]Princeton University

Abstract

Assistive agents should make humans' lives easier. Classically, such assistance is studied through the lens of inverse reinforcement learning, where an assistive agent (e.g., a chatbot, a robot) infers a human's intention and then selects actions to help the human reach that goal. This approach requires inferring intentions, which can be difficult in high-dimensional settings. We build upon prior work that studies assistance through the lens of empowerment: an assistive agent aims to maximize the influence of the human's actions such that they exert a greater control over the environmental outcomes and can solve tasks in fewer steps. We lift the major limitation of prior work in this area—scalability to high-dimensional settings—with contrastive successor representations. We formally prove that these representations estimate a similar notion of empowerment to that studied by prior work and provide a ready-made mechanism for optimizing it. Empirically, our proposed method outperforms prior methods on synthetic benchmarks, and scales to Overcooked, a cooperative game setting. Theoretically, our work connects ideas from information theory, neuroscience, and reinforcement learning, and charts a path for representations to play a critical role in solving assistive problems.[1]

1 Introduction

AI agents deployed in the real world should be helpful to humans. When we know the utility function of the humans an agent could interact with, we can directly train assistive agents through reinforcement learning with the known human objective as the agent's reward. In practice, agents rarely have direct access to a scalar reward corresponding to human preferences (if such a consistent model even exists) [1], and must infer them from human behavior [2, 3]. This inference can be challenging, as humans may act suboptimally with respect to their stated goals, not know their goals, or have changing preferences [4]. Optimizing a misspecified reward function can have poor consequences [5].

An alternative paradigm for assistance is to train agents that are *intrinsically* motivated to assist humans, rather than directly optimizing a model of their preferences. An analogy can be drawn to a parent raising a child. A good parent will empower the child to make impactful decisions and flourish, rather than proscribing an "optimal" outcome for the child. Likewise, AI agents might seek to *empower* the human agents they interact with, maximizing their capacity to change the environment [6]. In practice, concrete notions of empowerment can be difficult to optimize as an objective, requiring extensive modeling assumptions that don't scale well to the high-dimensional settings deep reinforcement learning agents are deployed in.

What is a good intrinsic objective for assisting humans that doesn't require these assumptions? We propose a notion of assistance based on maximizing the influence of the human's actions on the environment. This approach only requires one structural assumption: the AI agent is interacting with an environment where there is a notion of actions taken by the human agent—a more general setting than the case where we model the human actions as the outcome of some optimization procedure, as in inverse RL [7, 8] or preference-based RL [9].

[1]Code: `https://github.com/vivekmyers/empowerment_successor_representations`.

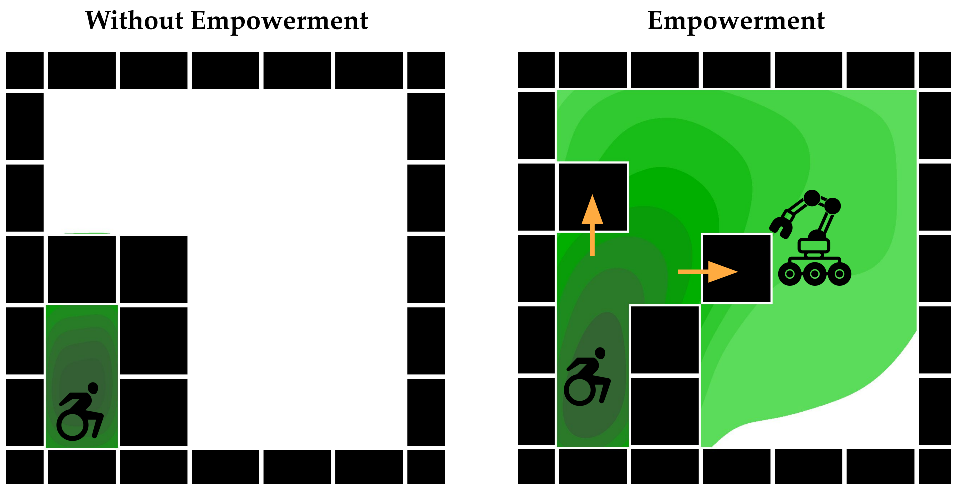


Figure 1: We propose an algorithm training assistive agents to empower human users—the assistant should take actions that enable human users to visit a wide range of future states, and the human's actions should exert a high degree of influence over the future outcomes. Our algorithm scales to high-dimensional settings, opening the door to building assistive agents that need not directly reason about human intentions.

Prior work has studied many effective objectives for empowerment. For instance, Du et al. [6] approximates human empowerment as the variance in the final states of random rollouts. Despite excellent results in certain settings, this approach can be challenging to scale to higher dimensional settings, and does not necessarily enable human users to achieve the goals the want to achieve. By contrast, our approach exclusively empowers the human with respect to the distribution of (useful) behaviors induced by their current policy, and can be implemented through a simple objective derived from contrastive successor features, which can then be optimized with scalable deep reinforcement learning (Fig. 1). We provide a theoretical framework connecting our objective to prior work on empowerment and goal inference, and empirically show that agents trained with this objective can assist humans in the Overcooked environment [10] as well as the obstacle gridworld assistance benchmark proposed by Du et al. [6].

Our core contribution is a novel objective for training agents that are intrinsically motivated to assist humans without requiring a model of the human's reward function. Our objective, Empowerment via Successor Representations (ESR), maximizes the influence of the human's actions on the environment, and, unlike past approaches for assistance without reward inference, is based on a scalable model-free objective that can be derived from learned successor features that encode which states the human is likely to want to reach given their current action. Our objective empowers the human to reach the desired states, not all states, without assuming a human model. We analyze this objective in terms of empowerment and goal inference, drawing novel mathematical connections between time-series representations, decision-making, and assistance. We empirically show that agents trained with our objective can assist humans in two benchmarks proposed by past work: the Overcooked environment [10] and an obstacle-avoidance gridworld [6].

2 Related Work

Our approach broadly connects ideas from contrastive contrastive representation learning and intrinsic motivation to the problem of assisting humans.

Assistive Agents. There are two lines of past work on assistive agents that are most relevant.

The first line of work focuses on the setting of an assistance game [2], where a robot (AI) agent tries to optimize a human reward of which it is initially unaware. Practically, inverse reinforcement learning (IRL) can be used in such a setting to infer the human's reward function and assist the human in achieving their goals [3]. The key challenge with this approach is that it requires modeling the human's reward function. This can be difficult in practice, especially if the human's behavior is not well-modeled by the reward architecture. Slightly mispecified reward functions can lead to catastrophic outcomes (i.e., directly harmful behavior in the assistance context) [11–13]. By contrast, our approach does not require modeling the human's reward function.

The second line of work focuses on empowerment-like objectives for assistance and shared autonomy. Empowerment generally refers to a measure of an agent's ability to influence the environment [14, 15]. In the context of assistance, Du et al. [6] show one such approximation of empowerment (AvE) can be approximated in simple environments through random rollouts to assist humans. Meanwhile, empowerment-like objectives have been used in shared autonomy settings to assist humans with teleoperation [16] and general assistive interfaces [17]. A key limitation of these approaches for general assistance is they only model empowerment over one time step. Our approach enables a more scalable notion of empowerment that can be computed over multiple time steps.

Intrinsic Motivation. Intrinsic motivation broadly refers to agents that accomplish behaviors in the absence of an externally-specified reward or task [18]. Common applications of intrinsic motivation in single-agent reinforcement learning include exploration and skill discovery [19–21], empowerment [15, 14], and surprise minimization [22, 23, 15]. When applied to settings with humans, these objectives may lead to antisocial behavior [5]. Our approach applies intrinsic motivation to the setting of assisting humans, where the agent's goal is an empowerment objective—to maximize the human's ability to change the environment.

Information-theoretic Decision Making. Information-theoretic approaches have seen broad applicability across unsupervised reinforcement learning [24, 15, 19]. These methods have been applied to goal-reaching [25], skill discovery [26, 27, 20, 28, 29], and exploration [21, 30, 31]. In the context of assisting humans, information-theoretic methods have primarily been used to reason about the human's goals or rewards [32–34].

Our approach is made possible by advances in contrastive representation learning for efficient estimation of the mutual information of sequence data [35]. While these methods have been widely used for representation learning [36, 37] and reinforcement learning [38–41], to the best of our knowledge prior work has not used these contrastive techniques for learning assistive agents.

3 The Information Geometry of Empowerment

We will first state a general notion of an assistive setting, then show how an empowerment objective based on learned successor representations can be used to assist humans without making assumptions about the human following an underlying reward function. In Section 5, we provide empirical evidence supporting these claims.

3.1 Preliminaries

Formally, we adapt the notation of Hadfield-Menell et al. [2], and assume a "robot" (**R**) and "human" (**H**) policy are training together in an MDP $M = (\mathcal{S}, \mathcal{A}_{\mathbf{H}}, \mathcal{A}_{\mathbf{R}}, R, P, \gamma)$. The states s consist of the joint states of the robot and the human; we do not have separate observations for the human and robot. At any state $s \in \mathcal{S}$, the robot policy selects actions distributed according to $\pi_R(a^{\mathbf{R}} \mid s)$ for $a^{\mathbf{R}} \in \mathcal{A}_{\mathbf{R}}$ and the human selects actions from $\pi_H(a^{\mathbf{H}} \mid s)$ for $a^{\mathbf{H}} \in \mathcal{A}_{\mathbf{H}}$. The transition dynamics are defined by a distribution $P(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{R}})$ over the next state $s' \in \mathcal{S}$ given the current state $s \in \mathcal{S}$ and actions $a^{\mathbf{H}} \in \mathcal{A}_{\mathbf{H}}$ and $a^{\mathbf{R}} \in \mathcal{A}_{\mathbf{R}}$, as well as an initial state distribution $P(s_0)$. For notational convenience, we will additionally define random variables $\mathfrak{s}_t$ to represent the state at time t, and $\mathfrak{a}_t^{\mathbf{R}} \sim \pi_R(\bullet \mid \mathfrak{s}_t)$ and $\mathfrak{a}_t^{\mathbf{H}} \sim \pi_H(\bullet \mid \mathfrak{s}_t)$ to represent the human and robot actions at time t, respectively.

Empowerment. Our work builds on a long line of prior methods that use information theoretic objectives for RL. Specifically, we adopt *empowerment* as an objective for training an assistive agent [6, 42, 43]. This section provides the mathematical foundations for empowerment, as developed in prior work. Our work will build on the prior work by (1) providing an information geometric interpretation of what empowerment does (Section 3.4) and (2) providing a scalable algorithm for estimating and optimizing empowerment, going well beyond the gridworlds studied in prior work.

The idea behind empowerment is to think about the changes that an agent can effect on a world; an agent is more empowered if it can effect a larger degree of change over future outcomes. Following prior work [25, 43, 42], we measure empowerment by looking at how much the actions taken *now* affect outcomes *in the future*. An agent with a high degree of empowerment exerts a high degree of control of the future states by simply changing the actions taken now. Like prior work, we measure this degree of control through the mutual information $I(\mathfrak{s}^+; a^{\mathbf{H}})$ between the current action $a^{\mathbf{H}}$ and the future states $\mathfrak{s}^+$. Note that these future states might occur many time steps into the future.

Empowerment depends on several factors: the environment dynamics, the choice of future actions, the current state, and other agents in the environment. Different problem settings involve maximizing empowerment using these different factors. In this work, we study the setting where a "human" agent and a "robot" agent collaborate in an environment; the robot will aim to maximize the empowerment of the human. This problem setting was introduced in prior work [6]. Compared with other mathematical frameworks for learning assistive agents [44], framing the problem in terms of empowerment means that the assistive agent need not infer the human's underlying intention, an inference problem that is typically challenging [45, 46].

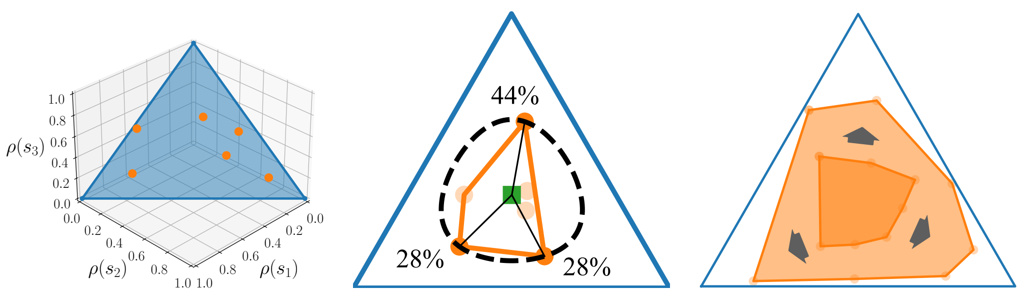
We now define our objective. To do this, we introduce random variable $\mathfrak{s}^+$, which corresponds to a state sampled $K \sim \mathrm{Geom}(1-\gamma)$ steps into the future under the behavior policies π_H and π_R. We will use $\rho(\mathfrak{s}^+ \mid s_t)$ to denote the density of this random variable; this density is sometimes referred to as the discounted state occupancy measure. We will use mutual information to measure how much the action a_t at time t changes this distribution:

$$I(a_t^{\mathbf{H}}; \mathfrak{s}^+ \mid s_t) \triangleq \mathbb{E}_{s_t, s_{t+k}, a_t^{\mathbf{H}}, a_t^{\mathbf{R}}}\left[\log \frac{p(\mathfrak{s}_{t+K} = s_{t+k} \mid \mathfrak{s}_t = s_t, \mathfrak{a}_t^{\mathbf{H}} = a_t)}{p(\mathfrak{s}_{t+K} = s_{t+k} \mid \mathfrak{s}_t = s_t)}\right]. \tag{1}$$

Our overall objective is *empowerment*, $\mathcal{E}(\pi_H, \pi_R)$: the mutual information between the human's actions and the future states $\mathfrak{s}^+$ while interacting with the robot:

$$\mathcal{E}(\pi_H, \pi_R) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t I(\mathfrak{a}_t^{\mathbf{H}}; \mathfrak{s}^+ \mid \mathfrak{s}_t)\Big]. \tag{2}$$

Note that this objective resembles an RL objective: we do not just want to maximize this objective greedily at each time step, but rather want the assistive agents to take actions now that help the human agent reach states where it will have high empowerment in the future.



(a) State marginal polytope (b) Mutual information (c) Maximizing empowerment

Figure 2: **The Information Geometry of Empowerment**, illustrating the analysis in Section 3.4. *(Left)* For a given state s_t and assistant policy π_R, we plot the distribution over future states for 6 choices of the human policy π_H. In a 3-state MDP, we can represent each policy as a vector lying on the 2-dimensional probability simplex. We refer to the set of all possible state distributions as the *state marginal polytope*. *(Center)* Mutual information corresponds to the distance between the center of the polytope and the vertices that are maximally far away. *(Right)* Empowerment corresponds to maximizing the size of this polytope. For example, when an assistive agent moves an obstacle out of a human user's way, the human user can spend more time at desired state.

3.2 Intuition and Geometry of Empowerment

Intuitively, the assistive agent should aim to maximize the number of future outcomes. We will mathematically quantify this in terms of the discounted state occupancy measure, $\rho^\pi(s^+ \mid s)$.

Intuitively, an agent has a large empowerment if the future states for one action are very different from the future actions after taking a different action; i.e., when $\rho(a_t = a_1; \mathfrak{s}^+ \mid s_t)$ is quite different from $\rho(a_t \mid s_2; \mathfrak{s}^+ \mid s_t)$ for actions $a_1 \neq a_2$. The mutual information (Eq. (1)) quantifies this degree of control: $I(a_t; \mathfrak{s}^+ \mid s_t)$.

One way of understanding this mutual information is through *information geometry* [47–50]. For a fixed current state s_t, assistant policy π_R and human policy π_H, each potential action a_t that the human takes induces a different distribution over future states: $\rho^{\pi_R,\pi_H}(\mathfrak{s}^+ \mid s_t, a_t)$. We can think about the set of these possible distributions: $\{\rho^{\pi_R,\pi_H}(\mathfrak{s}^+ \mid s_t, a_t) \mid a_t \in \mathcal{A}\}$. Figure 2 *(Left)* visualizes this distribution on a probability simplex for 6 choices of action a_t. If we look at any possible distribution over actions, then this set of possible future distributions becomes a polytope (see orange polygon in Fig. 2 *(Center)*).

Intuitively, the mutual information $I(a_t; \mathfrak{s}^+ \mid s_t)$ used to define our empowerment objective corresponds to the *size* or *volume* of this state marginal polytope. This intuition can be formalized by using results from information geometry [51–53]. The human policy $\pi_H(a_t \mid s_t)$ places probability mass on the different points in Figure 2 (Center). Maximizing the mutual information corresponds to "picking out" the state distributions that are maximally spread apart (see probabilities in Fig. 2 *(Center)*). To make this formal, define

$$\rho(\mathfrak{s}^+ \mid s_t) \triangle \mathbb{E}_{\pi(a_t \mid s_t)}[\rho(\mathfrak{s}^+ \mid s_t, a_t)] \tag{3}$$

as the *average* state distribution from taking the human's actions (see green square in Fig. 2 *(Center)*).

Lemma 1. *Mutual information corresponds to the distance between the average state distribution (Eq.* (3)*) and the furthest achievable state distributions:*

$$I(a_t; \mathfrak{s}^+ \mid s_t) = \max_{a_t} D_{KL}\big(\rho(a_t; \mathfrak{s}^+ \mid s_t) \,\big\|\, \rho(\mathfrak{s}^+ \mid s_t)\big) \triangleq d_{max}. \tag{4}$$

This distance is visualized as the black lines in Fig. 2. TODO: Annotate Fig. 2 to highlight some of the points above. Revise the caption to make sure that it is consistent with the description above. When we talk about the "size" of the state marginal polytope, we are specifically referring to the length of these black lines (as measured with a KL divergence).

This sort of mutual information is a way for measuring the degree of control that an agent exerts on an environment. This measure is well defined for any agent/policy; that agent need not be maximizing mutual information, and could instead be maximizing some arbitrary reward function. This point is important in our setting: this means that the assistive agent can estimate and maximize the empowerment of the human user *without having to infer what reward function the human is trying to maximize.*

Finally, we come back to our empowerment objective, which is a discounted sum of the mutual information terms that we have been analyzing above. This empowerment objective says that the human is more empowered when this set has a larger size – the human can visit a wider range of future state (distributions). The empowerment objective says that the assistive agent should act to try to maximize the size of this polytope. Importantly, this maximization problem is done sequentially: the assistive agent wants the size of this polytope to be large both at the current state and at future states; the human's actions should exert a high degree of influence over the future outcomes both now and in the future. Thus, our overall objective looks at a sum of these mutual informations.

Not only does this analysis provides a geometric picture for what empowerment is doing, it also lays the groundwork for formally relating empowerment to reward.

3.3 Relating Empowerment to Reward.

In this section we take aim at the question: does empowering the human guarantee that they will get higher rewards? Answering this question is important because, at the end of the day, it is reward maximization that the human cares above; while our method avoids the complex reward inference step required by prior assistive methods, our method would be useless if it did not actually help humans maximize their rewards.

size of this set of possible measures. We can formalize this intuition by employing a result from Eysenbach et al. [51, Lemma 6.2].

We start by writing a version of this lemma that additionally conditions on the current state:

$$I(z; a^+ \mid s_t) = \max_z D_{\text{KL}}\big(\rho(s^+ \mid z^*, s_t) \parallel \rho(s^+ \mid s_t)\big) \triangleq d_{\text{max}}(s_t), \tag{5}$$

which says that a human maximizing mutual information will only select those skills z that are maximally far away from the prior.

Lemma 2 (Lemma 6.2 from Eysenbach et al. [51])**.** *Let $\pi_H(z)$ be the human's skill distribution that maximizes mutual information. Then we have*

$$\pi_H^*(z) > 0 \implies D_{KL}\big(\rho(s \mid z)\big\|\rho(s)\big) = \max_{z^*} D_{KL}\big(\rho(s \mid z^*)\|\rho(s)\big). \tag{6}$$

Noting that the mutual information is the expected value of this KL divergence over $\pi_H^*(z)$, we have

$$I^{\pi_H^*}(\mathfrak{s}^+; z) = \max_{z^*} D_{KL}\big(\rho(s \mid z^*)\big\|\rho(s)\big) \triangleq d_{max}. \tag{7}$$

Thus, we can think about mutual information maximization as finding the set of skills with the maximal coverage – where skills are maximally far away from their center.

Now, by extension, a robot assistant that is maximizing this mutual information also aims to increase the size of this set:

$$\max_{\pi_R} I^{\pi_R, \pi_H^*}(\mathfrak{s}^+; z) = \max_{\pi_R} d_{\text{max}}. \tag{8}$$

In other words, an agent maximizing the empowerment of the human will aim to increase the support of goals the human can reach, conditioned on their intention, as illustrated in Fig. 2.

3.4 The Information Geometry of Empowerment

To build on this intuition, we will show that in the special case where the human is well-modeled as optimizing a reward function, we can relate empowerment maximization to reward maximization. Since a key advantage of empowerment is that it does not necessarily require this assumption to be a meaningful assistance objective, we can view our objective as a generalization of the assistance problem beyond the CIRL setting [2]. In particular, we will show that under certain assumptions maximizing empowerment corresponds to provably increasing their expected rewards.

Lemma 3. *Assume that a human has learned skills $\pi(a \mid s, z)$ by maximizing mutual information $I(\mathfrak{s}^+; z)$ and adapts to a reward function by minimizing the regularized regret:*

$$\min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{KL}\big(\rho^*(s)\|\rho(s)\big). \tag{9}$$

We assume that the human chooses the prior $\rho(s)$ that minimizes this regret for the worst-case choice of reward function (i.e., the minimax optimal prior). An assistive agent that maximizes $I^{\pi_R}(\mathfrak{s}^+; z)$ minimizes the worst-case (regularized) regret incurred by the human.

Letting $\pi_R^* \in \arg\max_{\pi_R} I^{\pi_R}(\mathfrak{s}^+; z)$, we have

$$\pi_R^* \in \arg\min_{\pi_R}\left(\min_{\rho(s) \in \mathcal{C}^{\pi_R}} \max_{r(s)} \min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}^{\pi_R}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{KL}(\rho^*(s)\|\rho(s))\right). \tag{10}$$

The proof is in Appendix B. To the best of our knowledge, this theoretical result provides the first formal link between empowerment maximization and reward maximization. This motivates us to develop a scalable algorithm for empowerment maximization, which we introduce in the following section.

4 Estimating and Maximizing Empowerment with Contrastive Representations

Directly computing equation 2 would require access to the human policy, which we don't have. Therefore, we want a tractable estimation that still performs well in large environments which are more difficult to model due to the exponentially increasing set of possible future states. To better-estimate empowerment, we learn contrastive representations that encode information about which future states are likely to be reached from the current state. These contrastive representations learn to model mutual information between the current state, action, and future state, which we then use to compute the empowerment objective.

4.1 Estimating Empowerment

To estimate this empowerment objective, we need a way of learning the probability ratio inside the expectation. Prior methods such as Du et al. [6] and Salge et al. [42] rollout possible future states and compute a measure of their variance as a proxy for empowerment, however this doesn't scale when the environment becomes complex. Other methods learn a dynamics model, which also doesn't scale when dynamics become challenging to model [27]. Modeling these probabilities directly is challenging in settings with high-dimensional states, so we opt for an indirect approach. Specifically, we will learn representations that encode two probability ratios. Then, we will be able to compute the desired probability ratio by combining these other probability ratios.

Our method will learn three representations:

1. $\phi(s, a^{\mathbf{R}}, a^{\mathbf{H}})$ – This representation can be understood as a sort of latent-space model, predicting the future representation given the current state s and the human's current action $a^{\mathbf{H}}$ as well as the robot's current action $a^{\mathbf{R}}$.
2. $\phi'(s, a^{\mathbf{R}})$ – This representation can be understood as an uncontrolled model, predicting the representation of a future state without reference to the current human action $a^{\mathbf{H}}$. This representation is analogous to a value function.
3. $\psi(s^+)$ – This is a representation of a future state.

We will learn these three representations with two contrastive losses, one that aligns $\psi(s, a^R, a^H) \leftrightarrow \psi(s^+)$ and one that aligns $\psi(s, a^R) \leftrightarrow \psi(s^+)$

$$\max_{\phi,\phi',\psi} \mathbb{E}_{\{(s_i,a_i,s_i')\sim p(\mathfrak{s}_t,\mathfrak{a}_t^{\mathbf{H}},\mathfrak{s}_{t+k})\}_{i=1}^N} \left[\mathcal{L}_{\mathsf{c}}(\{\phi(s_i,a_i)\},\{\psi(s_j')\}) + \mathcal{L}_{\mathsf{c}}(\{\phi'(s_i)\},\{\psi(s_j')\})\right],$$

where the contrastive loss $\mathcal{L}_{\mathsf{c}}$ is the symmetrized infoNCE objective [35]:

$$\mathcal{L}_{\mathsf{c}}(\{x_i\},\{y_j\}) \triangleq \sum_{i=1}^N \left[\log\left(\frac{e^{x_i^T y_i}}{\sum_{j=1}^N e^{x_i^T y_j}}\right) + \log\left(\frac{e^{x_i^T y_i}}{\sum_{j=1}^N e^{x_j{}^T y_i}}\right)\right]. \tag{11}$$

We have colored the index j for clarity. At convergence, these representations encode two probability ratios [24], which we will ultimately be able to use to estimate empowerment (Eq. (2)):

$$\phi(s, a^{\mathbf{R}}, a^{\mathbf{H}})^T\psi(g) = \log\left[\frac{p(\mathfrak{s}_{t+K} = g \mid \mathfrak{s}_t = s, \mathfrak{a}_t^{\mathbf{H}} = a^{\mathbf{H}}, \mathfrak{a}_t^{\mathbf{R}} = a^{\mathbf{R}})}{C_1 p(\mathfrak{s}_{t+K} = g)}\right] \tag{12}$$

$$\phi'(s, a^{\mathbf{R}})^T\psi(g) = \log\left[\frac{p(\mathfrak{s}_{t+K} = s_{t+k} \mid \mathfrak{s}_t = s_t, \mathfrak{a}_t^{\mathbf{R}} = a^{\mathbf{R}})}{C_2 p(\mathfrak{s}_{t+K} = g)}\right]. \tag{13}$$

Note that our definition of empowerment (Eq. (2)) is defined in terms of similar probability ratios. The constants C_1 and C_2 will mean that our estimate of empowerment may be off by an additive constant, but that constant will not affect the solution to the empowerment maximization problem.

4.2 Estimating Empowerment with the Learned Representations

To estimate empowerment, we will look at the difference between these two inner products:

$$\begin{aligned}
&\phi(s_{t+K}, a^{\mathbf{R}}, a^{\mathbf{H}})^T\psi(g) - \phi(s_{t+K}, a^{\mathbf{R}})^T\psi(g) \\
&= \log p(s_{t+K} \mid s, a^{\mathbf{H}}) - \log C_1 - \cancel{\log p(s_{t+K})} - \log p(s_{t+K} \mid s) + \log C_2 + \cancel{\log p(s_{t+K})} \\
&= \log\frac{p(s_{t+K} \mid s, a^{\mathbf{H}})}{p(s_{t+K} \mid s)} + \log\frac{C_2}{C_1}.
\end{aligned}$$

Note that the expected value of the first term is the *conditional* mutual information $I(s_{t+K}; a^{\mathbf{H}} \mid s)$. Our empowerment objective corresponds to averaging this mutual information across all the visited states. In other words, our objective corresponds to an RL problem, where empowerment corresponds to the expected discounted sum of these log ratios:

$$\mathcal{E}(\pi_H, \pi_R) = \mathbb{E}_{\pi_H,\pi_R}\left[\sum_{t=0}^\infty \gamma^t I(s^+; a_t^{\mathbf{H}} \mid s_t)\right]$$

Algorithm 1: Empowerment via Successor Representations (ESR)

Input: Human policy $\pi_H(a \mid s)$
Randomly initialize assistive agent policy $\pi_R(a \mid s)$, and representations $\phi(s, a^{\mathbf{R}}, a^{\mathbf{H}})$, $\psi(s, a^T)$, and $\psi(g)$.
Initialize replay buffer $\mathcal{B}$.
while not converged **do**
  Collect a trajectory of experience with human policy and assistive agent policy, store in replay buffer $\mathcal{B}$.
  Update representations $\phi(s, a^{\mathbf{R}}, a^{\mathbf{H}})$, $\psi(s, a^T)$, and $\psi(g)$ with the contrastive losses in Eq. (11).
  Update $\pi_R(a \mid s)$ with RL using reward function $r(s, a^{\mathbf{R}}, a^{\mathbf{H}}) = (\phi(s, a^{\mathbf{R}}, a^{\mathbf{H}}) - \phi'(s, a^{\mathbf{R}}))^T \psi(g)$.
Return: Assistive policy $\pi_R(a \mid s)$.

$$\approx \mathbb{E}_{\pi_H, \pi_R}\Big[\sum_{t=0}^{\infty} \gamma^t (\phi(s_t, a^{\mathbf{R}}, a^{\mathbf{H}}) - \phi(s_t, a^{\mathbf{R}}))^T \psi(g) - \log \frac{C_2}{C_1}\Big].$$

The approximation above comes from function approximation in learning the Bayes optimal representations. Again, note that the constants C_1 and C_2 do not change the optimization problem. Thus, to maximize empowerment we will apply RL to the assistive agent $\pi_R(a \mid s)$ using a reward function

$$r(s, a^{\mathbf{R}}) = (\phi(s_t, a^{\mathbf{R}}, a^{\mathbf{H}}) - \phi(s_t, a^{\mathbf{R}}))^T \psi(g). \tag{14}$$

4.3 Algorithm Summary

We propose an actor-critic method for learning the assistive agent. Our method will alternate between updating these contrastive representations and using them to estimate a reward function (Eq. (14)) that is optimized via RL. We summarize the algorithm in Algorithm 1. In practice, we use SAC [54] as our RL algorithm. In our experiments, we will also study the setting where the human user updates their policy alongside the assistive agent.

5 Experiments

We seek to answer two questions with our experiments. *First*, does our approach enable assistance in standard cooperation benchmarks? *Second*, does our approach scale to harder benchmarks where prior methods fail?

Our experiments will use two benchmarks designed by prior work to study assistance: the obstacle gridworld [6] and Overcooked [10]. Our main **baseline** is AvE [6], a prior empowerment-based method. Our conjecture is that both methods will perform well on the lower-dimensional gridworld task, and that our method will scale more gracefully to the higher dimensional Overcooked environment. We will also compare against a naïve baseline where the assistive agent acts randomly.

5.1 Do contrastive successor representations effectively estimate empowerment?

We test our approach in the assistance benchmark suggested in Du et al. [6]. The human (orange) is tasked with reaching a goal state (green) while avoiding the obstacles (purple). The AI assistant can move blocks one step at a time in any direction [6]. While the original benchmark used $N = 2$ obstacles, we will additionally evaluate on harder versions of this task with $N = 5, 7, 10$ obstacles. We show results in Fig. 3. On the easiest task, both our method and AvE achieve similar asymptotic reward, though our method learns more slowly than AvE. However, on the tasks with moderate and high degrees of complexity, our approach (ESR) achieves significantly higher rewards than AvE, which performs worse than a random controller. These experiments support our claim that contrastive successor representations provide an effective means for estimating empowerment, and hint that ESR might be well suited for solving higher dimensional tasks.

5.2 Does our approach scale to tasks with image-based observations?

Our second set of experiments look at scaling ESR to the image-based Overcooked environment. Since contrastive learning is often applied to image domains, we conjectured that ESR would scale gracefully to this setting. We will evaluate our approach in assisting a human policy trained with behavioral cloning taken from Laidlaw and Dragan [56]. The human prepares dishes by picking

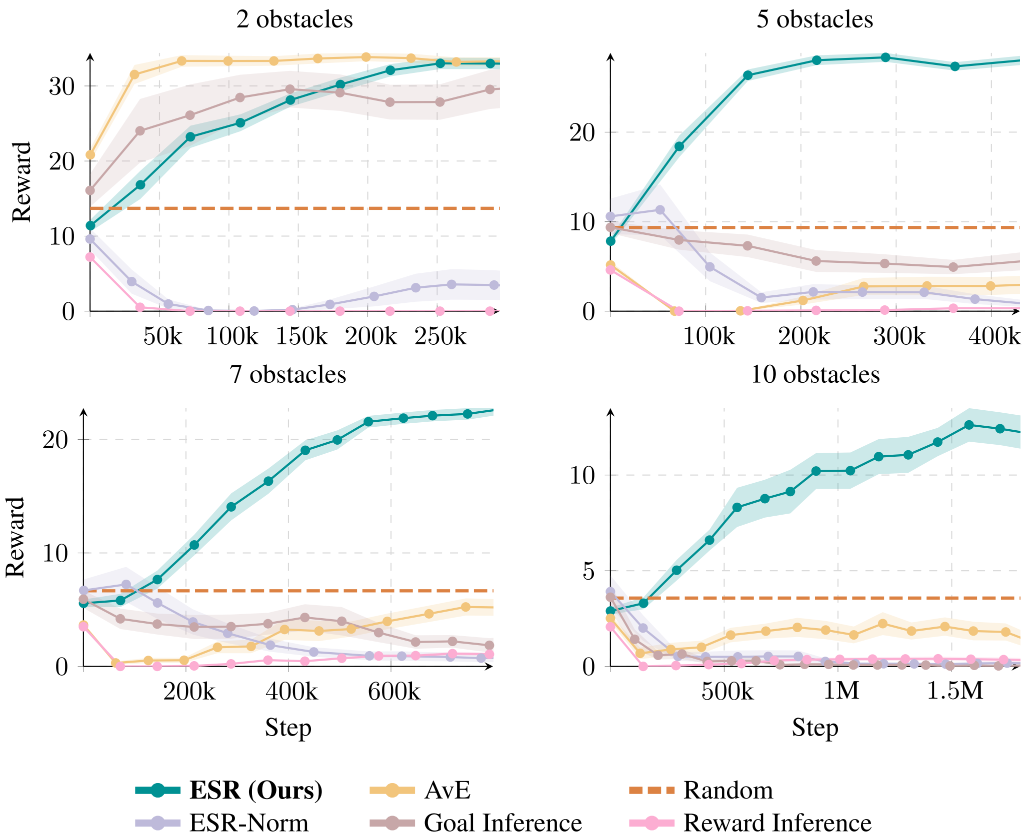
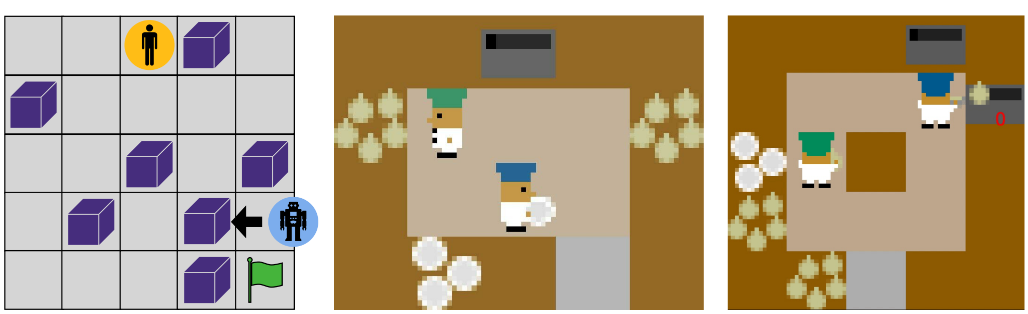


Figure 3: We apply our method to the benchmark proposed in prior work [6], visualized in Fig. 4a. The four subplots show variant tasks of increasing complexity (more blocks), (± 1 SE). We compare against AvE [6], the Goal Inference baseline from [6] which assumes access to a world model, and Reward Inference [55] where we recover the reward from a learned q-value. These prior approaches fail on all except the easiest task, highlighting the importance of scalability.



(a) Obstacle Gridworld (b) Cramped Room (c) Coordination Ring

Figure 4: *(a)* The modified environment from Du et al. [6] scaled to $N = 7$ blocks, and *(b, c)* the two layouts of the Overcooked environment [10].

up ingredients and cooking them on a stove, while the AI assistant moves ingredients and dishes around the kitchen. We focus on two environments within this setting: a cramped room where the human must pass ingredients and dishes through a narrow corridor, and a coordination ring where the human must pass ingredients and dishes around a ring-shaped kitchen (Figs. 4b and 4c). As before, we compare with AvE as well as a naïve random controller. We report results in Table 1. On both tasks, we observe that our approach achieves higher rewards than AvE baseline, which performs no better than a random controller. In Fig. 5, we show an example of one of the collaborative behaviors

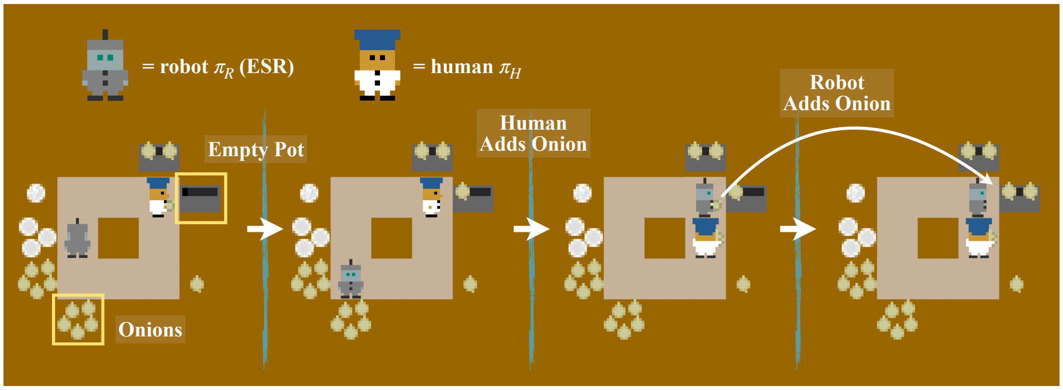


Figure 5: In Coordination Ring, our ESR agent learns to wait for the human to add an onion to the pot, and then adds one itself. There is another pot at the top which is nearly full, but the empowerment agent takes actions to maximize the impact of the human's actions, and so follows the lead of the human by filling the empty pot.

learned by ESR. Taken together with the results in the previous setting, these results highlight the scalability of ESR to higher dimensional problems.

Table 1: Overcooked Results

Layout	**ESR (Ours)**	Reward Inference	AvE	Random
Asymmetric Advantages	$\mathbf{72.00 \pm 5.37}$	60.33 ± 0.26	36.71 ± 1.71	59.36
Coordination Ring	$\mathbf{8.40 \pm 0.69}$	5.96 ± 0.20	5.69 ± 0.93	6.02
Cramped Room	$\mathbf{91.33 \pm 4.08}$	39.24 ± 0.35	5.13 ± 1.31	69.26

6 Discussion

One of the most important problems in AI today is equipping AI agents with the capacity to assist humans achieve their goals. While much of the prior work in this area requires inferring the human's intention, our work builds on prior work in studying how an assistive agent can *empower* a human user without inferring their intention. Relative to prior methods, we demonstrate how empowerment can be readily estimated using contrastive learning, paving the way for deploying these techniques on high-dimensional problems.

Limitations. One of the main limitations of our approach is the assumption that the assistive agent has access to the human's actions, which could be challenging to observe in practice. Automatically inferring the human's actions remains an important problem for future work. A second limitation is that the method is currently an on-policy method, in the sense that the assistive agent has to learn by trial and error. A third limitation is that the ESR formulation assumes that both agents share the same state space. In many cases the empowerment objective will still lead to desirable behavior, however, care must be taken in cases where the agent can restrict the information in its own observations, which could lead to reward hacking. Finally, our experiments do not test our method against real humans, whose policies may differ from the simulated policies. In the future, we plan to investigate techniques from off-policy evaluation and cooperative game theory to enable faster learning of assistive agents with fewer trials. We also plan to test the ESR objective in environments with partial observability over the human's state.

Safety risks. Perhaps the main risk involved with maximizing empowerment is that it may be at odds with a human's agents goal, especially in contexts where the pursuit of that goal limits the human's capacity to pursue other goals. For example, a family choosing to have a kid has many fewer options over where they can travel for vacation, yet we do not want assistive agents to stymie families from having children.

One key consideration is *whom* should be empowered. The present paper assumes there is a single human agent. Equivalently, this can be seen as maximizing the empowerment of all exogenous agents. However, it is easy to adapt the proposed method to maximize the empowerment of a single target individual. Given historical inequities in the distribution of power, practitioners must take care when considering whose empowerment to maximize. Similarly, while we focused on *maximizing* empowerment, it is trivial to change the sign so that an "assistive" agent minimizes empowerment. One could imagine using such a tool in policies to handicap one's political opponents.

Acknowledgments. We would like to thank Micah Carroll and Cameron Allen for their helpful feedback, as well as Niklas Lauffer for suggesting JaxMARL. This research was partly supported by ARL DCIST CRA W911NF-17-2-0181 and ONR N00014-22-1-2773, as well as NSF HCC 2310757, the Jump Cocosys Center, Princeton Research Computing, and the DoD through the NDSEG Fellowship Program.

References

[1] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. 2023.

[2] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. Cooperative Inverse Reinforcement Learning. *Advances in neural information processing systems*, 29, 2016.

[3] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J. Russell, and Anca Dragan. Inverse Reward Design. *Advances in neural information processing systems*, 30, 2017.

[4] Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. Estimating and Penalizing Preference Shift in Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 661–667. Association for Computing Machinery, 2021.

[5] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend to Seek Power. In *NeurIPS 2021*. arXiv, 2023.

[6] Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, and Anca Dragan. AvE: Assistance via Empowerment. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4560–4571. Curran Associates, Inc., 2020.

[7] Stuart Russell. Learning Agents for Uncertain Environments (Extended Abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 101–103. ACM, 1998.

[8] Saurabh Arora and Prashant Doshi. A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress. *Artificial Intelligence*, 297:103500, 2021.

[9] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A Survey of Preference-Based Reinforcement Learning Methods. *Journal of Machine Learning Research*, 2017.

[10] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the Utility of Learning about Humans for Human-AI Coordination. In *Conference on Neural Information Processing Systems*. arXiv, 2019.

[11] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*, arXiv:2201.03544. arXiv, 2022.

[12] Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. Causal Confusion and Reward Misidentification in Preference-Based Reward Learning. In *International Conference on Learning Representations*. arXiv, 2023.

[13] Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Preventing Reward Hacking with Occupancy Measure Regularization. 2024.

[14] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment – an Introduction. 2013.

[15] Ildefons Magrans de Abril and Ryota Kanai. A Unified Strategy for Implementing Curiosity and Empowerment Driven Reinforcement Learning. 2018.

[16] Sean Chen, Jensen Gao, Siddharth Reddy, Glen Berseth, Anca D. Dragan, and Sergey Levine. ASHA: Assistive Teleoperation via Human-in-the-Loop Reinforcement Learning. In *International Conference on Robotics and Automation*. 2022.

[17] Siddharth Reddy, Sergey Levine, and Anca Dragan. First Contact: Unsupervised Human-Machine Co-Adaptation via Mutual Information Maximization. *Advances in Neural Information Processing Systems*, 35:31542–31556, 2022.

[18] Andrew G. Barto. Intrinsic Motivation and Reinforcement Learning. In Gianluca Baldassarre and Marco Mirolli, editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47. Springer, 2013.

[19] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A Survey on Intrinsic Motivation in Reinforcement Learning. 2019.

[20] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity Is All You Need: Learning Skills without a Reward Function. In *International Conference on Learning Representations (ICLR)*. 2019.

[21] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by Random Network Distillation. 2018.

[22] Karl Friston. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[23] Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. SMiRL: Surprise Minimizing Reinforcement Learning in Unstable Environments. In *International Conference on Learning Representations*. arXiv, 2021.

[24] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

[25] Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational empowerment as representation learning for goal-conditioned reinforcement learning. In *International Conference on Machine Learning*, pp. 1953–1963. PMLR, 2021.

[26] Shakir Mohamed and Danilo Jimenez Rezende. Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[27] Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for Continuous Agent—Environment Systems. *Adaptive Behavior*, 19(1):16–39, 2011.

[28] Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery. 2022.

[29] Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-Constrained Unsupervised Skill Discovery. In *International Conference on Learning Representations*. 2021.

[30] Susanne Still and Doina Precup. An Information-Theoretic Approach to Curiosity-Driven Reinforcement Learning. *Theory in Biosciences*, 131(3):139–148, 2012.

[31] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-Directed Exploration for Deep Reinforcement Learning. 2019.

[32] Erdem Biyik, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences. 2021.

[33] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning Multimodal Rewards from Rankings. In *Conference on Robot Learning*, pp. 342–352. PMLR, 2021.

[34] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv preprint arXiv:1112.5745*, 2011.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. 2019.

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742. 2018.

[38] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pp. 5639–5650. PMLR, 2020.

[39] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R. Salakhutdinov. Contrastive Learning as Goal-Conditioned Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35603–35620. 2022.

[40] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993.

[41] Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. The successor representation in human reinforcement learning. *Nature human behaviour*, 1(9):680–692, 2017.

[42] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. *Guided Self-Organization: Inception*, pp. 67–114, 2014.

[43] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128–135. IEEE, 2005.

[44] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Shared Autonomy via Deep Reinforcement Learning. 2018.

[45] Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum Margin Planning. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pp. 729–736. ACM Press, 2006.

[46] Pieter Abbeel and Andrew Y. Ng. Apprenticeship Learning via Inverse Reinforcement Learning. In *Twenty-First International Conference on Machine Learning - ICML '04*, p. 1. ACM Press, 2004.

[47] Thomas M Cover. *Elements of Information Theory*. John Wiley & Sons, 1999.

[48] Shun-ichi Amari. Information Geometry. *Japanese Journal of Mathematics*, 16(1):1–48, 2021.

[49] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 34*. Springer International Publishing, Cham, 2017.

[50] Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10):1100, 2020.

[51] Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. The Information Geometry of Unsupervised Reinforcement Learning. In *International Conference on Learning Representations*. arXiv, 2022.

[52] Robert G. Gallager. Source Coding with Side Information and Universal Coding. 1979.

[53] Boris Yakovlevich Ryabko. Coding of a Source with Unknown but Ordered Probabilities. *Problems of Information Transmission*, 15(2):134–138, 1979.

[54] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[55] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse Soft-Q Learning for Imitation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4028–4039. Curran Associates, Inc., 2021.

[56] Cassidy Laidlaw and Anca Dragan. The Boltzmann Policy Distribution: Accounting for Systematic Suboptimality in Human Models. In *International Conference on Learning Representations*. arXiv, 2022.

[57] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 2017.

[58] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, et al. JaxMARL: Multi-Agent RL Environments in JAX. In *International Conference on Autonomous Agents and Multiagent Systems*. arXiv, 2024.

[59] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *International Conference on Machine Learning*, volume 30, p. 3. Atlanta, GA, 2013.

[60] Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *International Conference on Machine Learning*. arXiv, 2020.

[61] Benjamin Eysenbach, Vivek Myers, Ruslan Salakhutdinov, and Sergey Levine. Inference via Interpolation: Contrastive Representations Provably Enable Planning and Inference. 2024.

A Experimental Details

We ran all our experiments on NVIDIA RTX A6000 GPUs with 48GB of memory within an internal cluster. Each evaluation seed took around 5-10 hours to complete. Our losses (Eqs. (11) and (14)) were computed and optimized in JAX with Adam [57]. We used a hardware-accelerated version of the Overcooked environment from the JaxMARL package [58]. The experimental results described in Section 5 were obtained by averaging over 5 seeds for the Overcooked coordination ring layout, 15 for the cramped room layout, and 20 for the obstacle gridworld environment. Specific hyperparameter values can be found in our code, which is available at `https://github.com/vivekmyers/empowerment_successor_representations`.

A.1 Network Architecture

In the obstacle grid environment, we used a network with 2 convolutional and 2 fully connected layers and SiLU activations. In Overcooked, we adapted the policy architecture from past work [4], using 3 convolutional layers followed by 4 MLP layers with Leaky ReLU activations [59]. We concatenate in $a^{\mathbf{R}}$ and $a^{\mathbf{H}}$ to the state as one-hot encoded channels, i.e. if the action is 5, 6 additional channels will be concatenated to the state with all set to 0s except the 5th channel which is set to 1s.

B The Information Geometry of Empowerment

This section considers an objective that might be slightly different: $I^{\pi_R}(\mathfrak{s}^+; z)$, where z is a representation of the human's intention. In practice, this could be represented as a sequence of actions (as in the main doc above), but it also includes reactive and closed loop policies. This mutual information also depends on the human's policy π_H, but here we are interested in just the dependence on the robot.

Here's the primary question of interest: what actions/behaviors should the robot employ to maximize the mutual information between the human's intentions and the outcomes. Note that this is a standard mutual information skill learning objective. However, whereas prior work typically optimizes this objective w.r.t. the human's policy $\pi_H(a \mid s, z)$, here we aim to optimize this w.r.t. the robot's policy $\pi_R(a \mid s)$. Note that the robot is not conditioned on the human's intention z. We assume that this intention is not observed. [2] The objective can then be written as

$$\max_{\pi_R} I^{\pi_R}(\mathfrak{s}^+; z). \tag{15}$$

One way of thinking about this optimization problem is that we are modifying the MDP itself. However, rather than (say) changing the positions of clouds or changing the framerate, we will only consider changes that can be mediated by an interactive robot agent. These include changes such as pushing an object, opening a drawer, charging or discharging another robot.

B.1 Preliminaries: Relating Mutual Information to Reward Maximization

The mutual information is an information theoretic quantity, defined in terms of bits and probabilities. However, what we actually care about is the ability of an assisted human to achieve high rewards.

[2] One area for future work is to study whether actively inferring this intention improves assistance. Another area for future work is to study whether non-Markovian robot policies can perform better than their Markovian counterparts because they can accumulate information about the human's intentions across time.

So, we need a way of relating this mutual information objective to reward maximizing. We start by recalling the result from Eysenbach et al. [51], which provides one such relation:

$$\max_{\pi_H(z)} I(\mathfrak{s}^+; z) = \min_{\rho(s)\in\mathcal{C}} \max_{r(s)} \min_{\rho^*\in\mathcal{C}} \underbrace{\max_{\rho^+\in\mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)]}_{\text{regret}} + D_{\mathrm{KL}}\big(\rho^*(s) \,\|\, \rho(s)\big). \tag{16}$$

Unpacking this objective. There's a lot of math in that equation, so let's unpack it a bit. The LHS is about learning skills for the human policy π_H. We assume that all possible skills are enumerated, so the human simply has to select from this menu of skills by deciding how much more of each skill $\pi_H(z)$ to order from this menu. Of course, this menu is exponentially long, but it is finite and well defined, and practical algorithms won't actually attempt to enumerate this menu of skills. The optimization problem on the LHS is about selecting those skills that most readily maximize the mutual information – the skills that have a strong influence over the states visited in the future.

The RHS has a whole bunch of terms. For a given reward function $r(s)$, we care about how much reward a particular policy gets. The RHS studies this standard expected reward by using the dual of the RL problem, thinking about the states $\rho(s)$ visited by a policy and counting up the rewards at those states. The term $\mathbb{E}_{\rho^+(s)}[r(s)]$ is the expected reward for a policy with occupancy measure $\rho^+(s)$. Thus, $\max_{\rho^+(s)} \mathbb{E}_{\rho^+(s)}[r(s)]$ is the maximal reward that any policy can get on this particular reward function.

We are often given a policy (or its occupancy measure $\rho^*(s)$) and reward function $r(s)$ and want to know how good that policy is for that reward function. While we could directly measure the expected reward, we usually don't know whether this is a particularly good value or not. Instead, we might measure the *regret* of the policy: how much *lower* is its expected reward, as compared to the optimal policy for that reward function:

$$\textsc{Regret}(\rho(s)^*, r(s)) = \max_{\rho^+(s)} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)]. \tag{17}$$

This is the term that appears on the RHS on Eq. (16). Now, when we have limited data, we usually want to minimize a regularized notion of regret. This is what ρ^* is doing, using the KL divergence against a prior $\rho(s)$ as the regularization term:

$$\min_{\rho^*(s)} \textsc{Regret}(\rho^*(s), r(s)) + D_{\mathrm{KL}}\big(\rho^*(s) \,\|\, \rho(s)\big). \tag{18}$$

This objective above can be interpreted as the difficulty of learning to maximize reward function $r(s)$. But, in the unsupervised RL setting, which reward functions should we learn how to optimize? We could take an average case approach, but this runs into challenges because "average" depends on a choice of measure. Instead, we take a worst-case approach, selecting the reward function that is most challenging to adapt to:

$$\max_{r(s)} \min_{\rho^*(s)} \textsc{Regret}(\rho^*(s), r(s)) + D_{\mathrm{KL}}\big(\rho^* s \,\|\, \rho(s)\big). \tag{19}$$

Finally, when discussing adaptation, we had some prior $\rho(s)$ to which we were referring. The overall aim is to find the prior $\rho(s)$ that makes it easy to adapt to the most challenging reward function:

$$\min_{\rho(s)\in\mathcal{C}} \max_{r(s)} \textsc{AdaptationObjective}(\rho(s), r(s)). \tag{20}$$

Equation (16) tells us that the problem of finding this optimal prior is equivalent to maximizing mutual information.

B.2 Application to Empowerment

We can extend this result to the assistive setting, thinking about how an assistive robot should act to make it easier for a human to maximize their worst-case rewards. From the human's perspective, the robot is just another part of the MDP.[3] So, to apply the result from Equation (16) to empowerment, we just need to modify the definitions to depend on the choice of π_R.

[3]The reason we wanted to assume that the robot was Markovian was so that this remains a *Markov* decision process.

On the LHS, let's use $I^{\pi_R}(\mathfrak{s}^+; z)$ to denote the mutual information between the *human's* choice of skills $\pi_H(z)$ and the future states, when interacting in an environment alongside a robot $\pi_R(a \mid s)$. The RHS thinks about the state occupancy measure of the human, terms like $\rho(s), \rho^+(s), \rho^*(s)$. An effective assistive agent will enable a human to visit a wide distribution over states, or to spend more time visiting any given state. We will use $\mathcal{C}^{\pi_R}$ to denote the feasible occupancy measures when interacting alongside an assistive agent.

B.3 Assistive Agents Minimize Regret

We can now state our main result, which is a direct corollary of Eq. (16) Consider the human and the robot as one monolithic agent selecting actions $a_H, a_R \sim \pi_H(a_H \mid s, z)\pi_R(a_R \mid s)$. This policy is Markovian, so we can immediately apply Eq. (16).

We start with some intuition: we would like an assistive agent to help the human maximize rewards. The challenge is that the assistive agent doesn't know what reward function the human is trying to solve, and we would like to avoid this inverse RL problem. So, we will take a worst-case approach, thinking about how the assistive agent can help the human solve the hardest task. We will measure difficulty as a combination of (1) regret versus the optimal policy, and (2) divergence from a prior over policies.

Notation. Let $\mathcal{C}^{\pi_R}$ denote the set of feasible state marginal distributions with cooperating with assistive agent $\pi_R(a \mid s)$. We assume that this assistive agent does not know the human's intention. We will measure regret against an omniscient assistive agent, which knows the human's intent. Thus, we compare to an occupancy measure optimized within the larger set $\mathcal{C}$, which includes adaptive strategies.

Assume human $\pi_H(a \mid s)$ and robot $\pi_R(a \mid s)$ induce state occupancy measure $\rho^*(s)$. We define their regret, which is measured relative to the highest reward they could achieve with *any* assistive agent (hence, we use $\rho^+ \in \mathcal{C}$ rather than $\rho^+ \in \mathcal{C}^{\pi_R}$):

$$\max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)].$$

We will include an additional regularization term, so the overall objective becomes

$$\max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{\text{KL}}\big(\rho^*(s) \parallel \rho(s)\big).$$

Given a reward function $r(s)$, we assume that the human adapts by minimizing this regularized regret. We assume that the assistive agent does not adapt. Thus, the human is optimizing over the smaller set $\mathcal{C}^{\pi_R}$:

$$\min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{\text{KL}}\big(\rho^*(s) \parallel \rho(s)\big).$$

As before, the reward function is adversarially chosen. And, the human's job is to find the prior $\rho(s)$ that is minimax optimal:

$$\min_{\rho(s) \in \mathcal{C}^{\pi_R}} \max_{r(s)} \min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{\text{KL}}\big(\rho^*(s) \parallel \rho(s)\big).$$

Lemma 4. *Assume that a human has learned skills $\pi(a \mid s, z)$ by maximizing mutual information $I(\mathfrak{s}^+; z)$ and adapts to a reward function by minimizing the regularized regret:*

$$\min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{KL}\big(\rho^*(s) \parallel \rho(s)\big).$$

We assume that the human chooses the prior $\rho(s)$ that minimizes this regret for the worst-case choice of reward function (i.e., the minimax optimal prior). An assistive agent that maximizes $I^{\pi_R}(\mathfrak{s}^+; z)$ minimizes the worst-case (regularized) regret incurred by the human.

Letting $\pi_R^* \in \arg\max_{\pi_R} I^{\pi_R}(\mathfrak{s}^+; z)$, we have

$$\pi_R^* \in \arg\min_{\pi_R} \left(\min_{\rho(s) \in \mathcal{C}^{\pi_R}} \max_{r(s)} \min_{\rho^* \in \mathcal{C}^{\pi_R}} \max_{\rho^+ \in \mathcal{C}} \mathbb{E}_{\rho^+(s)}[r(s)] - \mathbb{E}_{\rho^*(s)}[r(s)] + D_{KL}\big(\rho^*(s) \parallel \rho(s)\big) \right). \tag{21}$$

C Simplifying the Objective

The reward function in Eq. (14) is itself a random variable because it depends on future states g. This subsection describes how this randomness can be removed. To do this, we follow prior work [60, 61] in arguing that the learned representations $\psi(g)$ follow a Gaussian distribution:

Assumption 1 (Based on Wang and Isola [60])**.** *The representations of future states $\psi(g)$ learned by contrastive learning have a marginal distribution that is Gaussian:*

$$p(\psi) = \int p(g)\delta(\psi = \psi(g))\,\mathrm{d}g \stackrel{d}{=} \mathcal{N}(0, I). \tag{22}$$

With this assumption, we can remove the random sampling of g from the reward function. We start by noting that the learned representations tell us the *relative* likelihood of seeing a future state Eq. (13)). Assumption 1 will allow us to convert these relative likelihoods into likelihoods.

$$\begin{aligned}
\mathbb{E}_{p(\mathfrak{s}^+|s,a^{\mathbf{R}},a^{\mathbf{H}})}[r(s,a^{\mathbf{R}})] &= \mathbb{E}_{p(\mathfrak{s}^+)}\left[\tfrac{p(\mathfrak{s}^+|s,a^{\mathbf{R}},a^{\mathbf{H}})}{p(\mathfrak{s}^+)} r(s,a^{\mathbf{R}})\right] \\
&= \mathbb{E}_{p(\mathfrak{s}^+)}\left[C_1 e^{\phi(s,a^{\mathbf{R}},a^{\mathbf{H}})^T\phi(\mathfrak{s}^+)} r(s,a^{\mathbf{R}})\right] \\
&= C_1 \mathbb{E}_{\psi\sim p(\phi(\mathfrak{s}^+))}\left[e^{\phi(s,a^{\mathbf{R}},a^{\mathbf{H}})^T\psi}(\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}) - \phi(s,a^{\mathbf{R}}))^T\psi\right] \\
&= C_1\big(\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}) - \phi(s,a^{\mathbf{R}})\big)^T \int \tfrac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\psi\|_2^2 + \phi(s,a^{\mathbf{R}},a^{\mathbf{H}})^T\psi}\psi\,\mathrm{d}\psi \\
&= C_1\big(\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}) - \phi(s,a^{\mathbf{R}})\big)^T e^{\frac{1}{2}\|\phi(s,a^{\mathbf{R}},a^{\mathbf{H}})\|_2^2} \\
&\qquad \int \tfrac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\psi\|_2^2 + \phi(s,a^{\mathbf{R}},a^{\mathbf{H}})^T\psi - \frac{1}{2}\|\phi(s,a^{\mathbf{R}},a^{H})\|_2^2}\psi\,\mathrm{d}\psi \\
&= C_1\big(\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}) - \phi(s,a^{\mathbf{R}})\big)^T e^{\frac{1}{2}\|\phi(s,a^{\mathbf{R}},a^{\mathbf{H}})\|_2^2}\mathbb{E}_{\psi\sim\mathcal{N}(\mu=\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}),\Sigma=I)}\left[\psi\right] \\
&= C_1 e^{\frac{1}{2}\|\phi(s,a^{\mathbf{R}},a^{\mathbf{H}})\|_2^2}\big(\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}) - \phi(s,a^{\mathbf{R}})\big)^T\phi(s,a^{\mathbf{R}},a^{\mathbf{H}}).
\end{aligned} \tag{23}$$

D Learning Representations without the Robot Action

In our estimation of empowerment Eq. (13)) we supply the robot action $a^{\mathbf{R}}$ when learning both ϕ and ϕ, however, the theoretical empowerment formulation in Section 3.4 does not require it.

To evaluate the impact of including $a^{\mathbf{R}}$, we run an additional ablation without it on the gridworld environment, shown in Fig. 6. This ablation shows that the inclusion of $a^{\mathbf{R}}$ is most impactful in more challenging (higher number of boxes) environments. We hypothesize that conditioning the representations on the robot action reduces the noise in the mutual information estimation, and also reduces the difficulty of classifying true future states.

E Greedy Empowerment Policy

All of our experiments have used Soft-Q learning to learn a policy from the empowerment estimation. Here, we additionally study a greedy empowerment policy which takes the most empowering action at each step. We model this by setting the q-learning gamma to 0 to fully discount future rewards.

Results for this ablation are shown in Fig. 7. Unsurprisingly, the greedy optimization vastly underperforms the policy with $\gamma = 0.9$.

F Measuring ESR Training

This section plots the mutual information during training of the ESR agent in the gridworld environment with 5 obstacles. In 8, the mutual information quickly becomes positive and remains so throughout training. As long as the mutual information is positive, the classifier is able to reward the agent for taking actions that empower the human.

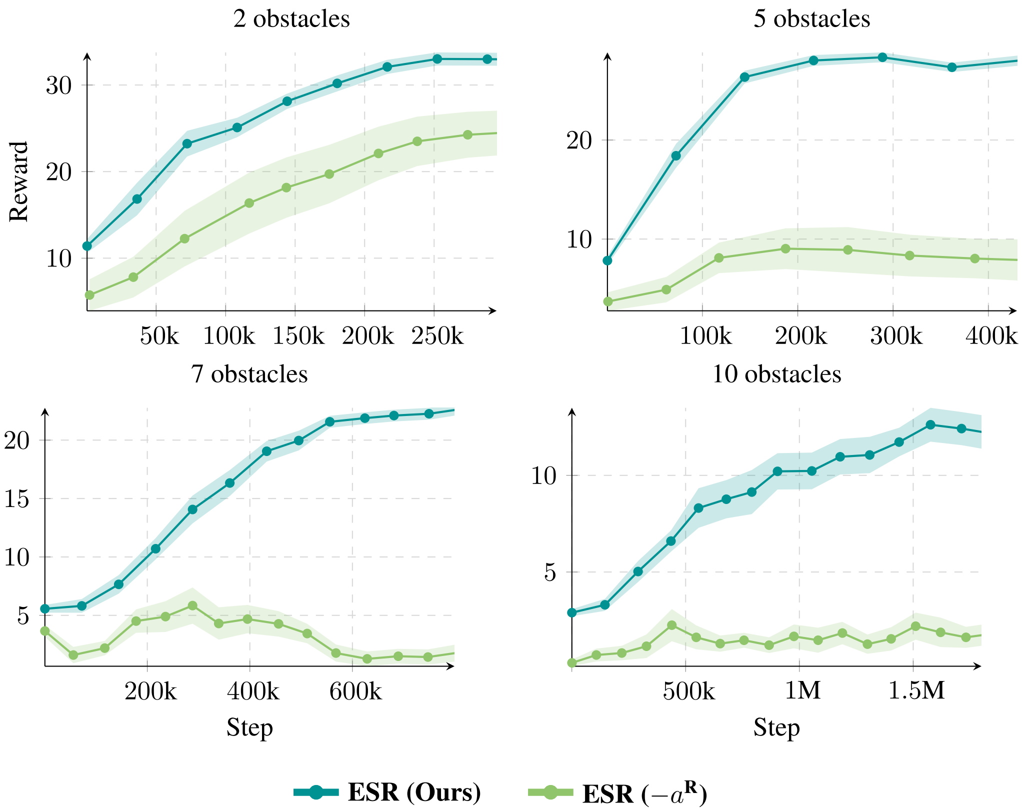


Figure 6: We evaluate our method with and without conditioning on the robot action $a^{\mathbf{R}}$. Conditioning aids learning significantly, which we theorize is because it removes uncertainty in the classification.

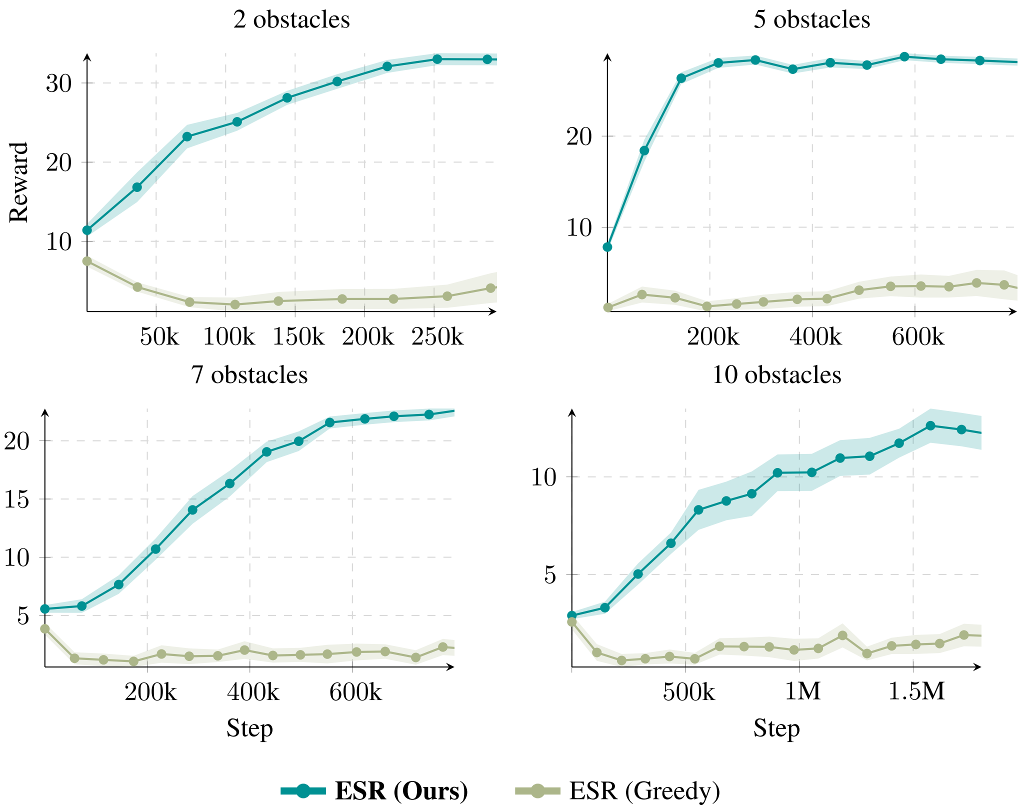


Figure 7: We compare a greedy policy ($\gamma = 0$) against our standard policy ($\gamma = 0.9$).

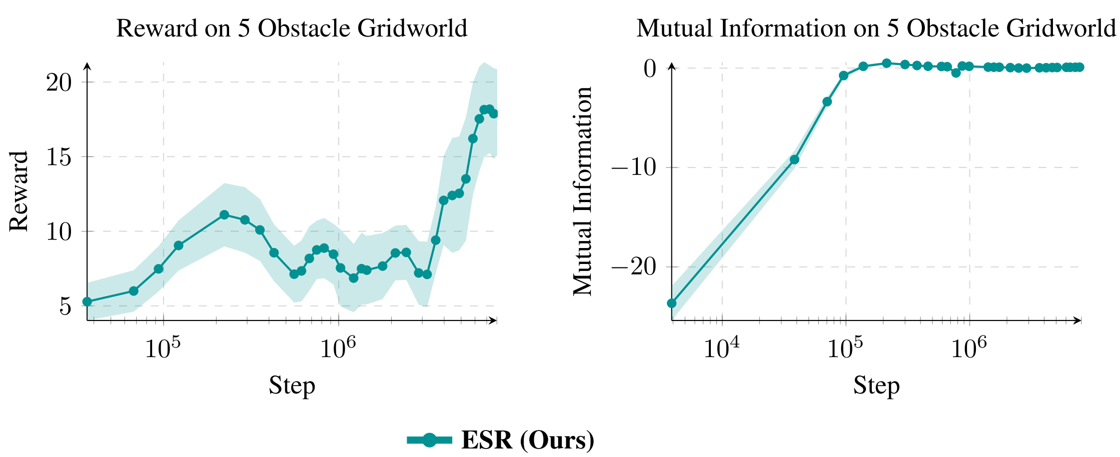


Figure 8: Reward and mutual information plots for ESR shown in the 5 obstacles gridworld environment. Mutual information increases at the start, and then levels off at a slightly positive value for the rest of training.