

EXPECTED RETURN SYMMETRIES

Anonymous authors

Paper under double-blind review

ABSTRACT

Symmetry is an important inductive bias that can improve model robustness and generalization across many deep learning domains. In multi-agent settings, *a priori known* symmetries have been shown to address a fundamental coordination failure mode known as *mutually incompatible symmetry breaking*; e.g. in a game where two independent agents can choose to move “left” or “right”, and where a reward of +1 or −1 is received when the agents choose the same action or different actions, respectively. However, the efficient and automatic *discovery* of environment symmetries, in particular for decentralized partially observable Markov decision processes, remains an open problem. Furthermore, environmental symmetry breaking constitutes only one type of coordination failure, which motivates the search for a more accessible and broader symmetry class. In this paper, we introduce such a broader group of previously unexplored symmetries, which we call *expected return symmetries*, which contains environment symmetries as a subgroup. We show that agents trained to be compatible under the group of expected return symmetries achieve better zero-shot coordination results than those using environment symmetries. As an additional benefit, our method makes minimal *a priori* assumptions about the structure of their environment and does not require access to ground truth symmetries.

1 INTRODUCTION

Incorporating the symmetries of an underlying problem into models has had demonstrable success in improving generalization and accuracy across many different machine learning domains (Bronstein et al., 2021; Krizhevsky et al., 2017; Cohen et al., 2019; Finzi et al., 2020; Van der Pol et al., 2020). As an important example, using data augmentations and equivariant networks has been shown to improve *zero-shot coordination* (ZSC), the ability of independently trained agents to coordinate in cooperative multi-agent settings at test time (Hu et al., 2020; Muglich et al., 2022a). Without accounting for symmetries, independently trained agents can converge onto equivalent yet mutually incompatible policies during training, leading to coordination failure at test time. For example, one team of agents might use the color “blue” to signal “play” in a cooperative card game like Hanabi, while a different team might use the equivalent color “red”. This issue, known as *mutually incompatible symmetry breaking*, can be mitigated by incorporating symmetries like color into the training process, as is done in other-play (OP) (Hu et al., 2020). OP addresses this by applying an independently sampled symmetry transformation to each agent during every training episode, ensuring compatibility amongst equivalent policies.

Environmental symmetry breaking constitutes a type of *over-coordination*, a situation where agents adopt arbitrary and obscure conventions that hinder the ability of previously unseen partners to adapt effectively within the scope of an episode. However, environmental symmetry breaking is far from a complete characterization of over-coordination, many other coordination failure modes exist. For instance, even in environments that lack non-trivial environmental symmetries, agents can still over-coordinate by adopting overly specific conventions (see Example 2). Another key limitation with current symmetry-based methods is that they assume *a priori* access to said symmetries, while their automatic discovery, especially in large-scale decentralized partially observable Markov decision processes (Dec-POMDPs (Oliehoek et al., 2007)), can be computationally infeasible (Narayanamurthy & Ravindran, 2008).

To address these two issues, in this paper, we define the group of *expected return symmetries* (ER symmetries), a relatively underexplored symmetry group that contains the environment symmetries of a Dec-POMDP as a subgroup. Since in most cases ER symmetries are a strict superset of the envi-



Figure 1: Mutually incompatible symmetry breaking between chess players is shown in the left side panels (Chess.com, 2021). Let π_1, π_2 represent the joint policies under which both players choose handshake, fist bump, respectively, and let the reward be ± 1 , depending on whether they match or not. The self-play score of both joint policies is 1, but the cross-play score between them is -1 . Thus, policies incompatibly break the symmetry between handshake and fist bump.

ronment symmetries, using them as the symmetry group for OP enforces training time compatibility with a greater number of policies.

We also introduce a scalable method for discovering approximate ER symmetries, which leverages gradient-based optimization to search for transformations that preserve expected returns across optimal policies. Furthermore, we show that ER symmetries better improve zero-shot coordination amongst independently trained agents than Dec-POMDP symmetries, while maintaining completely model-free assumptions. To summarize, our main contributions are:

1. We define the group of expected return symmetries and introduce novel algorithms for learning them.
2. We demonstrate that, when combined with the OP learning rule, expected return symmetries are significantly more effective at preventing over-coordination than Dec-POMDP symmetries. Furthermore, we demonstrate that our method is applicable in settings where both off-belief learning Hu et al. (2021) and cognitive hierarchies Cui et al. (2021) fail.
3. We empirically demonstrate that expected return symmetries can be used as a policy improvement operator at test time by using a *symmetrizer*; an operator that maps any policy to a policy that is invariant with respect to a given set of symmetries (see Section 4.4).
4. To the best of our knowledge, our method is the first symmetry-based method to improve zero-shot coordination without a priori/privileged environment information, such as of symmetries or dynamics.

2 BACKGROUND

2.1 DECENTRALIZED PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

We formalize the cooperative multi-agent setting as a decentralized partially observable Markov decision process (Dec-POMDP) Oliehoek et al. (2007), defined as a 9-tuple $(\mathcal{S}, \mathcal{N}, \{\mathcal{A}^i\}_{i=1}^n, \{\mathcal{O}^i\}_{i=1}^n, \mathcal{T}, \mathcal{R}, \{\mathcal{U}^i\}_{i=1}^n, H, \gamma)$, where:

- \mathcal{S} is the set of states, $\mathcal{N} = \{1, \dots, n\}$ is the set of agents.
- \mathcal{A}^i and \mathcal{O}^i are the action and observation sets for agent i , with joint action and observation sets $\mathcal{A} = \times_i \mathcal{A}^i$ and $\mathcal{O} = \times_i \mathcal{O}^i$.
- The state transition is governed by $s_{t+1} \sim \mathcal{T}(s_{t+1} | s_t, a_t)$, and observations are drawn from $o_{t+1} \sim \mathcal{U}(o_{t+1} | s_{t+1}, a_t)$.
- Rewards $r_{t+1} \sim \mathcal{R}(r_{t+1} | s_{t+1}, a_t)$, and the horizon is H , with a discount factor $\gamma \in [0, 1]$.

Each agent i selects actions based on its local action-observation history $\tau_t^i = (a_0^i, o_1^i, \dots, a_{t-1}^i, o_t^i)$, following a local policy $a_t^i \sim \pi^i(a_t^i | \tau_t^i)$. The joint policy $\pi = (\pi^1, \dots, \pi^n)$ depends on the joint history $\tau = (\tau^1, \dots, \tau^n)$. The self-play (SP) objective $J : \Pi \rightarrow \mathbb{R}$ is the expected return:

$$J(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{H-1} \gamma^t r_{t+1} \right].$$

2.2 ZERO-SHOT COORDINATION

The self-play objective is widely used in multi-agent reinforcement learning (MARL) (Samuel, 1959; Tesauro et al., 1995), where agents train together under a joint policy. While effective for coordination, it often results in arbitrary conventions that only work among agents trained together.

108 However, many real-world tasks require coordination with unknown partners (Mariani et al., 2021;
109 Resnick et al., 2018; Kakish et al., 2019), making this limitation problematic.

110 To address this, Hu et al. (2020) introduced zero-shot coordination (ZSC). In ZSC, agents first
111 agree on a learning rule, which they each implement independently (e.g., cannot agree on seeds).
112 Each agent then trains a joint policy in a Dec-POMDP environment, without communication or
113 coordination between agents during training. Finally, agents participate in *cross-play* (XP), where
114 joint policies trained by different agents are combined to evaluate the XP objective (defined here
115 only for $n = 2$, but it can be extended to $n > 2$):

$$116 \text{XP}(\pi_1, \pi_2) := \frac{1}{2} (J(\pi_1^1, \pi_2^2) + J(\pi_2^1, \pi_1^2)), \quad (1)$$

117 for independently learned joint policies π_1 and π_2 . ZSC aims for learning rules that optimize cross-
118 play with other rational partners using the same minimal set of assumptions (e.g., no access to be-
119 havioral data or coordination experience with specific groups of agents). ZSC presents a promising
120 approach to addressing real-world coordination challenges where relying on arbitrary conventions
121 is impractical. ZSC has become a key benchmark for human-AI coordination and is an important
122 step towards more generalized coordination capabilities (Ji et al., 2023; Hu et al., 2021).

124 2.3 SYMMETRY GROUPS AND OTHER-PLAY IN DEC-POMDPS

125 We consider symmetries that can be expressed as maps $\phi = (\phi_S, \phi_A, \phi_O)$, consisting of bijective
126 maps $\phi_S : \mathcal{S} \rightarrow \mathcal{S}$, $\phi_A : \mathcal{A} \rightarrow \mathcal{A}$, and $\phi_O : \mathcal{O} \rightarrow \mathcal{O}$, which satisfy $\phi_A(\mathcal{A}^i) = \mathcal{A}^i$ and $\phi_O(\mathcal{O}^i) = \mathcal{O}^i$
127 for all $i = 1, \dots, n$. The set of all such maps forms a group, which we denote by Ψ . Given $\phi \in \Psi$,
128 we slightly abuse notation and define $\phi(s) := \phi_S(s)$, $\phi(a) := \phi_A(a)$ and $\phi(o) := \phi_O(o)$, for
129 $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $o \in \mathcal{O}$. Given a joint action-observation history $\tau = (\tau^1, \dots, \tau^n)$, we define
130 $\phi(\tau) := (\phi(\tau^1), \dots, \phi(\tau^n))$, with $\phi(\tau^i) := (\phi(a_0^i), \phi(o_1^i), \dots, \phi(a_{i-1}^i), \phi(o_i^i))$. Furthermore, we let
131 a symmetry $\phi \in \Psi$ act on joint policies through the formula

$$132 \phi(\pi)(\phi(a) | \phi(\tau)) := \pi(a | \tau). \quad (2)$$

133 Any subgroup $\Phi \subset \Psi$ partitions the space of joint policies into disjoint equivalence classes: given a
134 joint policy π , we define its equivalence class $[\pi] := \{\phi(\pi) : \phi \in \Phi\}$.

135 **Definition 1** (Dec-POMDP Symmetries). A map $\phi \in \Psi$ is called a Dec-POMDP symmetry if for
136 all $(o', s, a, s', r) \in \mathcal{O} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R}$ it holds that

$$137 \begin{aligned} 138 \mathcal{T}(s' | s, a) &= \mathcal{T}(\phi(s') | \phi(s), \phi(a)), \\ 139 \mathcal{U}(o' | s, a) &= \mathcal{U}(\phi(o') | \phi(s), \phi(a)), \\ 140 \mathcal{R}(r | s, a) &= \mathcal{R}(r | \phi(s), \phi(a)). \end{aligned} \quad (3)$$

141 We denote the set of all the Dec-POMDP symmetries of a given Dec-POMDP by Φ^{MDP} .

142 Dec-POMDP symmetries form a subgroup of Ψ , which corresponds to relabelings of the state,
143 action, and observation spaces that leave all the transition and reward functions of the Dec-POMDP
144 unchanged (see Example 2). We note that Dec-POMDP symmetries can be extended to also include
145 permutations between players (Treutlein et al., 2021).

146 Hu et al. (2020) demonstrate that policies equivalent under Φ^{MDP} are prone to mutually incompat-
147 ible symmetry breaking: without biases like initialization or reward shaping, the learning rule may
148 converge to either π or its equivalent $\phi(\pi)$, as the learning process cannot distinguish between them
149 due to the symmetries of the Dec-POMDP. Although $J(\pi) = J(\phi(\pi))$, the cross-play return may
150 suffer ($J(\pi) > \text{XP}(\pi, \phi(\pi))$), meaning π and $\phi(\pi)$ were not trained to be compatible. See Figure 1.

151 To constrain policies to be compatible with policies that are equivalent with respect to the symmetry
152 group Φ^{MDP} , Hu et al. (2020) introduced the *other-play objective* (can be extended to $n > 2$):

153 **Definition 2** (Other-Play (OP) Objective). Given a Dec-POMDP and a symmetry group $\Phi \subset \Psi$, we
154 define the other-play (OP) objective $\text{OP}^\Phi : \Pi \rightarrow \mathbb{R}$ w.r.t. Φ by

$$155 \text{OP}^\Phi(\pi) := \mathbb{E}_{\tilde{\pi} \in [\pi]} [\text{XP}(\pi, \tilde{\pi})] = \frac{1}{2|[\pi]|} \sum_{\tilde{\pi} \in [\pi]} (J(\pi^1, \tilde{\pi}^2) + J(\tilde{\pi}^1, \pi^2)). \quad (4)$$

156 Hu et al. (2020) proposed the *OP learning rule* $\pi_* = \arg \sup_{\pi} \text{OP}^\Phi(\pi)$, for $\Phi = \Phi^{\text{MDP}}$. Agents
157 trained using the OP objective take into account modes of symmetry breaking resulting from the fact
158 that a test-time partner is unbiased in their choice between $\phi(\pi)$ and π , $\forall \phi \in \Phi$.

3 METHOD

In Section 3.1, we generalize the OP objective (Equation 4) to be defined over a general symmetry group Φ and highlight desirable properties that a symmetry group for OP should satisfy. In Section 3.2, we define the group of expected return symmetries, which we argue is better suited for OP than Dec-POMDP symmetries by way of the aforementioned desirable properties. In Section 3.3, we propose a method for learning expected return symmetries.

3.1 OTHER-PLAY OVER GENERAL SYMMETRY GROUPS

While Hu et al. (2020) introduced symmetry breaking w.r.t. Φ^{MDP} , we extend the definition for a general symmetry group Φ :

Definition 3 (Symmetry Breaking). Given a Dec-POMDP and a symmetry group Φ , we define a joint policy π to be incompatibly breaking symmetry w.r.t. $\phi \in \Phi$ if $J(\pi) > \text{XP}(\pi, \phi(\pi))$, and w.r.t. Φ if $J(\pi) > \text{OP}^\Phi(\pi)$.

The OP objective (Equation 4) evaluates the expected return when an agent from one policy is matched in a team with members of randomly chosen policies from the same equivalence class induced by Φ . Thus OP optimal policies are maximally compatible with policies within their equivalence class, as they best avoid incompatible symmetry breaking w.r.t. Φ . We note that different OP^Φ -optimal policies are not necessarily in the same equivalence class and can therefore be incompatible; for example, when $\Phi = \text{Id}$, OP^Φ reduces to SP, and each OP^Φ -optimal policy forms its own one-element equivalence class. We denote Π_*^Φ as the set of all optimal policies under OP^Φ .

Rational and independent ZSC agents would therefore choose a symmetry group Φ such that:

1. **(Diversity within Equivalence Classes)** The choice of Φ should ensure that for all $\pi \in \Pi_*^\Phi$, the equivalence class $[\pi]$ is meaningfully *diverse*. This diversity should be such that using OP to enforce π and $[\pi]$ to be compatible makes π broadly compatible with policies it could encounter at test-time, assuming other agents also adopt OP^Φ as their learning rule.
2. **(Optimality within Equivalence Classes)** Φ should separate poor policies from good policies, i.e. $\text{OP}^\Phi(\pi') \approx \text{OP}^\Phi(\pi)$, for any $\pi \in \Pi_*^\Phi$ and $\pi' \in [\pi]$, since otherwise there is no reason to constrain oneself to be compatible with π' (a rational test-time partner would not use π').

Example 1. For $\Phi = \{\text{Id}\}$, $\text{OP}^\Phi = \text{SP}$, which corresponds to perfectly preserving optimality, but not introducing any diversity to the equivalence class; i.e. perfect satisfaction of Item 2 but extremely poor satisfaction of Item 1. On the other extreme, we can consider Φ to be the set of all bijections on Π , which enforces compatibility with every possible test-time partner (i.e., rational partners), but also with all other possible policies (which are mostly poor), thus converging to the best response to a random player; i.e., perfect satisfaction of Item 1 but extremely poor satisfaction of Item 2.

The goal of ZSC is to find a learning rule that maximizes the expected XP score between independently trained test-time partners. For the learning rule OP^Φ the expected XP score is:

$$\text{XP}(\text{OP}_*^\Phi) := \mathbb{E}_{\pi_1, \pi_2 \sim \Pi_*^\Phi} [\text{XP}(\pi_1, \pi_2)]. \quad (5)$$

Thus one can interpret $\max_{\pi \in \Pi} \text{OP}^\Phi(\pi)$ as estimating $\text{XP}(\text{OP}_*^\Phi)$, but only through the cross-play scores *within* a given equivalence class $[\pi]$ induced by Φ :

$$\max_{\pi \in \Pi} \text{OP}^\Phi(\pi) = \mathbb{E}_{\pi \sim \Pi_*^\Phi} [\text{OP}^\Phi(\pi)] = \mathbb{E}_{\pi_1 \sim \Pi_*^\Phi} [\mathbb{E}_{\pi_2 \sim [\pi_1]} [\text{XP}(\pi_1, \pi_2)]] . \quad (6)$$

Since optimal OP^Φ policies are not trained to be compatible across *different* equivalence classes, we can always assume w.l.o.g. that $\text{OP}^\Phi(\pi) > \mathbb{E}_{\pi'' \in [\pi']} [\text{XP}(\pi, \pi'')]$ for any $\pi, \pi' \in \Pi_*^\Phi$ for which $\pi \notin [\pi']$ (else one just merges $[\pi]$ and $[\pi']$ by adding the transposition¹ between π and π' to Φ). Thus if Item 2 is satisfied perfectly, it follows that

$$\max_{\pi \in \Pi} \text{OP}^\Phi(\pi) \geq \text{XP}(\text{OP}_*^\Phi), \quad (7)$$

¹A transposition is a permutation that swaps exactly two elements.

with equality if and only if $[\pi] = \Pi_*^\Phi$ for any and thus all $\pi \in \Pi_*^\Phi$. This means that if one finds a symmetry group Φ which satisfies both of Items 1 and 2 perfectly, then $[\pi] = \Pi_*^\Phi$ for any $\pi \in \Pi_*^\Phi$ and $\text{OP}^\Phi(\pi) = \text{XP}(\text{OP}_*^\Phi)$. In other words, with such a choice of Φ , agents during training account for any potential test-time partner produced by OP^Φ , and only for such partners. Items 1 and 2 are thus desirable criteria for choosing a group Φ that makes OP^Φ a suitable learning rule for ZSC. However, these criteria alone are *not sufficient* for Φ to be optimal for ZSC, because agents in ZSC cannot choose a symmetry group Φ that is tailored to a specific Dec-POMDP. For example, suppose agents select Φ as the set of all bijections on Π that leave a particular SP optimal policy π unchanged. In this case, except for trivially simple Dec-POMDPs, we have $\Pi_*^\Phi = [\pi] = \{\pi\}$. This choice of Φ would trivially satisfy Item 1 (since $\{\pi\}$ is entirely representative of test-time policies) and Item 2 (since $\pi \in \Pi_*^\Phi$ and $\phi(\pi) = \pi, \forall \phi \in \Phi$). However, such a symmetry group is not permissible in ZSC because it is specifically constructed for a particular policy in a specific Dec-POMDP, violating the requirement for generality in ZSC. Therefore, while Items 1 and 2 are desirable properties, they are not sufficient on their own for choosing an appropriate symmetry group Φ for ZSC.

3.2 EXPECTED RETURN SYMMETRIES

We propose that the group Φ^{ER} of expected return (ER) symmetries, which can be learned with completely model-free assumptions, handles the above trade-off given by Items 1 and 2 favorably:

Definition 4 (Expected Return Symmetries). We denote by Π_* the set of joint policies that maximize the self-play objective, i.e. achieve the best possible expected return. We define Φ^{ER} as the subset of Ψ which leaves Π_* invariant, in the following sense:

$$\Phi^{\text{ER}} := \left\{ \phi \in \Psi \mid \forall \pi \in \Pi_* : \mathbb{E}_{\tau \sim \pi} [V_\pi(\tau)] = \mathbb{E}_{\tau' \sim \phi(\pi)} [V_{\phi(\pi)}(\tau')] \right\}, \quad (8)$$

where $V_\pi(\tau_t) = \mathbb{E}_\pi \left[\sum_{t'=t+1}^H \gamma^{t'-t} r_{t'} \mid \tau_t \right]$ is the expected return from τ_t onwards when following π , i.e. V_π is the value function of π .

It is easy to see that Φ^{ER} is a group under function composition (see Appendix B), and that for any Dec-POMDP symmetry $\phi \in \Phi^{\text{MDP}}$ and any joint policy π it holds that $\mathbb{E}_{\tau \sim \pi} [V_\pi(\tau)] = \mathbb{E}_{\tau' \sim \phi(\pi)} [V_{\phi(\pi)}(\tau')]$ (see Appendix B and Treutlein et al. (2021)). These two facts together imply that Φ^{MDP} is a subgroup of Φ^{ER} .

At a high level, Φ^{MDP} captures coordination differences based only on relabeling actions and observations, offering limited diversity. In contrast, self-play-optimal policies can differ significantly in coordination strategies, beyond mere label permutations. By grouping such diverse policies into the same equivalence class, Φ^{ER} better addresses Item 1 than Φ^{MDP} . While Φ^{ER} may not fully satisfy Item 2 (as Φ^{ER} is only required to preserve the expected return of self-play-optimal policies), we show in Section 4.4 that learned symmetries in Φ^{ER} approximately meet this criterion. Overall, this suggests ZSC agents using $\text{OP}^{\Phi^{\text{ER}}}$ will coordinate better at test time than those using $\text{OP}^{\Phi^{\text{MDP}}}$. Section 4 confirms that Φ^{ER} significantly improves (zero-shot) coordination across various environments.

The following example illustrates the advantage of ER symmetries over Dec-POMDP symmetries.

Example 2. Consider a cooperative communication game based on Hu et al. (2021). Alice observes a binary variable (pet) representing either “cat” or “dog”. Her actions include turning a light bulb on for a reward of 0.01, light bulb off for a reward of 0, bailing for a reward of 1 which ends the game, or removing a barrier at a cost of 5 to let Bob directly observe the pet. Bob can bail for 0.5 or guess the pet, receiving 10 for a correct “cat” guess, 11 for a correct “dog” guess, and losing 10 for an incorrect guess.

In this game, there are two ways for Alice to communicate with Bob: a “cheap-talk” channel (turning the light on or off) and a “grounded” channel (removing the barrier). There are exactly two self-play-optimal joint policies, and both use the cheap-talk channel. However, they differ: one policy associates “light on = dog, light off = cat,” while the other assigns the reverse. These policies are incompatible in cross-play, where coordination fails if each agent follows a different encoding.

This cat/dog game contains no non-trivial Dec-POMDP symmetries, which means that OP with Φ^{MDP} reduces to self-play, which, as just discussed, leads to coordination failure since independently trained agents will converge on either of the two self-play-optimal but cross-play-incompatible policies. Indeed, permuting “on/off” and “cat/dog” is not a Dec-POMDP symmetry because these environmental features have different reward dynamics.

270 However, these pairings *are* ER symmetries, as they transform the two self-play-optimal policies into
 271 one another and thus preserve their expected return. Therefore, these two policies are symmetric to
 272 each other w.r.t. Φ^{ER} , are put into the same equivalence class, and OP with Φ^{ER} is able to anticipate
 273 their coordination failure. OP with Φ^{ER} then chooses the optimal grounded policy, leading to the
 274 best possible cross-play score in this game for independent rational agents. This is demonstrated
 275 empirically in Section 4.2.

276 3.3 ALGORITHMIC APPROACHES

277 In practice, if Φ^{ER} is large, we find that in some settings it can be approximated by a few learned
 278 ϕ that are sufficiently diverse (see Section 3.1 for the importance of diversity). We therefore aim
 279 to develop an algorithm to learn a few ER symmetries (which we find in Section 4 is sufficient to
 280 significantly enhance agent coordination across various environments).
 281

282 Based on Equation 8, we formulate the following objective for learning ER symmetries:

$$283 \phi_{\theta^*} \text{ s.t. } \theta^* = \arg \inf_{\theta \in \Theta} \mathbb{E}_{\pi \sim \Pi'} \left| \mathbb{E}_{\tau \sim \pi} [V_{\pi}(\tau)] - \mathbb{E}_{\tau' \sim \phi_{\theta}(\pi)} [V_{\phi_{\theta}(\pi)}(\tau')] \right|, \quad (9)$$

286 where Θ is the parameter space for the family of transformations Ψ_{Θ} , and $\Pi' \subset \Pi_*$ is a training pool
 287 of (near) optimal SP policies. Equation 9 enforces that the learned transformation ϕ_{θ^*} preserves the
 288 expected return across the entire distribution of trajectories generated by each policy in Π'^2 . The
 289 broader the set Π' is, the more representative it will be of Π_* , and hence the less likely the learned
 290 ϕ_{θ^*} will overfit to a specific policy (i.e. not able to preserve expected return for other optimal policies
 291 outside the training set).

292 Note that since the policies in Equation 9 are approximately SP optimal, it suffices to consider

$$293 \phi_{\theta^*} \text{ s.t. } \theta^* = \arg \sup_{\theta \in \Theta} \mathbb{E}_{\pi \sim \Pi'} [\mathbb{E}_{\tau \sim \phi_{\theta}(\pi)} [V_{\phi_{\theta}(\pi)}(\tau)]] . \quad (10)$$

296 It is important to note that the optimization in Equation 10 focuses only on the ER symmetry. In
 297 this process, we train the transformation ϕ_{θ} within a reinforcement learning loop, but we keep the
 298 policies in Π' fixed (i.e., their weights remain unchanged). During training, we sample actions from
 299 π to ensure we explore the full range of possible actions. See Algorithm 1 in Appendix D for details.

300 Recall that $\phi_{\theta} = \{\phi_{\mathcal{S},\theta}, \phi_{\mathcal{O},\theta}, \phi_{\mathcal{A},\theta}\}$. Since we are interested in expected return symmetries insofar
 301 as they act on the policy space rather than the Dec-POMDP itself, we fix $\phi_{\mathcal{S},\theta} = \text{Id}$ ³. Further-
 302 more, since typically $|\mathcal{A}| \ll |\mathcal{O}|$, rather than learning both $\{\phi_{\mathcal{O},\theta}, \phi_{\mathcal{A},\theta}\}$ in a reinforcement learning
 303 loop, we can consider learning $\phi_{\mathcal{A},\theta}$ through search over transpositions of the available actions; i.e.,
 304 we initialize a fixed transposition on the actions as $\phi_{\mathcal{A},\theta}$ and learn $\phi_{\mathcal{O},\theta}$ as per Equation 10. See
 305 Algorithm 1 in Appendix D for an outline of this procedure.

306 Learning $\phi_{\mathcal{A},\theta}$ through search over the transposition space has several advantages: 1) transpositions
 307 can generate any permutation in the action space (Cayley, 1854), and in Section 4.4 we show that
 308 the learned transformations approximately preserve expected return; 2) if both $\phi_{\mathcal{O},\theta}$ and $\phi_{\mathcal{A},\theta}$ are
 309 learned together, regularizing training to avoid trivial identity transformations is difficult. Fixing a
 310 non-identity transposition in the action space allows $\phi_{\mathcal{O},\theta}$ to converge to the optimal observation
 311 transformation relative to the action transposition; 3) this approach facilitates enforcing invertibility
 312 in Equation 11, as $\phi_{\mathcal{A},\theta}^2 = \text{Id}$ by construction.

313 The action space allows for $\binom{|\mathcal{A}|}{2}$ distinct transpositions, representing the number of unique ways
 314 two actions can be permuted. Consequently, learning each observation transformation $\phi_{\mathcal{O},\theta}$ cor-
 315 responding to every possible action transposition requires $O(|\mathcal{A}|^2)$ optimizations of Equation 10 to
 316 perform an exhaustive search (i.e. $O(|\mathcal{A}|^2)$ iterations of the outer for loop in Algorithm 1). However,
 317 in Section 4.4 we show that a non-exhaustive search that undersamples the space of transpositions
 318 is still sufficient for learning symmetries that improve coordination amongst agents.

319 Note the objective in Equation 10 does not enforce *closure under composition* (i.e., $\phi_1 \circ \phi_2 \in \Phi^{\text{ER}}$)
 320 or *invertibility* (i.e., the existence of $\phi^{-1} \in \Phi^{\text{ER}}$ such that $\phi^{-1} \circ \phi = \text{Id}$), both necessary for Φ^{ER} to
 321

322 ²This is to ensure ϕ_{θ} explores the trajectory space in a way that respects the asymmetry of actions that might
 323 not otherwise be taken by a self-play-optimal policy; e.g. “reveal” and “bail” in Example 2.

³Under the OP objective, $\phi_{\theta}(\pi)$ only interacts with the original Dec-POMDP D .

form a group. To address this, we add regularization terms to the objective in Equation 10:

$$\begin{aligned} \phi_{\theta^*} \text{ s.t. } \theta^* = \arg \max_{\theta \in \Theta} & \mathbb{E}_{\pi \sim \Pi'} \mathbb{E}_{w \sim \text{Bernoulli}(\lambda_1)} \mathbb{E}_{\tau \sim \tilde{\Phi}_{\theta}(\pi)} \left(w \cdot V_{\phi_{\theta}(\pi)}(\tau) \right. \\ & + (1 - w) \cdot \mathbb{E}_{\phi_i, \phi_j \sim \Phi^{\text{ER}}} [V_{\phi_i \circ \phi_{\theta} \circ \phi_j}(\tau)] \\ & \left. - \lambda_2 D_{KL}(\pi(\cdot | \tau) \| \phi_{\theta}^2(\pi)(\cdot | \tau)) \right), \end{aligned} \quad (11)$$

where $\lambda_1 \in (0, 1)$ controls compositionality strength, and λ_2 enforces the learned transformation to be its own inverse. Since $\phi_{\mathcal{A}, \theta}$ is a fixed transposition, $\phi_{\mathcal{A}, \theta}^2 = \text{Id}$ by design, so we can easily enforce $\phi_{\mathcal{O}, \theta}^2 = \text{Id}$. As shown in Appendix C, transformations learned via Equation 10 / Algorithm 1 largely satisfy closure under composition without explicit enforcement, making them reasonable proxies for ϕ_i, ϕ_j when optimizing Equation 11. The objective is made stochastic to avoid having to compute multiple policy gradients per update. This is detailed in Algorithm 2 in Appendix D.

We also propose an alternative objective for learning symmetries through cross-play maximization:

$$\phi_{\theta^*} \text{ s.t. } \theta^* = \arg \sup_{\theta \in \Theta} \mathbb{E}_{\tau' \sim (\pi_i^1, \phi_{\theta}(\pi_j^2))} [V_{(\pi_i^1, \phi_{\theta}(\pi_j^2))}(\tau')], \quad (12)$$

where $\pi_i, \pi_j \in \Pi'$ are a pair of SP optimal policies chosen from the training pool. If π_i and π_j belong to the same equivalence class induced by Φ^{ER} , then by definition there exists an ER symmetry ϕ that maximizes Equation 12 to the self-play optimum value of $\mathbb{E}_{\tau \sim \pi_i} [V_{\pi}(\tau)]$. Therefore, for each pair of optimal policies $\pi_i, \pi_j \in \Pi'$, we optimize Equation 12 over ϕ_{θ} , and save the ϕ_{θ} that optimize Equation 12 to the highest value. We outline this approach in Algorithm 3 of Appendix D. We highlight a trade-off between the objectives of Equation 11 and Equation 12: while the former more directly optimizes for an ER symmetry, it assumes $\phi_{\mathcal{A}, \theta}$ to be of a certain form, while the latter assumes no such form but tacitly assumes some pair in Π' belong to the same equivalence class.

4 EXPERIMENTS

We evaluate our method in four different environments, focusing on how ER symmetries impact zero-shot coordination (ZSC) compared to self-play and other-play with Dec-POMDP symmetries. Specifically, we train independent agent populations that take advantage of ER symmetries and compare their cross-play performance within the population to baseline populations. The goal is to assess whether the use of ER symmetries leads to better coordination between agents than self-play or Dec-POMDP-symmetry-based training.

Populations of agents using ER symmetries for ZSC are formed as follows: each agent i chooses k seeds at random to train k different optimal policies, Π'_i . Agent i then independently performs ER symmetry discovery with their specific Π'_i by optimizing Equation 10, Equation 11 or Equation 12, and among their learned transformations uses the l that best preserve expected return as their ER symmetries. Agent i then chooses m seeds at random and uses their learned symmetries to train m policies $\{\pi_{i,k}\}_{k=1}^m$ with reinforcement learning constrained by the learning rule in Equation 4; multiple policies ($m > 1$) are trained to mitigate the effect of a seed that sub-optimally explores the space. The agent then selects $\pi_i := \arg \max_{k=1, \dots, m} J(\pi_{i,k})$, and deploys π_i for cross-play.

Aside from the environments in Sections 4.1 and 4.2, we parameterize $\phi_{\mathcal{O}, \theta}$ as a feed-forward neural network with two hidden layers. The experiments in Sections 4.3 and 4.4 use the JaxMARL environment and implementations (Rutherford et al., 2023). For details on our setup and hyperparameters, refer to Appendix A. See Appendix E for plots of interpretability of agent play. If accepted, we will release our full working code for all four environments.

4.1 ITERATED THREE-LEVER GAME

We consider a Dec-POMDP inspired by Treutlein et al. (2021), where two agents simultaneously choose one of three levers, receiving a reward of +1 if their choices match and -1 otherwise. This repeats for 2 rounds, with each agent observing the other’s previous action. Thus, $|\mathcal{S}| = 1$, and $\mathcal{A} = \mathcal{O} = \{1, 2, 3\} \times \{1, 2, 3\}$. Rewards are $r_{t+1} = -1 + 2 \cdot \mathbb{1}_{a_t^1 = a_t^2}$, and observations are $o_{t+1}^1 = a_t^2, o_{t+1}^2 = a_t^1$.

There are 6 Dec-POMDP symmetries in this game, which correspond to the 6 permutations of the three interchangeable levers. The optimal $\text{OP}^{\Phi^{\text{MDP}}}$ policy first chooses a lever uniformly at random in

the first round, and then if the agents match, they stick with that lever; otherwise, they coordinate on the unique unchosen lever. This joint policy achieves an expected return of $2/3$, optimal for ZSC as first-round success always only has probability $1/3$. In this game, $\Phi^{\text{MDP}} = \Phi^{\text{ERS}}$, and our Algorithm 1 successfully learns these symmetries to find an (approximately) optimal ZSC policy.

We train 10 ERS agents, each training $k = 20$ self-play optimal policies using IQL with shared Q-values. For each fixed action permutation, ERS agents select the observation permutation maximizing the objective in Equation (10). The agents then use those $l = 6$ learned ER symmetries in the OP objective, to train $m = 1$ policies to be employed for cross-play. The 10 ERS agents achieve an average SP score of 0.7 ± 0.059 and an average XP score of 0.672 ± 0.092 . To compare, we train 10 $\text{OP}^{\Phi^{\text{MDP}}}$ agents, who achieve mean SP score 0.654 ± 0.143 and mean XP score 0.671 ± 0.099 .

This game highlights a setting where OP is advantaged over OBL (Hu et al., 2021). OBL fails in this game (even after transforming it into a turn-based version) because agents assuming uniform randomness in others’ past actions cannot prefer one lever over another, leading OBL policies to converge onto the uniform distribution. We note however that in the variant of this game with two levers, OP with Φ^{MDP} also fails Treutlein et al. (2021), and since $\Phi^{\text{MDP}} = \Phi^{\text{ERS}}$, OP with Φ^{ERS} fails as well. This is because when there are just two levers, then there is no uniquely identifiable lever to choose in the second round, whether the agents matched in the first round or not. Thus choosing to repeat or to switch are both $\text{OP}^{\Phi^{\text{MDP}}}$ optimal policies, but they are in separate equivalence classes and incompatible. This provides an example of insufficient expressivity in the symmetry group.

4.2 CAT/DOG ENVIRONMENT

We take the cat/dog game from Example 2. We build a population of 5 independent Q-learning agents (IQL) (Tan, 1993) as a self-play baseline. We build a population of 5 ER symmetry agents, for which each agent trains $k = 10$ optimal self-play policies as Π' , and then optimizes Equation 10 with vanilla policy gradient (Sutton et al., 1999), where $\phi_{\mathcal{O},\theta}$ is parameterized as a probability distribution over all possible permutations of the observations (there being $2!$ for Alice and $4!$ for Bob). Equation 10 / Algorithm 1 suffices because permutations already satisfy compositionality and invertibility. Each ER symmetry agent uses $l = 3$ ER symmetries to then train $m = 1$ $\text{OP}^{\Phi^{\text{ER}}}$ policy.

IQL agents achieve a mean within-population cross-play score of -2.11 ± 0.2 , whereas the ER symmetry agents converge onto optimal grounded communication and achieve 5.50 ± 0.02 . Thus, ER symmetries can prevent over-coordination in settings where non-trivial Dec-POMDP symmetries do not even exist. It is worth mentioning that approaches based on cognitive hierarchies fail to find the optimal grounded communication protocol in this setting (Hu et al., 2021; Cui et al., 2021; Camerer et al., 2004), since they assume other agents follow random or lower-level strategies, and instead consistently converge onto “bailing” for a return of 1.

4.3 OVERCOOKED V2

Overcooked V2 is a recent AI benchmark for ZSC [Anonymous, under submission], which improves on the cooperative multi-agent benchmark Overcooked (Carroll et al., 2019), by introducing asymmetric information and increased stochasticity, creating more nuanced coordination challenges.

For ZSC, we train a population of 5 IPPO (Yu et al., 2022) policies as a SP baseline, where each policy uses an RNN coupled with a CNN to process the observations. The population of ER symmetry agents each train $k = 12$ IPPO SP policies, to then use Equation 12 / Algorithm 3 to obtain $l = 16$ ER symmetries. Each agent trains $m = 2$ $\text{OP}^{\Phi^{\text{ER}}}$ policies using their learned symmetries.

The IPPO baseline population exhibits a highly bi-modal distribution of cross-play (XP) scores, where agents are either largely compatible or largely incompatible. In Figure 2, we plot the cross-play score distribution of both agent populations, and can see the SP-optimal IPPO baseline population achieves a mean cross-play score of 6.74, whereas the $\text{OP}^{\Phi^{\text{ER}}}$ -optimal population achieves a mean cross-play score of 15.8. In particular, the ER symmetry population is able to significantly close the self-play to cross-play (SP-XP) gap compared to the baseline population, demonstrating more consistent coordination across different agents. We emphasize that Overcooked V2 presents a challenging environment where agents move simultaneously, which renders standard methods like off-belief Learning (Hu et al., 2021) less applicable. This further highlights the effectiveness of ER symmetries in improving ZSC.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

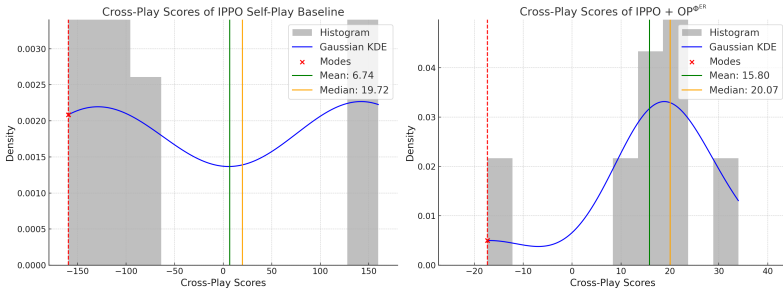


Figure 2: Cross-play score distribution of the IPPO self-play baseline population and the ER symmetry agent population in Overcooked V2. The baseline population achieves mean SP scores of 162.33 ± 0.14 , and the ER symmetry population achieves mean SP scores of 27.81 ± 0.3 .

4.4 HANABI

Hanabi (see Appendix F for details) is a challenging AI benchmark, and has served as the primary test bed for many algorithms designed for zero-shot coordination, ad-hoc teamplay, and other cooperative tasks (Bard et al., 2020; Cui et al., 2021; Nekoei et al., 2021; 2023; Muglich et al., 2022a;b).

Preserves OP Optimality

Since ER symmetries contain Dec-POMDP symmetries, and capture equivalences beyond just relabelling, they are clearly more diverse than Dec-POMDP symmetries, and hence better satisfy Item 1 from Section 3.1. We verify the ER symmetries also approximately satisfy Item 2.

We take the 8 learned regularized ER symmetries from above, denoting this set as $\hat{\Phi}^{ER}$. We find that $\mathbb{E}_{\pi \sim \Pi_{\hat{\Phi}^{ER}}} [OP^{\hat{\Phi}^{ER}}(\pi)] = 23.59 \pm 0.04$ and $\mathbb{E}_{\pi \sim \Pi_{\hat{\Phi}^{ER}}} [\mathbb{E}_{\phi \sim \hat{\Phi}^{ER}} [OP^{\hat{\Phi}^{ER}}(\phi(\pi))]] = 23.34 \pm 0.05$, where we train 5 $OP^{\hat{\Phi}^{ER}}$ policies. Thus, Item 2 is approximately satisfied by $\hat{\Phi}^{ER}$.

Zero-Shot Coordination

For ZSC, we train as baselines 1) a population of 5 IPPO agents, and 2) a population of 5 IPPO agents with access to all Dec-POMDP symmetries constrained by the OP objective in Equation 4. We train a population of ER symmetry agents that each independently discover ER symmetries for the OP objective. Each agent in the ER symmetry population uses $k = 6$ seeds to learn ER symmetries, amongst which they save the $l = 8$ that best preserve expected return. Each population (baseline and return symmetry populations alike) has a total of 5 agents, where each agent trains $m = 3$ policies and deploys the one achieving higher return.

Inspired by the symmetrizer in Treutlein et al. (2021); Muglich et al. (2022a); Van der Pol et al. (2020), we consider $S(\pi)(a | \tau) := \frac{1}{|\Phi|} \sum_{\phi \in \Phi} \phi(\pi)(a | \tau)$, where Φ can be taken to be either Φ^{ER} or Φ^{MDP} . Agents from any population can thus transform their policy with S before deploying for cross-play, ensuring invariance of the deployed policy w.r.t. Φ (since $\phi(S(\pi)) = S(\pi), \forall \phi \in \Phi$). As per the empirical results below, the symmetrizer functions as a policy improvement operator for cross-play across multiple policy populations.

Table 1 shows that the agents using ER symmetries improve in ZSC over both baselines; this is even in spite of the Dec-POMDP agent population assuming access to environment symmetries. As well, even with just using a subset of transformations from Φ^{ER} , the return symmetry agents are able to converge on policies that generalize well in cross-play. We also notice that symmetrization with respect to Φ^{ER} or Φ^{MDP} improves coordination amongst agents of all populations considered; we can see that Φ^{ER} better improves the optimal self-play policy population, and Φ^{MDP} better improves the other two, which aligns with expectation since Φ^{ER} is only explicitly enforced to maintain invariance over optimal self-play policies, whereas Φ^{MDP} maintains invariance over any policy type.

5 RELATED WORK

Extensive research exists on coordination in multi-agent systems, particularly in zero-shot coordination. Methods like Hu et al. (2020); Muglich et al. (2022a) use Dec-POMDP symmetries to avoid incompatible policies, while Hu et al. (2021) rely on environment dynamics for grounded policies. Diversity-based approaches also leverage known symmetries and simulator access (Cui et al., 2023; Lupu et al., 2021). In contrast, ER symmetries can be learned from agent-environment interactions

Table 1: Self-play, within-population mean cross-play (XP) and median cross-play (XP(*)) scores are reported. The $OP^{\Phi^{MDP}}$ population used all 120 Dec-POMDP symmetries, whereas the $OP^{\Phi^{ER}}$ population used 8 ER symmetries. “MDP” indicates the population was symmetrized with Dec-POMDP symmetries at test time, and “ER” analogously indicates symmetrization with expected return symmetries. The ER symmetrizer uses 8 expected return symmetries.

Model	Self-Play	XP	XP(*)	XP(*)+MDP	XP(*)+ER
IPPO	24.04 ± 0.02	4.02 ± 0.17	0.12 ± 0.03	0.14 ± 0.03	0.10 ± 0.03
IPPO + $OP^{\Phi^{MDP}}$	23.81 ± 0.03	8.61 ± 0.17	8.14 ± 0.15	8.70 ± 0.16	9.91 ± 0.14
IPPO + $OP^{\Phi^{ER}}$	23.74 ± 0.03	21.64 ± 0.07	22.03 ± 0.05	22.50 ± 0.06	22.25 ± 0.05

without privileged information, enabling grounded signaling and effective coordination in concurrent environments (see Sections 4.2 and 4.3).

In single-agent settings, symmetry has been shown to reduce sample complexity in reinforcement learning (Van der Pol et al., 2020; Zhu et al., 2022; Nguyen et al., 2024). In multi-agent systems, symmetries reduce policy space complexity and help agents identify equivalent strategies (van der Pol et al., 2021; Muglich et al., 2022a). However, many methods require explicit knowledge of symmetries, or rely on predefined groups (Abreu et al., 2023; Yu et al., 2024; Nguyen et al., 2024). Our work generalizes these approaches by introducing ER symmetries, which do not require prior symmetry knowledge or equivariant networks, and can be learned directly through environment interactions.

Our work relates to value-based abstraction, which groups states or observations with similar value functions. Rezaei-Shoshtari et al. (2022) use lax-bisimulation to learn MDP homomorphisms, while Grimm et al. (2021) learn a model of the underlying MDP for value-based planning. In contrast, we focus on symmetries in the policy space that preserve expected return. ER symmetries are conceptually related to Q^* -irrelevance abstractions (Li et al., 2006) in that both aim to preserve the optimal value function of an MDP. However, whereas Q^* -irrelevance abstractions reduce complexity by aggregating states, ER symmetries form a group that acts bijectively on the policy space, transforming optimal policies into other policies with the same expected return.

6 CONCLUSION

This paper defined expected return symmetries—a group whose action preserves policy expected return. We demonstrated that the symmetries in this group can be learned purely from interactions with the environment and without requiring privileged environment information. We demonstrated that this symmetry class significantly enhances zero-shot coordination, significantly outperforming traditional Dec-POMDP symmetries, which are a subset of this group. Importantly, we showed that expected return symmetries are effective in challenging settings where state-of-the-art ZSC methods, such as off-belief learning (Hu et al., 2021) in Hanabi or approaches based on cognitive hierarchies, either fail completely (e.g., in the lever game and cat/dog environments) or face difficulties in their application (e.g., Overcooked V2).

One major limitation of our approach is that we constrain the search for symmetries to bijections over the action and observation spaces. While this works well in many settings, as shown in our experiments, there are environments, e.g. the two-lever game, in which this limited expressivity cannot provide enough diversity within the equivalence classes of policies that are optimal w.r.t. OP with Φ^{ERS} , to prevent coordination failure in ZSC. Another potential limitation is that our method relies on access to a set of approximately self-play-optimal policies to identify ER symmetries. Its performance in scenarios where these self-play policies are significantly suboptimal remains unclear.

Our work opens several avenues for future research. One direction is to explore the use of expected return symmetries in ad-hoc teamwork or single-agent settings. Another is to investigate broader classes of symmetries, beyond the ones that arise from bijections on the action and observation spaces. A further possibility is to iterate the symmetry learning process: training symmetries that preserve the optimal policies under OP with Φ^{ERS} , and using these new symmetries in subsequent other-play objectives. We would expect this to further improve optimality and diversity within equivalence classes, and thus improve coordination outcomes.

REFERENCES

- 540
541
542 Miguel Abreu, Luis Paulo Reis, and Nuno Lau. Addressing imperfect symmetry: a novel symmetry-
543 learning actor-critic extension. *arXiv preprint arXiv:2309.02711*, 2023.
- 544 Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio
545 Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A
546 new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.
- 547 Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning:
548 Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.
- 549 Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The*
550 *Quarterly Journal of Economics*, 119(3):861–898, 2004.
- 551 Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca
552 Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural*
553 *information processing systems*, 32, 2019.
- 554 Arthur Cayley. Vii. on the theory of groups, as depending on the symbolic equation $\theta^n = 1$. *The*
555 *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 7(42):40–47,
1854.
- 556 Chess.com. The worst and the best chess handshakes of world cup 2021. [https://www.](https://www.youtube.com/watch?v=6fS7bDyNYHI)
557 [youtube.com/watch?v=6fS7bDyNYHI](https://www.youtube.com/watch?v=6fS7bDyNYHI), 2021. Accessed: 2024-09-30.
- 558 Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolu-
559 tional networks and the icosahedral cnn. In *International conference on Machine learning*, pp.
560 1321–1330. PMLR, 2019.
- 561 Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. K-level reasoning for zero-shot
562 coordination in hanabi. *Advances in Neural Information Processing Systems*, 34:8215–8228,
563 2021.
- 564 Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Fo-
565 erster. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning*
566 *Representations*, 2023.
- 567 Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolu-
568 tional neural networks for equivariance to lie groups on arbitrary continuous data. In *International*
569 *Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020.
- 570 Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value
571 equivalence. *Advances in neural information processing systems*, 34:7773–7786, 2021.
- 572 Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. “other-play” for zero-shot
573 coordination. In *International Conference on Machine Learning*, pp. 4399–4410. PMLR, 2020.
- 574 Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief
575 learning. In *International Conference on Machine Learning*, pp. 4369–4379. PMLR, 2021.
- 576 Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan,
577 Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv*
578 *preprint arXiv:2310.19852*, 2023.
- 579 Zahi M Kakish, F Rodríguez-Lera, D Bischel, A Mosquera, R Boumghar, S Kaczmarek,
580 T Seabrook, P Metzger, and JL Galanche. Open-source ai assistant for cooperative multi-agent
581 systems for lunar prospecting missions. In *8th European Conference for Aeronautics and Space*
582 *Sciences (EUCASS)*, 2019.
- 583 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convo-
584 lutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- 585 Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction
586 for mdps. *AI&M*, 1(2):3, 2006.

- 594 Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot
595 coordination. In *International conference on machine learning*, pp. 7204–7213. PMLR, 2021.
- 596
- 597 Stefano Mariani, Giacomo Cabri, and Franco Zambonelli. Coordination of autonomous vehicles:
598 Taxonomy and survey. *ACM Computing Surveys (CSUR)*, 54(1):1–33, 2021.
- 599
- 600 Darius Muglich, Christian Schroeder de Witt, Elise van der Pol, Shimon Whiteson, and Jakob Foer-
601 ster. Equivariant networks for zero-shot coordination. *Advances in Neural Information Processing*
602 *Systems*, 35:6410–6423, 2022a.
- 603 Darius Muglich, Luisa M Zintgraf, Christian A Schroeder De Witt, Shimon Whiteson, and Jakob Fo-
604 erster. Generalized beliefs for cooperative ai. In *International Conference on Machine Learning*,
605 pp. 16062–16082. PMLR, 2022b.
- 606
- 607 Shravan Matthur Narayanamurthy and Balaraman Ravindran. On the hardness of finding symmetries
608 in markov decision processes. In *Proceedings of the 25th international conference on Machine*
609 *learning*, pp. 688–695, 2008.
- 610
- 611 Hadi Nekoei, Akilesh Badrinaaraayanan, Aaron Courville, and Sarath Chandar. Continuous coor-
612 dination as a realistic scenario for lifelong learning. In *International Conference on Machine*
Learning, pp. 8016–8024. PMLR, 2021.
- 613
- 614 Hadi Nekoei, Xutong Zhao, Janarthanan Rajendran, Miao Liu, and Sarath Chandar. Towards few-
615 shot coordination: Revisiting ad-hoc teamplay challenge in the game of hanabi. In *Conference on*
Lifelong Learning Agents, pp. 861–877. PMLR, 2023.
- 616
- 617 Hai Nguyen, Tadashi Kozuno, Cristian C Beltran-Hernandez, and Masashi Hamaya. Symmetry-
618 aware reinforcement learning for robotic assembly under partial observability with a soft wrist.
619 *arXiv preprint arXiv:2402.18002*, 2024.
- 620
- 621 Frans A Oliehoek, Matthijs TJ Spaan, Nikos Vlassis, et al. Dec-pomdps with delayed communica-
622 tion. In *The 2nd Workshop on Multi-agent Sequential Decision-Making in Uncertain Domains*.
Citeseer, 2007.
- 623
- 624 Cinjon Resnick, Ilya Kulikov, Kyunghyun Cho, and Jason Weston. Vehicle communication strate-
625 gies for simulated highway driving. *arXiv preprint arXiv:1804.07178*, 2018.
- 626
- 627 Sahand Rezaei-Shoshtari, Rosie Zhao, Prakash Panangaden, David Meger, and Doina Precup. Con-
628 tinuous mdp homomorphisms and homomorphic policy gradient. *Advances in Neural Information*
Processing Systems, 35:20189–20204, 2022.
- 629
- 630 Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ing-
631 varsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi
632 Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson,
633 Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. Jaxmarl:
634 Multi-agent rl environments in jax. *arXiv preprint arXiv:2311.10090*, 2023.
- 635
- 636 Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of*
research and development, 3(3):210–229, 1959.
- 637
- 638 Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient meth-
639 ods for reinforcement learning with function approximation. *Advances in neural information*
processing systems, 12, 1999.
- 640
- 641 Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings*
642 *of the tenth international conference on machine learning*, pp. 330–337, 1993.
- 643
- 644 Gerald Tesauro et al. Temporal difference learning and td-gammon. *Communications of the ACM*,
645 38(3):58–68, 1995.
- 646
- 647 Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism,
method and open issues for zero-shot coordination. In *International Conference on Machine*
Learning, pp. 10413–10423. PMLR, 2021.

648 Elise Van der Pol, Daniel Worrall, Herke van Hoof, Frans Oliehoek, and Max Welling. Mdp homo-
649 morphic networks: Group symmetries in reinforcement learning. *Advances in Neural Information*
650 *Processing Systems*, 33:4199–4210, 2020.

651
652 Elise van der Pol, Herke van Hoof, Frans A Oliehoek, and Max Welling. Multi-agent mdp homo-
653 morphic networks. In *International Conference on Learning Representations*, 2021.

654 Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The
655 surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information*
656 *Processing Systems*, 35:24611–24624, 2022.

657
658 Xin Yu, Rongye Shi, Pu Feng, Yongkai Tian, Simin Li, Shuhao Liao, and Wenjun Wu. Leveraging
659 partial symmetry for multi-agent reinforcement learning. In *Proceedings of the AAAI Conference*
660 *on Artificial Intelligence*, volume 38, pp. 17583–17590, 2024.

661 Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample
662 efficient grasp learning using equivariant models. *arXiv preprint arXiv:2202.09468*, 2022.

664 A EXPERIMENTAL SETUP

665
666 For expected return symmetry discovery in the three-lever game, For expected return symmetry
667 discovery in cat/dog, we use a temperature of $T = 2.667$ for Boltzmann exploration to promote
668 sufficient exploration of different actions. We use a constant baseline function with value 9.5 for the
669 policy gradient. We use a learning rate of 0.01 and 2000 episodes for each inner loop of Algorithm
670 1.

671 For expected return symmetry discovery in Overcooked V2 and Hanabi, we parameterize $\hat{\phi}_{\mathcal{O},\theta}$ as a
672 two hidden layer, feedforward neural network, with each linear layer initialized as a $|\mathcal{O}^i|$ -dimensional
673 identity matrix; this choice of initialization is necessary as the symmetry discovery is highly initial-
674 ization sensitive. We apply ReLU to the final output of the network to promote sparsity in the repre-
675 sentation. We build on top of the environment implementation and baseline algorithms in JaxMARL
676 (Rutherford et al., 2023).

677 For Hanabi, we run each inner loop of Algorithm 2 for 1.5e9 timesteps across vectorized Hanabi
678 environments. As per Equation 11 we use $\lambda_1 = 0.65$, $\lambda_2 = 2.5e-9$. $\phi_{\mathcal{A},\theta}$ is a fixed action transpo-
679 sition for each learned symmetry. We use a temperature of $T = 1$ for Boltzmann exploration.

680 For Overcooked V2, we run each inner loop of Algorithm 3 for 1.5e8 timesteps. $\phi_{\mathcal{A},\theta}$ is a learned
681 affine map. We use a temperature of $T = 1.1$ for Boltzmann exploration.

682
683 For both Hanabi and Overcooked V2, we use PPO and Generalized Advantage Estimation. For Han-
684 abi, we use 4 epochs, 1024 environments per pretrained policy, 128 environment steps per update, 4
685 minibatches, $\gamma = 0.99$, GAE Lambda = 0.95, CLIP EPS = 0.2, VF COEFF = 0.5, MAX GRAD
686 NORM = 0.5, a learning rate of $1e-5$ and a linear learning rate annealing schedule. For Over-
687 cooked V2, we use 4 epochs, 256 environments, 256 environment steps per update, 64 minibatches,
688 $\gamma = 0.99$, GAE Lambda = 0.95, CLIP EPS = 0.2, VF COEFF = 0.5, MAX GRAD NORM = 0.25,
689 a learning rate of $1e-5$ with no annealing.

690 Methods for Hanabi and Overcooked V2 were ran on A40 and L40S GPUs.

691 If accepted, we will release our full working code for all four environments.

693 B PROOFS

694
695 **Theorem.** $\Phi^{\text{ER}} := \{ \phi \in \Psi \mid \forall \pi \in \Pi_* : \mathbb{E}_{\tau \sim \pi} [V_{\pi}(\tau)] = \mathbb{E}_{\tau' \sim \phi(\pi)} [V_{\phi(\pi)}(\tau')] \}$ forms a group under
696 function composition.

697
698 *Proof.* To show that Φ^{ER} forms a group under function composition, we verify the group axioms:
699 closure, associativity, identity, and inverses.

700 **Closure:** For any $\phi_1, \phi_2 \in \Phi^{\text{ER}}$, we need to show that $\phi_1 \circ \phi_2 \in \Phi^{\text{ER}}$. For the composition $\phi_1 \circ \phi_2$,
701 we need to check:

$$\mathbb{E}_{\tau \sim \pi} [V_{\pi}(\tau)] = \mathbb{E}_{\tau'' \sim (\phi_1 \circ \phi_2)(\pi)} [V_{(\phi_1 \circ \phi_2)(\pi)}(\tau'')].$$

Using the preservation of expected returns under both ϕ_1 and ϕ_2 , we get:

$$\mathbb{E}_{\tau \sim \pi}[V_\pi(\tau)] = \mathbb{E}_{\tau' \sim \phi_2(\pi)}[V_{\phi_2(\pi)}(\tau')] = \mathbb{E}_{\tau'' \sim \phi_1(\phi_2(\pi))}[V_{\phi_1(\phi_2(\pi))}(\tau'')].$$

Hence, $\phi_1 \circ \phi_2 \in \Phi^{\text{ER}}$, proving closure.

Associativity: Function composition is associative, so for any $\phi_1, \phi_2, \phi_3 \in \Phi^{\text{ER}}$:

$$(\phi_1 \circ \phi_2) \circ \phi_3 = \phi_1 \circ (\phi_2 \circ \phi_3).$$

Thus, associativity holds.

Identity: The identity function $\text{Id} \in \Psi$ satisfies $\text{Id}(\pi) = \pi$ for all $\pi \in \Pi$, and thus preserves the expected return of optimal policies $\pi \in \Pi_*^\Phi$:

$$\mathbb{E}_{\tau \sim \pi}[V_\pi(\tau)] = \mathbb{E}_{\tau' \sim \text{Id}(\pi)}[V_{\text{Id}(\pi)}(\tau')] = \mathbb{E}_{\tau' \sim \pi}[V_\pi(\tau')].$$

Therefore, $\text{Id} \in \Phi^{\text{ER}}$ and acts as the identity element.

Inverses: For each $\phi \in \Phi^{\text{ER}}$, there exists $\phi^{-1} \in \Psi$ such that $\phi \circ \phi^{-1} = \phi^{-1} \circ \phi = \text{Id}$. For any $\pi \in \Pi_*^\Phi$ it then holds that

$$\mathbb{E}_{\tau' \sim \phi(\pi)}[V_{\phi(\pi)}(\tau')] = \mathbb{E}_{\tau \sim \pi}[V_\pi(\tau)] = \mathbb{E}_{\tau'' \sim \phi^{-1}(\phi(\pi))}[V_{\phi^{-1}(\phi(\pi))}(\tau'')].$$

Since for any $\tilde{\pi} \in \Pi_*^\Phi$ there exists $\pi \in \Pi_*^\Phi$ such that $\tilde{\pi} = \phi(\pi)$, we see that $\phi^{-1} \in \Phi^{\text{ER}}$.

Since Φ^{ER} satisfies closure, associativity, identity, and inverses, it forms a group under function composition. \square

Theorem. If π is a joint policy over a Dec-POMDP and $\phi \in \Phi^{\text{MDP}}$ is a Dec-POMDP symmetry, then $\mathbb{E}_{\tau \sim \pi}[V_\pi(\tau)] = \mathbb{E}_{\tau' \sim \phi(\pi)}[V_{\phi(\pi)}(\tau')]$.

Proof. Let $V^\pi(\tau)$ denote the value function for π . Then due to Bellman,

$$V^\pi(\tau) = \sum_{s \in \mathcal{S}} P(s | \tau) \sum_{a \in \mathcal{A}} \pi(a | \tau) [\mathbb{E}_{\mathcal{R}}[R(s, a)] + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' | s, a) \sum_{o' \in \mathcal{O}} \mathcal{U}(o' | s, a) V^\pi(\tau')],$$

where $\tau' = (\tau^{1'}, \dots, \tau^{n'})$ is the updated joint trajectory, where $\tau^{i'} = (\tau^i, o', a^i)$.

Denoting $\tilde{\pi} = \phi(\pi)$, we have

$$\begin{aligned} V^{\tilde{\pi}}(\phi(\tau)) &= \sum_{s \in \mathcal{S}} P(\phi(s) | \phi(\tau)) \sum_{a \in \mathcal{A}} \tilde{\pi}(\phi(a) | \phi(\tau)) [\mathbb{E}_{\mathcal{R}}[R(\phi(s), \phi(a))] \\ &\quad + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(\phi(s') | \phi(s), \phi(a)) \sum_{\phi(o') \in \mathcal{O}} \mathcal{U}(\phi(o') | \phi(s), \phi(a)) V^{\tilde{\pi}}(\phi(\tau'))] \\ &= \sum_{s \in \mathcal{S}} P(s | \tau) \sum_{a \in \mathcal{A}} \pi(a | \tau) [\mathbb{E}_{\mathcal{R}}[R(s, a)] + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' | s, a) \sum_{o' \in \mathcal{O}} \mathcal{U}(o' | s, a) V^{\tilde{\pi}}(\phi(\tau'))], \end{aligned}$$

where the second equality follows from Equations 3 and 2. The result follows via induction on the steps from the terminal state. \square

C LEARNED TRANSFORMATIONS SATISFY GROUP PROPERTIES

This section analyzes the learned ER symmetries for their group properties.

We first train six (near-)optimal IPPO policies with independent seeds as Π' , obtaining a mean expected return of $\mathbb{E}_{\pi \sim \Pi'}[J(\pi)] = 24.04 \pm 0.02$. Next, we randomly select 64 action-space transpositions to fix as $\phi_{\mathcal{A}, \theta_l}_{l=1}^{64}$, and learn the corresponding $\phi_{\mathcal{O}, \theta_l}_{l=1}^{64}$ via optimizing Equation 10 / Algorithm 1. This is a significant undersampling of the 190 possible transpositions, yet as we show below we still learn effective ER symmetries for coordination. We then save the 8 best transformations (those that maximize Equation 10) as unregularized ER symmetries. These are used to maximize Equation 11 / Algorithm 2 on another set of 64 random transpositions, now enforcing compositional closure and invertibility. The 8 best are saved as regularized ER symmetries.

Table 2: Comparison of 8 unregularized and 8 regularized ER symmetries applied to 6 unseen optimal policies ($|\Pi_{\text{unseen}}| = 6$). Regularization enforces compositionality and invertibility. For $k = 1, 2, 3$, let $J_k = \mathbb{E}_{\phi_i \sim \Phi} \mathbb{E}_{\pi \sim \Pi_{\text{unseen}}}[J((\phi_1 \circ \dots \circ \phi_k)(\pi))]$ denote the expected return after composing k randomly sampled transformations. We report Single Transform. (J_1), Double Comp. (J_2), and Triple Comp. (J_3). Relative Reconstruction Loss measures approximate invertibility (lower is better): $\mathbb{E}_{\pi \sim \Pi_{\text{unseen}}} \mathbb{E}_{\tau \sim \pi}[\frac{\|\tau - \phi_{\mathcal{O}}^2(\tau)\|}{\|\tau\|}]$, using the ℓ_2 norm for trajectory vectors. Recall $\mathbb{E}_{\pi \sim \Pi'}[J(\pi)] = 24.04 \pm 0.02$.

	Single Transform.	Double Comp.	Triple Comp.	Rel. Rec. Loss
Unreg.	22.88 ± 0.07	21.10 ± 0.09	20.36 ± 0.11	31.4% ± 1.8%
Reg.	23.32 ± 0.05	22.16 ± 0.07	20.94 ± 0.12	16.7% ± 0.18%

Table 2 shows that up to minor deviations in expected return preservation, the learned ER symmetries still approximately satisfy closure under composition. In addition, when invertibility is enforced, the relative reconstruction loss decreases substantially (a lower relative reconstruction loss tells us applying the transformation twice brings us closer to the original trajectory, suggesting approximate invertibility). We conclude that the learned ER symmetries, especially the regularized ones, approximately satisfy the desired group-theoretic properties.

D ALGORITHMS

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

Algorithm 1 Learning Expected Return Symmetries with Policy Gradients (without enforcing compositionality nor invertibility) ... Optimization of Equation 10

- 1: **Input:** A set of (near-)optimal policies Π' , parameter space Θ for transformations, action space \mathcal{A} , learning rate $\eta > 0$, l for top transformations to save
- 2: Initialize list of top l average expected values: $\bar{V}_{\text{top}} = [-\infty, \dots, -\infty]$ (length l)
- 3: Initialize list of top l transformations: $\phi_{\text{top}} = [\emptyset, \dots, \emptyset]$ (length l)
- 4: **for** each action transposition $\phi_{\mathcal{A},\theta} \in \text{Transpositions}(\mathcal{A})$ **do**
- 5: Initialize $\phi_{\mathcal{O},\theta}$ with random parameters $\theta \in \Theta$
- 6: **while** not converged **do**
- 7: **for** each policy $\pi \in \Pi'$ **do**
- 8: Sample a batch of trajectories τ using the transformed policy $\phi_{\theta}(\pi)$
- 9: Compute the return $R(\tau)$ for each trajectory τ
- 10: Compute advantage $A^{\phi_{\theta}(\pi)}(\tau_t, a_t)$ using any advantage function (e.g., TD, GAE)
- 11: Compute policy gradient:

$$\nabla_{\theta} J(\phi_{\theta}(\pi)) = \mathbb{E}_{\tau \sim \phi_{\theta}(\pi)} \left[\sum_{t=0}^H \nabla_{\theta} \log \phi_{\theta}(\pi)(a_t | \tau_t) A^{\phi_{\theta}(\pi)}(\tau_t, a_t) \right]$$

- 12: Update parameters: $\theta \leftarrow \theta + \eta \nabla_{\theta} J(\phi_{\theta}(\pi))$
- 13: **end for**
- 14: Compute average expected return over trajectories:

$$\bar{V}_{\phi_{\theta}} = \frac{1}{|\Pi'|} \sum_{\pi \in \Pi'} \mathbb{E}_{\tau \sim \pi} [V_{\phi_{\theta}(\pi)}(\tau)]$$

- 15: **end while**
 - 16: Find the index of the lowest return in \bar{J}_{top} , say i_{\min}
 - 17: **if** $\bar{V}_{\phi_{\theta}} > \bar{V}_{\text{top}}[i_{\min}]$ **then**
 - 18: Replace the lowest return: $\bar{V}_{\text{top}}[i_{\min}] \leftarrow \bar{V}_{\phi_{\theta}}$
 - 19: Replace the corresponding transformation: $\phi_{\text{top}}[i_{\min}] \leftarrow \{\phi_{\mathcal{O},\theta}, \phi_{\mathcal{A},\theta}\}$
 - 20: **end if**
 - 21: **end for**
 - 22: **Output:** Set of transformations $\phi_{\text{top}} = \{\phi_{\theta^*} = \{\phi_{\mathcal{O},\theta^*}, \phi_{\mathcal{A},\theta^*}\}\}$
-

Algorithm 2 Learning Expected Return Symmetries (enforcing compositionality and invertibility)
 ... Optimization of Equation 11

864
 865
 866
 867 1: **Input:** Policies Π' , parameter space Θ , action space \mathcal{A} , learning rate η , transformations
 868 $\{\phi_1, \dots, \phi_m\}$ obtained from Algorithm 1, l for top transformations to save, regularization
 869 weights λ_1, λ_2
 870 2: Initialize list of top l average expected values: $\bar{V}_{\text{top}} = [-\infty, \dots, -\infty]$ (length l)
 871 3: Initialize list of top l transformations: $\phi_{\text{top}} = [\emptyset, \dots, \emptyset]$ (length l)
 872 4: **for** each transposition $\phi_{\mathcal{A},\theta} \in \text{Transpositions}(\mathcal{A})$ **do**
 873 5: Initialize $\phi_{\mathcal{O},\theta}$ with random $\theta \in \Theta$
 874 6: **while** not converged **do**
 875 7: **for** each policy $\pi \in \Pi'$ **do**
 876 8: Sample ordinary or compositional trajectory based on λ_1
 877 9: **if** ordinary trajectory **then**
 878 10: Set $\tilde{\phi}_\theta = \phi_\theta$
 879 11: **else**
 880 12: Sample $\phi_i, \phi_j \in \{\phi_1, \dots, \phi_m\}$
 881 13: Set $\tilde{\phi}_\theta = \phi_i \circ \phi_\theta \circ \phi_j$
 882 14: **end if**
 883 15: Sample a batch of trajectories τ using transformed policy $\tilde{\phi}_\theta(\pi)$, where π is Boltz-
 884 mann exploratory
 885 16: Compute return $R(\tau)$ and advantage $A^{\tilde{\phi}_\theta(\pi)}(\tau_t, a_t)$
 886 17: Compute invertibility regularizer: $L(\theta) = \|\phi_{\mathcal{O},\theta}^2 - \mathbf{Id}\|^2$
 887 18: Compute policy gradient:
 888
$$\nabla_\theta J(\tilde{\phi}_\theta(\pi)) = \mathbb{E}_{\tau \sim \tilde{\phi}_\theta(\pi)} \left[\sum_{t=0}^H \nabla_\theta \log \tilde{\phi}_\theta(\pi)(a_t | \tau_t) A^{\tilde{\phi}_\theta(\pi)}(\tau_t, a_t) \right]$$

 889
 890 19: Update $\theta \leftarrow \theta + \eta(\nabla_\theta J(\tilde{\phi}_\theta(\pi)) - \lambda_2 \nabla_\theta L(\theta))$
 891 20: **end for**
 892 21: Compute average expected return over trajectories:
 893
$$\bar{V}_{\phi_\theta} = \frac{1}{|\Pi'|} \sum_{\pi \in \Pi'} \mathbb{E}_{\tau \sim \pi} V_{\phi_\theta(\pi)}(\tau)$$

 894
 895
 896 22: **end while**
 897 23: Find the index of the lowest return in \bar{V}_{top} , say i_{\min}
 898 24: **if** $\bar{V}_{\phi_\theta} > \bar{V}_{\text{top}}[i_{\min}]$ **then**
 899 25: Replace the lowest return: $\bar{V}_{\text{top}}[i_{\min}] \leftarrow \bar{V}_{\phi_\theta}$
 900 26: Replace the corresponding transformation: $\phi_{\text{top}}[i_{\min}] \leftarrow \{\phi_{\mathcal{O},\theta}, \phi_{\mathcal{A},\theta}\}$
 901 27: **end if**
 902 28: **end for**
 903 29: **Output:** Set of transformations $\phi_{\text{top}} = \{\phi_{\theta^*} = \{\phi_{\mathcal{O},\theta^*}, \phi_{\mathcal{A},\theta^*}\}\}$

917

Algorithm 3 Learning Expected Return Symmetries for Pairs of Policies ... Optimization of Equation 12

-
- 1: **Input:** A set of (near-)optimal policies Π' , parameter space Θ for transformations, action space \mathcal{A} , observation space \mathcal{O} , learning rate $\eta > 0$, l for top transformations to save
 - 2: Initialize list of top l average expected values: $\bar{V}_{\text{top}} = [-\infty, \dots, -\infty]$ (length l)
 - 3: Initialize list of top l transformations: $\phi_{\text{top}} = [\emptyset, \dots, \emptyset]$ (length l)
 - 4: **for** each pair of policies $(\pi_i^1, \pi_j^2) \in \Pi' \times \Pi'$ **do**
 - 5: Initialize $\phi_{\mathcal{O}, \theta}$ and $\phi_{\mathcal{A}, \theta}$ with random parameters $\theta \in \Theta$
 - 6: **while** not converged **do**
 - 7: Sample a batch of trajectories τ' using the transformed pair $(\pi_i^1, \phi_{\theta}(\pi_j^2))$
 - 8: Compute the return $R(\tau')$ for each trajectory τ'
 - 9: Compute advantage $A^{(\pi_i^1, \phi_{\theta}(\pi_j^2))}(\tau'_t, a_t)$ using any advantage function (e.g., TD, GAE)
 - 10: Compute policy gradient:

$$\nabla_{\theta} J(\pi_i^1, \phi_{\theta}(\pi_j^2)) = \mathbb{E}_{\tau' \sim (\pi_i^1, \phi_{\theta}(\pi_j^2))} \left[\sum_{t=0}^H \nabla_{\theta} \log \phi_{\theta}(\pi_j^2)(a_t | \tau'_t) A^{(\pi_i^1, \phi_{\theta}(\pi_j^2))}(\tau'_t, a_t) \right]$$

- 11: Update parameters: $\theta \leftarrow \theta + \eta \nabla_{\theta} J(\pi_i^1, \phi_{\theta}(\pi_j^2))$
- 12: Compute average expected return over trajectories:

$$\bar{V}_{(\pi_i^1, \phi_{\theta}(\pi_j^2))} = \frac{1}{|\Pi'|} \sum_{(i,j)} \mathbb{E}_{\tau' \sim (\pi_i^1, \phi_{\theta}(\pi_j^2))} [V_{(\pi_i^1, \phi_{\theta}(\pi_j^2))}(\tau')]$$

- 13: **end while**
 - 14: Find the index of the lowest return in \bar{J}_{top} , say i_{\min}
 - 15: **if** $\bar{V}_{(\pi_i^1, \phi_{\theta}(\pi_j^2))} > \bar{V}_{\text{top}}[i_{\min}]$ **then**
 - 16: Replace the lowest return: $\bar{V}_{\text{top}}[i_{\min}] \leftarrow \bar{V}_{(\pi_i^1, \phi_{\theta}(\pi_j^2))}$
 - 17: Replace the corresponding transformation: $\phi_{\text{top}}[i_{\min}] \leftarrow \{\phi_{\mathcal{O}, \theta}, \phi_{\mathcal{A}, \theta}\}$
 - 18: **end if**
 - 19: **end for**
 - 20: **Output:** Set of transformations $\phi_{\text{top}} = \{\phi_{\theta^*} = \{\phi_{\mathcal{O}, \theta^*}, \phi_{\mathcal{A}, \theta^*}\}\}$
-

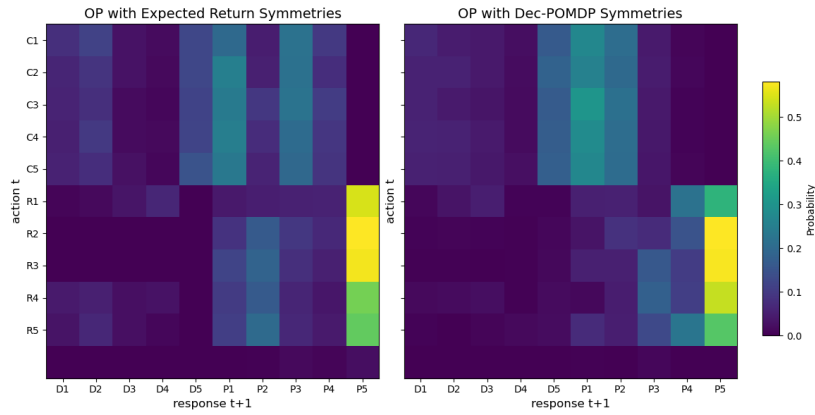


Figure 3: Conditional action matrices of $OP^{\Phi^{MDP}}$ -optimal and $OP^{\Phi^{ER}}$ -optimal policies; i.e., $P(a_t^i | a_{t-1}^j)$. We select the agent from both respective populations achieving the highest cross-play scores. We can see the $OP^{\Phi^{ER}}$ -optimal policy more consistently uses a rank hint to signal playing the fifth card, whereas the $OP^{\Phi^{MDP}}$ -optimal policy uses a similar convention but less consistently.

E INTERPRETABILITY OF HANABI OP AGENTS

1026 F HANABI

1027

1028 Hanabi is a cooperative card game that can be played with 2 to 5 people. Hanabi is a popular game,
 1029 having been crowned the 2013 “Spiel des Jahres” award, a German industry award given to the
 1030 best board game of the year. Hanabi has been proposed as an AI benchmark task to test models of
 1031 cooperative play that act under partial information Bard et al. (2020). To date, Hanabi has one of the
 1032 largest state spaces of all Dec-POMDP benchmarks.

1033 The deck of cards in Hanabi is comprised of five colors (white, yellow, green, blue and red), and five
 1034 ranks (1 through 5), where for each color there are three 1’s, two each of 2’s, 3’s and 4’s, and one
 1035 5, for a total deck size of fifty cards. Each player is dealt five cards (or four cards if there are 4 or 5
 1036 players). At the start, the players collectively have eight information tokens and three fuse tokens,
 1037 the uses of which shall be explained presently.

1038 In Hanabi, players can see all other players’ hands but their own. The goal of the game is to play
 1039 cards to collectively form five consecutively ordered stacks, one for each color, beginning with a
 1040 card of rank 1 and ending with a card of rank 5. These stacks are referred to as fireworks, as playing
 1041 the cards in order is meant to draw analogy to setting up a firework display.

1042 We call the player whose turn it is the active agent. The active agent must conduct one of three
 1043 actions:

1044

1045 • **Hint** - The active agent chooses another player to grant a hint to. A hint involves the active
 1046 agent choosing a color or rank, and revealing to their chosen partner all cards in the partner’s
 1047 hand that satisfy the chosen color or rank. Performing a hint exhausts an information token.
 1048 If the players have no information tokens, a hint may not be conducted and the active agent
 1049 must either conduct a discard or a play.

1050 • **Discard** - The active agent chooses one of the cards in their hand to discard. The identity
 1051 of the discarded card is revealed to the active agent and becomes public information.
 1052 Discarding a card replenishes an information token should the players have less than eight.

1053 • **Play** - The active agent attempts to play one of the cards in their hand. The identity of
 1054 the played card is revealed to the active agent and becomes public information. The active
 1055 agent has played successfully if their played card is the next in the firework of its color to
 1056 be played, and the played card is then added to the sequence. If a firework is completed,
 1057 the players receive a new information token should they have less than eight. If the player
 1058 is unsuccessful, the card is discarded, without replenishment of an information token, and
 1059 the players lose a fuse token.

1060 The game ends when all three fuse tokens are spent, when the players successfully complete all five
 1061 fireworks, or when the last card in the deck is drawn and all players take one last turn. If the game
 1062 finishes by depletion of all fuse tokens (i.e. by “bombing out”), the players receive a score of 0.
 1063 Otherwise, the score of the finished game is the sum of the highest card ranks in each firework, for
 1064 a highest possible score of 25.

1065 More facts about Hanabi:

1066

1067 1. The Dec-POMDP symmetries correspond to permutations of the five card colors ($5! =$
 1068 120).

1069 2. In two-player Hanabi, there are 20 possible actions per turn, organized into four types:
 1070 Play, Discard, Color Hint, and Rank Hint. These yield 190 distinct action transpositions.

1071 3. A perfect score is 25, though some deck permutations make this score unreachable, so no
 1072 policy can guarantee an expected return of 25.

1073

1074

1075

1076

1077

1078

1079