Appendix

In Appendix A, we briefly compare our approach to Harsanyi & Selten (1988). In Appendix B, we provide additional details about our implemented algorithms, cross-play evaluation, and further results. Everything afterwards, starting with Appendix C, is a self-contained rigorous treatment of the results that were informally stated in the paper. The two main theorems are subject of Appendices E and F. In the following, we give a brief outline of Appendices C–F.

In Appendix C, we make rigorous the definition of an LFC game and LFC problem from Section 4.2, and we provide auxiliary results required to prove our main theorems. Among those, we show in Section C.7 that any optimal symmetric profile of learning algorithms in an LFC game is a Nash equilibrium. In addition, in Appendix C.8, we briefly discuss a condition under which the objective in the LFC problem is equivalent to the formulation used in our experiments as outlined in Section 6.

In Appendix D, we provide characterizations of both the OP objective and the payoff in an LFC game, in terms of equivalence classes of policies under random permutations by automorphisms. This notion of equivalence makes it possible to analyze the OP-optimal policies in terms of representatives of equivalence classes that are invariant to automorphisms. We will use our results from this section for the proofs about the LFC problem, and for a proof about the existence of random tie-breaking functions.

In Appendix E, we then turn to stating and proving a rigorous version of Theorem 7. To that end, we show that there are two distinct OP-optimal equivalence classes in the two-stage lever game (Appendix E.1), and then prove that any algorithm that learns both of these is an OP learning algorithm, but not optimal in the LFC problem of that game (Appendix E.2).

Lastly, in Appendix F, we provide theoretical results about OP with tie-breaking and state and prove a rigorous version of Theorem 8. First, we define OP with tie-breaking and discuss to what degree the formal definition is satisfied by our method (Appendix F.1). Second, we show that OP with tie-breaking is optimal in the LFC problem and that all principals using OP with tie-breaking is an optimal symmetric Nash equilibrium of any LFC game (Appendix F.2). Third, we prove that a modification of the tie-breaking function introduced in Section 5 satisfies our formal requirements (Appendix F.3).

List of Symbols

General mathematical notation

$\mathbb{N}$	natural numbers excluding 0
$\mathbb{R}$	real numbers
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$
$\mathcal{P}(\mathcal{X})$	power set of the set $\mathcal{X}$
$\prod_{i=1}^{N} \mathcal{X}_i$	Cartesian product $\mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ of sets $\mathcal{X}_1, \ldots, \mathcal{X}_N$
$\lvert\mathcal{X}\rvert$	cardinality of the set $\mathcal{X}$

$\mathcal{X} \setminus \mathcal{Y}$	set of elements of $\mathcal{X}$ that are not in $\mathcal{Y}$
$X := Y$	X is defined as Y
$f \circ g$	composition of two composable functions f and g
$\text{proj}_i(x)$	projection on the ith component of the vector $x = (x_i)_{i \in \mathcal{X}}$
x_{-i}	vector with the ith component removed, $x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_N)$
$(\tilde{x}_i, x_{-i})$	vector with $\tilde{x}_i$ as ith component, $(\tilde{x}_i, x_{-i}) := (x_1, \ldots x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots x_N)$
δ_{ij}	Kronecker delta
$\mathbb{1}_{\mathcal{X}}$	indicator function of set $\mathcal{X}$
$\mathbb{P}$	probability measure
$\Delta(\mathcal{X})$	set of probability mass functions or measures over the set $\mathcal{X}$
$\mathcal{E}_1 \otimes \mathcal{E}_2$	product-σ-Algebra of $\mathcal{E}_1$ and $\mathcal{E}_2$
$\mu_1 \otimes \mu_2$	product measure of μ_1 and μ_2
$\mathbb{E}[X]$	expectation of the random variable X
$\mathbb{E}_{x \sim \nu}[f(x)]$	integral of f with respect to the measure ν
$\mathcal{U}(\mathcal{X})$	uniform distribution over the set $\mathcal{X}$
δ_x	Dirac measure
θ	parameter value
ξ	neural network
$\mathcal{L}$	loss function

Dec-POMDPs

D, E, F	Dec-POMDPs
$\mathcal{X}^D$	set belonging to the Dec-POMDP D (e.g., Π^D is the set of policies of D)
$\mathcal{N}$	set of agents or principals $\mathcal{N} = \{1, \ldots, N\}$
i	agent or principal
$\mathcal{S}$	set of states
S_t	random variable for the state at step t
s	state
$\mathcal{A}$	set of joint actions
$\mathcal{A}_i$	set of actions of player i
$A_{i,t}$	random variable for the action of agent i at step t
a	joint action
a_i	action of agent i
$\mathcal{O}$	set of observations
$\mathcal{O}_i$	set of observations of agent i
$O_{i,t}$	random variable for the observation of agent i at step t
o	joint observation
o_i	observation of agent i
R_t	random variable for the reward at step t

r	reward
$P(s' \mid s, a)$	transition probability
$O(o \mid s, a)$	observation probability
$\mathcal{R}(a, s)$	reward given joint action a and state s
T	horizon
$\overline{\mathcal{AO}}$	set of joint action-observation histories
$\overline{\mathcal{AO}}_{i,t}$	set of local action-observation history of agent i of length t
$\overline{AO}_{i,t}$	random variable for action-observation history of agent i of length t
$\tau_{i,t}$	local action-observation history of agent i of length t
$\mathcal{H}$	set of histories
H	random variable for the history
τ	history
$\mathcal{H}_t$	set of histories of length t
H_t	random variable for the history of length t
τ_t	history of length t
Π	set of joint policies
Π_i	set of local policies of agent i
π	joint policy
π_i	local policy of agent i
Π^0	set of joint deterministic policies
Π_i^0	set of local deterministic policies of agent i
$(\Omega, \mathcal{P}(\Omega), \mathbb{P}_\pi)$	measure space for a Dec-POMDP environment induced by policy π
$\mathbb{E}_\pi$	expectation with respect to $\mathbb{P}_\pi$
$J^D(\pi)$	expected return of policy π in Dec-POMDP D

Label-free coordination and other-play

$\mathrm{Aut}(D)$	set of automorphisms
$\mathrm{Iso}(D, E)$	set of isomorphisms from D to E
$\mathrm{Sym}(D)$	set of labelings of D
f	isomorphism or labeling
f^*D	relabeled Dec-POMDP
$f^*\pi$	pushforward policy
$\mathbf{f}$	profile of isomorphisms
$\mathrm{Aut}(D)$	set of automorphisms of D
g	automorphism
e	identity automorphism
$\mathbf{g}$	profile of automorphisms
$J^D_{\mathrm{OP}}(\pi)$	other-play value of π in Dec-POMDP D
$\mathcal{D}, \mathcal{C}$	sets of Dec-POMDPs
$\mathcal{F}_i$	σ-Algebra over Π_i

$\mathcal{F}$	product-σ-Algebra over Π
ν	distribution over policies
$\Sigma^{\mathcal{D}}$	set of learning algorithms for $\mathcal{D}$
σ	learning algorithm
$\boldsymbol{\sigma}$	profile of learning algorithms
$U^D(\boldsymbol{\sigma})$	payoff in the label-free coordination game for D given strategy profile $\boldsymbol{\sigma}$
$U^D(\sigma)$	value of σ in the label-free coordination problem for D
μ	distribution over policies with independent local policies
Z_i	latent variable for the policy of agent i
$\mathcal{Z}$	measurable set of policies
π^μ	policy corresponding to the distribution μ
$\Psi(\pi)$	policy corresponding to the other-play distribution of π
$[\pi]$	equivalence class of policies
$\#$	hash function
χ	tie-breaking function

A. Comparison to the solution by Harsanyi & Selten

Here, we compare the solution to the equilibrium selection problem provided by Harsanyi & Selten (1988) to our approach. It is unclear how to apply Harsanyi & Selten (1988)'s solution to Dec-POMDPs, and this would be an interesting area for future work. However, we can translate Dec-POMDPs into Harsanyi & Selten (1988)'s formalism of standard-form games, using similar constructions as the ones for normal-form games and extensive-form games by Oliehoek et al. (2006), and apply Harsanyi & Selten (1988)'s solution to such a problem. We can then compare it to OP with tie-breaking as an optimal solution to the LFC problem.

Below, we give an example in which OP with tie-breaking is equivalent to any OP algorithm in theory, as there is only one OP-optimal policy (ignoring differences between policies that do not matter under OP; see Appendix D). We also consider that policy as good solution to ZSC *in spirit*. However, applying Harsanyi & Selten (1988)'s procedure to a corresponding standard-form game leads to a policy in which agents cannot coordinate and which thus leads to a lower payoff. We restrict ourselves to an informal exposition and leave a more rigorous analysis to future work.

Consider a version of the two-stage lever game with 10 instead of 2 levers. As in the two-stage lever game, pulling the same lever gives a reward of 1, non-coordination gives a reward of -1, and the game is fully observable. Note that, like in the game with two levers, an OP-optimal policy uniformly randomizes between all levers in the first round. If no coordination was achieved in the first round, then in the second round, an optimal policy randomizes between the two levers that have been played in the first round by both players, similarly to the two-lever variant. There is a difference, however, if players coordinated on one lever in the first round. Clearly, in one optimal policy, players repeat their action from the first round, as was the case in the two-stage lever game. However, unlike in the two-lever case, here, there is no second optimal policy. It is not possible for the players to consistently switch to a different lever, as there are now not one but 9 other levers to choose from. Hence, the only optimal policy is one that chooses the unique lever that was chosen in the first round. This appears to us as a good solution to ZSC in this case.

Now consider a corresponding standard-form game (Harsanyi & Selten, 1988, ch. 2). It is sufficient for us here to note that in this game, each player $i = 1, 2$ is split into agents $j_{\tau_{i,t}}$ with distinct sets of actions $\mathcal{A}_{\tau_{i,t}}$ (corresponding to the 10 levers) for each possible action-observation history $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$ in the corresponding Dec-POMDP. The payoff for a strategy for all agents of all players is then the expected return that the corresponding policy would receive. Importantly, symmetries as introduced by Harsanyi & Selten (1988, ch. 3.4) can permute each of the individual action sets $\mathcal{A}_{\tau_{i,t}}$ separately, as long as this does not change the payoffs (there are more rules for how symmetries can permute actions, players, and agents, but these do not matter for us here). For instance, one symmetry may leave the actions of both players $i = 1, 2$ in the first round unchanged, while it may apply one permutation to the action sets $\mathcal{A}_{\tau_{i,1}}$ corresponding to action-observation histories $\tau_{i,1} \in \overline{\mathcal{AO}}_{i,1}$ of all agents of both players in the second round. Since rewards in the second round do not depend on actions in the first round, and permuting all actions of all second-round agents in the same way does not change the rewards for actions, such a permutation is a symmetry of the game.

As a result, in the first and the second round, all individual actions are symmetric, and, unlike in the corresponding Dec-POMDP, the symmetries for both rounds can be applied independently of each other. Hence, a symmetry-invariant strategy needs to play all actions with equal probability in both rounds. Since Harsanyi & Selten (1988)'s solution always chooses a strategy that is invariant to symmetries (Harsanyi & Selten, 1988, ch. 3.4), it follows that the strategy chosen by their procedure is a uniform distribution. Clearly, this strategy yields a lower return than the OP-optimal policy described above. In particular, since this applies independently of labelings, it follows from Theorem 8 that the solution must be suboptimal in the associated LFC problem.

A similar argument could be made about cheap-talk: in a standard-form game, players using a symmetry-invariant policy would never be able to use cheap-talk, as they could not learn the meanings of each others' messages over time. Transforming a Dec-POMDP into a standard-form game thus yields too many symmetries, precluding players from coordinating based on the structure of the Dec-POMDP, even if it was possible to uniquely do so. In contrast, OP exhibits in a sense the opposite failure mode in the two-stage lever game, allowing players to coordinate arbitrarily due to too few symmetries between policies.

B. Further experimental details

Here, we provide additional details about the experiments outlined in Section 6. We describe our implementation of OP (Appendix B.1), our implementation other OP with tie-breaking (Appendix B.2), and discuss our cross-play evaluation procedure as well as some further results (Appendix B.3).

B.1. Other-play implementation

Our implementation of the OP learning algorithm is based on the PyMARL framework (Samvelyan et al., 2019). We use recurrent neural networks to parameterize the policies of agents and a policy gradient algorithm to train agents' policies. Given that our toy problems are very small, they could also be solved by simple tabular methods. Nevertheless, we choose to employ this framework to demonstrate that our results transfer to state-of-the-art methods, even if the problems do not require them.

PyMARL is based on the PyTorch deep learning framework (Paszke et al., 2019). Neural network layers are implemented using the PyTorch module `nn.Linear` and the recurrent neural network uses a single `nn.GRUCell`, with the input encoding being one layer with ReLU activation functions. Hidden states are transformed into probabilities by a single `nn.Linear` layer followed by a softmax. The dimension of the hidden state is 64. Agent parameters are optimized using the `RMSProp` module, with a learning rate of 0.0005, an alpha of 0.99 and epsilon of 0.00001. These hyperparameters were all adopted as default values from the PyMARL framework.

Algorithm 1 Other-play learning algorithm based on vanilla policy gradient

Input: Dec-POMDP D
    Number of training steps L
    Episode batch-size K
    Gradient-based optimizer
Output: Joint policy $\pi \in \Pi^D$
Initialize θ
for $l = 1$ to L **do**
  for $k = 1$ to K **do**
    Sample profile of automorphisms $\mathbf{g}^{(k)} \sim \mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})$
    Sample history $\tau^{(k)} \sim \mathbb{P}^D_{\mathbf{g}^{(k)*}\pi_\theta}$ using joint policy $\mathbf{g}^{(k)*}\pi_\theta$
    for $t = 1$ to T **do**
      $G_t^{(k)} \leftarrow \sum_{t'=t}^T r_{t'}^{(k)}$
    end for
  end for
  Compute loss $\mathcal{L}(\theta) \leftarrow -\frac{1}{KTN} \sum_{k=1}^K \sum_{t=1}^T G_t^{(k)} \sum_{i=1}^N \log \pi_{\theta, \mathbf{g}_i^{(k)-1} i}(\mathbf{g}_i^{(k)-1} a_{i,t}^{(k)} \mid \mathbf{g}_i^{(k)-1} \tau_{i,t}^{(k)})$
  Update θ using $\nabla_\theta \mathcal{L}(\theta)$ to minimize $\mathcal{L}$
end for
Return π_θ

Since our generalization of the OP objective requires policies that can randomize, and since it cannot be implemented as the SP objective in a modified Dec-POMDP (see Appendix D.5), it is not clear how to use multi-agent methods based on value functions (e.g. Sunehag et al., 2018; Foerster et al., 2018). For this reason, we use a vanilla multi-agent policy gradient algorithm without baseline (Nguyen et al., 2017; Williams, 1992; Sutton & Barto, 2018, ch. 13.1), which can easily be applied to our generalization of the OP objective (see Algorithm 1).

We use *weight sharing*, that is, all agents use the same neural network and receive an additional observation specifying their agent-ID. In the two-stage lever game, where agents are symmetric, we omit this agent-ID and thus force the resulting joint policy to be symmetric, $\pi_1 = \pi_2$. We can do this as a symmetric policy is optimal under OP in this case (see Theorem 70). As a benefit, we do not have to implement permutations of agents for OP. In the asymmetric lever game, since agents are not symmetric, agent-IDs are added to observations as one-hot vectors.

Lastly, we add a penalty for the negative entropy of a policy to the loss-function (Mnih et al., 2016; Schulman et al., 2017). The entropy of the probability distribution $\pi_{\theta,i}(\cdot \mid \tau_{i,t})$ is defined as

$$H(\pi_{\theta,i}(\cdot \mid \tau_{i,t})) := -\sum_{a_i \in \mathcal{A}_i} \pi_{\theta,i}(a_i \mid \tau_{i,t}) \log \pi_{\theta,i}(a_i \mid \tau_{i,t}).$$

The loss-function is then

$$\tilde{\mathcal{L}}(\theta) = -\frac{1}{KTN} \sum_{k=1}^K \sum_{t=0}^T \left(G_t^{(k)} \sum_{i=1}^N \log \pi_{\theta,i}(a_{i,t}^{(k)} \mid \tau_{i,t}^{(k)}) + \alpha \sum_{i \in \mathcal{N}} H(\pi_{\theta,i}(\cdot \mid \tau_{i,t}^{(k)})) \right), \tag{15}$$

where α is a hyperparameter.

We choose $\alpha = 0.5$, as the highest α at which we still got fast convergence to an approximately optimal policy, after testing $\alpha = 1, 0.5, 0.1$, and 0.05. First, this encourages exploration and avoids a policy prematurely converging to a local minimum. Without this term, a small percentage of the learned policies in the asymmetric lever game converged to suboptimal equilibria. Second, we did this to make sure that agents learn to play a unique uniform distribution where actions do not matter for the OP value of a policy. Since it can be the case that actions do not matter but each choice of distribution still creates a different policy, this helps reduce the number of different policies that are learned, and thus facilitates tie-breaking between the remaining policies. Additionally, in the lever game with asymmetric players, it ensures that one can always infer from a distribution over histories the Dec-POMDP that these histories belong to. It hence suffices to let our hash function depend only on histories (see Appendix F.3).

Finally, we briefly outline how sampling from a policy $f^*\pi$ is implemented, in the two-stage lever game where agents are symmetric and use the same policy network. Actions and observations are encoded as one-hot vectors, i.e., as elements of the canonical basis $\{e_1, \ldots, e_k\}$, where k is the cardinality of the respective set. For a given episode, one profile of automorphisms $\mathbf{g}_1, \mathbf{g}_2 \sim \mathcal{U}(\mathrm{Aut}(D))$ is sampled. At time step t, the observation input of the agent i is $\mathbf{g}_i^{-1}(O_{i,t}, A_{i,t-1})$. Then an action $\tilde{A}_{i,t}$ is sampled from the agent policy, and that action is permuted by applying $\mathbf{g}_i$, i.e., $A_{i,t} := \mathbf{g}_i \tilde{A}_{i,t}$. Otherwise, the Dec-POMDP model proceeds as normal. One can easily see that this results in a history $H \sim \mathbb{P}_{\mathbf{g}^*\pi}$.

B.2. Deep tie-breaking

We implement OP with tie-breaking as described in Section 5 (see Algorithm 2).

Algorithm 2 OP with tie-breaking

Input: Dec-POMDP E
OP learning algorithm σ^{OP}
Tie-breaking-function χ
Number of seeds K
Output: Joint policy $\pi^* \in \Pi^E$
for $k = 1$ to K **do**
  Train policy $\pi^{(k)} \sim \sigma^{\mathrm{OP}}(E)$
  Calculate tie-breaking value $x^{(k)} \leftarrow \chi(E, \pi^{(k)})$
end for
$k_{\max} \leftarrow \arg\max_{k=1}^{K} x^{(k)}$
$\pi^* \leftarrow \pi^{(k_{\max})}$
Return π^*

Turning to the tie-breaking function, the function is only applied to actions of both agents and to rewards, but not to observations or states. This is because in our problems, states are always the same, and observations are completely determined by actions. The hash network has four hidden layers with ReLu activation functions, a hidden dimension of 32, and weights and biases are initialized uniformly in $[-1, 1]$. We chose these hyperparameters mostly based on prior considerations, but we did compare neural network depths 2 to 5 and hidden-layer dimensions 8, 16, 32 and 64 to determine hyperparameters for which OP with tie-breaking performed well. To calculate the tie-breaking function, we use 2048 episode samples. See Algorithm 3 for pseudo code.

Algorithm 3 Tie-breaking function

Input: Dec-POMDP E
Policy $\pi \in \Pi^E$
Neural network architecture $\xi_\#$
Random seed n
Number of episode samples K
Tie-breaking value $\chi(E, \pi)$
Initialize $\xi_\#$ using random seed n
for $k = 1$ to K **do**
  Sample profile of automorphisms $\mathbf{g}^{(k)} \sim \mathcal{U}(\mathrm{Aut}(E)^{\mathcal{N}})$
  Sample history $\tau^{(k)} \sim \mathbb{P}^E_{\mathbf{g}^{(k)*}\pi}$
end for
$\chi(E, \pi) \leftarrow \frac{1}{N!K} \sum_{k=1}^K \sum_{f_N \in \mathrm{Bij}(\mathcal{N})} \xi_\#(f_N(\iota(\tau^{(k)}))$
Return $\chi(D, \pi)$

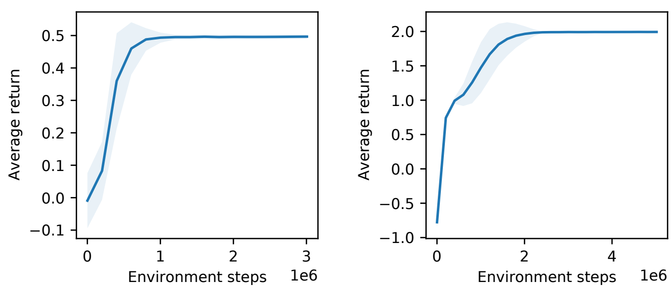

Figure 7. Learning curves for 320 independent training runs of the OP algorithm, with shaded standard deviation.

B.3. Cross-play evaluation

Recall that, to evaluate algorithms in the LFC problems for our environments, we simplify the objective to

$$\tilde{U}^D(\sigma) := \mathbb{E}_{\pi^{(i)} \sim \sigma(D),\, i=1,2}\Big[\mathbb{E}_{\mathbf{g}_i \sim \mathcal{U}(\mathrm{Aut}(D)),\, i=1,2}\Big[J^D((\mathbf{g}_1^*\pi^{(1)})_1, (\mathbf{g}_2^*\pi^{(2)})_2)\Big]\Big]. \tag{16}$$

We discuss this simplification further in Appendix C.8.

For each game, we train 320 joint policies in total, using OP with different seeds for network initialization and environmental randomness. In the two-stage lever game, we train each policy for three million environment steps, and in the variant with asymmetric players for five million steps (see Figure 7 for learning curves). The training for both environments took around three days on a MacBook Pro laptop from 2017. We split the 320 policies into 10 sets of 32 policies each. The policies in each set are used for application of the tie-breaking method, while the 10 different sets represent independent runs which can be used to calculate cross-play values.

Given a list of 10 policies $\pi^{(1)}, \ldots, \pi^{(10)}$ produced by independent runs of an algorithm, we estimate the cross-play value $G_{k,l} \approx \mathbb{E}_{\mathbf{g}_i \sim \mathcal{U}(\mathrm{Aut}(D)),\, i=1,2}\left[J((\mathbf{g}_1^*\pi^{(k)})_1, (\mathbf{g}_2^*\pi^{(l)})_2)\right]$ for any two indices of policies

Table 2. Average off-diagonal cross-play value and standard deviation for OP with tie-breaking with $K = 1$, 2, 4, 8, 16, and 32, in the two-stage lever game (TSLG) and asymmetric lever game (ALG).

Problem	1	2	4	8	16	32
TSLG	-0.03 (±0.00)	0.14 (±0.13)	0.48 (±0.06)	0.50 (±0.00)	0.50 (±0.00)	0.50 (±0.00)
ALG	0.90 (±0.00)	1.23 (±0.08)	1.75 (±0.22)	1.90 (±0.16)	1.96 (±0.11)	1.97 (±0.10)

Table 3. Percentage of policies learned corresponding to different classes of mutually compatible policies, for two-stage lever game (TSLG) and asymmetric lever game (ALG), each time out of 320 seeds for training.

Problem	Class 1	Class 2	Class 3	Class 4
TSLG	50.94%	49.06%	/	/
ALG	43.75%	20.94%	19.38%	15.94%

$k, l \in \{1, \dots, 10\}$. These values are used to print cross-play matrices (Figure 4) The average off-diagonal cross-play value is calculated as $G := \frac{1}{10(10-1)} \sum_{k \neq l} G_{k,l}$, where we leave out values on the diagonal because these do not represent cross-play between independent runs. Note that in the two-stage lever game, off-diagonal values can be higher than the optimal OP value. This is because in cross-play, agents can use different joint policies that accidentally work better in cross-play than a symmetric policy. It is counterbalanced in expectation by other entries of the matrix, in which unsuitable agents are matched. We provide in Table 2 the average off-diagonal cross-play values used to create the graph in Figure 5.

Lastly, we categorize policies into classes of mutually compatible policies. To do so, we calculate a cross-play value for each combination of two out of the 320 trained joint policies, using 256 episodes each. We then dynamically build classes of policies by comparing a policy's expected return to the cross-play value with a policy from a given class, and assigning the policy to that class if the difference between the values is below a threshold of 0.6. If no class is compatible with a policy, the method creates a new class containing that policy. In that way, all 320 policies are assigned to a class. In Table 3 we list the relative sizes of the different classes.

While in the two-stage lever game, both classes are represented approximately equally, a clear majority of policies belongs to one class in the asymmetric lever game. This class corresponds to the strategy in which player 2 switches to a different action upon non-coordination in the first round, and in which both players repeat their action given successful coordination in the first round. The least frequent class was the one in which player 2 switches to a different action upon non-coordination, but where both players switch to a different action if they were successful in the first round. The imbalance of classes in this case could be used to implement a tie-breaking rule that chooses the policy that is learned more often.

We think that a reasonable policy in the two-stage lever game is one in which a player repeats their action upon coordination. Hence, a tie-breaking function that chooses the "repeat"-policy is preferable. Interestingly, out of 20 seeds for the hash function, only 4 of the resulting tie-breaking functions gave higher values to that policy.

We used new random seeds for the final experiments that have not been used to improve the method.

C. Formalization of a label-free coordination game and problem

In this section, we formalize label-free coordination (LFC) games and the LFC problem, and we provide auxiliary results required to prove our main theorems.

In Appendix C.1, we recall the definition of Dec-POMDPs and introduce some additional notation. In Appendix C.2, we recall the definition of isomorphisms and automorphisms, provide more comprehensive examples, and prove first lemmas about these concepts. Afterwards, in Appendix C.3, we discuss the pushforward and prove, among other things, that a pushforward policy has the same expected return as the original policy. In Appendix C.4, we show that automorphisms define group actions on joint actions, policies, etc. and introduce the concept of an orbit. In Appendix C.5, we then introduce Dec-POMDP labels to construct the set of relabeled Dec-POMDPs used to define an LFC game. In Appendix C.6, we recall the definition of LFC games and of the LFC problem, and we provide different expressions for the payoff in LFC games. We also prove that LFC games for isomorphic Dec-POMDPs are equivalent, up to a permutation of the principals in the game. In Appendix C.7, we prove that any strategy profile that is optimal among those that respects symmetries between principals is a Nash equilibrium in the game, making use of group actions and orbits. This theorem is needed later to prove that all principals using OP with tie-breaking is a Nash equilibrium. Lastly, in Appendix C.8, we briefly discuss a condition under which the objective in the LFC problem is equivalent to the formulation without relabeled Dec-POMDPs used in our experiments as outlined in Section 6.

C.1. Recapitulation of Dec-POMDPs

Before we turn to isomorphisms and automorphisms, we briefly recapitulate Dec-POMDPs and introduce some additional notation that we will use throughout the following. That is, we also introduce a history of length t, $\tau_t \in \mathcal{H}_t$, the set of deterministic policies Π^0 and we define the measure space $(\Omega, \mathcal{P}(\Omega), \mathbb{P}_\pi)$ corresponding to a Dec-POMDP in which agents follow the joint policy π. We also define the notation $x_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots x_N)$ and the projection operator $\mathrm{proj}_i(x) = x_i$. Apart from this, the definitions in this section are a more elaborate version of those in Section 3 of the main text.

To begin, recall the definition of a Dec-POMDP.

Definition 9. A (finite-horizon) Dec-POMDP is a tuple

$$D = \Big(\mathcal{N}, \mathcal{S}, \mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i, P, \mathcal{R}, \mathcal{O} = \prod_{i \in \mathcal{N}} \mathcal{O}_i, O, b_0, T\Big)$$

where

- $\mathcal{N} = \{1, \dots, N\}$, $N \in \mathbb{N}$ is a finite set of agents.
- $\mathcal{S}$ is a finite set of states.
- $\mathcal{A}_i$ is a finite set of actions for player $i \in \mathcal{N}$.
- $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability kernel (where $\Delta(\mathcal{S})$ denotes the set of probability mass functions over $\mathcal{S}$).
- $\mathcal{R}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the joint reward function.
- $\mathcal{O}_i$ is a finite set of observations for player $i \in \mathcal{N}$.

- $O\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{O})$ is the observation probability kernel.
- $b_0 \in \Delta(\mathcal{S})$ is a distribution over the initial state.
- $T \in \mathbb{N}_0$ is the horizon of the problem.

When considering different Dec-POMDPs D, E at the same time, we write $\mathcal{A}^D, \mathcal{A}^E$, etc., to indicate to which Dec-POMDP the set belongs. Similarly, we do this for transition, observation, and reward functions. We omit the index D if it is clear which Dec-POMDP is meant.

Given a Dec-POMDP D, define the set of *local action-observation histories* of length $t \in \{0, \dots, T\}$ for player $i \in \mathcal{N}$ as

$$\overline{\mathcal{AO}}_{i,t} := (\mathcal{A}_i \times \mathcal{O}_i)^t ,$$

(where $(\mathcal{A} \times \mathcal{O})^0 := \{\emptyset\}$) and the set of local action-observation histories for player i as

$$\overline{\mathcal{AO}}_i := \bigcup_{0 \le t \le T} (\mathcal{A}_i \times \mathcal{O}_i)^t .$$

Moreover, we define the set of histories of length t as

$$\mathcal{H}_t := \mathcal{S} \times \mathcal{A} \times \mathcal{R}(\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{O} \times \mathcal{A} \times \mathcal{R}(\mathcal{S} \times \mathcal{A}))^t .$$

and $\mathcal{H} := \mathcal{H}_T$ as the set of histories. Histories $\tau \in \mathcal{H}$ are also called episodes, when one is talking about a particular sample of the stochastic process induced by agents following a policy in the Dec-POMDP, as defined below.

At step $0 \le t \le T$, agent $i \in \mathcal{N}$ chooses a distribution over actions, conditional on a past action-observation history $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$. This choice is described by a *local (stochastic) policy* for player i, which is a mapping $\pi_i\colon \overline{\mathcal{AO}}_i \to \Delta(\mathcal{A}_i)$. Since actions can be chosen stochastically, a policy and past observations do not imply which past actions have been taken, so policies are also able to condition on past actions (i.e., agents remember their past actions). We write $\pi_i(a_i \mid \tau_{i,t})$ for the probability of action $a_i \in \mathcal{A}_i$ given the action-observation history $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$. A *(joint stochastic) policy* is a tuple $\pi = (\pi_1, \dots, \pi_N)$ with a local policy for each player. We denote Π^D as the set of joint stochastic policies for a Dec-POMDP D, and Π_i^D as the set of local policies for player i. A *local deterministic policy* for player $i \in \mathcal{N}$ is defined as a policy π_i such that $\pi_i(\cdot \mid \tau_{i,t})$ is concentrated on exactly one action (i.e., there exists $a_i \in \mathcal{A}_i$ such that $\pi_i(a_i \mid \tau_{i,t}) = 1$) for each $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$. A joint deterministic policy is defined analogously to joint stochastic policies, and $(\Pi^0)^D$ is the set of joint deterministic policies for D.[4]

A Dec-POMDP D together with a joint policy $\pi \in \Pi^D$ specify the stochastic process according to which an episode evolves, yielding a distribution over histories. Formally, we define a discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P}_\pi)$ with random variables for states, actions, observations, and rewards at all time steps.

The distributions of these random variables are defined inductively in the following way. First, $S_0 \sim b_0$, that is, the first state is an independent random variable with values in $\mathcal{S}$ and image distribution b_0. Similarly, the

[4]Normally, a deterministic policy is defined as having values in $\mathcal{A}_i$, but we can identify each element $a_i \in \mathcal{A}_i$ with a probability mass function in $\Delta(\mathcal{A}_i)$ with support $\{a_i\}$, so these definitions are interchangeable. Since a deterministic policy outputs only one action for each action-observation history, it there is only one possible action for each observation history $(o_{i,1}, \dots, o_{i,t})$, and we could hence write a deterministic policy as a map from observation histories to actions. We omit this here for simplicity.

first action of agent i is distributed according to $A_{i,0} \sim \pi_i(\cdot \mid \emptyset)$ (note that $\emptyset \in \overline{\mathcal{AO}}_i$ by definition), and the first reward is $R_0 := \mathcal{R}(S_0, A_0)$, where we define $A_0 := (A_{i,0})_{i\in\mathcal{N}}$.

Now assume there are random variables for all states, actions, and observations until step $0 \le t \le T$ (in the case $t = 0$, there are no observations defined yet). We can summarize them into a random variable for the history of length t as

$$H_t := (S_0, A_0, (S_{t'}, (O_{i,t'}, A_{i,t'})_{i\in\mathcal{N}}, R_{t'})_{1\le t'\le t})$$

(and we let $H := H_T$). At step $t+1$, a new state $S_{t+1} \sim P(\cdot \mid S_t, A_t)$ and a new joint observation $O_{t+1} \sim O(\cdot \mid S_{t+1}, A_t)$ are sampled. Note that, in a slight abuse of notation, we use O for both observation probabilities and observation random variable. We can define a random variable for the action-observation history of agent i at time $t+1$ as $\overline{AO}_{i,t+1} := (A_{i,0}, O_{i,1}, A_{i,1}, \dots, A_{i,t}, O_{i,t+1})$. Conditioning on this action-observation history, agent i samples an action $A_{i,t+1} \sim \pi_i(\cdot \mid \overline{AO}_{i,t+1})$. This yields a new joint action A_{t+1}. Finally, the new reward is $R_{t+1} := \mathcal{R}(S_{t+1}, A_{t+1})$, which concludes the definition.

We will sometimes make use of a simplified notation for a tuple excluding a particular player. Let $(x_i)_{i\in\mathcal{N}}$ be a vector indexed by agents, with elements $x_i \in \mathcal{X}_i$ where $\mathcal{X}_i, i \in \mathcal{N}$ are some sets. Then we define $x_{-j} := (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_N)$ for $j \in \mathcal{N}$. Moreover, for any $x'_i \in X_i$, we then write $(x'_i, x_{-i}) := (x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_N)$. We will also sometimes use the projection operator $\mathrm{proj}_i(x) := x_i$ to clearly refer to a particular element of x when x is a more complicated expression.

Now, given a problem D and a joint policy $\pi \in \Pi^D$, the measure $\mathbb{P}_\pi$ defined above specifies a distribution over rewards $R_0, \dots, R_T$. We can use this fact to define the expected return of the policy. To that end, let $\mathbb{E}_\pi := \mathbb{E}_{\mathbb{P}_\pi}$ denote the expectation with respect to $\mathbb{P}_\pi$.

Definition 10 (Self-play objective)**.** Let D be a Dec-POMDP. Define the *self-play (SP) objective* $J^D : \Pi^D \to \mathbb{R}$ for D via

$$J^D(\pi) := \mathbb{E}_\pi\left[\sum_{t=0}^{T} R_t\right]$$

for $\pi \in \Pi^D$. Here, $J^D(\pi)$ is called the expected return of a joint policy π.

C.2. Dec-POMDP isomorphisms and automorphisms

Here, we recall the definitions of isomorphisms and automorphisms from Sections 4.1 and 4.3 and give more comprehensive examples than in the main text. Then we prove some elementary results.

Let D, E be two Dec-POMDPs. Consider a tuple of bijective maps

$$f := (f_N, f_S, (f_{A_i})_{i\in\mathcal{N}}, (f_{O_i})_{i\in\mathcal{N}}),$$

where

$$f_N\colon\ \mathcal{N}^D \to \mathcal{N}^E \tag{17}$$

$$f_S\colon\ \mathcal{S}^D \to \mathcal{S}^E \tag{18}$$

$$\forall i \in \mathcal{N}\colon\quad f_{A_i}\colon \mathcal{A}_i^D \to \mathcal{A}_{f_N(i)}^E \tag{19}$$

$$\forall i \in \mathcal{N}\colon\quad f_{O_i}\colon \mathcal{O}_i^D \to \mathcal{O}_{f_N(i)}^E. \tag{20}$$

Given a joint action $a \in \mathcal{A}^D$ and such tuple f, we define

$$f_A(a) := \left(f_{A_{f_N^{-1}(i)}} a_{f_N^{-1}(i)} \right)_{i \in \mathcal{N}^E} \in \mathcal{A}^E, \tag{21}$$

and analogously for $o \in \mathcal{O}^D$

$$f_O(o) := \left(f_{O_{f_N^{-1}(i)}} o_{f_N^{-1}(i)} \right)_{i \in \mathcal{N}^E} \in \mathcal{O}^E. \tag{22}$$

That is, in the joint action $f_A(a) \in \mathcal{A}^E$, the agent $j = f_N(i) \in \mathcal{N}^E$ (where $i \in \mathcal{N}^D$) plays the action $f_{A_i}(a_i) \in \mathcal{A}^E_j$. Note that it is really $f_A(a) \in \mathcal{A}^E$, as for any $a \in \mathcal{A}^D$ and $j := f_N^{-1}(i) \in \mathcal{N}^D$, by definition it is

$$f_{A_j}(a_j) \in \mathcal{A}^E_{f_N(j)} = \mathcal{A}^E_{f_N(f_N^{-1}(i))} = \mathcal{A}^E_i.$$

The analogous holds for observations.

Definition 11 (Dec-POMDP isomorphism)**.** Let D, E be Dec-POMDPs such that both have the same horizon $T^D = T^E$, and let

$$f := (f_N, f_S, (f_{A_i})_{i \in \mathcal{N}}, (f_{O_i})_{i \in \mathcal{N}})$$

be a tuple of bijective maps as defined in equations (17)–(20). Then f is an isomorphism from D to E if for any $a \in \mathcal{A}^D$, $s, s' \in \mathcal{S}^D$ and $o \in \mathcal{O}^D$, it is

$$P^D(s' \mid s, a) = P^E(f_S(s') \mid f_S(s), f_A(a)) \tag{23}$$

$$O^D(o \mid s, a) = O^E(f_O(o) \mid f_S(s), f_A(a)) \tag{24}$$

$$\mathcal{R}^D(s, a) = \mathcal{R}^E(f_S(s), f_A(a)) \tag{25}$$

$$b_0^D(s) = b_0^E(f_S(s)). \tag{26}$$

If such an isomorphism exists, D and E are called isomorphic. We denote $\text{Iso}(D, E)$ for the set of isomorphisms from D to E.

Remark 12. In a Dec-POMDP, distributions over histories and policies can be ranked by their associated expected return. If a reward function is multiplied with a positive constant $\alpha \in \mathbb{R}_{>0}$ and shifted by a constant $\beta \in \mathbb{R}$, then the new reward function $\mathcal{R}' := \alpha\mathcal{R} + \beta$ still induces the same ranking. For this reason, one could consider such transformations as part of an isomorphism between two problems. For instance, Harsanyi & Selten (1988, p. 72) make it part of their definition of an isomorphism in the framework of standard-form games. For simplicity, we ignore this complication here and consider only isomorphic problems with reward functions that have the same range.

As in the main text, we write fa instead of $f_A(a)$ and fa_i instead of $f_{A_i}a_i$, and we do the same for observations, states, etc. We also write $f\tau_{i,t}$ for actions of isomorphisms on action-observation histories, which is defined as the element-wise application of f, and letting $fr := r$ for rewards, we define $f\tau$ for entire histories $\tau \in \mathcal{H}$.

Now we prove a first basic result about actions of isomorphisms. Note that an isomorphism $f \in \text{Iso}(D, E)$ is a bijective map

$$f \colon \mathcal{N}^D \times \mathcal{S}^D \times \prod_{i \in \mathcal{N}^D} \mathcal{A}_i^D \times \prod_{i \in \mathcal{N}^D} \mathcal{O}_i^D \to \mathcal{N}^E \times \mathcal{S}^E \times \prod_{i \in \mathcal{N}^E} \mathcal{A}^E_{fi} \times \prod_{i \in \mathcal{N}^E} \mathcal{O}^E_{fi}, \tag{27}$$

which can be inverted and composed with other maps. The following result shows that actions of isomorphisms on joint actions and joint observations are compatible with function composition and with taking inverses of functions. It follows that also the element-wise application to histories and action-observation histories is compatible in this way. In particular, this justifies omitting brackets when applying isomorphisms. We will use this lemma liberally in the following.

Lemma 13. *Let D, E, F be Dec-POMDPs, $a \in \mathcal{A}^D, o \in \mathcal{O}^D$, let $e \in \mathrm{Iso}(D, D)$ be the identity, and let $f \in \mathrm{Iso}(D, E)$, $\tilde{f} \in \mathrm{Iso}(E, F)$. Then it is*

(i) $ea = a$ *and* $eo = o$.

(ii) $\tilde{f}(fa) = (\tilde{f} \circ f)a$ *and* $\tilde{f}(fo) = (\tilde{f} \circ f)o$.

In particular, actions of isomorphisms f on joint actions and joint observations can be inverted using f^{-1}.

Proof. First, it is $ea = (ea_{e^{-1}i})_{i\in\mathcal{N}} = e$, and analogously for o. Second,

$$\tilde{f}(fa) = \tilde{f}(fa_{f^{-1}i})_{i\in\mathcal{N}} = (\tilde{f}(fa_{f^{-1}\tilde{f}^{-1}i}))_{i\in\mathcal{N}} = (\tilde{f}fa_{(\tilde{f}f)^{-1}i})_{i\in\mathcal{N}} = (\tilde{f} \circ f)a$$

as function composition is associative and the inverse of $\tilde{f} \circ f$ is $f^{-1} \circ \tilde{f}^{-1}$.

Next, it follows from the previous that $f^{-1}(fa) = (f^{-1} \circ f)a = ea = a$. The analogous holds for o, which concludes the proof. □

The following corollary states that the action of f on histories $\tau_t \in \mathcal{H}_t$ is bijective. An analogous corollary also hold for action-observation histories.

Corollary 14. *Let D, E be isomorphic Dec-POMDPs with isomorphism $f \in \mathrm{Iso}(D, E)$ and let $t \in \{0, \ldots, T\}$. Then $f_H \colon \mathcal{H}_t^D \to \mathcal{H}_t^E, \tau_t \mapsto f\tau_t$ is a bijective map.*

Proof. Let $\tau_t = (s_0, a_0, s_1, o_1, a_1, r_1, \ldots, s_t, o_t, a_t, r_t) \in \mathcal{H}_t^D$. Define f_H as above and $f_H^{-1} \colon \tau \mapsto f^{-1}\tau$. Then

$$\begin{aligned} f^{-1}(f(\tau_t)) &= f^{-1}((fs_0, fa_0, fs_1, fo_1, fa_1, fr_1, \ldots, fs_t, fo_t, fa_t, fr_t)) \\ &= (f^{-1}fs_0, f^{-1}fa_0, f^{-1}fs_1, f^{-1}fo_1, f^{-1}fa_1, f^{-1}fr_1, \ldots, f^{-1}fr_t)) \\ &\overset{\text{Lemma 13}}{=} (s_0, a_0, s_1, o_1, a_1, r_1, \ldots, s_t, o_t, a_t, r_t) = \tau_t. \end{aligned} \tag{28}$$

□

Recall the definition of an *automorphism*, which can be thought of as describing a symmetry of the Dec-POMDP.

Definition 15 (Dec-POMDP automorphism)**.** An isomorphism $f \in \mathrm{Iso}(D, D)$ from D to itself is called an automorphism. We define $\mathrm{Aut}(D) := \mathrm{Iso}(D, D)$ as the set of all automorphisms of D.

Now we give an example of an isomorphism and automorphism, using the lever coordination game.

Example 16 (Lever coordination game)**.** Consider the lever coordination game introduced in Section 1. This can be formalized as a Dec-POMDP D with only one state, one observation for each agent, and in which $T = 0$. The agents are $\mathcal{N} = \{1, 2\}$ and actions are $\mathcal{A}_1 = \mathcal{A}_2 = \{1, \dots, 10\}$. One possible reward function is

$$\mathcal{R}(a_1, a_2, s) = \delta_{a_1,a_2} \left(1.0 \cdot \mathbb{1}_{\{1,\dots,9\}}(a_1) + 0.9 \cdot \mathbb{1}_{\{10\}}(a_1)\right),$$

where

$$\delta_{a_1,a_2} := \begin{cases} 1 & \text{if } a_1 = a_2 \\ 0 & \text{otherwise} \end{cases}$$

is the Kronecker delta.

As an isomorphic problem E, consider a problem with the same sets of actions for both players, but where the reward function is defined as

$$\mathcal{R}^E(a_1, a_2, s) = \delta_{a_1,a_2} \left(1.0 \cdot \mathbb{1}_{\{2,\dots,10\}}(a_1) + 0.9 \cdot \mathbb{1}_{\{1\}}(a_1)\right)$$

for $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$ and the trivial state s. Note that this is formally a different Dec-POMDP. Nevertheless, we can define an isomorphism $f \in \mathrm{Iso}(D, E)$ in the following way. For $i = 1, 2$, define $f_{A_i} \colon \mathcal{A}_i \to \mathcal{A}_i$ such that $f_{A_i}(10) = 1$ and $f_{A_i}(1) = 10$, and let $f_{A_1} = f_{A_2}$ be arbitrary otherwise. Let the remaining components of f be the identity map. Then for any joint action a and the trivial state s, it is

$$\begin{aligned}\mathcal{R}^E(fs, fa) = \delta_{fa_1,fa_2} &\left(1.0 \cdot \mathbb{1}_{\{2,\dots,10\}}(fa_1) + 0.9 \cdot \mathbb{1}_{\{1\}}(fa_1)\right) \\ &= \delta_{a_1,a_2} \left(1.0 \cdot \mathbb{1}_{\{1,\dots,9\}}(a_1) + 0.9 \cdot \mathbb{1}_{\{10\}}(a_1)\right) = \mathcal{R}^D(s, a)\end{aligned} \quad (29)$$

and hence D and E are isomorphic (as observation and transition probabilities as well as the initial state distribution are trivial here).

Note that we could have used any two (potentially different) permutations $\hat{f}_1, \hat{f}_2$ of the two sets $\mathcal{A}_1, \mathcal{A}_2$ and defined a new reward function $\mathcal{R}'(s, a) := \mathcal{R}^D(s, \hat{f}_1^{-1}a_1, \hat{f}_2^{-1}a_2)$. This reward function would then define a new Dec-POMDP D', and the isomorphism from D to D' would be exactly f defined by $f_{A_1} = \hat{f}_1, f_{A_2} = \hat{f}_2$ and the identity in the other components, as $\mathcal{R}'(fs, fa) = \mathcal{R}^D(s, f_{A_1}\hat{f}_1^{-1}a_1, f_{A_2}\hat{f}_2^{-1}a_2) = \mathcal{R}^D(s, \hat{f}_1\hat{f}_1^{-1}a_1, \hat{f}_2\hat{f}_2^{-1}a_2) = \mathcal{R}^D(s, a)$. We will use this idea in Section C.5 to define relabeled Dec-POMDPs.

Next, consider the automorphisms of the lever coordination game. Note that the agents in this game are symmetric. For instance, we can define g via $g_N(1) = 2$ and $g_N(2) = 1$, and such that $g_{A_1} = g_{A_2} = \hat{g}$, where $\hat{g}$ is any permutation of $\{1, \dots, 10\}$ such that $\hat{g}(10) = 10$. Then one can easily check that $\mathcal{R}^D(gs, ga) = \mathcal{R}^D(s, a)$ for any joint action a and the state s.

Next, we introduce the automorphisms of the two-stage lever game (Example 3), which we will need later to prove that OP is suboptimal in the corresponding LFC problem.

Example 17. Recall that in the two-stage lever game, there are two agents, $\mathcal{N} = \{1, 2\}$, and the problem has two rounds, so $T = 1$. Each round, each agent has to pull one of two levers, $\mathcal{A}_1 = \mathcal{A}_2 = \{1, 2\}$. If both agents choose the same lever, they get a reward of 1. Otherwise, the reward is -1. There is again only one state, but there are two observations, $\mathcal{O}_1 = \mathcal{O}_2 = \{1, 2\}$. In the second stage ($t = 1$), each player observes the previous action of the other player, so $O_{i,1} = A_{-i,0}$ for $i = 1, 2$. The reward and observation probabilities are given in Table 4.

Table 4. Reward function and observation probabilities in the two-stage lever game (Example 3).

(a) Reward function $\mathcal{R}(s, a)$ for each joint action a

a	$\mathcal{R}(s, a)$
(1,1)	1
(1,2)	-1
(2,1)	-1
(2,2)	1

(b) Observation probabilities $O(o \mid s, a)$ for joint observations o and joint actions a.

$a \setminus o$	(1,1)	(1,2)	(2,1)	(2,2)
(1,1)	1	0	0	0
(1,2)	0	0	1	0
(2,1)	0	1	0	0
(2,2)	0	0	0	1

Table 5. Visualization of precomposition of both reward function and observation probability kernel with g^{-1}, by applying g to the index column and header row of the tables from Table 4. Note that apart from a permutation of rows and columns, the tables are identical to the ones in Table 4, showing that g is an automorphism.

(a) Reward function $\mathcal{R}(g^{-1}s, g^{-1}a)$

a	$\mathcal{R}(s, a)$
(2,2)	1
(1,2)	-1
(2,1)	-1
(1,1)	1

(b) Observation probabilities $O(g^{-1}o \mid g^{-1}s, g^{-1}a)$

$a \setminus o$	(2,2)	(1,2)	(2,1)	(1,1)
(2,2)	1	0	0	0
(1,2)	0	0	1	0
(2,1)	0	1	0	0
(1,1)	0	0	0	1

Using the table for reward function and observation probabilities, we can easily visualize isomorphisms and automorphisms. Consider any isomorphism $f \in \mathrm{Iso}(D, E)$ where E is some other Dec-POMDP. Considering, for instance, the reward function, we know that $\mathcal{R}^D(s, a) = \mathcal{R}^E(fs, fa)$. This means that if we want to check the value of the reward function of E when agents choose action fa, we can look it up in the table corresponding to $\mathcal{R}^D$ in the row for a. So we can visualize an isomorphism by applying f_A to the index (i.e., first) column of this table, but leaving the other cells unchanged. This creates a new table with the reward function for E, or, equivalently, this new table corresponds to the reward function $\mathcal{R}^D$ precomposed with f^{-1}, i.e., $\mathcal{R}^D(f_S^{-1}\cdot, f_A^{-1}\cdot) = \mathcal{R}^E$. Analogously, we can apply f_O to the header row and f_A to the index column of the table with observation probabilities, yielding $O^D(f_O^{-1}\cdot \mid f_S^{-1}\cdot, f_A^{-1}\cdot) = O^E$. An automorphism is then simply an isomorphism such that applying it to index column and header of the tables does not change the table, other than permuting rows and columns.

Now let g be an automorphism. g_N can either be the identity or it can switch both agents. In either case, it must be $g_{A_1} = g_{A_2}$, which can also either be the identity or the map that switches the actions. The observation permutation then has to be equal to that of the actions, $g_{O_1} = g_{O_2} = g_{A_1}$. There is trivially only one option for the state permutation.

We visualize the application of the automorphism that switches agents as well as actions and thus also observations in Table 5. For instance, using the definition of actions of automorphisms on joint actions in Equations (21) and (22), we have $f_A(1,1) = (f_{A_{f_N 1}} 1, f_{A_{f_N 2}} 1) = (f_{A_2} 1, f_{A_1} 1) = (2,2)$, and $f_A(1,2) = (f_{A_2} 2, f_{A_1} 1) = (1,2)$.

Before we turn to applying isomorphisms and automorphisms to policies, we briefly provide two basic results about isomorphisms and the relationship between isomorphisms and automorphisms.

First, we prove that function composition and inversion preserve isomorphisms.

Lemma 18. *Let D, E, F be Dec-POMDPs and let $f \in \mathrm{Iso}(D, E)$ and $\tilde{f} \in \mathrm{Iso}(E, F)$. Then $f^{-1} \in \mathrm{Iso}(E, D)$ and $\tilde{f} \circ f \in \mathrm{Iso}(D, F)$.*

Proof. Let $s, s' \in \mathcal{S}^D$, $a \in \mathcal{A}^D$. First, using the definition of an isomorphism and associativity of function composition, it is

$$P^D(s' \mid s, a) = P^E(fs' \mid fs, fa) = P^F(\tilde{f}fs' \mid \tilde{f}fs, \tilde{f}fa) = P^F((\tilde{f} \circ f)s' \mid (\tilde{f} \circ f)s, (\tilde{f} \circ f)a).$$

An analogous calculation applies to observation probabilities, reward functions and initial state distribution. This shows that $\tilde{f} \circ f \in \mathrm{Iso}(D, F)$.

Next, let $s', s \in \mathcal{S}^E$. Using Lemma 13 and the definition of an isomorphism, it is

$$P^E(s' \mid s, a) = P^E(ff^{-1}s' \mid ff^{-1}s, ff^{-1}a) = P^D(f^{-1}s' \mid f^{-1}s, f^{-1}a).$$

Again, an analogous calculation applies to observation probabilities and the other relevant functions. This shows that $f^{-1} \in \mathrm{Iso}(E, D)$. ☐

The next lemma shows that one can decompose isomorphisms into any isomorphism composed with an automorphism. Essentially, an isomorphism is a map from one Dec-POMDP to another, composed with a symmetry of that Dec-POMDP. This will be important later when we apply these concepts to define the LFC problem and show how OP relates to it.

Lemma 19. *Let D, E be Dec-POMDPs and $f, \tilde{f} \in \mathrm{Iso}(D, E)$. Then there exists exactly one $g \in \mathrm{Aut}(E)$ such that $g \circ f = \tilde{f}$. Analogously, there exists exactly one $g \in \mathrm{Aut}(D)$ such that $f \circ g = \tilde{f}$. In particular, it is*

$$\mathrm{Iso}(D, E) = f \circ \mathrm{Aut}(E) = \mathrm{Aut}(D) \circ f,$$

where $f \circ \mathrm{Aut}(E) := \{f \circ g \mid g \in \mathrm{Aut}(E)\}$ and $\mathrm{Aut}(D) \circ f$ is defined analogously.

Proof. Existence: By Lemma 18, $g := \tilde{f} \circ f^{-1}$ is an isomorphism in $\mathrm{Iso}(E, E) = \mathrm{Aut}(E)$, and it is $g \circ f = (\tilde{f} \circ f^{-1}) \circ f = \tilde{f}$.

Uniqueness: Assume $g \circ f = \tilde{g} \circ f = \tilde{f}$ for automorphisms $g, \tilde{g} \in \mathrm{Aut}(E)$. Then it follows that

$$g = g \circ (f \circ f^{-1}) = (g \circ f) \circ f^{-1} = (\tilde{g} \circ f) \circ f^{-1} = \tilde{g}.$$

The proof for the second part of the lemma is exactly analogous, but using $g := f^{-1}\tilde{f}$.

Turning to the "in particular" statement, the above shows that

$$\mathrm{Iso}(D, E) \subseteq f \circ \mathrm{Aut}(E)$$

and

$$\mathrm{Iso}(D, E) \subseteq \mathrm{Aut}(D) \circ f.$$

The two inclusions in the other direction follow directly from Lemma 18. ☐

C.3. Pushforward policies

Recall the definition of pushforward policies.

Definition 20 (Pushforward policy)**.** Let D, E be isomorphic Dec-POMDPs, let $f \in \mathrm{Iso}(D, E)$, and let $\pi \in \Pi^D$. Then we define the pushforward $f^*\pi \in \Pi^E$ of π by f via

$$(f^*\pi)_i(a_i \mid \tau_{i,t}) := \pi_{f^{-1}i}(f^{-1}a_i \mid f^{-1}\tau_{i,t})$$

for all $i \in \mathcal{N}^E, a_i \in \mathcal{A}_i^E, t \in \{0, \dots, T\}$, and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}^E$. That is, in the joint policy $f^*\pi$, agent $j \in \mathcal{N}^E$ gets assigned the local policy π_i of agent $i := f^{-1}j \in \mathcal{N}^D$, precomposed with f^{-1}.

One can easily see that $f^*\pi$ is a policy for the Dec-POMDP E. Hence, when f is an automorphism, $f^*\pi$ is a policy for the same Dec-POMDP as π.

Like actions of isomorphisms on joint actions and observations, the pushforward is compatible with function composition, and it can be inverted using the inverse map f^{-1}.

Lemma 21. *Let D, E and F be Dec-POMDPs. Let $\pi \in \Pi^D$, $e \in \mathrm{Aut}(D, D)$ be the identity and $f \in \mathrm{Iso}(D, E)$, $\tilde{f} \in \mathrm{Iso}(E, F)$. Then it is*

(i) $e^*\pi = \pi$.

(ii) $\tilde{f}^*(f^*\pi) = (\tilde{f} \circ f)^*\pi$.

In particular, the pushforward can be inverted using the inverse map f^{-1}.

Proof. First, let $i \in \mathcal{N}^D, a_i \in \mathcal{A}_i^D, t \in \{0, \dots, T\}$, and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}^D$. Then

$$(e^*\pi)_i(a_i \mid \tau_{i,t}) = \pi_{ei}(ea_i \mid e\tau_{i,t}) = \pi_i(a_i \mid \tau_{i,t}),$$

which shows that $e^*\pi = \pi$.

Second, let $i \in \mathcal{N}^F, a_i \in \mathcal{A}_i^F$ and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}^F$. Using Lemma 13, it is

$$\begin{aligned}(\tilde{f}^*(f^*\pi))_i(a_i \mid \tau_{i,t}) &= (f^*\pi)_{\tilde{f}^{-1}i}(\tilde{f}^{-1}a_i \mid \tilde{f}^{-1}\tau_{i,t}) = \pi_{f^{-1}(\tilde{f}^{-1}i)}(f^{-1}(\tilde{f}^{-1}a_i) \mid f^{-1}(\tilde{f}^{-1}\tau_{i,t})) \\ &\overset{\text{Lemma 13}}{=} \pi_{(f\circ\tilde{f})^{-1}i}((\tilde{f}\circ f)^{-1}a_i \mid (\tilde{f}\circ f)^{-1}\tau_{i,t}) = ((\tilde{f}\circ f)^*\pi)_i(a_i \mid \tau_{i,t}), \qquad (30)\end{aligned}$$

which proves that $\tilde{f}^*(f^*\pi) = (\tilde{f} \circ f)^*\pi$.

Regarding the invertibility of the pushforward, note that by (ii), it is

$$(f^{-1})^*(f^*\pi) \overset{\text{(ii)}}{=} (f^{-1} \circ f)^*\pi = \pi.$$

This concludes the proof. □

One may wonder how pushforward policies and isomorphisms are related. In particular, how is the distribution over histories induced by a pushforward policy related to the distribution induced by the original policy? This question is answered by the following theorem, which we will use throughout this paper. The theorem demonstrates how the isomorphism from D to E preserves the structure of the problem. It is similar to Kang & Kim (2012)'s Theorem 3, which applies to automorphisms in POSGs and makes a statement about the expected returns of all agents under the pushforward policy. We will provide a result about expected returns as a corollary.

Theorem 22. *Let D, E be isomorphic Dec-POMDPs, let $f \in \mathrm{Iso}(D, E)$, and let $\pi \in \Pi^D$. Then for any $\tau \in \mathcal{H}^D$, it is*

$$\mathbb{P}_\pi(H^D = \tau) = \mathbb{P}_{f^*\pi}(H^E = f\tau).$$

In particular, for any $i \in \mathcal{N}, t \in \{0, \dots, T\}$ and $\tau_{i,t} \in \overline{\mathcal{AO}}_i$, it is

$$\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t}) = \mathbb{P}_{f^*\pi}(\overline{AO}_{fi,t} = f\tau_{i,t}).$$

Proof. Proof by induction over $t \in \{0, \dots, T\}$.

To start the induction, let $s \in \mathcal{S}^D, a \in \mathcal{A}, r \in \mathbb{R}$ and $\tau_0 := (s, a, r) \in \mathcal{H}_0^D$. Then it is

$$\mathbb{P}_\pi(H_0^D = \tau_0) = b_0^D(s) \prod_{i \in \mathcal{N}^D} \pi_i(a_i \mid \emptyset) \delta_{\mathcal{R}^D(s,a),r} \tag{31}$$

$$= b_0^E(fs) \prod_{i \in \mathcal{N}^D} \pi_{f^{-1}fi}(f^{-1}fa_i \mid \emptyset) \delta_{\mathcal{R}^E(fs,fa),r} \tag{32}$$

$$= b_0^E(fs) \prod_{i \in \mathcal{N}^D} (f^*\pi)_{fi}(fa_i \mid \emptyset) \delta_{\mathcal{R}^E(fs,fa),r} \tag{33}$$

$$= b_0^E(fs) \prod_{i \in \mathcal{N}^E} (f^*\pi)_i(fa_{f^{-1}i} \mid \emptyset) \delta_{\mathcal{R}^E(fs,fa),r} \tag{34}$$

$$= \mathbb{P}_{f^*\pi}(H_0^E = f\tau). \tag{35}$$

Here, in (32), we use the definition of an isomorphism, in (33) we use the definition of the pushforward policy, and in (35) we use the definition of fa from Equation (21).

Next, let $t > 0$ and assume that it is $\mathbb{P}_\pi(\tau_{t-1}^D = \tau_{t-1}) = \mathbb{P}_{f^*\pi}(\tau^E = f\tau_{t-1})$ for any $\tau_{t-1} \in \mathcal{H}_{t-1}$. Let

$$\tau_t = (\dots, s_{t-1}, o_{t-1}, a_{t-1}, r_{t-1}, s_t, o_t, a_t, r_t) \in \mathcal{H}_t^D$$

arbitrary, define $a_{i,t}$ for $i \in \mathcal{N}$ such that $(a_{i,t})_{i \in \mathcal{N}^D} = a_t$, and define $\tau_{i,t}$ as the action-observation history for player $i \in \mathcal{N}^D$ corresponding to τ_t.

If $\mathbb{P}_\pi(H_{t-1}^D = \tau_{t-1}) = 0$, then also $\mathbb{P}_{f^*\pi}(H_{t-1}^E = f\tau_{t-1}) = 0$, and thus also

$$\mathbb{P}_\pi(H_t^D = \tau_t) = 0 = \mathbb{P}_{f^*\pi}(H_t^E = f\tau_t).$$

Thus, assume now that $\mathbb{P}_\pi(H^D_{t-1} = \tau_{t-1}) > 0$. Then it is

$$\mathbb{P}_\pi(S^D_t = s_t, O^D_t = o_t, A^D_t = a_T, R^D_t = r_t \mid H^D_{t-1} = \tau_{t-1}) \tag{36}$$

$$= P^D(s_t \mid s_{t-1}, a_{t-1}) O^D(o_t \mid s_t, a_{t-1}) \prod_{i \in \mathcal{N}^D} \pi_i(a_{i,t} \mid \tau_t) \delta_{\mathcal{R}^D(s_t,a_t),r} \tag{37}$$

$$= P^E(fs_t \mid fs_{t-1}, fa_{t-1}) O^E(fo_t \mid fs_t, fa_{t-1}) \tag{38}$$

$$\prod_{i \in \mathcal{N}^D} \pi_{f^{-1}fi}(f^{-1}fa_{i,t} \mid f^{-1}f\tau_t) \delta_{\mathcal{R}^E(fs_t,fa_t),r} \tag{39}$$

$$= P^E(fs_t \mid fs_{t-1}, fa_{t-1}) O^E(fo_t \mid fs_t, fa_{t-1}) \tag{40}$$

$$\prod_{i \in \mathcal{N}^D} (f^*\pi)_{fi}(fa_{i,t} \mid f\tau_t) \delta_{\mathcal{R}^E(fs_t,fa_t),r} \tag{41}$$

$$= P^E(fs_t \mid fs_{t-1}, fa_{t-1}) O^E(fo_t \mid fs_t, fa_{t-1}) \tag{42}$$

$$\prod_{i \in \mathcal{N}^E} (f^*\pi)_i(fa_{f^{-1}i,t} \mid f\tau_{f^{-1}i,t}) \delta_{\mathcal{R}^E(fs_t,fa_t),r} \tag{43}$$

$$= \mathbb{P}_{f^*\pi}(S^E_t = fs_t, O^E_t = fo_t, A^E_t = fa_t, R^E_t = r_t \mid H^E_{t-1} = f\tau_{t-1}). \tag{44}$$

Again, we have used the definitions of isomorphism, pushforward policy, and Equations (21) and (22) in lines (38), (40), and (44), respectively. Using the inductive hypothesis, it follows that

$$\mathbb{P}_\pi(H^D_t = \tau_t) \tag{45}$$

$$= \mathbb{P}_\pi(S^D_t = s_t, O^D_t = o_t, A^D_t = a_T, R^D_t = r_t \mid H^D_{t-1} = \tau_{t-1}) \mathbb{P}_\pi(H^D_{t-1} = \tau_{t-1}) \tag{46}$$

$$= \mathbb{P}_{f^*\pi}(S^E_t = fs_t, O^E_t = fo_t, A^E_t = fa_t, R^E_t = r_t \mid H^E_{t-1} = f\tau_{t-1}) \mathbb{P}_\pi(H^D_{t-1} = \tau_{t-1}) \tag{47}$$

$$\overset{\text{I.H.}}{=} \mathbb{P}_{f^*\pi}(S^E_t = fs_t, O^E_t = fo_t, A^E_t = fa_t, R^E_t = r_t \mid H^E_{t-1} = f\tau_{t-1}) \mathbb{P}_{f^*\pi}(H^E_{t-1} = f\tau_{t-1}) \tag{48}$$

$$= \mathbb{P}_{f^*\pi}(H^E_t = f\tau_t). \tag{49}$$

This concludes the induction and thus proves that $\mathbb{P}_\pi(H^D = \tau) = \mathbb{P}_{f^*\pi}(H^E = f\tau)$ for any history $\tau := \tau_T \in \mathcal{H}^D_T$.

Turning to the "in particular" part of the proposition, let $i \in \mathcal{N}, t \in \{0, \ldots, T\}$, and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$. Let $\tau_t \in \mathcal{H}^D_t$ such that $\mathbb{P}_\pi(H^D_t = \tau_t) > 0$ and thus also $\mathbb{P}_{f^*\pi}(H^E_t = f\tau_t) > 0$.

First, assume that actions and observations of agent i in this history equal those in $\tau_{i,t}$. Then it is $\{H^D_t = \tau_t\} \subseteq \{\overline{AO}^D_{i,t} = \tau_{i,t}\}$ and hence $\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t} \mid H_t = \tau_t) = 1$. Moreover, by Equations (21) and (22), it follows for any $t' \leq t, a_{t'} \in \mathcal{A}^D$ and $o_{t'} \in \mathcal{O}^D$ that $\mathrm{proj}_{fi}(fa_{t'}) = fa_{i,t'}$ and analogously $\mathrm{proj}_{fi}(fo_{t'}) = fo_{i,t'}$. Hence, it follows that actions of observations of agent fi in $f\tau_t$ are equal to those in $f\tau_{i,t}$, and thus it is $\{H^E_t = f\tau_t\} \subseteq \{\overline{AO}^E_{fi,t} = f\tau_{i,t}\}$, which implies $\mathbb{P}_{f^*\pi}(\overline{AO}^E_{fi,t} = f\tau_{i,t} \mid H_t = f\tau_t) = 1$.

If, on the other hand, a history τ_t disagrees with $\tau_{i,t}$ in any way, then trivially $\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t} \mid H_t = \tau_t) = 0$ and thus by same argument as before also $\mathbb{P}_{f^*\pi}(\overline{AO}_{fi,t} = f\tau_{i,t} \mid H_t = f\tau_t) = 0$.

It follows that for any $\tau_t \in \mathcal{H}^D_t$ such that $\mathbb{P}_\pi(H^D_t = \tau) > 0$, it is

$$\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t} \mid H_t = \tau_t) = \mathbb{P}_{f^*\pi}(\overline{AO}^E_{fi,t} = f\tau_{i,t} \mid H_t = f\tau_t), \tag{50}$$

and hence

$$\begin{aligned}
\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t}) &= \sum_{\tau_t \in \mathcal{H}_t^D} \mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t} \mid H_t = \tau_t)\mathbb{P}_\pi(H_t^D = \tau_t) && (51)\\
&\overset{(45)}{=} \sum_{\tau_t \in \mathcal{H}_t^D} \mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t} \mid H_t = \tau_t)\mathbb{P}_{f^*\pi}(H_t^D = f\tau_t) && (52)\\
&\overset{(50)}{=} \sum_{\tau_t \in \mathcal{H}_t^D} \mathbb{P}_{f^*\pi}(\overline{AO}_{fi,t} = f\tau_{i,t} \mid H_t = f\tau_t)\mathbb{P}_{f^*\pi}(H_t^D = f\tau_t) && (53)\\
&= \sum_{\tau_t \in f^{-1}(\mathcal{H}_t^D)} \mathbb{P}_{f^*\pi}(\overline{AO}_{fi,t} = f\tau_{i,t} \mid H_t = \tau_t)\mathbb{P}_{f^*\pi}(H_t^D = \tau_t) && (54)\\
&= \mathbb{P}_{f^*\pi}(\overline{AO}_{fi,t} = f\tau_{i,t}). && (55)
\end{aligned}$$

In line (55), we have used that, by Corollary 14, f is a bijective map when applied to histories, and thus $f^{-1}\left(\mathcal{H}_t^D\right) = \mathcal{H}_t^E$. This concludes the proof. □

It is an immediate corollary that the expected return of a policy is not changed by the pushforward.

Corollary 23. *Let D, E be Dec-POMDPs, let $f \in \mathrm{Iso}(D, E)$, and let $\pi \in \Pi^D$. Then*

$$J^D(\pi) = J^E(f^*\pi).$$

Proof. By Corollary 14 and Theorem 22, it is

$$\begin{aligned}
\mathbb{P}_\pi(f\left(H^D\right) = \tau^E) &\overset{\text{Corollary 14}}{=} \mathbb{P}_\pi(H^D = f^{-1}\tau^E)\\
&\overset{\text{Theorem 22}}{=} \mathbb{P}_{f^*\pi}(H^E = f(f^{-1}\tau^E)) = \mathbb{P}_{f^*\pi}(H^E = \tau^E) \qquad (56)
\end{aligned}$$

for any history $\tau^E \in \mathcal{H}^E$. This shows that the random variable $f\left(H^D\right)$ has the same image distribution under $\mathbb{P}_\pi$ as the variable H^E under $\mathbb{P}_{f^*\pi}$. In particular, this means that for any $t = 0, \ldots, T$, the variables $f(R_t^D) = R_t^D$ and R_t^E have the same distribution in the respective probability spaces (*). Using the definition of the expected return, it follows that

$$J^D(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{T} R_t^D\right] \overset{(*)}{=} \mathbb{E}_{f^*\pi}\left[\sum_{t=0}^{T} R_t^E\right] = J^E(f^*\pi). \qquad (57)$$

□

C.4. Relation to group theory

The study of symmetries is a focus of group theory, and the concepts introduced above hence correspond to group-theoretic notions. For instance, as we show below, $\mathrm{Aut}(D)$ is a group, and its elements do act on the elements of a Dec-POMDP in the sense of group actions. We discuss this here as we will need these results later, in the discussion of symmetric profiles of learning algorithms in Appendix C.7, as well as in the discussion of random tie-breaking functions in Appendix F.3. For a reference on the group-theoretic concepts discussed here, see Rotman (2012, ch. 3).

We begin by showing that $\mathrm{Aut}(D)$ is a group.

Proposition 24. *Let D be a Dec-POMDP. Then $(\mathrm{Aut}(D), \circ)$ is a group, where $\circ$ is the function composition.*

Proof. First, we show that the binary operation

$$\circ\colon \mathrm{Aut}(D) \times \mathrm{Aut}(D) \to \mathrm{Aut}(D), (g, \tilde{g}) \mapsto g \circ \tilde{g}$$

is well-defined. By Equation 27, an automorphism $g \in \mathrm{Aut}(D)$ is a bijective self-map, so we can compose any two automorphisms $g, \tilde{g} \in \mathrm{Aut}(D)$. Moreover, by Lemma 18, for any $g \in \mathrm{Aut}(D), \tilde{g} \in \mathrm{Aut}(D)$, we also have $\tilde{g} \circ g \in \mathrm{Iso}(D, D) = \mathrm{Aut}(D)$. This shows that $\mathrm{Aut}(D)$ is closed under function composition.

Second, note that $\circ$ is an associative operation as function composition is associative. Moreover, for the identity map e, it is $e \circ g = g$ for any $g \in \mathrm{Aut}(D)$, so $\mathrm{Aut}(D)$ has a neutral element. Lastly, by Lemma 18, it is also $g^{-1} \in \mathrm{Aut}(D)$, and since $g^{-1} \circ g = e$, this implies that g has an inverse in $\mathrm{Aut}(D)$. This concludes the proof. ☐

Next, we turn to group actions, which formalize the idea that elements of groups can be applied to sets. In the case of symmetry groups, this connects the abstract group elements with their role as transformation of an underlying set. For instance, consider the set X of the vertices of an equilateral triangle in $\mathbb{R}^2$ and the cyclic group $\mathbb{Z}/3\mathbb{Z}$. Each element of $\mathbb{Z}/3\mathbb{Z}$ can be regarded as a rotation of the vertices of the triangle, mapping one vertex to another.

Definition 25 (Group action)**.** Let $(G, \cdot)$ be a group with identity e and let X be any set. A group action is defined as a map $\alpha\colon G \times X \to X$ such that

(i) Identity: $\alpha(e, x) = x$ for any $x \in X$

(ii) Compatibility: $\alpha(f, \alpha(g, x)) = \alpha(f \cdot g, x)$ for any $f, g \in G$, $x \in X$.

It is common to write $gx := \alpha(g, x)$ for $g \in G$, $x \in X$, if it is clear which group action is referred to.

We have already proven these two properties for isomorphisms and both their actions on joint actions and observations, as well as the pushforward of policies, in Lemma 13 and Lemma 21, respectively. Hence, it follows that also $\mathrm{Aut}(D)$ acts on these sets in the sense of group actions.

Corollary 26. *Let D be a Dec-POMDP. The actions of $\mathrm{Aut}(D)$ on $\mathcal{A}$ and $\mathcal{O}$, defined as $\alpha_A\colon (g, a) \mapsto g_A(a)$ respectively $\alpha_O\colon (g, o) \mapsto g_O(o)$ as in Equations 21 and 22 are group actions. Similarly, the pushforward of policies by automorphisms $(g, \pi) \mapsto g^*\pi$ as defined in Definition 20 is an action of $\mathrm{Aut}(D)$ on Π^D.*

Proof. This follows directly from Lemma 13 and Lemma 21. ☐

Of course, automorphisms also act on states and agents, and one can also easily see that they act on histories and action-observation histories.

Some further results immediately follow from this, such as the fact that $\mathcal{N}$ decomposes into equivalence classes of *orbits* under $\mathrm{Aut}(D)$. The orbit of agent i is defined as the set of all agents j that can be obtained from i by applying automorphisms.

Definition 27 (Orbit)**.** Let D be a Dec-POMDP and assume that $\mathrm{Aut}(D)$ acts on the set X. Then for $x \in X$, the set

$$\mathrm{Aut}(D)x := \{gx \mid g \in \mathrm{Aut}(D)\}$$

is called the orbit of x under $\mathrm{Aut}(D)$.

For instance, for $i \in \mathcal{N}$, the set $\mathrm{Aut}(D)i := \{gi \mid g \in \mathrm{Aut}(D)\}$ is the orbit of agent i under $\mathrm{Aut}(D)$. It is a standard result from group theory that orbits form a partition $\{\mathrm{Aut}(D)i \mid i \in \mathcal{N}\} \subseteq \mathcal{P}(\mathcal{N})$ of the set. This follows from the fact that since group actions have an identity and are invertible, belonging to the same orbit is an equivalence relation. For example, in the case of the triangle in $\mathbb{R}^2$, all the vertices in V can be reached from any other vertex by rotations, so they all belong to the same orbit. In general, though, the orbits may form any other partition of the set.

C.5. Dec-POMDP labelings and relabeled Dec-POMDPs

In the following, assume that a Dec-POMDP D is given. Here, we want to define a set of isomorphic Dec-POMDPs $\mathcal{D}$ as described in Section 4.2, in which the sets of states, actions, etc. are of the form $\{1, 2, \dots, k-1, k\} \subseteq \mathbb{N}, k \in \mathbb{N}$. This set can then be used to define the LFC game for D in a way that does not depend on labels.

We begin by defining a *labeling* of D. A labeling f is a special Dec-POMDP isomorphism from D to another, relabeled Dec-POMDP, that can be constructed using f.

Definition 28 (Dec-POMDP labeling)**.** A Dec-POMDP labeling is a tuple of bijective maps

$$f := (f_N, f_S, (f_{A_i})_{i\in\mathcal{N}}, (f_{O_i})_{i\in\mathcal{N}}),$$

where

$$\begin{aligned} & f_N\colon\ \mathcal{N} \to \{1, \dots, |\mathcal{N}|\} && (58)\\ & f_S\colon\ \mathcal{S} \to \{1, \dots, |\mathcal{S}|\} && (59)\\ \forall i \in \mathcal{N}\colon\quad & f_{A_i}\colon \mathcal{A}_i \to \{1, \dots, |\mathcal{A}_i|\} && (60)\\ \forall i \in \mathcal{N}\colon\quad & f_{O_i}\colon \mathcal{O}_i \to \{1, \dots, |\mathcal{O}_i|\}. && (61) \end{aligned}$$

We denote $\mathrm{Sym}(D)$ for the set of labelings of D.

Note that if X is some set, then $\mathrm{Sym}(X)$ usually denotes the *symmetric group* of X. The symmetric group is the set of permutations of X, together with the operation of function composition. We use the same notation, as $\mathrm{Sym}(D)$ can be understood of as containing all the permutations of the different sets that D consists of, with the caveat that we first map those sets to subsets of the first k natural numbers. This is done for simplification, especially regarding the treatment of the individual action and observation sets of different agents.

Next, we introduce the pushforward Dec-POMDP f^*D for a labeling $f \in \mathrm{Sym}(D)$. This is a Dec-POMDP that is isomorphic to D, with isomorphism f. In the following, we let $f \in \mathrm{Sym}(D)$ act on joint actions, observations, etc., in the same way as before for isomorphisms. For instance, for $a \in \mathcal{A}$, it is $fa := (f_{A_{f_N^{-1}(i)}}(a_{f_N^{-1}(i)}))_{i\in\{1,\dots,|\mathcal{N}|\}}$. The compatibility of these actions with function composition and function inversion trivially still hold.

Definition 29 (Relabeled Dec-POMDP)**.** Let $f \in \mathrm{Sym}(D)$. Let $N \in \mathbb{N}$ such that $\{1, \dots, N\} = \mathcal{N}$. The pushforward of D by f, called a relabeled Dec-POMDP, is the Dec-POMDP

$$f^*D := (\hat{\mathcal{N}}, \hat{\mathcal{S}}, \hat{\mathcal{A}}, \hat{P}, \hat{\mathcal{R}}, \hat{\mathcal{O}}, \hat{O}, \hat{b}_0, \hat{T}),$$

where

- $\hat{\mathcal{N}} := \{1, \ldots, N\} = \mathcal{N}$.
- $\hat{\mathcal{S}} := \{1, \ldots, |\mathcal{S}|\}$.
- $\hat{\mathcal{A}}_i := \{1, \ldots, |\mathcal{A}_{f^{-1}i}|\}$ for $i \in \hat{\mathcal{N}}$.
- $\hat{P}(s' \mid s, a) := P(f^{-1}s' \mid f^{-1}s, f^{-1}a)$ for $s', s \in \hat{\mathcal{S}}, a \in \hat{\mathcal{A}}$.
- $\hat{\mathcal{R}}(s, a) := \mathcal{R}(f^{-1}s, f^{-1}a)$ for $s \in \hat{\mathcal{S}}, a \in \hat{\mathcal{A}}$.
- $\hat{\mathcal{O}}_i := \{1, \ldots, |\mathcal{O}_{f^{-1}i}|\}$ for $i \in \hat{\mathcal{N}}$.
- $\hat{O}(o \mid s, a) := O(f^{-1}o \mid f^{-1}s, f^{-1}a)$ for $o \in \hat{\mathcal{O}}, s \in \hat{\mathcal{S}}, a \in \hat{\mathcal{A}}$.
- $\hat{b}_0(s) := b_0(f^{-1}s)$ for $s \in \hat{\mathcal{S}}$.
- $\hat{T} := T$.

First, we have to check that this is well-defined, e.g., that $f^{-1}a \in \mathcal{A}$ for any $a \in \hat{\mathcal{A}}$. For states, it is clear from the definition of $\mathrm{Sym}(D)$ that $f^{-1}(\hat{\mathcal{S}}) = f^{-1}(\{1, \ldots, |\mathcal{S}|\}) = \mathcal{S}$. Moreover, the same applies to agents, i.e., $f^{-1}i \in \mathcal{N}$ for any $i \in \{1, \ldots, |\mathcal{N}|\}$. This leaves joint actions and observations.

Proposition 30. *For $\hat{\mathcal{A}}, \hat{\mathcal{O}}$ as defined above, it is $f^{-1}(\hat{\mathcal{A}}) = \mathcal{A}$ and $f^{-1}(\hat{\mathcal{O}}) = \mathcal{O}$.*

Proof. Let $\hat{a} \in \hat{\mathcal{A}}$. Then for any $i \in \hat{\mathcal{N}}$, we can define $j \in \mathcal{N}$ and $a_j \in \mathcal{A}_j$ such that $fj = i$ and $fa_j = \hat{a}_i$. Then

$$f^{-1}\hat{a} = (f^{-1}\hat{a}_{fj}))_{j\in\mathcal{N}} = (f^{-1}fa_j))_{j\in\mathcal{N}} = (a_j)_{j\in\mathcal{N}} = a \in \mathcal{A}.$$

The same argument works for $\hat{o} \in \hat{\mathcal{O}}$. □

Importantly, it can be $f^*D \neq D$ for a labeling $f \in \mathrm{Sym}(D)$. Nevertheless, it is easy to see from the definitions that f^*D is actually isomorphic to D, with isomorphism f. This also implies that it is $f^*D = D$ if and only if f is an automorphism.

Lemma 31. *For any $f \in \mathrm{Sym}(D)$, it is $f \in \mathrm{Iso}(D, f^*D)$.*

Proof. This follows directly from the definition of an isomorphism, together with Lemma 13. For instance, considering transition probabilities, it is

$$P(s' \mid s, a) = P(f^{-1}fs' \mid f^{-1}fs, f^{-1}fa) = \hat{P}(fs' \mid fs \mid fa)$$

for any $s', s \in \mathcal{S}, a \in \mathcal{A}$. Similar calculations apply to all the other relevant functions. □

It follows as a corollary that $f^*\pi$ is a policy for the Dec-POMDP f^*D, where $\pi \in \Pi^D$, $f \in \mathrm{Sym}(D)$. Note also that the results in Lemma 21 still apply to the pushforward by labelings.

Corollary 32. *Let $f \in \mathrm{Sym}(D)$ and $\pi \in \Pi^D$. Then $f^*\pi \in \Pi^{f^*D}$.*

Proof. Follows from the definition of $f^*\pi$ and Lemma 31. □

Lastly, we provide some further useful results about labelings. First, the set $\mathrm{Sym}(D)$ already contains all the isomorphisms in $\mathrm{Iso}(D, f^*D)$. We will need this result later to relate results about isomorphisms and automorphisms to the relabeled Dec-POMDPs used in an LFC game.

Lemma 33. *Let $f \in \mathrm{Sym}(D)$. Then*

$$\mathrm{Iso}(D, f^*D) = \{\tilde{f} \in \mathrm{Sym}(D) \mid \tilde{f}^*D = f^*D\}$$

Proof. "$\supseteq$": for any $\tilde{f} \in \mathrm{Sym}(D)$ such that $\tilde{f}^*D = f^*D$, it is also $\mathrm{Iso}(D, \tilde{f}^*D) = \mathrm{Iso}(D, f^*D)$, and thus it follows from Lemma 31 that $\tilde{f} \in \mathrm{Iso}(D, \tilde{f}^*D) = \mathrm{Iso}(D, f^*D)$.

"$\subseteq$": Let $\hat{f} \in \mathrm{Iso}(D, f^*D)$. Note that the set of agents, states, and the individual action and observation sets in f^*D are all of the form $\{1, \ldots, k\}$ where $k \in \mathbb{N}$ depends on the respective set. Now consider, for instance, the map $\hat{f}_{A_i}$ for $i \in \mathcal{N}$. Then by the definition of an isomorphism, $\hat{f}_{A_i}$ is a bijective map, and its domain and codomain are $\mathcal{A}_i$ and $\{1, \ldots, k\}$ for some $k \in \mathbb{N}$. Moreover, since $\hat{f}_{A_i}$ is bijective, it must be $k = |\mathcal{A}_i|$. *Mutatis mutandis*, the same applies to all of the other maps that are part of the tuple $\hat{f}$. Hence, $\hat{f}$ satisfies the definition of a Dec-POMDP labeling, so $\hat{f} \in \mathrm{Sym}(D)$.

Next, it follows that $\hat{f} \in \mathrm{Iso}(D, \hat{f}^*D)$ by Lemma 31, and thus $e = \hat{f} \circ \hat{f}^{-1} \in \mathrm{Iso}(f^*D, \hat{f}^*D)$ by the assumption and Lemma 18. Hence, using the definition of an isomorphism, it follows that also $f^*D = \hat{f}^*D$. This shows that

$$\hat{f} \in \{\tilde{f} \in \mathrm{Sym}(D) \mid \tilde{f}^*D = f^*D\},$$

which concludes the proof. □

Second, we show that labelings and pushforward are compatible with composition with isomorphisms.

Lemma 34. *Let D, E be isomorphic Dec-POMDPs with $f \in \mathrm{Iso}(D, E)$. Then*

$$\mathrm{Sym}(D) = \mathrm{Sym}(E) \circ f.$$

Moreover, it is $\tilde{f}^*E = (\tilde{f} \circ f)^*D$ for any $\tilde{f} \in \mathrm{Sym}(E)$.

Proof. First, let $\tilde{f} \in \mathrm{Sym}(D)$ and define $\hat{f} := \tilde{f} \circ f^{-1}$. Note that $\hat{f}$ has as components bijective maps with a domain and codomain that satisfies the definition of a labeling of E. Hence, $\hat{f} \in \mathrm{Sym}(E)$. Next, let $\tilde{f} \in \mathrm{Sym}(E)$. Then similarly, $\tilde{f} \circ f$ fulfills the requirements for a labeling in $\mathrm{Sym}(D)$.

To prove the second statement, let again $\tilde{f} \in \mathrm{Sym}(E)$. Note that since D and E are isomorphic, they must have the same set of players and sets of states with the same cardinalities. Now let $i \in \mathcal{N}^D$. Using the definition of an isomorphism and of a labeling, it is then

$$\mathcal{A}_i^{\tilde{f}^*E} = \{1, \ldots, |\mathcal{A}^E_{\tilde{f}^{-1}i}|\} = \{1, \ldots, |\mathcal{A}^D_{f^{-1}(\tilde{f}^{-1}i)}|\} = \{1, \ldots, |\mathcal{A}^D_{(\tilde{f}\circ f)^{-1}i}|\} = \mathcal{A}_i^{(\tilde{f}\circ f)^*D}.$$

A similar argument applies to the sets $\mathcal{O}_i$ for $i \in \mathcal{N}$. Finally, let $s, s' \in \mathcal{S}^{\tilde{f}^*E}, a \in \mathcal{A}^{\tilde{f}^*E}$. Using again the definition of an isomorphism and a labeling, it follows that

$$\begin{aligned} P^{\tilde{f}^*E}(s' \mid s, a) &= P^E(\tilde{f}^{-1}s' \mid \tilde{f}^{-1}s, \tilde{f}^{-1}a) \\ &= P^D(f^{-1}\tilde{f}^{-1}s' \mid f^{-1}\tilde{f}^{-1}s, f^{-1}\tilde{f}^{-1}a) = P^{(\tilde{f}\circ f)^*D}(s' \mid s, a). \end{aligned} \tag{62}$$

Again, an analogous argument applies to the observation probability kernel and reward function, as well as the initial state distribution. This concludes the proof. □

C.6. The label-free coordination game and problem

Here, we recall the definitions of an LFC game and of the LFC problem. To begin, we define a measure space of policies and recall the definition of a learning algorithm. For any Dec-POMDP D, let $\Delta(\Pi^D)$ be the set of measures on the space $(\Pi^D, \mathcal{F}^D)$ where $\mathcal{F}^D := \otimes_{i\in\mathcal{N}}\mathcal{F}_i^D$ is a product σ-Algebra and $\mathcal{F}_i^D \subseteq \mathcal{P}(\Pi_i^D)$ are σ-Algebras that make the random variables and sets discussed in this paper measurable. For instance, for $i \in \mathcal{N}$, this could be the Borel σ-Algebra with respect to the standard topology on Π_i^D that comes from regarding Π_i^D as a subset of $[0,1]^{\mathcal{A}_i^D\times\overline{\mathcal{AO}}_i^D}$. Although we do not investigate this here, all the relevant functions and sets should be measurable in that sense.

Definition 35 (Learning algorithm)**.** Let $\mathcal{D}$ be a finite set of Dec-POMDPs. A learning algorithm for $\mathcal{D}$ is a map

$$\sigma\colon \mathcal{D} \to \bigcup_{D\in\mathcal{D}} \Delta(\Pi^D)$$

such that $\sigma(D) \in \Delta(\Pi^D)$ for all $D \in \mathcal{D}$. We write $\Sigma^{\mathcal{D}}$ for the set of learning algorithms for $\mathcal{D}$.

Note that this definition is general enough so as to include planning algorithms that construct a policy directly from the environment dynamics, instead of incrementally updating a policy from experience. Nevertheless, here, we imagine that $\sigma(D)$ is a policy that was trained by an RL algorithm, using a simulator of D. Note also that a learning algorithm can learn different joint policies in different training runs, which we formalize as outputting a measure over joint policies.

Similarly to the case of policies, for a distribution $\nu \in \Delta(\Pi^D)$ and an isomorphism $f \in \mathrm{Iso}(D,E)$, we can define a pushforward distribution $f^*\nu := \nu \circ (f^*)^{-1} \in \Delta(\Pi^E)$, which is the image measure of ν under f^*. It is apparent that for two isomorphisms $f \in \mathrm{Iso}(D,E), \tilde{f} \in \mathrm{Iso}(E,F)$, it is $\tilde{f}^*(f^*\nu) = (\tilde{f}\circ f)^*\nu$.

In the following, for some distributions $\nu^{(i)} \in \Delta(\Pi^D)$ for $i \in \mathcal{N}$ and bounded measurable function $\eta\colon \Pi^D \times \dots \times \Pi^D \to \mathbb{R}$, we will use the notational shorthands

$$\mathbb{E}_{\pi^{(i)}\sim\nu^{(i)},\, i\in\mathcal{N}}\left[\eta(\pi^{(1)},\dots,\pi^{(N)})\right] := \mathbb{E}_{\pi^{(1)}\sim\nu^{(1)}}\left[\dots\left[\mathbb{E}_{\pi^{(N)}\sim\nu^{(N)}}\left[\eta(\pi^{(1)},\dots,\pi^{(N)})\right]\right]\dots\right]$$

and

$$\mathbb{E}_{\pi^{(i)}\sim\nu^{(i)}}\left[\eta(\pi^{(1)},\dots,\pi^{(N)})\right] := \int_{\Pi^D} \eta(\pi^{(1)},\dots,\pi^{(N)})\mathrm{d}\nu^{(i)}(\pi^{(i)}).$$

Note that by Fubini's theorem (see Williams, 1991, ch. 8), it is

$$\mathbb{E}_{\pi^{(i)}\sim\nu^{(i)},\, i\in\mathcal{N}}\left[\eta(\pi^{(1)},\dots,\pi^{(N)})\right] = \int_{\Pi^D\times\dots\times\Pi^D} \eta(\pi^{(1)},\dots,\pi^{(N)})\mathrm{d}\otimes_{i\in\mathcal{N}} \nu^{(i)}.$$

Now we define the LFC game for a Dec-POMDP.

Definition 36 (Label-free coordination game)**.** Let D be a Dec-POMDP and define $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$. The *label-free coordination (LFC) game* for D is defined as a tuple $\Gamma^D := (\mathcal{N}^D, (\Sigma^{\mathcal{D}})_{i\in\mathcal{N}}, (U^D)_{i\in\mathcal{N}})$ where

- $\mathcal{N}^D$ is the set of players, called principals.
- $\Sigma^{\mathcal{D}}$ is the set of strategies for all principals $i \in \mathcal{N}^D$.

- the common payoff for the strategy profile $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N \in \Sigma^D$ is

$$U^D(\boldsymbol{\sigma}) := \mathbb{E}_{D_i \sim \mathcal{U}(\mathcal{D}),\, i \in \mathcal{N}}\Big[\mathbb{E}_{f_j \sim \mathcal{U}(\mathrm{Iso}(D_j, D)),\, j \in \mathcal{N}}\Big[\mathbb{E}_{\pi^{(k)} \sim f_k^* \boldsymbol{\sigma}_k(D_k),\, k \in \mathcal{N}}\Big[J^D\big((\pi_l^{(l)})_{l \in \mathcal{N}}\big)\Big]\Big]\Big], \quad (63)$$

where $\mathcal{U}(\mathcal{D})$ is a uniform distribution over $\mathcal{D}$ and $\mathcal{U}(\mathrm{Iso}(D_j, D))$ a uniform distribution over $\mathrm{Iso}(D_j, D)$.

Remark 37. Note that set of strategies Σ^D is continuous. The game Γ^D could hence be considered a continuous game, which is a generalization of the concept of a normal-form game to continuous strategy spaces (see Glicksberg, 1952). In continuous games, it is usually assumed that the set of strategies is compact and that the payoffs are continuous functions, which we believe does apply in our case.

Moreover, we believe that there is some other normal-form game with finite strategy space such that the *mixed strategies* in that game correspond to the set of strategies $\Sigma^{\mathcal{D}}$ in Γ^D (for a reference on these concepts from game theory, see Osborne & Rubinstein, 1994; Gibbons, 1992). In particular, one can see that the set of strategies $\Sigma^{\mathcal{D}}$ is already convex.

We do not need any further characterization of an LFC game in the following, so we do not investigate issues such as compactness or convexity of the set of strategies. The formalism at hand was chosen primarily to work well with an intuitive formulation of the LFC problem and to suit our discussion of the OP algorithm.

Next, to recall the definition of the LFC problem, let any set $\mathcal{C}$ of Dec-POMDPs be given, and denote $\overline{\mathcal{C}} := \bigcup_{D \in \mathcal{C}} \mathcal{D}^D$ where $\mathcal{D}^D := \{f^* D \mid f \in \mathrm{Sym}(D)\}$ is the set of all relabeled problems of D. The LFC problem for $\mathcal{C}$ is then defined as the problem of finding one learning algorithm $\sigma \in \Sigma^{\overline{\mathcal{C}}}$ to be used by principals in a randomly drawn game Γ^D for $D \sim \mathcal{U}(\mathcal{C})$.

Definition 38 (Label-free coordination problem)**.** Let $\mathcal{C}$ be any set of Dec-POMDPs. Define the objective $U^{\mathcal{C}} : \Sigma^{\overline{\mathcal{C}}} \to \mathbb{R}$ via

$$U^{\mathcal{C}}(\sigma) := \mathbb{E}_{E \sim \mathcal{U}(\mathcal{C})}\left[U^E(\sigma, \ldots, \sigma)\right] \quad (64)$$

for $\sigma \in \Sigma^{\overline{\mathcal{C}}}$. Then we define the *Label-free coordination (LFC) problem* for $\mathcal{C}$ as the optimization problem

$$\max_{\sigma \in \Sigma^{\overline{\mathcal{C}}}} U^{\mathcal{C}}(\sigma) \quad (65)$$

and we call $U^{\mathcal{C}}(\sigma)$ the value of σ in the LFC problem for $\mathcal{C}$. If $\mathcal{C} = \{D\}$, we write $U^D := U^{\{D\}}$ in a slight abuse of notation and refer to this as the LFC problem for E.

Remark 39. The aim of the LFC problem is to find a general learning algorithm to recommended to principals in any LFC game. For this reason, we defined the problem here for a distribution over LFC games. However, a learning algorithm is optimal in the problem for a set of Dec-POMDPs if and only if it is optimal in the problem for each Dec-POMDP in that set. That is because, as one can easily see, the sets $\mathcal{D}^D, \mathcal{D}^E$ do never overlap for two non-isomorphic Dec-POMDPs D, E, and we will show in Corollary 43 that the LFC problems for two isomorphic Dec-POMDPs are identical. So to evaluate a learning algorithm in the LFC problem for a set of Dec-POMDPs, we can decompose the set into equivalence classes of isomorphic Dec-POMDPs and evaluate the learning algorithm separately for each of these classes. In the following, we will thus simplify our analysis and restrict ourselves entirely to problems defined for single Dec-POMDPs. Note that the objective in the LFC problem is then simply

$$U^D(\sigma) = U^{\{D\}}(\sigma) = \mathbb{E}_{E \sim \mathcal{U}(\{D\})}\left[U^E(\sigma, \ldots, \sigma)\right] = U^D(\sigma, \ldots, \sigma). \quad (66)$$

If we then prove, e.g., that a learning algorithm is optimal in any such problem, it follows that it is also optimal for the problem defined for any set of Dec-POMDPs.

Now we will provide two different expressions of the payoff in an LFC game. To that end, let D be a Dec-POMDP and define $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$. First, we provide an expression in terms of labelings.

Lemma 40. *Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$. Then*

$$U^{\mathcal{D}}(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N) = \mathbb{E}_{\mathbf{f} \sim \mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})} \left[\mathbb{E}_{\pi^{(i)} \sim (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i (\mathbf{f}_i^* D),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \right] \tag{67}$$

Proof. It follows from Lemma 33 that it is $\mathrm{Sym}(D) = \bigcup_{E \in \mathcal{D}} \mathrm{Iso}(D, E)$, where one can easily see that the union is disjoint (i). Moreover, it follows from Lemma 19 that $|\mathrm{Iso}(D, E)| = |\mathrm{Iso}(D, F)|$ for any Dec-POMDPs $E, F \in \mathcal{D}$, so there exists $M \in \mathbb{N}$ such that $M = |\mathrm{Iso}(D, E)|$ for any $E \in \mathcal{D}$, and from (i) it follows that $|\mathrm{Sym}(E)| = |\mathcal{D}||M|$ (ii).

Next, by Lemma 18, for any $f \in \mathrm{Iso}(D, D_j)$, it is $f^{-1} \in \mathrm{Iso}(D_j, D)$ for any $f \in \mathrm{Iso}(D_j, D)$. In addition, by the same Lemma, for $f \in \mathrm{Iso}(D_j, D)$, it is $f = (f^{-1})^{-1}$ and $f^{-1} \in \mathrm{Iso}(D, D_j)$ and thus $f \in \{\tilde{f}^{-1} \mid \tilde{f} \in \mathrm{Iso}(D, D_j)\}$. Hence, it is

$$\{\tilde{f} \mid \tilde{f} \in \mathrm{Iso}(D_j, D)\} = \{\tilde{f}^{-1} \mid \tilde{f} \in \mathrm{Iso}(D, D_j)\}. \tag{68}$$

Using the above, it follows that

$$U^{\mathcal{D}}(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N) \tag{69}$$
$$= \mathbb{E}_{D_i \sim \mathcal{U}(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\mathbf{f}_j \sim \mathrm{Iso}(D_j, D),\, j \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(k)} \sim \mathbf{f}_k^* \boldsymbol{\sigma}_k (D_k),\, k \in \mathcal{N}} \left[J^D((\pi_l^{(l)})_{l \in \mathcal{N}}) \right] \right] \right] \tag{70}$$
$$\overset{(68)}{=} \mathbb{E}_{D_i \sim \mathcal{U}(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\mathbf{f}_j \in \mathrm{Iso}(D, D_j),\, j \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(k)} \sim (\mathbf{f}_k^{-1})^* \boldsymbol{\sigma}_k (D_k),\, k \in \mathcal{N}} \left[J^D((\pi_l^{(l)})_{l \in \mathcal{N}}) \right] \right] \right] \tag{71}$$
$$\overset{\text{(i), (ii)}}{=} \mathbb{E}_{\mathbf{f}_i \sim \mathcal{U}(\mathrm{Sym}(E)),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim (\mathbf{f}_j^{-1})^* \boldsymbol{\sigma}_j (\mathbf{f}_j^* D),\, j \in \mathcal{N}} \left[J^D((\pi_k^{(k)})_{k \in \mathcal{N}}) \right] \right] \tag{72}$$
$$= \mathbb{E}_{\mathbf{f} \sim \mathcal{U}(\mathrm{Sym}(E)^{\mathcal{N}})} \left[\mathbb{E}_{\pi^{(i)} \sim (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i (\mathbf{f}_i^* D),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \right]. \tag{73}$$

This concludes the proof. □

Second, we can prove a useful decomposition of the payoff in an LFC game into isomorphisms and automorphisms. We can already see here the connection to the OP objective (we will recall the OP objective in Appendix D.1). Recall the projection operator, $\mathrm{proj}_i(x) := x_i$ for $x = (x_i)_{i \in \mathcal{N}}$.

Lemma 41. *Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$. For any $E, F \in \mathcal{D}$, choose $f_{E,F} \in \mathrm{Iso}(E, F)$ arbitrarily. Then*

$$U^{\mathcal{D}}(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N) \tag{74}$$
$$= \mathbb{E}_{D_i \sim \mathcal{U}(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f_{D_j, D}^* \boldsymbol{\sigma}_j (D_j),\, j \in \mathcal{N}} \left[\mathbb{E}_{\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}} \left[J^D \left(\left(\mathrm{proj}_k (\mathbf{g}_k^* \pi^{(k)}) \right)_{k \in \mathcal{N}} \right) \right] \right] \right]. \tag{75}$$

Proof. Let $i \in \mathcal{N}, D_i \in \mathcal{D}$. We know from Lemma 19 that for any $f \in \mathrm{Iso}(D_i, D)$ there is a unique $g \in \mathrm{Aut}(D)$ such that $f = g \circ f_{D_i, D}$. Also, for the pushforward measure, it is $(g \circ f_{D_i, D})^* \boldsymbol{\sigma}_i (D_i) =$

$g^* f^*_{D_i,D}\boldsymbol{\sigma}_i(D_i)$ for any $g \in \mathrm{Aut}(D)$. Using this, it follows that

$$\mathbb{E}_{D_i\sim\mathcal{U}(\mathcal{D}),\, i\in\mathcal{N}}\left[\mathbb{E}_{\mathbf{f}_j\sim\mathcal{U}(\mathrm{Iso}(D_j,D)),\, j\in\mathcal{N}}\left[\mathbb{E}_{\pi^{(k)}\sim\mathbf{f}^*_k\boldsymbol{\sigma}_k(D_k),\, k\in\mathcal{N}}\left[J^D((\pi^{(l)}_l)_{l\in\mathcal{N}})\right]\right]\right] \tag{76}$$

$$= \mathbb{E}_{D_i\sim\mathcal{U}(\mathcal{D}),\, i\in\mathcal{N}}\left[\mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(k)}\sim\mathbf{g}^*_k f^*_{D_k,D}\boldsymbol{\sigma}_k(D_k),\, k\in\mathcal{N}}\left[J^D((\pi^{(j)}_j)_{j\in\mathcal{N}})\right]\right]\right] \tag{77}$$

$$= \mathbb{E}_{D_i\sim\mathcal{U}(\mathcal{D}),\, i\in\mathcal{N}}\left[\mathbb{E}_{\pi^{(k)}\sim f^*_{D_k,D}\boldsymbol{\sigma}_k(D_k),\, k\in\mathcal{N}}\left[\mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J^D\left(\left(\mathrm{proj}_j(\mathbf{g}^*_j\pi^{(j)})\right)_{j\in\mathcal{N}}\right)\right]\right]\right], \tag{78}$$

where we have used a change of variables for pushforward measures in the last line. □

Finally, we can show that the payoff in an LFC game is equal for isomorphic Dec-POMDPs, up to a possible permutation of principals. Intuitively, this means that the game does not depend on labels for the problem.

Theorem 42. *Let D, E be isomorphic and $f \in \mathrm{Iso}(D,E)$ arbitrary. Define $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$ and $\mathcal{C} := \{f^*E \mid f \in \mathrm{Sym}(E)\}$. Then $\mathcal{D} = \mathcal{C}$, and for any profile of algorithms $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_N) \in \Sigma^{\mathcal{D}}$, it is*

$$U^D(\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_N) = U^E(\boldsymbol{\sigma}_{f^{-1}1},\dots,\boldsymbol{\sigma}_{f^{-1}N}).$$

Proof. First, note that by Lemma 34, it is

$$\mathcal{C} = \{\tilde f^*E \mid \tilde f \in \mathrm{Sym}(E)\} \overset{\text{Lemma 34}}{=} \{(\tilde f \circ f)^*D \mid \tilde f \in \mathrm{Sym}(E)\}$$
$$= \{\hat f^*D \mid \hat f \in \mathrm{Sym}(E)\circ f\} \overset{\text{Lemma 34}}{=} \{\hat f^*D \mid \hat f \in \mathrm{Sym}(D)\} = \mathcal{D}. \tag{79}$$

Now let $\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$ arbitrary. Then, using the expression of U^D from Lemma 40, it is

$$U^D(\boldsymbol{\sigma}_1,\dots,\boldsymbol{\sigma}_N) \tag{80}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^D((\pi^{(j)}_j)_{j\in\mathcal{N}})\right]\right] \tag{81}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^E(f^*(\pi^{(j)}_j)_{j\in\mathcal{N}})\right]\right] \tag{82}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^E((\pi^{(f^{-1}j)}_{f^{-1}j}(f^{-1}\cdot \mid f^{-1}\cdot))_{j\in\mathcal{N}})\right]\right] \tag{83}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(f^*\pi^{(f^{-1}j)}))_{j\in\mathcal{N}})\right]\right] \tag{84}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim f^*(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(\pi^{(f^{-1}j)}))_{j\in\mathcal{N}})\right]\right] \tag{85}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim((\mathbf{f}_i\circ f^{-1})^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iD),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(\pi^{(f^{-1}j)}))_{j\in\mathcal{N}})\right]\right] \tag{86}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim((\mathbf{f}_i\circ f^{-1})^{-1})^*\boldsymbol{\sigma}_i((\mathbf{f}_i\circ f^{-1})^*E),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(\pi^{(f^{-1}j)}))_{j\in\mathcal{N}})\right]\right] \tag{87}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(E)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}^*_iE),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(\pi^{(f^{-1}j)}))_{j\in\mathcal{N}})\right]\right] \tag{88}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(E)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_{f^{-1}i}^{-1})^*\boldsymbol{\sigma}_{f^{-1}i}(\mathbf{f}^*_{f^{-1}i}E),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_j(\pi^{(j)}))_{j\in\mathcal{N}})\right]\right] \tag{89}$$

$$= \mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(E)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(i)}\sim(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_{f^{-1}i}(\mathbf{f}^*_iE),\, i\in\mathcal{N}}\left[J^E((\mathrm{proj}_i(\pi^{(i)}))_{i\in\mathcal{N}})\right]\right] \tag{90}$$

$$= U^E(\boldsymbol{\sigma}_{f^{-1}1},\dots,\boldsymbol{\sigma}_{f^{-1}N}). \tag{91}$$

Here, in (82), we use Theorem 22; in (83) and (84), we use the definition of the pushforward policy; in (85), we apply a change of variables for pushforward measures, applied to each of the measures $(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}_i^*D)$ separately; in (86), we use the associativity of the pushforward measure as well as the fact that $(\mathbf{f}_i \circ f^{-1})^{-1} = f \circ \mathbf{f}_i^{-1}$; in (87) and in (88), we use the second respectively first part of Lemma 34; in (89), we again use a change of variables for pushforward measures, this time applied to the joint measure $\otimes_{i\in\mathcal{N}}(\mathbf{f}_i^{-1})^*\boldsymbol{\sigma}_i(\mathbf{f}_i^*D)$; and in (90) we use the symmetry of the set $\mathrm{Sym}(E)^{\mathcal{N}}$ with respect to player permutations, concluding the proof. □

First, this result implies that the LFC problems for two isomorphic problems are identical.

Corollary 43. *Let D, E be two isomorphic Dec-POMDPs and let $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$ and $\mathcal{C} := \{f^*E \mid f \in \mathrm{Sym}(E)\}$. Then $\mathcal{D} = \mathcal{C}$ and $U^D(\sigma) = U^E(\sigma)$ for any $\sigma \in \Sigma^{\mathcal{D}} = \Sigma^{\mathcal{C}}$.*

Proof. By Theorem 42, we have $\mathcal{D} = \mathcal{C}$. Now let $\sigma \in \Sigma^{\mathcal{D}} = \Sigma^{\mathcal{C}}$. Then again by Theorem 42, it is

$$U^D(\sigma) = U^D(\sigma, \dots, \sigma) \overset{\text{Theorem 42}}{=} U^E(\sigma, \dots, \sigma) = U^E(\sigma).$$

□

Second, the theorem shows that symmetries between the agents in a Dec-POMDP are also symmetries between principals in Γ^D.

Corollary 44. *Let D be a Dec-POMDP. Then it is*

$$U^D(\sigma_1, \dots, \sigma_N) = U^D(\sigma_{g^{-1}1}, \dots, \sigma_{g^{-1}N}).$$

for any $g \in \mathrm{Aut}(D)$.

Proof. This follows from Theorem 42, using that $\mathrm{Aut}(D) = \mathrm{Iso}(D, D)$. □

C.7. Optimal symmetric strategy profiles

Above, we have shown that the payoff in an LFC game is invariant with respect to symmetries of the agents in D. Similarly to the case of Dec-POMDPs and their symmetries, we can also apply the concept of symmetry to profiles of learning algorithms. We can then ask whether a profile of learning algorithms is invariant to symmetries of the principals, in which case we say that the profile is symmetric. In the following, we will show that optimal profiles among the ones that are symmetric are Nash equilibria of an LFC game. Since we will show in Appendix F that a profile in which all principals choose OP with tie-breaking is an optimal symmetric profile, it will result as a corollary that all principals using OP with tie-breaking is a Nash equilibrium of the game.

To begin, we define symmetric principals and profiles of learning algorithms. In the following, let again D be a Dec-POMDP and $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$.

Definition 45 (Symmetric principals and strategy profiles)**.** We say that two principals $i, j \in \mathcal{N}$ are symmetric if there exists an automorphism $g \in \mathrm{Aut}(D)$ such that $i = gj$. A profile of learning algorithms $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^D$ is called symmetric if it is $\sigma_i = \sigma_{g^{-1}i}$ for any automorphism $g \in \mathrm{Aut}(D)$ and principal $i \in \mathcal{N}$.

An optimal symmetric profile is then defined as a symmetric profile $\boldsymbol{\sigma}$ such that for all other symmetric profiles $\tilde{\boldsymbol{\sigma}}$, it is $U^D(\boldsymbol{\sigma}) \geq U^D(\tilde{\boldsymbol{\sigma}})$. Note that if there are non-symmetric principals in $\mathcal{N}$, then for a single learning algorithm $\sigma \in \Sigma^{\mathcal{D}}$, the property that $\sigma, \dots, \sigma$ is an optimal symmetric profile in the LFC *game* for D is stronger than the property that the algorithm σ is optimal in the LFC *problem* for D. The latter only requires that $U^D(\sigma, \dots, \sigma) \geq U^D(\sigma', \dots, \sigma')$ for all $\sigma' \in \Sigma^{\mathcal{D}}$, while $\sigma, \dots, \sigma$ being an optimal symmetric profile means that $U^D(\sigma, \dots, \sigma) \geq U^D(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N)$ for all symmetric profiles $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$, where a profile could potentially include different learning algorithms for non-symmetric principals.

For a simple characterization of symmetric profiles, consider the orbit $\mathrm{Aut}(D)i$ of a principal $i \in \mathcal{N}$, as defined in Appendix C.4. Clearly, saying that a profile is symmetric can equivalently be expressed as saying that all principals from the same orbit are assigned the same learning algorithm.

Lemma 46. *A profile $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^D$ is symmetric if and only if it is $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_j$ for any two symmetric principals $i, j \in \mathcal{N}$.*

Proof. This can be easily seen from the definition of the orbit and the properties of actions of automorphisms on agents and principals. ☐

Lastly, we define a Nash equilibrium of an LFC game.

Definition 47 (Nash equilibrium of an LFC game)**.** A profile of learning algorithms $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$ is a Nash equilibrium of the LFC game for D if, for any principal $i \in \mathcal{N}$ and learning algorithm $\boldsymbol{\sigma}'_i \in \Sigma^{\mathcal{D}}$, it is

$$U^D(\boldsymbol{\sigma}) \geq U^D(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{-i}).$$

Now we show that any optimal symmetric profile is a Nash equilibrium. An analogous result for normal-form games was proven in Emmons et al. (2021). Our proof closely follows that proof, adapted to our setting.

Theorem 48. *Any optimal symmetric strategy profile in an LFC game is a Nash equilibrium.*

Proof. In the following, fix a Dec-POMDP D and let $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$. Let $U := U^D$. Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N$ be an optimal symmetric strategy profile. Towards a contradiction, assume that $\boldsymbol{\sigma}$ is not a Nash equilibrium, i.e., that there is $i \in \mathcal{N}$ and $\tilde{\boldsymbol{\sigma}}_i \in \Sigma^{\mathcal{D}}$ such that $U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) > U(\boldsymbol{\sigma})$. We show that then there is another symmetric strategy profile $\hat{\boldsymbol{\sigma}}$ that achieves a higher payoff than $\boldsymbol{\sigma}$, $U(\hat{\boldsymbol{\sigma}}) > U(\boldsymbol{\sigma})$, contradicting the assumption that $\boldsymbol{\sigma}$ was optimal among the symmetric profiles.

To that end, for arbitrary $p \in (0, 1]$ define the profile $\hat{\boldsymbol{\sigma}}_j := p\tilde{\boldsymbol{\sigma}}_i + (1-p)\boldsymbol{\sigma}_i$ for any $j \in \mathrm{Aut}(D)i$ and $\hat{\boldsymbol{\sigma}}_j := \boldsymbol{\sigma}_j$ for $j \in \mathcal{N} \setminus \mathrm{Aut}(D)i$. Note that, since we jointly change all learning algorithms in one orbit $\mathrm{Aut}(D)i$, the remaining profile $\hat{\boldsymbol{\sigma}}$ is symmetric by Lemma 46. Next, let $K := |\mathrm{Aut}(D)i|$, choose

Dec-POMDPs $D_1, \ldots, D_N \in \mathcal{D}$, and measurable sets $\mathcal{Z}_1, \ldots, \mathcal{Z}_N \subseteq \Pi^D$. Then it is

$$
\begin{aligned}
&(\otimes_{j\in\mathcal{N}}\hat{\boldsymbol{\sigma}}(D_j))(\prod_{l\in\mathcal{N}}\mathcal{Z}_l) && (92)\\
&= \prod_{j\in\mathcal{N}\setminus\mathrm{Aut}(D)i}\boldsymbol{\sigma}(D_j)(\mathcal{Z}_j)\prod_{l\in\mathrm{Aut}(D)i}(p\tilde{\boldsymbol{\sigma}}_i(D_l)(\mathcal{Z}_l)+(1-p)\boldsymbol{\sigma}_i(D_l)(\mathcal{Z}_l)) && (93)\\
&= (1-p)^K\prod_{j\in\mathcal{N}}\boldsymbol{\sigma}_j(D_j)(\mathcal{Z}_j)\\
&\quad+\sum_{j\in\mathrm{Aut}(D)i}p(1-p)^{K-1}\tilde{\boldsymbol{\sigma}}_i(D_j)(\mathcal{Z}_j)\prod_{l\in\mathcal{N}\setminus\{j\}}\boldsymbol{\sigma}_l(D_l)(\mathcal{Z}_l)\\
&\quad+\sum_{k=2,\ldots,K}p^k(1-p)^{K-k}\mu^{j,k,D_j}\left(\prod_{j\in\mathcal{N}}\mathcal{Z}_j\right) && (94)\\
&= (1-p)^K\otimes_{j\in\mathcal{N}}\boldsymbol{\sigma}_j(D_j)\left(\prod_{j\in\mathcal{N}}\mathcal{Z}_j\right)\\
&\quad+\sum_{j\in\mathrm{Aut}(D)i}p(1-p)^{K-1}\left(\tilde{\boldsymbol{\sigma}}_i(D_j)\otimes_{l\in\mathcal{N}\setminus\{j\}}\boldsymbol{\sigma}_l(D_l)\right)\left(\prod_{j\in\mathcal{N}}\mathcal{Z}_j\right)\\
&\quad+\sum_{k=2,\ldots,K}p^k(1-p)^{K-k}\mu^{j,k,D_j}\left(\prod_{j\in\mathcal{N}}\mathcal{Z}_j\right) && (95)
\end{aligned}
$$

where $\mu^{k,D_1,\ldots,D_N}$ is some measure (not necessarily a probability measure) on the space $\Pi^D\times\cdots\times\Pi^D$ that depends on k and $D_1,\ldots,D_N$, but not on p. This tells us that we can also decompose the integral with respect to the measure $\otimes_{j\in\mathcal{N}}\hat{\boldsymbol{\sigma}}(D_j)$ as in (95). It follows that

$$
\begin{aligned}
&U(\hat{\boldsymbol{\sigma}}_1,\ldots,\hat{\boldsymbol{\sigma}}_N) && (96)\\
&=\mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(j)}\sim(\mathbf{f}_j^{-1})^*\hat{\boldsymbol{\sigma}}_j(\mathbf{f}_j^*D),\,j\in\mathcal{N}}\left[J^D((\pi_l^{(l)})_{l\in\mathcal{N}})\right]\right] && (97)\\
&=\mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\pi^{(j)}\sim\hat{\boldsymbol{\sigma}}_j(\mathbf{f}_j^*D),\,j\in\mathcal{N}}\left[J^D\left(\left(\mathrm{proj}_l((\mathbf{f}_l^{-1})^*\pi^{(l)})\right)_{l\in\mathcal{N}}\right)\right]\right] && (98)\\
&\overset{(95)}{=}\mathbb{E}_{\mathbf{f}\sim\mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}\Bigg[(1-p)^K\mathbb{E}_{\pi^{(j)}\sim\boldsymbol{\sigma}_j(\mathbf{f}_j^*D),\,j\in\mathcal{N}}\left[J^D\left(\left(\mathrm{proj}_l((\mathbf{f}_l^{-1})^*\pi^{(l)})\right)_{l\in\mathcal{N}}\right)\right]\\
&\quad+\sum_{j\in\mathrm{Aut}(D)i}p(1-p)^{K-1}\mathbb{E}_{\pi^{(m)}\sim\boldsymbol{\sigma}_m(\mathbf{f}_m^*D),\,m\in\mathcal{N}\setminus j}\left[\mathbb{E}_{\pi^{(j)}\sim\tilde{\boldsymbol{\sigma}}_i(\mathbf{f}_j^*D)}\left[J^D\left(\left(\mathrm{proj}_l((\mathbf{f}_l^{-1})^*\pi^{(l)})\right)_{l\in\mathcal{N}}\right)\right]\right]\\
&\quad+\sum_{k=2,\ldots,K}p^k(1-p)^{K-k}\int J^D\left(\left(\mathrm{proj}_l((\mathbf{f}_l^{-1})^*\pi^{(l)})\right)_{l\in\mathcal{N}}\right)\mathrm{d}\mu^{k,\mathbf{f}_1^*D,\ldots,\mathbf{f}_N^*D}\left(\pi^{(1)},\ldots,\pi^{(N)}\right)\Bigg] && (99)\\
&=(1-p)^KU(\boldsymbol{\sigma})+\sum_{j\in\mathrm{Aut}(D)i}p(1-p)^{K-1}U(\boldsymbol{\sigma}_1,\ldots,\boldsymbol{\sigma}_{j-1},\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{j+1},\ldots,\boldsymbol{\sigma}_N)\\
&\quad+\sum_{k=2,\ldots,K}p^k(1-p)^{K-k}C_k && (100)
\end{aligned}
$$

$$= B(m = 0, p)U(\boldsymbol{\sigma}) + B(m = 1, p)U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) + \sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k \tag{101}$$

for some constants $C_k \in \mathbb{R}$ for $k = 2, \dots, K$, and where we write $B(k = 0, p)$ to denote the probability that a binomial distribution with K trials and success chance p has 0 successful trials (and $B(k = 1, p)$ analogously). Here, in (98), we use a change of variables for pushforward measures, separately for each measure $\hat{\boldsymbol{\sigma}}_j(\mathbf{f}_j^* D)$; in (100), we use the linearity of the expectation; and in (101) we use that U and $\boldsymbol{\sigma}$ are both invariant to symmetries between principals, and thus for $g \in \mathrm{Aut}(D)$ with $g^{-1}j = i$, it is

$$U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) = U(\tilde{\boldsymbol{\sigma}}_{g^{-1}1}, \dots, \boldsymbol{\sigma}_{g^{-1}(j-1)}, \tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{g^{-1}(j+1)}, \dots, \boldsymbol{\sigma}_{g^{-1}N})$$
$$= U(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_{j-1}, \tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{j+1}, \dots, \boldsymbol{\sigma}_N). \tag{102}$$

Now note that if we can show that

$$(1 - B(m = 0, p))U(\boldsymbol{\sigma}) < B(m = 1, p)U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) + \sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k,$$

then it would also follow that

$$U(\hat{\boldsymbol{\sigma}}) \overset{(96)–(101)}{=} B(m = 0, p)U(\boldsymbol{\sigma}) + B(m = 1, p)U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) + \sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k$$
$$> B(m = 0, p)U(\boldsymbol{\sigma}) + (1 - B(m = 0, p))U(\boldsymbol{\sigma}) = U(\boldsymbol{\sigma}). \tag{103}$$

Thus, this would show that $\hat{\boldsymbol{\sigma}}$ is a symmetric profile with higher payoff than $\boldsymbol{\sigma}$, proving the required contradiction.

In the following, we show the equivalent condition

$$U(\boldsymbol{\sigma}) < \frac{B(1, p)}{B(m > 0, p)}U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) + \frac{\sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k}{B(m > 0, p)},$$

where $B(m > 0, p) = \sum_{k=1}^{K} B(m = k, p) = 1 - B(m = 0, p)$. To that end, note that since $U(\boldsymbol{\sigma}) < U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i})$ by assumption, we can choose some small $\epsilon > 0$ such that still

$$U(\boldsymbol{\sigma}) < U(\tilde{\boldsymbol{\sigma}}_i, \boldsymbol{\sigma}_{-i}) - \epsilon.$$

Moreover, note that $B(m > 0, p)$ is a polynomial in p, and the degree of its nonzero term with lowest degree is 1. Similarly, for $\sum_{k=2,\dots,K} p^{K-k}(1-p)^2 C_k$, that lowest degree is 2. Hence, it is

$$\frac{\sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k}{B(m > 0, p)} = \frac{\sum_{k=2,\dots,K} p^{k-1}(1-p)^{K-k}C_k}{C + Q(p)}$$

for some constant $C \neq 0$ and polynomial Q in p, and it follows that

$$\lim_{p \to 0} \frac{\sum_{k=2,\dots,K} p^k(1-p)^{K-k}C_k}{B(m > 0, p)} = \lim_{p \to 0} \frac{\sum_{k=2,\dots,K} p^{k-1}(1-p)^{K-k}C_k}{C + Q(p)} = 0.$$

With the same argument, it is also $\lim_{p\to 0}\frac{B(m>1,p)}{B(m>0,p)}=0$ and thus

$$\lim_{p\to 0}\frac{B(m=1,p)}{B(m>0,p)}=1-\lim_{p\to 0}\frac{B(m>1,p)}{B(m>0,p)}=1.$$

Hence, we can find some $p>0$ that is small enough such that both

$$\frac{\sum_{k=2,\dots,K}p^k(1-p)^{K-k}C_k}{B(m>0,p)}<\frac{\epsilon}{2}$$

and

$$1+\frac{\epsilon}{2|U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})|}>\frac{B(m=1,p)}{B(m>0,p)}>1-\frac{\epsilon}{2|U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})|}.$$

It follows that

$$\begin{aligned}U(\boldsymbol{\sigma})<U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})-\epsilon&=\left(1-\frac{\epsilon}{2U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})}\right)U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})-\frac{\epsilon}{2}\\&<\frac{B(m=1,p)}{B(m>0,p)}U(\tilde{\boldsymbol{\sigma}}_i,\boldsymbol{\sigma}_{-i})+\frac{\sum_{k=2,\dots,K}p^k(1-p)^{K-k}C_k}{B(m>0,p)},\end{aligned}\tag{104}$$

which is what we wanted to show. This concludes the proof. □

C.8. Evaluating equivariant learning algorithms

In the following, let D be a Dec-POMDP and define $\mathcal{D}:=\{f^*D\mid f\in\mathrm{Sym}(D)\}$. In the special case in which a learning algorithm is in a sense independent from the used labels, we can find a simpler form of the algorithm's value in the LFC problem for D. To do so, we define the notion of an equivariant learning algorithm.

Definition 49 (Equivariant learning algorithms)**.** Let $\sigma\in\Sigma^{\mathcal{D}}$. Then σ is called *equivariant* if for any two labelings $f,\tilde{f}\in\mathrm{Sym}(D)$, it is

$$(f^{-1})^*\sigma(f^*D)=(\tilde{f}^{-1})^*\sigma(\tilde{f}^*D).$$

Remark 50. We believe that a learning algorithm implemented via neural networks and one-hot encodings, as used in our experiments, should be equivariant. To see this, note that by a symmetry argument, the distribution over functions corresponding to a randomly initialized neural network is invariant with respect to coordinate permutations. Assume that actions, observations, and agents of a given problem are implemented as one-hot vectors, i.e., elements of a canonical basis $\{e_1,\dots,e_k\}\in\mathbb{R}^k$ where $k\in\mathbb{N}$ is the cardinality of the respective set. Then the distribution over randomly initialized neural network policies will also not depend on particular assignments of actions, etc., to one-hot vectors. We conjecture that, if the used optimizer is equivariant with respect to coordinate permutations (i.e., if the parameter dimensions are permuted, then the prescribed updates to the parameters are equally permuted), then the resulting learning algorithm is equivariant. We leave a rigorous exploration of this issue to future work.

An equivariant algorithm can be evaluated in the LFC problem for D by evaluating its cross-play value in any Dec-POMDP $E\in\mathcal{D}$. The resulting policies can be permuted by random automorphisms or they can be evaluated as they are.

Proposition 51. *Let $\sigma \in \Sigma^{\mathcal{D}}$ be equivariant. Then for any $f \in \mathrm{Sym}(D)$ and $E = f^*D \in \mathcal{D}$, it is*

$$U^D(\sigma) = \mathbb{E}_{\pi^{(i)} \sim \sigma(E),\, i \in \mathcal{N}} \left[\mathbb{E}_{\mathbf{g} \in \mathcal{U}(\mathrm{Aut}(E)^{\mathcal{N}})} \left[J^E \left(\left(\mathrm{proj}_j(\mathbf{g}_j^* \pi^{(j)}) \right)_{j \in \mathcal{N}} \right) \right] \right] \tag{105}$$

$$= \mathbb{E}_{\pi^{(i)} \sim \sigma(E),\, i \in \mathcal{N}} \left[J^E \left(\left(\pi_j^{(j)} \right)_{j \in \mathcal{N}} \right) \right]. \tag{106}$$

Proof. First, let $f \in \mathrm{Sym}(D)$ and $E := f^*D$. Using the expression of the payoff in the LFC game for D from Lemma 40, it is

$$U^D(\sigma) = U^D(\sigma, \ldots, \sigma) = \mathbb{E}_{\mathbf{f} \sim \mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})} \left[\mathbb{E}_{\pi^{(i)} \sim (\mathbf{f}_i^{-1})^* \sigma(\mathbf{f}_i^* D),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \right] \tag{107}$$

$$= \mathbb{E}_{\pi^{(i)} \sim (f^{-1})^* \sigma(E),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \tag{108}$$

$$= \mathbb{E}_{\pi^{(i)} \sim \sigma(E),\, i \in \mathcal{N}} \left[J^D((f^{-1})^*(\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \tag{109}$$

$$= \mathbb{E}_{\pi^{(i)} \sim \sigma(E),\, i \in \mathcal{N}} \left[J^E((\pi_j^{(j)})_{j \in \mathcal{N}}) \right], \tag{110}$$

where we use equivariance in (108), a change of variables for pushforward measures in (109), and Theorem 22 in (110).

Second, by Lemma 33, it is $\mathrm{Iso}(D, E) \subseteq \mathrm{Sym}(D)$. Thus, using Lemma 19, it is $\mathrm{Aut}(E) \circ f = \mathrm{Iso}(D, E) \subseteq \mathrm{Sym}(D)$. Hence, it follows that

$$U^D(\sigma) = U^D(\sigma, \ldots, \sigma) = \mathbb{E}_{\pi^{(i)} \sim (f^{-1})^* \sigma(E),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \tag{111}$$

$$= \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(E)^{\mathcal{N}}} \left[\mathbb{E}_{\pi^{(i)} \sim (f^{-1} \circ \mathbf{g}_i)^* \sigma(E),\, i \in \mathcal{N}} \left[J^D((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \right] \tag{112}$$

$$= \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(E)^{\mathcal{N}}} \left[\mathbb{E}_{\pi^{(i)} \sim \mathbf{g}_i^* \sigma(E),\, i \in \mathcal{N}} \left[J^E((\pi_j^{(j)})_{j \in \mathcal{N}}) \right] \right] \tag{113}$$

$$= \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(E)^{\mathcal{N}}} \left[\mathbb{E}_{\pi^{(i)} \sim \sigma(E),\, i \in \mathcal{N}} \left[J^E((\mathrm{proj}_j(\mathbf{g}_i^* \pi^{(j)}))_{j \in \mathcal{N}}) \right] \right], \tag{114}$$

where we again use equivariance in (112), a change of variables and Theorem 22 in (113), and another change of variables in (114). This concludes the proof. □

If a learning algorithm is equivariant, we can use either of the expressions above to evaluate it in the LFC problem. One may choose the first expression, i.e., apply random automorphisms to policies, since this transformation maps different policies that are equivalent under OP to a unique automorphism-invariant policy and thus reduces the variance of the cross-play values across different samples $\pi \sim \sigma(D)$. We will introduce these notions of equivalence and automorphism-invariant policies in Appendix D.

Note that an equivariant learning algorithm is not necessarily one that performs well in the LFC problem. For instance, a SP algorithm may be equivariant. However, if an algorithm is equivariant and it does well in cross-play, then the preceding shows that it will also do well in the LFC problem.

D. Characterization of other-play and of label-free coordination games

In this section, we define for a given policy π a policy $\Psi(\pi)$ that corresponds to agents choosing local policies that are randomly permuted by automorphisms, and that is itself invariant to pushforward by automorphisms.

We then use this self-map on policies Ψ, called the *symmetrizer*, to characterize both the OP objective and the payoff in an LFC game. This characterization helps us to analyze the OP-optimal policies in the two-stage lever game in Appendix E, and it allows us to prove a stronger result, showing that any OP algorithm that is not concentrated on only one equivalence class in the two-stage lever game is suboptimal in the corresponding LFC problem. We also use this notion of equivalence classes of policies to define OP with tie-breaking in Appendix F, which allows us to then show that random tie-breaking functions exist.

In the following, in Appendix D.1, we recall the definition of our generalization of OP, and we define policies that are invariant to automorphism. Afterwards, in Appendix D.2, we introduce the concept of a policy corresponding to a distribution over policies. In Appendix D.3, we introduce the other-play distribution, in which each agent's local policy is chosen as pushforward by a random automorphism, and we define the symmetrizer. Using these concepts, in Appendix D.4, we give a new expression for both the OP objective and the payoff in an LFC game. The OP objective can be understood of as transforming a policy into one that is invariant to automorphisms, and evaluating that policy in SP. Finally, in Appendix D.5, we show that our generalized OP objective can in general not be understood of as the SP objective in a modified Dec-POMDP.

D.1. Generalization of other-play

In the following, fix a Dec-POMDP D. Recall that for a profile of automorphisms $\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}$ and a joint policy $\pi \in \Pi^D$, we define the joint policy $\mathbf{g}^*\pi := \hat{\pi}$, where the local policy $\hat{\pi}_i$ of agent $i \in \mathcal{N}$ is given by the local policy of agent i in the pushforward policy $\mathbf{g}_i^*\pi$. That is, for $i \in \mathcal{N}$, we define

$$\hat{\pi}_i := \mathrm{proj}_i(\mathbf{g}_i^*\pi) = \pi_{\mathbf{g}_i^{-1}i}(\mathbf{g}_i^{-1}\cdot \mid \mathbf{g}_i^{-1}\cdot).$$

Using this, we define the OP objective as the expected return of a policy that is randomly permuted by such profiles of automorphisms.

Definition 52 (Other-play objective)**.** Define $J_{\mathrm{OP}}^D \colon \Pi^D \to \mathbb{R}$ via

$$J_{\mathrm{OP}}^D(\pi) := \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J^D(\mathbf{g}^*\pi)\right] \tag{115}$$

for $\pi \in \Pi^D$, where $\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})$ is a uniform distribution over $\mathrm{Aut}(D)^{\mathcal{N}}$. We say that J_{OP}^D is the *other-play (OP) objective* of D, and $J_{\mathrm{OP}}^D(\pi)$ is the *OP value* of $\pi \in \Pi^D$.

Remark 53. It is clear that this objective always admits a maximum. For instance, we can consider Π^D as a subset of $\prod_{i\in\mathcal{N}}[0,1]^{\mathcal{A}_i\times\overline{\mathcal{AO}}_i}$ with its standard topology. As Π^D is a Cartesian product of simplices, it is a compact subset of this space. Moreover, one can check that the objective $J^D(\pi)$ is continuous in the policy π, and that for $\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}$, the map $\pi \mapsto \mathbf{g}^*\pi$ is continuous as well. Thus, also $J_{\mathrm{OP}}^D(\pi)$ is continuous in π, as it is a finite linear combination of continuous functions. By the extreme value theorem, it follows that the function always attains a maximum.

Next, recall our formal definition of an OP learning algorithm as any algorithm that achieves an optimal OP value in expectation.

Definition 54 (Other-play learning algorithm)**.** Let $\mathcal{D}$ be a finite set of Dec-POMDPs. A learning algorithm $\sigma \in \Sigma^{\mathcal{D}}$ is called an *OP learning algorithm* if for any $D \in \mathcal{D}$, it is

$$\mathbb{E}_{\pi\sim\sigma^{\mathrm{OP}}(D)}[J_{\mathrm{OP}}^D(\pi)] = \max_{\pi\in\Pi^D} J_{\mathrm{OP}}^D(\pi).$$

Now fix again a Dec-POMDP D and consider the notion of invariance to automorphism.

Definition 55 (Invariance to automorphism)**.** A policy $\pi \in \Pi$ is called *invariant to automorphism* if $f^*\pi = \pi$ for any $f \in \mathrm{Aut}(D)$.

Clearly, if a policy is invariant to automorphism, then it has the same OP value and expected return.

Proposition 56. *Let $\pi \in \Pi$ be invariant to automorphism. Then*

$$J_{\mathrm{OP}}(\pi) = J(\pi).$$

Proof. It is

$$J_{\mathrm{OP}}(\pi) = \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J(\mathbf{g}^*\pi)\right] = \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J((\mathrm{proj}_i(\mathbf{g}_i^*\pi))_{i\in\mathcal{N}})\right]$$
$$\overset{(*)}{=} \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J((\mathrm{proj}_i(\pi))_{i\in\mathcal{N}})\right] = J(\pi), \quad (116)$$

where we have used invariance to automorphism in (*). □

In the following, we will show that if a policy π is not already invariant to automorphism, then one can understand the OP objective as first transforming the policy into a policy $\Psi(\pi)$ that is invariant to automorphism by applying the symmetrizer Ψ, and then evaluating the expected return of that policy. In that way, the OP objective ensures that policies cannot make use of arbitrary symmetry-breaking.

D.2. Policies corresponding to distributions over policies

In this section, let some Dec-POMDP D be fixed. As a first step towards defining the symmetrizer Ψ, we will define policies corresponding to distributions over policies for general distributions. Afterwards, we will turn to the particular policy $\Psi(\pi)$ that corresponds to the *OP distribution* of π.

Recall that we introduced the set of distributions over policies as $\Delta(\Pi)$, the set of measures on the space $(\Pi, \mathcal{F})$. Let $\nu \in \Delta(\Pi)$. For a given distribution $\nu \in \Delta(\Pi)$ and agent $i \in \mathcal{N}$, the marginal distribution ν_i is defined as $\nu_i(\mathcal{Z}_i) := \nu(\mathrm{proj}_i^{-1}(\mathcal{Z}_i))$ for any measurable set of local policies $\mathcal{Z}_i \subseteq \Pi_i$. We say that ν has *independent local policies*, if $\nu = \otimes_i \nu_i$, i.e., ν decomposes into independent marginal distributions over local policies for each agent. We denote such distributions by μ.

Now let $\mu \in \Delta(\Pi)$ be a distribution with independent local policies. We want to construct a policy π^μ that represents each agent sampling a local policy $\pi_i \sim \mu_i$ in the beginning of an episode, and then choosing actions according to that policy until the end of the episode. This policy should be equivalent to μ in the sense that it should yield the same expected return as μ, where we define the expected return of μ as

$$J(\mu) := \mathbb{E}_{\pi\sim\mu}\left[J(\pi)\right] = \mathbb{E}_{\pi_i\sim\mu_i,\, i\in\mathcal{N}}\left[J(\pi)\right]. \quad (117)$$

The statement that such a policy exists is analogous and more general than a result by Kuhn (1953), which says that in an extensive-form game, given some conditions on the game, for every mixed strategy there is an equivalent behavior strategy. Kuhn's theorem is relevant to Dec-POMDPs since there is a correspondence between Dec-POMDPs and extensive-form games (Oliehoek et al., 2006). For Dec-POMDPs, the analogous result states that for every distribution over deterministic policies, there is an equivalent stochastic policy.

We cannot directly apply Kuhn's theorem here, as we require a result for distributions over stochastic policies instead of deterministic policies. This is because such a result fits better with our remaining setup—for instance, the domain of the OP objective is the set of stochastic policies, and learning algorithms are defined as distributions over stochastic policies. Nevertheless, our proof is based on similar ideas as, for instance, the proof of Kuhn's theorem in Maschler et al. (2013), after translating between the different formal frameworks of extensive-form games and Dec-POMDPs.

To define a policy π^μ that is equivalent to μ, we begin by defining a new measure space $(\Pi \times \Omega, \mathcal{F} \otimes \mathcal{P}(\Omega))$, which is the product space of the space of policies $(\Pi, \mathcal{F})$ with the Dec-POMDP environment $(\Omega, \mathcal{P}(\Omega))$. On this space, for a given distribution μ, we define a probability measure $\mathbb{P}_\mu$ which represents the procedure outlined above, i.e., in which a policy π is chosen according to μ, and then samples in Ω are distributed according to $\mathbb{P}_\pi$. Formally, we define $\mathbb{P}_\mu$ as the unique measure such that

$$\mathbb{P}_\mu(\mathcal{Z} \times \mathcal{Q}) := \mathbb{E}_{\pi\sim\mu}\left[\mathbb{1}_{\mathcal{Z}}\mathbb{P}_\pi(\mathcal{Q})\right]^{5} \tag{118}$$

for any measurable sets $\mathcal{Z} \subseteq \Pi$ and $\mathcal{Q} \subseteq \Omega$. Note that the product sets $\mathcal{Z} \times \mathcal{Q}$ for measurable $\mathcal{Z} \subseteq \Pi$ and $\mathcal{Q} \subseteq \Omega$ are a π-system and generate the product σ-Algebra $\mathcal{F} \otimes \mathcal{P}(\Omega)$. Hence, by Carathéodory's extension theorem, there is a unique measure satisfying this definition (see Williams, 1991, ch. 1).

On this new product space $\Pi \times \Omega$, define the random variable Z as the projection onto Π. We can define histories H and all other random variables defined on the space $(\Omega, \mathcal{P}(\Omega))$ by composing them with the projection onto Ω (for notational convenience, we denote these random variables using the same symbols in both spaces).

Remark 57. Note that the conditional probability of a particular trajectory τ given the policy Z is just the probability of that history under $\mathbb{P}_Z$, that is,

$$\mathbb{P}_\mu(H = \tau \mid Z) = \mathbb{P}_Z(H = \tau) \tag{119}$$

for any $\tau \in \mathcal{H}$. Here, the conditional probability is a random variable, defined via the conditional expectation

$$\mathbb{P}_\mu(H = \tau \mid Z) := \mathbb{E}_\mu\left[\mathbb{1}_{H=\tau} \mid Z\right].$$

Intuitively, for a given sample (π, ω), the value of that random variable is the best estimate of the probability of $\{H = \tau\}$ given $Z = \pi$, ignoring ω. As, in general, $\{Z = \pi\}$ may have zero probability, it is impossible to define the conditional probability $\mathbb{P}(H = \tau \mid Z = \pi)$ via $\mathbb{P}(H = \tau \mid Z = Z) := \frac{\mathbb{P}(H=\tau, Z=\pi)}{\mathbb{P}(Z=\pi)}$. It is still possible to define the *conditional expectation* $\mathbb{E}_\mu\left[\mathbb{1}_{H=\tau} \mid Z\right]$, though.

Here, we briefly give the definition of the conditional expectation and show that (119) is correct. For a reference on conditional expectations, refer to (Williams, 1991, ch. 9). Applied to our setup, the conditional expectation of $\mathbb{1}_{H=\tau}$ given Z is any random variable (which can be shown to be almost surely unique), denoted by $\mathbb{E}_\mu\left[\mathbb{1}_{H=\tau} \mid Z\right]$, that is measurable with respect to

$$\sigma(Z) := \{Z^{-1}(\mathcal{Z}) \mid \mathcal{Z} \in \mathcal{F}\} = \{\mathcal{Z} \times \Omega \mid \mathcal{Z} \in \mathcal{F}\}$$

such that

$$\mathbb{E}_\mu\left[\mathbb{1}_{\mathcal{X}}\mathbb{E}_\mu\left[\mathbb{1}_{H=\tau} \mid Z\right]\right] = \mathbb{E}_\mu\left[\mathbb{1}_{\mathcal{X}}\mathbb{1}_{H=\tau}\right]. \tag{120}$$

for all $\mathcal{X} \in \sigma(Z)$. The fact that the random variable is measurable with respect to $\sigma(Z)$ can be equivalently expressed as saying that it can be written as a function of Z. Moreover, Equation 120 says that the conditional expectation should represent correct averages of the random variable $\mathbb{1}_{H=\tau}$ over the level-sets in $\sigma(Z)$.

[5] $\mathbb{P}_\mu$ is the semidirect product of μ with the Markov kernel $\kappa(\cdot \mid \pi) := \mathbb{P}_\pi(\cdot)$.

To show that (119) is correct, let $\mathcal{X} \in \sigma(Z)$ arbitrary. Since $\mathcal{X}$ is of the form $\mathcal{Z} \times \Omega$, and $\{H = \tau\} = \Pi \times \mathcal{Q}$ for some $\mathcal{Q} \subseteq \Omega$, it is $(\mathcal{Z} \times \Omega) \cap \{H = \tau\} = \mathcal{Z} \times \mathcal{Q}$. Using Equation 118, it follows that

$$\mathbb{E}_\mu [\mathbb{1}_\mathcal{X} \mathbb{P}_Z(H = \tau)] = \mathbb{E}_\mu [\mathbb{1}_{\mathcal{Z}\times\Omega} \mathbb{P}_Z(\mathcal{Q})] = \mathbb{E}_{\pi\sim\mu} [\mathbb{1}_\mathcal{Z}(\pi)\mathbb{P}_\pi(\mathcal{Q})] \stackrel{(118)}{=} \mathbb{P}_\mu(\mathcal{Z} \times \mathcal{Q}) = \mathbb{E}_\mu [\mathbb{1}_{\Pi\times\Omega}\mathbb{1}_{\mathcal{Z}\times\mathcal{Q}}] = \mathbb{E}_\mu [\mathbb{1}_\mathcal{X}\mathbb{1}_{H=\tau}] . \quad (121)$$

Hence, letting $\mathbb{E}_\mu[\mathbb{1}_{H=\tau} \mid Z] := \mathbb{P}_Z(H = \tau)$ satisfies condition (120). $\mathbb{P}_Z(H = \tau)$ is also $\sigma(Z)$-measurable, since it is just a function of Z.

We will repeatedly make use of (119) in the following, together with the tower property, which, applied to our case, says that

$$\mathbb{P}_\mu(H = \tau) = \mathbb{E}_\mu [\mathbb{P}_\mu (H = \tau \mid Z)] \stackrel{(119)}{=} \mathbb{E}_\mu [\mathbb{P}_Z (H = \tau)] \quad (122)$$

for any history $\tau \in \mathcal{H}$ (see Williams, 1991, ch. 9.7).

Using the measure space defined above, we now define a local policy $\pi_i^{\mu_i}$ for an agent $i \in \mathcal{N}$, corresponding to a distribution $\mu_i \in \Delta(\Pi_i)$. For an action $a_i \in \mathcal{A}_i$ and an action-observation history $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$, we define $\pi_i^{\mu_i}(a_i \mid \tau_{i,t})$ as the probability that agent i, who follows a policy that is sampled from μ_i, plays action a_i, conditional on $\{\overline{AO}_{i,t} = \tau_{i,t}\}$.

Definition 58. Let $i \in \mathcal{N}$ and let $\mu_i \in \Delta(\Pi_i)$ be a distribution over local policies of agent i. We define the local policy $\pi_i^{\mu_i}$ corresponding to μ_i in the following way. For $a_i \in \mathcal{A}_i$, $t \in \{0, \dots, T\}$ and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$, let

$$\pi_i^{\mu_i}(a_i \mid \tau_{i,t}) := \mathbb{P}_{\mu_i\otimes\mu_{-i}}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}), \quad (123)$$

where μ_{-i} is any distribution over Π_{-i} with independent local policies such that

$$\mathbb{P}_{\mu_i\otimes\mu_{-i}}(\overline{AO}_{i,t} = \tau_{i,t}) > 0.$$

If no such distribution exists, we let $\pi_i^{\mu}(a_i \mid \tau_{i,t}) := \frac{1}{|\mathcal{A}_i|}$.

Note that if $\mathbb{P}_{\mu_i\otimes\mu_{-i}}(\overline{AO}_{i,t} = \tau_{i,t}) = 0$ for all the distributions $\mu_{-i} \in \Delta(\Pi_{-i})$ over opponent policies, then agent i's action-observation history $\tau_{i,t}$ is almost never reached, independent of the other agents' policies. In that case, we can define the policy arbitrarily and this will never matter for the distribution over histories under $\mathbb{P}_{\mu_i\otimes\mu_{-i}}$, and as we will see, neither for the distribution under $\mathbb{P}_{\pi_i^{\mu_i},\pi_{-i}}$ for arbitrary π_{-i}.

First, we need to make sure that $\pi_i^{\mu_i}$ is well-defined, i.e., it does not depend on the chosen distribution μ_{-i}. Unfortunately, the proof for the following Lemma is somewhat technical.

Lemma 59. *Let $i \in \mathcal{N}$, $\mu_i \in \Delta(\Pi_i)$, $t \in \{0, \dots, T\}$, $\tau_{i,t} \in \overline{\mathcal{AO}}_i$, and $a_i \in \mathcal{A}_i$. Let $\mu_{-i}, \mu'_{-i} \in \Delta(\Pi_{-i})$ be any two distributions with independent local policies such that $\mathbb{P}_{\mu_i\otimes\mu_{-i}}(\overline{AO}_{i,t} = \tau_{i,t}) > 0$ and $\mathbb{P}_{\mu_i\otimes\mu'_{-i}}(\overline{AO}_{i,t} = \tau_{i,t}) > 0$. Then it is*

$$\mathbb{P}_{\mu_i\otimes\mu_{-i}}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \mathbb{P}_{\mu_i\otimes\mu'_{-i}}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}).$$

Proof. In the following, let $i \in \mathcal{N}, t \in \{0, \dots, T\}$ be fixed. Our goal is to find an expression for the distribution of Z_i given $\overline{AO}_{i,t}$ that only depends on μ_i. If we can do that, we can also show that the probability of a particular action chosen by an agent is independent of the distributions over other agents' policies. To begin, we analyze the joint distribution of Z_i and $\overline{AO}_{i,t}$ for an arbitrary distribution $\mu \in \Delta(\Pi)$ with

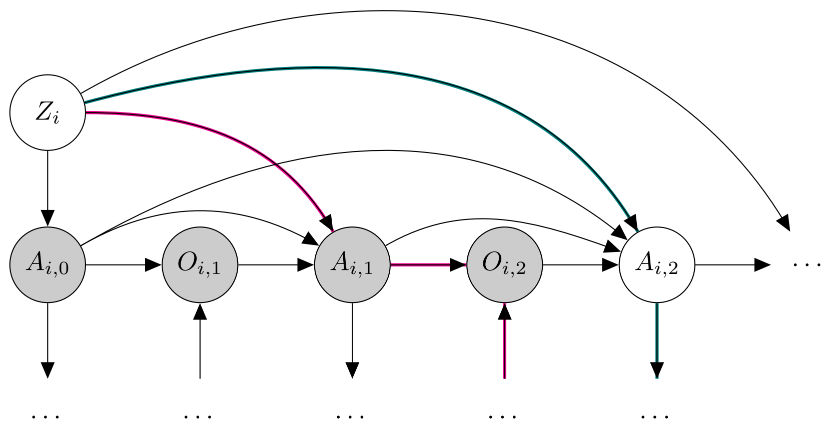

Figure 8. Part of a Bayesian graph of the random variables on the space $\Pi \times \Omega$. Here, the gray-marked nodes are part of the action-observation history $\overline{AO}_{i,2}$. Conditional on $\overline{AO}_{i,2}$, there is an unblocked path, marked in teal, from Z_i to $A_{i,2}$ and to nodes below $A_{i,2}$ not displayed here, making them d-connected and thus dependent. The path marked in a lighter magenta, on the other hand, is blocked, illustrating that the part of the graph below $O_{i,2}$ may be independent of Z_i.

independent local policies. Note that by assumption, the Z_j are independent for $j \in \mathcal{N}$ under μ. We now show that conditioning on $\overline{AO}_{i,t}$ still leaves Z_i independent from Z_{-i}.

To see this, one can consider a Bayesian graph of the random variables on the space $\Pi \times \Omega$. For a reference on Bayesian graphs and the concepts discussed below, refer to Pearl (2009, ch. 1.2). In Figure 8, we have displayed a part of this Bayesian graph with nodes for the agent i, indicating left-out parts of the graph with dots. The arrows in the graph illustrate the dependence relationships between the different variables: if there is an arrow from one node to another, this means that the other node depends on that node. Nodes belonging to the action-observation history $\overline{AO}_{i,2}$ of agent i are marked in gray.

Given such a Bayesian graph, the d-separation criterion tells us which variables are dependent after conditioning on a set $\mathcal{V}$ of variables. The criterion specifies valid, "unblocked" paths in the graph, depending on the graph structure and the nodes that are being conditioned on. The d-separation criterion says that two variables are dependent conditional on the variables in $\mathcal{V}$ if and only if there is an unblocked path in the graph between them; the variables are then said to be d-connected. If there is no path, then the variables are d-separated.

An unblocked path can contain a *chain* $X \to W \to Y$ or a *fork* $X \leftarrow W \to Y$ if W is not being conditioned on. If it contains a *collider* $X \to W \leftarrow Y$, on the other hand, i.e., the two incident edges to W in the path are both directed towards the node, then the path is blocked, unless W or a descendant of W in the graph is in $\mathcal{V}$. For instance, the path that is marked in a lighter magenta in Figure 8 is blocked—the path cannot go from Z_i over $A_{i,1}$ to $O_{i,2}$. An unblocked path is marked in teal.

Using the d-separation criterion, one can tell that Z_i may be d-separated from left-out parts of the graph in Figure 8 indicated by the dots below $A_{i,0}$, for instance, but that it is d-connected to some variables in the part below $A_{i,2}$. We do not work this out here completely, but once considering the entire graph, one can see that there is no unblocked path from Z_i to the Z_{-i}, because such a path would inevitably have to traverse a collider that is not being conditioned on and that has no descendants that are being conditioned on (for instance, the observations of all other agents are colliders that connect the Z_{-i} with Z_i). Generalizing from the example of $\tau_{i,2}$, we can conclude that Z_i and Z_{-i} are independent given $\overline{AO}_{i,t}$.

Now let $\mathcal{Z}_j \subseteq \Pi_j$ be measurable sets for $j \in \mathcal{N}$, and let $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$ arbitrary. Then, using the above, it follows that

$$\mathbb{E}_{\pi_{-i}\sim\mu_{-i}}[\mathbb{1}_{\pi_{-i}\in\mathcal{Z}_{-i}}\mathbb{E}_{Z_i\sim\mu_i}[\mathbb{1}_{\pi_i\in\mathcal{Z}_i}\mathbb{P}_\pi(\overline{AO}_{i,t}=\tau_{i,t})]] \tag{124}$$

$$= \mathbb{P}_\mu(Z_i \in \mathcal{Z}_i, \overline{AO}_{i,t}=\tau_{i,t}, Z_{-i}\in\mathcal{Z}_{-i}) \tag{125}$$

$$= \mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t}, Z_{-i}\in C_{-i})\mathbb{P}_\mu(\overline{AO}_{i,t}=\tau_{i,t}, Z_{-i}\in\mathcal{Z}_{-i}) \tag{126}$$

$$= \mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t})\mathbb{P}_\mu(\overline{AO}_{i,t}=\tau_{i,t}, Z_{-i}\in\mathcal{Z}_{-i}) \tag{127}$$

$$= \mathbb{E}_{\pi_{-i}\sim\mu_{-i}}[\mathbb{1}_{\pi_{-i}\in\mathcal{Z}_{-i}}\mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t})\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_\pi(\overline{AO}_{i,t}=\tau_{i,t})]], \tag{128}$$

where we use the argument about conditional independence in (127) and the definition of $\mathbb{P}_\mu$ from Equation 118 in (125) and (128).

Since the sets $\mathcal{Z}_j$ for $j \in \mathcal{N}\setminus\{i\}$ were arbitrary, it follows that μ_{-i}-almost surely, it is

$$\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{1}_{\pi_i\in\mathcal{Z}_i}\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})] = \mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t})\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]. \tag{129}$$

Moreover, since μ was arbitrary, Equation 129 holds for any distribution with independent local policies. If we divide by the term $\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]$, this becomes

$$\mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t}) = \frac{\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{1}_{\pi_i\in\mathcal{Z}_i}\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]}{\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]}, \tag{130}$$

which gives us a formula for the distribution of Z_i given $\overline{AO}_{i,t}$ under the measure $\mathbb{P}_\mu$ that is independent of the distributions μ_{-i}. But to be able to do so, we have to find a value for Z_{-i} such that $\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_{\pi_i,Z_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]$ is nonzero under μ.

Now let $\mu_i \in \Delta(\Pi_i)$ and let $\mu_{-i}, \mu'_{-i} \in \Delta(\Pi_{-i})$ be any two distributions with independent local policies such that $\mathbb{P}_{\mu_i\otimes\mu_{-i}}(\overline{AO}_{i,t}=\tau_{i,t}) > 0$ and $\mathbb{P}_{\mu_i\otimes\mu'_{-i}}(\overline{AO}_{i,t}=\tau_{i,t}) > 0$. Define $\mu'_i := \mu_i$, $\mu := \mu_i\otimes\mu_{-i}$, and $\mu' := \mu_i\otimes\mu'_{-i}$. Our goal is to show that

$$\mathbb{P}_\mu(A_{i,t}=a_i \mid \overline{AO}_{i,t}=\tau_{i,t}) = \mathbb{P}_{\mu'}(A_{i,t}=a_i \mid \overline{AO}_{i,t}=\tau_{i,t}).$$

To that end, we define a third distribution $\hat{\mu} := \otimes_{j\in\mathcal{N}}(\frac{1}{2}\mu_j + \frac{1}{2}\mu'_j)$. Apparently, it is then $\mu_i = \hat{\mu}_i = \mu'_i$ and also $\hat{\mu}$ has independent local policies. We now prove separately for μ and μ' that the distribution of Z_i given $\{\overline{AO}_{i,t}=\tau_{i,t}\}$ under μ respectively μ' is equal to the one under $\hat{\mu}$. To do so, we use that μ and μ' are absolutely continuous with respect to $\hat{\mu}$ to find desired values for Z_{-i} such that we can apply Equation 130.

First, let $\mathcal{Z}_i \in \Pi_i$ be an arbitrary measurable set. Using Equation 129, we can find measurable sets $\mathcal{Z}_{-i}, \hat{\mathcal{Z}}_{-i} \subseteq \Pi_{-i}$ such that $\mu_{-i}(\mathcal{Z}_{-i}) = 1 = \hat{\mu}_{-i}(\hat{\mathcal{Z}}_{-i})$, and such that

$$\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{1}_{\pi_i\in\mathcal{Z}_i}\mathbb{P}_{\pi_i,\pi_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})] = \mathbb{P}_\mu(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t}=\tau_{i,t})\mathbb{E}_{\pi_i\sim\mu_i}[\mathbb{P}_{\pi_i,\pi_{-i}}(\overline{AO}_{i,t}=\tau_{i,t})]. \tag{131}$$

for any $\pi_{-i} \in \mathcal{Z}_{-i}$ and

$$\mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{1}_{\pi_i \in \mathcal{Z}_i} \mathbb{P}_{\pi_i, \hat{\pi}_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})] \\ = \mathbb{P}_{\hat{\mu}}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}) \mathbb{E}_{\pi_i \sim \mu_i}[\mathbb{P}_{\pi_i, \hat{\pi}_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]. \quad (132)$$

for any $\hat{\pi}_{-i} \in \hat{\mathcal{Z}}_{-i}$.

Next, it follows that $\mathbb{P}_{\mu}(\{Z_{-i} \in \mathcal{Z}_{-i}\} \cap \{\overline{AO}_{i,t} = \tau_{i,t}\}) > 0$, which by definition of $\hat{\mu}$ implies $\mathbb{P}_{\hat{\mu}}(\{Z_{-i} \in \mathcal{Z}_{-i}\} \cap \{\overline{AO}_{i,t} = \tau_{i,t}\}) > 0$ and thus by definition of $\hat{\mathcal{Z}}_{-i}$ also $\mathbb{P}_{\hat{\mu}}(\{Z_{-i} \in \mathcal{Z}_{-i} \cap \hat{\mathcal{Z}}_{-i}\} \cap \{\overline{AO}_{i,t} = \tau_{i,t}\}) > 0$. Since

$$\mathbb{P}_{\hat{\mu}}(\{Z_{-i} \in \mathcal{Z}_{-i} \cap \hat{\mathcal{Z}}_{-i}\} \cap \{\overline{AO}_{i,t} = \tau_{i,t}\}) \\ = \mathbb{E}_{\pi_{-i} \sim \hat{\mu}_{-i}}\left[\mathbb{1}_{\mathcal{Z}_{-i} \cap \hat{\mathcal{Z}}_{-i}}(\pi_{-i}) \mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]\right], \quad (133)$$

there must be $\pi_{-i} \in \text{proj}_{\Pi_{-i}}(\{Z_{-i} \in \mathcal{Z}_{-i} \cap \hat{\mathcal{Z}}_{-i}\} \cap \{\overline{AO}_{i,t} = \tau_{i,t}\})$ such that

$$\mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})] > 0.$$

Lastly, using that $\hat{\mu}_i = \mu_i$, it follows that

$$\mathbb{E}_{\pi_i \sim \mu_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})] = \mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})] > 0. \quad (134)$$

Hence, we can use Equations 131 and 132 to conclude that

$$\mathbb{P}_{\mu}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}) \stackrel{(131)}{=} \frac{\mathbb{E}_{\pi_i \sim \mu_i}[\mathbb{1}_{\pi_i \in \mathcal{Z}_i} \mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]}{\mathbb{E}_{\pi_i \sim \mu_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]} \quad (135)$$

$$\stackrel{\mu_i = \hat{\mu}_i}{=} \frac{\mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{1}_{\pi_i \in \mathcal{Z}_i} \mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]}{\mathbb{E}_{\pi_i \sim \hat{\mu}_i}[\mathbb{P}_{\pi_i, \pi_{-i}}(\overline{AO}_{i,t} = \tau_{i,t})]} \quad (136)$$

$$\stackrel{(132)}{=} \mathbb{P}_{\hat{\mu}}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}). \quad (137)$$

Now note that one can make an exactly analogous argument for μ' and $\hat{\mu}$, potentially using a different π_{-i}. Hence, it follows that

$$\mathbb{P}_{\mu}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \mathbb{P}_{\hat{\mu}}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \mathbb{P}_{\mu'}(Z_i \in \mathcal{Z}_i \mid \overline{AO}_{i,t} = \tau_{i,t}). \quad (138)$$

Since $\mathcal{Z}_i \in \Pi_i$ was arbitrary, it follows that the distribution of Z_i given $\{\overline{AO}_{i,t} = \tau_{i,t}\}$ is equal under $\mathbb{P}_{\mu}$ and $\mathbb{P}_{\mu'}$.

To conclude the proof, we can use this to show that also the distribution of actions under $\mathbb{P}_{\mu}$ and $\mathbb{P}_{\mu'}$ are equal, conditional on $\{\overline{AO}_{i,t} = \tau_{i,t}\}$. For $a_i \in \mathcal{A}_i$, it is

$$\begin{aligned}
\mathbb{P}_\mu(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) &= \mathbb{E}_\mu\left[\mathbb{P}_\mu(A_{i,t} = a_i \mid Z, \overline{AO}_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (139)\\
&= \mathbb{E}_\mu\left[\mathbb{P}_Z(A_{i,t} = a_i \mid \overline{AO}_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (140)\\
&= \mathbb{E}_\mu\left[Z_i(a_i \mid \tau_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (141)\\
&\overset{(138)}{=} \mathbb{E}_{\mu'}\left[Z_i(a_i \mid \tau_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (142)\\
&= \mathbb{E}_{\mu'}\left[\mathbb{P}_Z(A_{i,t} = a_i \mid \overline{AO}_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (143)\\
&= \mathbb{E}_{\mu'}\left[\mathbb{P}_{\mu'}(A_{i,t} = a_i \mid Z, \overline{AO}_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] && (144)\\
&= \mathbb{P}_{\mu'}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}), && (145)
\end{aligned}$$

where we have used the tower property in (139) and (145), and Equation 119 in (140) and (144). This is what we wanted to show. □

In the following, we write $\pi^\mu := (\pi_i^{\mu_i})_{i\in\mathcal{N}}$ for the joint policy corresponding to a distribution μ with independent local policies. Our next goal is to prove that the distribution over histories is the same under $\mathbb{P}_{\pi^\mu}$ as under $\mathbb{P}_\mu$.

Proposition 60. *Consider any distribution $\mu \in \Delta(\Pi)$ with independent local policies. Let π^μ be the joint policy corresponding to μ, as defined above. Then the history H has the same distribution under $\mathbb{P}_\mu$ as under $\mathbb{P}_{\pi^\mu}$. In particular, it is $J^D(\mu) = J^D(\pi^\mu)$.*

Proof. Fix a distribution $\mu \in \Delta(\Pi)$ with independent local policies. We show by induction that for all $t \in \{0, \dots, T\}$, it is $\mathbb{P}_\mu(H_t = \tau_t) = \mathbb{P}_{\pi^\mu}(H_t = \tau_t)$ for any $\tau_t \in \mathcal{H}_t$.

First, note that by Definition 58, for any $i \in \mathcal{N}, a_i \in \mathcal{A}_i, t \in \{0, \dots, h\}$, and $\tau_{i,t} \in \overline{\mathcal{AO}}_i$ such that $\mathbb{P}_\mu(\overline{AO}_{i,t} = \tau_{i,t}) > 0$, it is

$$\mathbb{P}_{\pi^\mu}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \pi_i^{\mu_i}(a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \mathbb{P}_\mu(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}). \tag{146}$$

In particular, this holds for $t = 0$, in which case it is $\overline{AO}_{i,0} = \emptyset$ for $i \in \mathcal{N}$. Hence, for $a \in \mathcal{A}, s \in \mathcal{S}$, it is

$$\begin{aligned}
\mathbb{P}_\mu(S_0 = s, A_0 = a) &\overset{(122)}{=} \mathbb{E}_\mu[\mathbb{P}_Z(S_0 = s, A_0 = a)] = \mathbb{E}_\mu[\mathbb{P}_Z(S_0 = s)\mathbb{P}_Z(A_0 = a)]\\
&\overset{\text{(i)}}{=} b_0(s)\mathbb{E}_\mu[\mathbb{P}_Z(A_0 = a)] = b_0(s)\mathbb{E}_\mu[\prod_{i\in\mathcal{N}} Z_i(A_{i,0} \mid \emptyset)] \overset{\text{(ii)}}{=} b_0(s)\prod_{i\in\mathcal{N}} \mathbb{E}_\mu[Z_i(a_i \mid \emptyset)]\\
&\overset{(122)}{=} b_0(s)\prod_{i\in\mathcal{N}} \mathbb{P}_\mu(A_{i,0} = a_i) = b_0(s)\prod_{i\in\mathcal{N}} \mathbb{P}_\mu(A_{i,0} = a_i) \overset{(146)}{=} b_0(s)\prod_{i\in\mathcal{N}} \mathbb{P}_{\pi^\mu}(A_{i,0} = a_i)\\
&= b_0(s)\mathbb{P}_{\pi^\mu}(A_0 = a) = \mathbb{P}_{\pi^\mu}(S_0 = s, A_0 = a). \qquad (147)
\end{aligned}$$

Here, we also use (i) the fact that that the initial state distribution does not depend on the policy, and (ii) the fact that μ has independent local policies and thus also the Z_i are independent in the probability space $\mathbb{P}_\mu$.

Next, assume that $0 \le t-1 \le T$ and $\mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) = \mathbb{P}_{\pi^\mu}(H_{t-1} = \tau_{t-1})$ for any $\tau_{t-1} \in \mathcal{H}_{t-1}$. Let $\tau_{t-1} = (s_0, a_0, r_0, s_1, \dots, r_{t-1}) \in \mathcal{H}_{t-1}$ arbitrary such that $\mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) = \mathbb{P}_{\pi^\mu}(H_{t-1} = \tau_{t-1}) > 0$. As in the proof of Lemma 59, it follows from considering the d-separation criterion on a Bayesian graph of the random variables defined on $\Pi \times \Omega$ that conditioning on H_{t-1} does not make the Z_i dependent (see Pearl, 2009, ch. 1.2). The same criterion also says that, after conditioning on $A_{i,t-1}$ and $\overline{AO}_{i,t-1}$, one cannot gain additional information about Z_i from the other components of H_{t-1} or from $O_{i,t}$ (that is, $A_{i,t-1}$ and $\overline{AO}_{i,t-1}$ d-separate Z_i from the rest of the history). Using this in (151), it follows for arbitrary s_t, o_t, a_t, r_t that

$$
\begin{aligned}
&\mathbb{P}_\mu(S_t = s_t, O_t = o_t, A_t = a_t, R_t = r_t \mid H_{t-1} = \tau_{t-1}) && (148)\\
&= \mathbb{E}_\mu[\mathbb{P}_\mu(S_t = s_t, O_t = o_t, A_t = a_t, R_t = r_t \mid H_{t-1}, Z) \mid H_{t-1} = \tau_{t-1}] && (149)\\
&= \mathbb{E}_\mu[P(s_t \mid s_{t-1}, a_{t-1}) O(o_t \mid s_t, a_{t-1}) \prod_{i\in\mathcal{N}} Z_i(a_{i,t} \mid \tau_{i,t}) \mathbb{1}_{\mathcal{R}(s,a)=r} \mid H_{t-1} = \tau_{t-1}] && (150)\\
&= P(s_t \mid s_{t-1}, a_{t-1}) O(o_t \mid s, a_{t-1}) \\
&\quad \prod_{i\in\mathcal{N}} \mathbb{E}_\mu[Z_i(a_{i,t} \mid \tau_{i,t}) \mid A_{i,t-1} = a_{i,t-1}, \overline{AO}_{i,t-1} = \tau_{i,t-1}] \mathbb{1}_{\mathcal{R}(s_t,a_t)=r_t} && (151)\\
&= P(s_t \mid s_{t-1}, a_{t-1}) O(o_t \mid s, a_{t-1}) \prod_{i\in\mathcal{N}} \mathbb{P}_\mu(A_{i,t} = a_{i,t} \mid \overline{AO}_{i,t} = \tau_{i,t}) \mathbb{1}_{\mathcal{R}(s_t,a_t)=r_t} && (152)\\
&= P(s_t \mid s_{t-1}, a_{t-1}) O(o_t \mid s, a_{t-1}) \prod_{i\in\mathcal{N}} \mathbb{P}_{\pi^\mu}(A_{i,t} = a_{i,t} \mid \overline{AO}_{i,t} = \tau_{i,t}) \mathbb{1}_{\mathcal{R}(s_t,a_t)=r_t} && (153)\\
&= \mathbb{P}_{\pi^\mu}(S_t = s_t, O_t = o_t, A_t = a_t, R_t = r_t \mid H_{t-1} = \tau_{t-1}). && (154)
\end{aligned}
$$

In (153), we again use Equation 146, which is possible since $\mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) > 0$ implies that also $\mathbb{P}_\mu(\overline{AO}_{i,t-1} = \tau_{i,t-1}) > 0$, where $\tau_{i,t-1}$ is defined as the projection of τ_{t-1} onto $\overline{\mathcal{AO}}_{i,t-1}$.

To conclude the inductive step, let $\tau_t \in \mathcal{H}_t$ and choose τ_{t-1} as the projection of τ_t onto $\mathcal{H}_{t-1}$. If

$$\mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) = \mathbb{P}_{\pi^\mu}(H_{t-1} = \tau_{t-1}) = 0,$$

necessarily also $\mathbb{P}_\mu(H_t = \tau_t) = \mathbb{P}_{\pi^\mu}(H_t = \tau_t) = 0$ and there is nothing more to show. Assume now that this is not the case. Using the inductive hypothesis and Equations (148)–(154) in (*), it then follows that

$$
\begin{aligned}
\mathbb{P}_\mu(H_t = \tau_t) &= \mathbb{P}_\mu(H_t = \tau_t \mid H_{t-1} = \tau_{t-1}) \mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) \\
&= \mathbb{P}_\mu(S_t = s_t, O_t = o_t, A_t = a_t, R_t = r_t \mid H_{t-1} = \tau_{t-1}) \mathbb{P}_\mu(H_{t-1} = \tau_{t-1}) \\
&\overset{(*)}{=} \mathbb{P}_{\pi^\mu}(S_t = s_t, O_t = o_t, A_t = a_t, R_t = r_t \mid H_{t-1} = \tau_{t-1}) \mathbb{P}_{\pi^\mu}(H_{t-1} = \tau_{t-1}) \\
&= \mathbb{P}_{\pi^\mu}(H_t = \tau_t). \qquad (155)
\end{aligned}
$$

This concludes the induction. In particular, it follows that

$$\mathbb{P}_\mu(H_T = \tau_T) = \mathbb{P}_{\pi^\mu}(H_T = \tau_T).$$

Hence, $H = H_T$ has the same distribution under both $\mathbb{P}_\mu$ and $\mathbb{P}_{\pi^\mu}$.

Turning to the “in particular” statement, we can use the tower property (i) and Equation 119 (ii) to follow that

$$
\begin{aligned}
J(\pi^\mu) &= \mathbb{E}_{\pi^\mu}[\sum_{t=1}^h R_t] = \mathbb{E}_\mu[\sum_{t=1}^h R_t] \\
&\overset{(i)}{=} \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\sum_{t=1}^h R_t \,\middle|\, Z\right]\right] \overset{(ii)}{=} \mathbb{E}_\mu\left[\mathbb{E}_Z\left[\sum_{t=1}^h R_t\right]\right] \\
&= \mathbb{E}_{\pi\sim\mu}[J(\pi)] = J(\mu). \qquad (156)
\end{aligned}
$$

□

Before we turn to the OP distribution, we prove another useful Lemma, stating that the mapping from distributions to corresponding joint policies and the pushforward by isomorphisms commute. That is, the policy corresponding to the pushforward of a distribution and the pushforward of the policy corresponding to that distribution are the same.

Lemma 61. *Let D, E be isomorphic Dec-POMDPs with isomorphism $f \in \mathrm{Iso}(D, E)$. Let $\pi \in \Pi^D$ and let $\mu \in \Delta(\Pi^D)$ be any distribution with independent local policies. Then*

$$f^*\pi^\mu = \pi^{f^*\mu}.$$

Proof. First, we have to find an expression for the marginal distributions $(f^*\mu)_i$ and prove that $f^*\mu$ is a distribution with independent local policies. To that end, consider measurable sets $\mathcal{Z}_i \subseteq \Pi_i^E$ for $i \in \mathcal{N}^E$ and define $\mathcal{Z} := \prod_{i\in\mathcal{N}^E} \mathcal{Z}_i$. In the following, we adopt the notation $\pi_i \circ f := \pi_i(f\cdot \mid f\cdot)$ and $\mathcal{Z}_i \circ f := \{\pi_i \circ f \mid \pi_i \in \mathcal{Z}_i\}$. Note that

$$\begin{aligned}(f^*)^{-1}(\mathcal{Z}) &= \{\pi \mid f^*\pi \in \mathcal{Z}\} = \{\pi \mid \forall i \in \mathcal{N}^E \colon \pi_{f^{-1}i} \circ f^{-1} \in \mathcal{Z}_i)\} \\ &= \{\pi \mid \forall j \in \mathcal{N}^D \colon \pi_j \in \mathcal{Z}_{fj} \circ f\} = \prod_{j\in\mathcal{N}^D} \mathcal{Z}_{fj} \circ f. \end{aligned} \tag{157}$$

Hence, using in (i) that μ has independent local policies, it follows that

$$(f^*\mu)(\mathcal{Z}) = \mu((f^*)^{-1}(\mathcal{Z})) \overset{(157)}{=} \mu(\prod_{j\in\mathcal{N}^D} \mathcal{Z}_{fj} \circ f) \overset{\text{(i)}}{=} \prod_{j\in\mathcal{N}^D} \mu_j(\mathcal{Z}_{fj} \circ f) = \prod_{i\in\mathcal{N}^E} \mu_{f^{-1}i}(\mathcal{Z}_i \circ f). \tag{158}$$

Now let $i \in \mathcal{N}^E$ arbitrary. With the choice of $\hat{\mathcal{Z}}_i := \mathcal{Z}_i$ and $\hat{\mathcal{Z}}_k := \Pi_k^E$ for all $k \in \mathcal{N}^E \setminus \{i\}$, it is

$$\begin{aligned}(f^*\mu)_i(\mathcal{Z}_i) &= f^*\mu(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) = f^*\mu(\hat{\mathcal{Z}}_1 \times \cdots \times \hat{\mathcal{Z}}_N)) \\ &\overset{(158)}{=} \prod_{k\in\mathcal{N}^E} \mu_{f^{-1}k}(\hat{\mathcal{Z}}_k \circ f) = \mu_{f^{-1}i}(\mathcal{Z}_i \circ f), \end{aligned} \tag{159}$$

as $\mu_{f^{-1}k}(\hat{\mathcal{Z}}_k \circ f) = \mu_{f^{-1}k}(\Pi_k \circ f) = \mu_{f^{-1}k}(\Pi_{f^{-1}k}) = 1$ for any $k \in \mathcal{N}^E \setminus \{i\}$.

Since i was arbitrary, this shows that $f^*\mu(\mathcal{Z}) = \prod_{i\in\mathcal{N}^E} (f^*\mu)_i(\mathcal{Z}_i)$. Since the sets $\mathcal{Z}_i$ were arbitrary and the Cartesian products of these sets are a π-system and generate the product σ-Algebra $\mathcal{F}$, this shows that $f^*\mu$ has independent local policies.

Next, let $i \in \mathcal{N}^E$, $t \in \{0, \ldots, T\}$ and $\tau_{i,t} \in \overline{\mathcal{AO}}^E_{i,t}$. Note that $\mathrm{proj}_i(f^*\pi^\mu) = \pi_{f^{-1}i}^{\mu_{f^{-1}i}} \circ f^{-1}$ and $\mathrm{proj}_i(\pi^{f^*\mu}) = \pi_i^{(f^*\mu)_i}$. Hence, it remains to prove that $\pi_{f^{-1}i}^{\mu_{f^{-1}i}} \circ f^{-1} = \pi_i^{(f^*\mu)_i}$.

To that end, let $j \in \mathcal{N}^D$ such that $fj = i$. By Theorem 22, it is

$$\mathbb{P}_\pi(\overline{\mathcal{AO}}_{j,t} = f^{-1}\tau_{i,t}) = \mathbb{P}_{f^*\pi}(\overline{\mathcal{AO}}_{i,t} = \tau_{i,t}) \tag{160}$$

for any $\pi \in \Pi^D$. Letting, $\mu_{-j} \in \Delta(\Pi^D_{-j})$ with independent local policies arbitrary and defining $\mu := \mu_i \otimes \mu_{-i}$, this means that also

$$\mathbb{P}_\mu(\overline{AO}_{j,t} = f^{-1}\tau_{i,t}) \overset{\text{(ii)}}{=} \mathbb{E}_\mu\left[\mathbb{P}_Z(\overline{AO}_{j,t} = f^{-1}\tau_{i,t})\right] = \mathbb{E}_\mu\left[\mathbb{P}_{f^*Z}(\overline{AO}_{i,t} = \tau_{i,t})\right]$$
$$= \mathbb{E}_{f^*\mu}\left[\mathbb{P}_Z(\overline{AO}_{i,t} = \tau_{i,t})\right] \overset{\text{(ii)}}{=} \mathbb{P}_{f^*\mu}(\overline{AO}_{i,t} = \tau_{i,t}), \quad (161)$$

where we use Equation 122 in (ii). Since f^* is a bijection on the space of policies, it is

$$\{f^*(\mu_j \otimes \mu_{-j}) \mid \mu_{-j} \in \Delta(\Pi^D_{-j})\} = \{(f^*\mu)_i \otimes \hat{\mu}_{-i} \mid \hat{\mu}_{-i} \in \Delta(\Pi^E_{-i})\}.$$

Hence, (161) implies that there is some $\mu_{-j} \in \Delta(\Pi^D_{-j})$ with independent local policies such that $\mathbb{P}_{\mu_j\otimes\mu_{-j}}(\overline{AO}_{j,t} = f^{-1}\tau_{i,t}) > 0$ if and only if there is $\hat{\mu}_{-i} \in \Delta(\Pi^E_{-i})$ with independent local policies such that $\mathbb{P}_{(f^*\mu)_i\otimes\hat{\mu}_{-i}}(\overline{AO}_{i,t} = \tau_{i,t}) > 0$.

Using this fact, it suffices to distinguish the two cases where such a distribution does exist and where it does not exist. First, assume that it does not exist. Then by Definition 58, both $\pi_j^{\mu_j}(\cdot \mid f^{-1}\tau_{i,t})$ and $\pi_i^{(f^*\mu)_i}(\cdot \mid \tau_{i,t})$ are uniform distributions. Since f^{-1} is a bijection on $\mathcal{A}_i^E$, also $\pi_j^{\mu_j}(f^{-1}\cdot \mid f^{-1}\tau_{i,t})$ is a uniform distribution, and hence

$$\pi_j^{\mu_j}(f^{-1}\cdot \mid f^{-1}\tau_{i,t}) = \pi_i^{(f^*\mu)_i}(\cdot \mid \tau_{i,t}).$$

Second, consider the case in which a distribution $\mu_{-j} \in \Delta(\Pi^D_{-j})$ with independent local policies exists such that

$$\mathbb{P}_{\mu_j\otimes\mu_{-j}}(\overline{AO}_{j,t} = f^{-1}\tau_{i,t}) > 0,$$

and define $\mu := \mu_j \otimes \mu_{-j}$. Then for $a_i \in \mathcal{A}_i^E$, it is

$$\pi_j^{\mu_j}(f^{-1}a_i \mid f^{-1}\tau_{i,t}) = \mathbb{P}_\mu(A_{j,t} = f^{-1}a_i \mid \overline{AO}_{j,t} = f^{-1}\tau_{i,t}) \quad (162)$$
$$= \mathbb{E}_\mu[\mathbb{P}_\mu(A_{j,t} = f^{-1}a_i \mid \overline{AO}_{j,t}, Z) \mid \overline{AO}_{j,t} = f^{-1}\tau_{i,t}] \quad (163)$$
$$= \mathbb{E}_\mu[Z_j(f^{-1}a_i \mid f^{-1}\tau_{i,t}) \mid \overline{AO}_{j,t} = f^{-1}\tau_{i,t}] \quad (164)$$
$$= \frac{\int_{\Pi^D}\int_\Omega (f^*Z)_i(a_i \mid \tau_{i,t})\mathbb{1}_{\overline{AO}_{j,t}=f^{-1}\tau_{i,t}}\mathrm{d}\mathbb{P}_Z\mathrm{d}\mu}{\int_{\Pi^D}\int_\Omega \mathbb{1}_{\overline{AO}_{j,t}=f^{-1}\tau_{i,t}}\mathrm{d}\mathbb{P}_Z\mathrm{d}\mu} \quad (165)$$
$$= \frac{\int_{\Pi^D}(f^*Z)_i(a_i \mid \tau_{i,t})\mathbb{P}_Z(\overline{AO}_{j,t} = f^{-1}\tau_{i,t})\mathrm{d}\mu}{\int_{\Pi^D}\mathbb{P}_Z(\overline{AO}_{j,t} = f^{-1}\tau_{i,t})\mathrm{d}\mu} \quad (166)$$
$$= \frac{\int_{\Pi^D}(f^*Z)_i(a_i \mid \tau_{i,t})\mathbb{P}_{f^*Z}(\overline{AO}_{i,t} = \tau_{i,t})\mathrm{d}\mu}{\int_{\Pi^D}\mathbb{P}_{f^*Z}(\overline{AO}_{i,t} = \tau_{i,t})\mathrm{d}\mu} \quad (167)$$
$$= \frac{\int_{\Pi^E} Z_i(a_i \mid \tau_{i,t})\mathbb{P}_Z(\overline{AO}_{i,t} = \tau_{i,t})\mathrm{d}\mu\circ(f^*)^{-1}}{\int_{\Pi^E}\mathbb{P}_Z(\overline{AO}_{i,t} = \tau_{i,t})\mathrm{d}\mu\circ(f^*)^{-1}} \quad (168)$$
$$= \mathbb{E}_{f^*\mu}[Z_i(a_i \mid \tau_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}] \quad (169)$$
$$= \mathbb{P}_{f^*\mu}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) \quad (170)$$
$$= \pi_i^{(f^*\mu)_i}(a_i \mid \tau_{i,t}), \quad (171)$$

where we have used Definition 58 in (162) and (171), and Theorem 22 in (167). This concludes the second case and thus the proof.

□

D.3. The other-play distribution and the symmetrizer

Using the idea of a policy corresponding to a distributions over policies introduced above, we can now define a policy corresponding to the distribution over policies used in the OP objective. In the following, fix again a Dec-POMDP D.

Definition 62 (Other-play distribution)**.** Let $\pi \in \Pi$. We define the *OP distribution* of π as the distribution

$$\mu := |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi}, \tag{172}$$

where δ is the Dirac measure, i.e., for any measurable set $\mathcal{Z} \subseteq \Pi$, it is

$$\delta_{\mathbf{g}^*\pi}(\mathcal{Z}) = \begin{cases} 1 & \text{if } \mathbf{g}^*\pi \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, agent i chooses one of the automorphisms $\mathbf{g}_i \in \mathrm{Aut}(D)$ uniformly at random in the beginning of an episode and then follows the local policy $\mathrm{proj}_i(\mathbf{g}_i^*\pi)$. It can easily be shown that this distribution has independent local policies.

Lemma 63. *Let $\pi \in \Pi$ and let μ be the OP distribution of π. Then μ has independent local policies.*

Proof. Let $\mathcal{Z}_i \subseteq \Pi_i$ measurable for $i \in \mathcal{N}$ and let $\mathcal{Z} := \prod_{i\in\mathcal{N}} \mathcal{Z}_i$. Note that for any $i \in \mathcal{N}$ and $\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}$, it is

$$\delta_{\mathbf{g}^*\pi}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) = \prod_{j\in\mathcal{N}\setminus\{i\}} \delta_{\mathrm{proj}_j(\mathbf{g}_j^*\pi)}(\Pi_j)\delta_{\mathrm{proj}_i(\mathbf{g}_i^*\pi)}(\mathcal{Z}_i) = \delta_{\mathrm{proj}_i(\mathbf{g}_i^*\pi)}(\mathcal{Z}_i). \tag{173}$$

Hence, it follows that

$$\begin{aligned} \mu(\mathcal{Z}) = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}_i\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi}(\mathcal{Z}) &= \prod_{i\in\mathcal{N}} \sum_{\mathbf{g}_i\in\mathrm{Aut}(D)} |\mathrm{Aut}(D)|^{-1}\delta_{\mathrm{proj}_i(\mathbf{g}_i^*\pi)}(\mathcal{Z}_i)) \\ &= \prod_{i\in\mathcal{N}} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} |\mathrm{Aut}(D)|^{-(N-1)}|\mathrm{Aut}(D)|^{-1}\delta_{\mathrm{proj}_i(\mathbf{g}_i^*\pi)}(\mathcal{Z}_i)) \\ &\overset{(173)}{=} \prod_{i\in\mathcal{N}} |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) = \prod_{i\in\mathcal{N}} \mu_i(\mathcal{Z}_i). \end{aligned} \tag{174}$$

This shows that $\mu = \otimes_{i\in\mathcal{N}}\mu_i$. □

Using the OP distribution of a policy, we can define the symmetrizer Ψ^D for D, which maps a policy π to a policy $\Psi^D(\pi)$ that corresponds to the OP distribution of π. If it is clear which Dec-POMDP is considered, we also write $\Psi(\pi)$.

Definition 64 (Symmetrizer)**.** We define the *symmetrizer* for the Dec-POMDP D as the map $\Psi^D : \Pi^D \to \Pi^D$ such that for any policy $\pi \in \Pi$ and OP distribution μ of π, it is

$$\Psi^D(\pi) := \pi^\mu.$$

It is clear that if a policy is already invariant to automorphism, then $\pi_i = \Psi_i(\pi)$ for $i \in \mathcal{N}$, excluding action-observation histories that can never be reached under π_i. We formulate a slightly weaker proposition below, which is easier to prove.

Proposition 65. *Let $\pi \in \Pi$ be invariant to automorphism, and assume that, for all $t \in \{0, \ldots, T\}$ and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$, it is $\mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t}) > 0$. Then it is $\Psi(\pi) = \pi$.*

Proof. Let μ be the OP distribution of π. Since π is invariant to automorphism, it is

$$\mu = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi} = \delta_\pi.$$

Now let $i \in \mathcal{N}$, $t \in \{0, \ldots, T\}$, $a_i \in \mathcal{A}_i$ and $\tau_{i,t} \in \overline{\mathcal{AO}}_{i,t}$ arbitrary. Using Equation 122, it follows that

$$\mathbb{P}_{\delta_\pi}(\overline{AO}_{i,t} = \tau_{i,t}) \overset{(122)}{=} \mathbb{E}_{\delta_\pi}\left[\mathbb{P}_Z(\overline{AO}_{i,t} = \tau_{i,t})\right] = \mathbb{P}_\pi(\overline{AO}_{i,t} = \tau_{i,t}) > 0$$

and

$$\mathbb{P}_{\delta_\pi}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) \overset{(122)}{=} \mathbb{E}_{\delta_\pi}\left[\mathbb{P}_Z(A_{i,t} = a_i \mid \overline{AO}_{i,t}) \mid \overline{AO}_{i,t} = \tau_{i,t}\right] = \mathbb{P}_\pi(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}). \quad (175)$$

Hence, we can apply Definition 58 and conclude that

$$\Psi_i(\pi)(a_i \mid \tau_{i,t}) = \pi_i^{\mu_i}(a_i \mid \tau_{i,t}) \overset{\text{Definition 58}}{=} \mathbb{P}_{\delta_\pi}(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) \overset{(175)}{=} \mathbb{P}_\pi(A_{i,t} = a_i \mid \overline{AO}_{i,t} = \tau_{i,t}) = \pi_i(a_i \mid \tau_{i,t}). \quad (176)$$

□

It will be helpful to refer to policies as equivalent if they have the same image under Ψ.

Definition 66. Let $\pi, \pi' \in \Pi^D$. We say that π and π' are equivalent, denoted as $\pi \equiv_D \pi$, if $\Psi^D(\pi) = \Psi^D(\pi')$. Moreover, we write $[\pi] := \{\pi' \mid \pi' \equiv_D \pi\}$ for the equivalence class of π.

It is clear that $\equiv_D$ is an equivalence relation, since it is induced by the function Ψ^D. It follows that under $\equiv_D$, Π^D decomposes into a partition of equivalence classes, denoted by $\Pi^D/\equiv_D$.

Applying Lemma 61 to the symmetrizer in particular, we can show that it commutes with isomorphisms, and that the policy $\Psi(\pi)$ is invariant to automorphism.

Corollary 67. *Let $f \in \mathrm{Iso}(D, E)$ and $\pi \in \Pi^D$. Then it is*

$$f^*\Psi^D(\pi) = \Psi^E(f^*\pi).$$

If $E = D$, then

$$f^*\Psi^D(\pi) = \Psi^D(\pi),$$

i.e., $\Psi^D(\pi)$ is invariant to automorphism.

Proof. Let μ be the other-play distribution of π and $\hat{\mu}$ the other-play distribution corresponding to $f^*\pi$. Then, using the associativity of function composition and pushforward proven in Lemma 21, and using the "in particular" part of Lemma 19, it follows that

$$\begin{aligned}\hat{\mu} = |\mathrm{Aut}(E)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(E)^{\mathcal{N}}} \delta_{\mathbf{g}^*(f^*\pi)} = |\mathrm{Aut}(E)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(E)^{\mathcal{N}}} \delta_{(\mathrm{proj}_i(\mathbf{g}_i^* f^*\pi))_{i\in\mathcal{N}}} \\ \overset{\text{Lemma 19}}{=} |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{(\mathrm{proj}_i(f^*\mathbf{g}_i^*\pi))_{i\in\mathcal{N}}} = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{f^*(\mathbf{g}^*\pi)} \\ = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi} \circ (f^*)^{-1} = \mu \circ (f^*)^{-1}. \end{aligned} \tag{177}$$

Thus, using Lemma 61, it is

$$f^*\Psi^D(\pi) = f^*\pi^{\mu} \overset{\text{Lemma 61}}{=} \pi^{f^*\mu} = \pi^{\mu\circ(f^*)^{-1}} \overset{(177)}{=} \pi^{\hat{\mu}} = \Psi^E(f^*\pi) \tag{178}$$

Finally, assume $E = D$. Then f is an automorphism and $\mathrm{Aut}(E) = \mathrm{Aut}(D) = \mathrm{Aut}(D) \circ f$ by Lemma 19. Hence,

$$\begin{aligned}\mu \circ (f^*)^{-1} = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi} \circ (f^*)^{-1} \\ = |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{(\mathrm{proj}_i(f^*\mathbf{g}_i^*\pi))_{i\in\mathcal{N}}} \\ \overset{\text{Lemma 19}}{=} |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \delta_{\mathbf{g}^*\pi} = \mu. \end{aligned} \tag{179}$$

By (178), it follows that

$$f^*\Psi^D(\pi) \overset{(178)}{=} \pi^{\mu\circ(f^*)^{-1}} \overset{(179)}{=} \pi^{\mu} = \Psi^D(\pi),$$

which concludes the proof. □

A direct corollary is that we can define the pushforward purely in terms of equivalence classes of policies. This will also be useful later.

Definition 68. Let D, E be isomorphic Dec-POMDPs with $f \in \mathrm{Iso}(D, E)$. Let $[\pi] \in \Pi^D/\equiv_D$. We define the pushforward equivalence class $f^*[\pi] \in \Pi_E/\equiv_E$ via

$$f^*[\pi] := [f^*\pi].$$

The following corollary show that this is well-defined, that the pushforward of an equivalence class does not depend on the particular chosen isomorphism, and that it is compatible with function composition.

Corollary 69. *(i) The pushforward of an equivalence class is well-defined, i.e., for any $\pi, \pi' \in \Pi^D$ such that $\pi \equiv_D \pi'$, it is $[f^*\pi] = [f^*\pi']$.*

(ii) Any two isomorphisms $f, f' \in \mathrm{Iso}(D, E)$ induce the same pushforward.

(iii) Analogous results to those in Lemma 21 apply to the pushforward of equivalence classes.

Proof. First, let $\pi \equiv_D \pi' \in \Pi^D$ and $f \in \mathrm{Iso}(D, E)$ arbitrary. Then, using Corollary 67 and the definition of $\equiv_D$, it is

$$\Psi^E(f^*\pi) = f^*\Psi^D(\pi) = f^*\Psi^D(\pi') = \Psi^E(f^*\pi').$$

Thus, $[f^*\pi] = [f^*\pi']$, which proves the first part.

Second, let $f, \tilde{f} \in \mathrm{Iso}(D, E)$ and $\pi \in \Pi^D$ arbitrary. By Lemma 19, there then exists $g \in \mathrm{Aut}(E)$ such that $\tilde{f} = g \circ f$. Hence, using the second and first part of Corollary 67 and Lemma 21, it is

$$\Psi^E(f^*\pi) \overset{\text{Corollary 67}}{=} g^*\Psi^E(f^*\pi) \overset{\text{Corollary 67}}{=} \Psi^E(g^*(f^*\pi)) \overset{\text{Lemma 21}}{=} \Psi^E((g \circ f)^*\pi) = \Psi^E(\hat{f}^*\pi). \tag{180}$$

Finally, it follows that

$$f^*[\pi] = [f^*\pi] \overset{180}{=} [\tilde{f}^*\pi] = \tilde{f}^*[\pi],$$

which concludes the second part.

The third part follows directly from Lemma 21 by using the definition of the pushforward of equivalence classes. □

In the following, we say that two equivalence classes $[\pi], [\pi']$ for $\pi \in \Pi^D, \pi' \in \Pi^E$ correspond to each other if there exists an isomorphism $f \in \mathrm{Iso}(D, E)$ such that $f^*[\pi] = [\pi']$. In that case, in a slight abuse of the terms, we also say that π and π' are equivalent, extending the equivalence between policies defined above to policies for different Dec-POMDPs. Using Corollary 69, one can see that two policies $\pi, \pi' \in \Pi^D$ are equivalent in the sense that $[\pi] = [\pi']$ if and only if there exists an isomorphism $f \in \mathrm{Iso}(D, D)$ such that $f^*[\pi] = [\pi']$, so this extended notion is equivalent to the old one for two policies $\pi, \pi' \in \Pi^D$. We continue to reserve the notation $[\pi]$ and $\equiv$ for policies from the same Dec-POMDP.

D.4. Main characterizations

Having defined the symmetrizer Ψ, we can now characterize the OP objective as transforming a policy π into an invariant policy $\Psi(\pi)$ and evaluating the expected return of that policy. This means that we can "pass to the quotient" and consider the OP objective as a map $\tilde{J}_{\mathrm{OP}}$ of equivalence classes, $\tilde{J}_{\mathrm{OP}}([\pi]) := J_{\mathrm{OP}}(\pi)$ for $[\pi] \in \Pi/\equiv$, using the equivalence relation on policies introduced above. The result is essentially a rigorous version of Hu et al. (2020)'s Proposition 1 in our setup. It will help us later to analyze the OP-optimal policies in a given example, as it implies that we can restrict ourselves to considering representatives $\Psi(\pi)$ of equivalence classes.

Theorem 70. *Let D be a Dec-POMDP, let $\pi \in \Pi^D$, and let Ψ be the symmetrizer for D. Then $\Psi(\pi)$ is invariant to automorphism, and it is*

$$J^D_{\mathrm{OP}}(\pi) = J^D(\Psi(\pi)). \tag{181}$$

In particular, we can consider the OP objective as a function of equivalence classes $[\pi] \in \Pi^D/\equiv_D$, and if there exists an optimal policy for the OP objective, then there also exists an optimal policy that is invariant to automorphism.

Proof. In the following, fix a Dec-POMDP D. Let $\pi \in \Pi$ and let μ be the OP distribution of π, such that $\Psi(\pi) = \pi^\mu$. Then by the second part of Corollary 67, $\Psi(\pi)$ is invariant to automorphism. Moreover, using

Proposition 60, it is

$$\begin{aligned}
J_{\mathrm{OP}}(\pi) = \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J(\mathbf{g}^*\pi)\right] &= \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} |\mathrm{Aut}(D)|^{-N} J(\mathbf{g}^*\pi) \\
&= \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} |\mathrm{Aut}(D)|^{-N} \int_{\pi'\in\Pi} J(\pi')\mathrm{d}\left(\delta_{\mathbf{g}^*\pi}\right) \\
&= \int_{\pi'\in\Pi} J(\pi')\mathrm{d}\left(\sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} |\mathrm{Aut}(D)|^{-N}\delta_{\mathbf{g}^*\pi}\right) \\
&\overset{(172)}{=} \int_{\pi'\in\Pi} J(\pi')\mathrm{d}\mu \overset{(117)}{=} J(\mu) \overset{\text{Proposition 60}}{=} J(\Psi(\pi)), \qquad (182)
\end{aligned}$$

which proves Equation 181.

Turning to the "in particular" statement, let $\pi' \equiv \pi$ for a second policy $\pi' \in \Pi$. By Definition 66, this means that $\Psi(\pi') = \Psi(\pi)$. Hence, by Equation 182, it follows that $J_{\mathrm{OP}}(\pi) = J_{\mathrm{OP}}(\pi')$, which shows that the function $\tilde{J}_{\mathrm{OP}} \colon \Pi/\!\equiv \,\to \mathbb{R}, [\pi] \mapsto J_{\mathrm{OP}}(\pi)$ is well-defined.

Lastly, assume that there is $\pi \in \arg\max_{\pi'\in\Pi} J_{\mathrm{OP}}(\pi')$ (see Remark 53 regarding the existence of such a policy). Then by Equation 182, it is also $\Psi(\pi) \in \arg\max_{\pi'\in\Pi} J_{\mathrm{OP}}(\pi')$, so $\Psi(\pi)$ is an OP-optimal policy that is invariant to automorphism. □

As a corollary, we can show that isomorphisms do not affect the OP value of a policy (we already know this about SP from Corollary 23). In the following, we define $\Pi^D_{\mathrm{OP}} := \arg\max_{\pi\in\Pi^D} J^D_{\mathrm{OP}}(\pi)$ for a Dec-POMDP D.

Corollary 71. *Let D, E be isomorphic Dec-POMDPs with $f \in \mathrm{Iso}(D,E)$, and let $\pi \in \Pi^D$. Then it is*

$$J^D_{\mathrm{OP}}(\pi) = J^E_{\mathrm{OP}}(f^*\pi).$$

In particular, if $\pi \in \Pi^D_{\mathrm{OP}}$, then also $f^*\pi \in \Pi^E_{\mathrm{OP}}$.

Proof. Using Theorem 70, Corollary 67, and Corollary 23, it is

$$\begin{aligned}
J^E_{\mathrm{OP}}(f^*\pi) \overset{\text{Theorem 70}}{=} J^E(\Psi^E(f^*\pi)) \overset{\text{Corollary 67}}{=} J^E(f^*\Psi^D(\pi)) \\
\overset{\text{Corollary 23}}{=} J^D(\Psi^D(\pi)) \overset{\text{Theorem 70}}{=} J^D_{\mathrm{OP}}(\pi). \qquad (183)
\end{aligned}$$

Turning to the "in particular" statement, assume that $\pi \in \Pi^D_{\mathrm{OP}}$. By Lemma 18, it is $f^{-1} \in \mathrm{Iso}(E,D)$. Hence, for any $\tilde{\pi} \in \Pi^E$, it follows from the preceding that

$$J^E_{\mathrm{OP}}(\tilde{\pi}) = J^D_{\mathrm{OP}}((f^{-1})^*\tilde{\pi}) \leq J^D_{\mathrm{OP}}(\pi) = J^E_{\mathrm{OP}}(f^*\pi). \qquad (184)$$

This shows that $f^*\pi \in \Pi^E_{\mathrm{OP}}$. □

Finally, we turn to the connection between OP and the payoff in an LFC game. The following result will be helpful in both showing the inadequacy of OP and in proving that OP with tie-breaking is optimal. It shows that equivalent policies in $[\pi] \in \Pi/\!\equiv$ are all compatible when played against each other by different principals in the LFC game. The proof is based on Lemma 41 and Proposition 60.

Theorem 72. *Let D be a Dec-POMDP, let Ψ be the symmetrizer for D, and define $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$. Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$. For any $E \in \mathcal{D}$, choose $f_{D,E} \in \mathrm{Iso}(D, E)$ arbitrarily. Then it is*

$$U^D(\boldsymbol{\sigma}) = \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[J^D\left(\left(\Psi_k(\pi^{(k)})\right)_{k\in\mathcal{N}} \right) \right] \right]. \tag{185}$$

Proof. First, consider arbitrary joint policies $\pi^{(1)}, \dots, \pi^{(N)} \in \Pi^D$ and let $\mu^{(i)}$ be the OP distribution of $\pi^{(i)}$, so that $\pi^{\mu^{(i)}} = \Psi(\pi^{(i)})$ for $i \in \mathcal{N}$. Define the distribution

$$\hat{\mu}(\pi^{(1)}, \dots, \pi^{(N)}) := |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g} \in \mathrm{Aut}(D)^{\mathcal{N}}} \otimes_{i\in\mathcal{N}} \delta_{\mathrm{proj}_i(\mathbf{g}_i^* \pi^{(i)})} \tag{186}$$

as a function of $\pi^{(1)}, \dots, \pi^{(N)}$. It can easily be seen that $\hat{\mu}_(\pi^{(1)}, \dots, \pi^{(N)}) \in \Delta(\Pi^D)$ and that it has independent local policies. Moreover, $\hat{\mu}(\pi^{(1)}, \dots, \pi^{(N)})_i = \mu_i^{(i)}$, i.e., the marginal distribution for agent $i \in \mathcal{N}$ is equal in $\hat{\mu}(\pi^{(1)}, \dots, \pi^{(N)})$ and $\mu^{(i)}$. Hence, also the corresponding local policies are identical, that is,

$$\pi_i^{\hat{\mu}(\pi^{(1)},\dots,\pi^{(N)})_i} = \pi_i^{\mu_i^{(i)}} = \Psi_i(\pi^{(i)}) \tag{187}$$

for $i \in \mathcal{N}$.

It follows that

$$
\begin{aligned}
& U^D(\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_N) && (188)\\
&= \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[\mathbb{E}_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \left[J^D((\mathrm{proj}_k(\mathbf{g}_k^*\pi^{(k)}))_{k\in\mathcal{N}}) \right]\right]\right] && (189)\\
&= \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \Bigg[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \Bigg[&& (190)\\
&\qquad |\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \int_{\pi\in\Pi^D} J^D(\pi) \mathrm{d}\left(\otimes_{k\in\mathcal{N}} \delta_{\mathrm{proj}_k(\mathbf{g}_k^*\pi^{(k)})} \right) \Bigg]\Bigg] && (191)\\
&= \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \Bigg[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \Bigg[&& (192)\\
&\qquad \int_{\pi\in\Pi^D} J^D(\pi) \mathrm{d}\left(|\mathrm{Aut}(D)|^{-N} \sum_{\mathbf{g}\in\mathrm{Aut}(D)^{\mathcal{N}}} \otimes_{k\in\mathcal{N}} \delta_{\mathrm{proj}_k(\mathbf{g}_k^*\pi^{(k)})} \right) \Bigg]\Bigg] && (193)\\
&\overset{(186)}{=} \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[\int_{\pi\in\Pi^D} J^D(\pi) \mathrm{d}\hat{\mu}(\pi^{(1)}, \dots, \pi^{(N)}) \right]\right] && (194)\\
&\overset{(117)}{=} \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[J^D(\hat{\mu}(\pi^{(1)}, \dots, \pi^{(N)})) \right]\right] && (195)\\
&= \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[J^D((\pi_k^{\hat{\mu}(\pi^{(1)},\dots,\pi^{(N)})_k})_{k\in\mathcal{N}}) \right]\right] && (196)\\
&\overset{(187)}{=} \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j,D}\boldsymbol{\sigma}_j(D_j),\, j\in\mathcal{N}} \left[J^D\left(\left(\Psi_k(\pi^{(k)})\right)_{k\in\mathcal{N}} \right) \right]\right], && (197)
\end{aligned}
$$

where we have used Lemma 41 in (189) and Proposition 60 in (196). This concludes the proof. □

Table 6. Rewards for each joint action in Example 73.

	$a_{2,1}$	$a_{2,2}$
$a_{1,1}$	$-\frac{1}{2}$	1
$a_{1,2}$	1	-1

Based on this result, the LFC game for D can be understood in the following way. Principal $i \in \mathcal{N}$ observes a randomly relabeled problem $D_i \in \mathcal{D}$ and trains a joint policy $\pi^{(i)} \sim \boldsymbol{\sigma}_i(D_i)$ on this problem. The resulting policy $\pi^{(i)}$ is then translated back into a policy $f^*_{D_i,D}\pi^{(i)}$ for the original problem, using any isomorphism $f_{D_i,D} \in \mathrm{Iso}(D_i, D)$. Finally, this joint policy is made invariant to automorphism by applying the symmetrizer, and agent i in the original problem D is assigned the local policy $\Psi_i(f^*_{D_i,D}\pi^{(i)})$.

D.5. Other-play is not self-play in a different Dec-POMDP

Hu et al. (2020) show that the OP objective $\tilde{J}^D_{\mathrm{OP}}$ introduced by them can be understood of as the SP objective in a special Dec-POMDP. That is, for every Dec-POMDP D, there is a second Dec-POMDP E with $\Pi^E = \Pi^D$ such that for any $\pi \in \Pi^D$, it is $\tilde{J}^D_{\mathrm{OP}}(\pi) = \max_{\pi' \in \Pi^D} \tilde{J}^D_{\mathrm{OP}}(\pi')$ if and only if $J^E(\pi) = \max_{\pi' \in \Pi^E} J^E(\pi')$. Interestingly, when including player permutations, this is not the case anymore. We will prove this here, using the characterization of OP from the last section.

Intuitively, if agents are symmetric, then under OP, they will always act according to the same local policy in the environment. In some Dec-POMDPs, this means that it is optimal for the agents to randomize their actions, to end up with different actions some of the time. For instance, consider the following game:

Example 73. There are two players with two actions $\mathcal{A}_i := \{a_{i,1}, a_{i,2}\}$ for $i = 1, 2$ each, and an episode lasts only one step, making this a simple normal-form game. Rewards for each joint action are displayed in Table 6.

This example demonstrates that sometimes there does not exist a deterministic policy that is optimal under the OP objective.

Lemma 74. *In Example 73, for any deterministic policy π, it is $J_{\mathrm{OP}}(\pi) < \max_{\pi' \in \Pi^D} J_{\mathrm{OP}}(\pi')$.*

Proof. Let $R \in \mathbb{R}^{2,2}$ denote a matrix containing rewards as in Table 6.

First, note that in this game, players are symmetric, but actions are not. Moreover, there are no observations and only one state. Hence, $\mathrm{Aut}(D) = \{g, e\}$ where $g_N 1 = 2, g_N 2 = 1$ and e is the identity.

Now consider any deterministic policy $\pi = (\pi_1, \pi_2)$, corresponding to two vectors of action-probabilities

$$x, y \in \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

for the two players. Due to the symmetry of both players, the OP distribution μ of π assigns each policy π_1, π_2 to either player with probability $\frac{1}{2}$, so in $\Psi(\pi) = \pi^\mu$, both players play the distribution $z := \frac{1}{2}x + \frac{1}{2}y$ and receive a reward of

$$J^{\mathrm{OP}}(\pi) \overset{\text{Theorem 70}}{=} J(\Psi(\pi)) = z^\top R z.$$

It follows by the definition of x, y that

$$z \in \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \right\}.$$

Clearly, then $z^\top Rz$ is maximized at $z = [\frac{1}{2}, \frac{1}{2}]^\top$, yielding a reward of

$$J^{\mathrm{OP}}(\pi) = z^\top Rz = \frac{1}{8}. \tag{198}$$

Next, define $x = [\frac{4}{7}, \frac{3}{7}]^\top$ and let π^* be a policy such that $\pi_1^* = \pi_2^*$ and the two action-probabilities of both local policies are given by the vector x. Note that since $g^*\pi^* = (\pi^*_{gi})_{i=1,2} = \pi^*$, it follows from the above that π^* is invariant to automorphism.

It follows by Proposition 56 that we can evaluate the OP value of π^* by evaluating its expected return. That is,

$$J^{\mathrm{OP}}(\pi^*) \overset{\text{Proposition 56}}{=} J(\pi^*) = x^\top Rx = -\frac{1}{2}\left(\frac{4}{7}\right)^2 + 2\frac{4}{7}\frac{3}{7} - 1\left(\frac{3}{7}\right)^2 = \frac{1}{7}. \tag{199}$$

It follows that

$$J_{\mathrm{OP}}(\pi) \overset{(198)}{=} \frac{1}{8} < \frac{1}{7} = J^{\mathrm{OP}}(\pi^*) \leq \max_{\pi' \in \Pi^D} J_{\mathrm{OP}}(\pi'), \tag{200}$$

which concludes the proof. □

The fact that there is no deterministic optimal policy in Example 73 is in conflict with a canonical result about Dec-POMDPs.

Theorem 75 (Oliehoek et al., 2008, sec. 2.4.4)**.** *In every Dec-POMDP D, there is a deterministic policy $\pi \in (\Pi^0)^D$ such that $J^D(\pi) = \max_{\pi' \in \Pi^D} J^D(\pi')$.*

As a result, we can prove the following.

Proposition 76. *There exists a Dec-POMDP D such that for any other Dec-POMDP E with $\Pi^E = \Pi^D$, there exists a policy $\pi \in \Pi^E$ that is optimal for the SP objective of E, but not optimal for the OP objective of D.*

Proof. Let D be the Dec-POMDP as described in Example 73. Assume, towards a contradiction, that there exists a Dec-POMDP E with $\Pi^E = \Pi^D$ such that any optimal policy in that Dec-POMDP is optimal under the OP objective of D. Then by Theorem 75, there exists a deterministic policy $\pi \in (\Pi^0)^E$ such that $J^E(\pi)$ is maximal. Hence, by the assumption, also $J^D_{\mathrm{OP}}(\pi)$ is maximized. But by Lemma 74, it must be $\max_{\tilde{\pi} \in \Pi^D} J^D_{\mathrm{OP}}(\tilde{\pi}) > J^D_{\mathrm{OP}}(\pi)$. This is a contradiction, which means that D is an example of a Dec-POMDP that has the desired properties. □

This shows that to optimize the OP objective, we have to directly consider that objective and we cannot simply apply an RL algorithm to a different Dec-POMDP. Also, the fact that we need stochastic policies means that it is not immediately clear how to apply a Bellman equation to the objective.

E. Other-play is not optimal in the label-free coordination problem

In this section, our goal is to prove a rigorous version of Theorem 7 from the main text.

Recall that in Appendix D, we introduced the symmetrizer $\Psi \colon \Pi \to \Pi$, which maps a joint policy π to the policy corresponding to agents following randomly permuted local policies $(\mathbf{g}_i^*\pi)_i$ where $\mathbf{g}_i \sim \mathcal{U}(\mathrm{Aut}(D))$. This represents the random permutations employed in the OP objective, and hence by Theorem 70, it is

$J_{\mathrm{OP}}(\pi) = J(\Psi(\pi))$, i.e., the OP value of π is equal to the SP value of $\Psi(\pi)$. Moreover, we defined the equivalence classes $[\pi] = \Psi^{-1}(\{\Psi(\pi)\})$ of policies that get mapped to the same policy under Ψ.

As defined in Appendix D.1, an OP learning algorithm is any learning algorithm such that the policies that it learns achieve optimal OP value in expectation. In particular, it can be a learning algorithm that learns different policies in different training runs, as long as it chooses OP-optimal policies with probability 1. As we have seen in Theorem 72, in an LFC game, it does not matter which policy from an equivalence class $[\pi]$ is chosen. Unfortunately, though, there can also be different OP-optimal policies that are not equivalent. In this case, if an OP learning algorithm is not concentrated on only compatible policies, it is not optimal in the corresponding LFC problem.

In the remainder of this section, we will prove this statement. Concretely, in Appendix E.1, we will recall the two policies π^R, π^S in the two-stage lever game that we introduced in Section 4.4. We will show that both are optimal under OP, but that their cross-play value is inferior to the optimal OP value. In Appendix E.2, we will then formally state and prove the result that, if an OP algorithm is not concentrated on only one of the two incompatible equivalence classes of policies $[\pi^R], [\pi^S]$ in the two-stage lever game, then the algorithm is suboptimal in the LFC problem for that game.

E.1. Two incompatible optimal policies in the two-stage lever game

We begin by mapping out the space of OP-optimal policies in the two-stage lever game. Recall that this was a game with two agents, which proceeds in two rounds. Both agents have two actions, and their goal in both rounds is to choose the same action, for a reward of 1. Failure of coordination leads to a reward of -1. Moreover, in the second round, agents observe the actions of the other agent from the first round. In the following, let D stand for the Dec-POMDP associated to this game as described in Example 3.

Recall the two policies π^R and π^S introduced in Section 4.4. In both policies, agents randomize uniformly between both levers in the first round. They also both randomize in the second round if coordination was unsuccessful in the first one. If coordination in the first round was successful, there are two different strategies: in π^R, both agents repeat their respective actions from round one. In π^S, both agents switch to the action they did not play in round one.

The following lemma shows that both policies are optimal under OP.

Lemma 77. *Both π^R and π^S as described above are invariant to automorphism, and they maximize the OP objective, where*

$$J_{\mathrm{OP}}(\pi^R) = J_{\mathrm{OP}}(\pi^S) = \max_{\pi \in \Pi^D} J_{\mathrm{OP}}(\pi) = \frac{1}{2}.$$

Proof. Let $\tilde{\pi} \in \Pi^D$ arbitrary and define $\pi := \Psi(\tilde{\pi})$. By Theorem 70, π must be invariant to automorphism, and it must be

$$J_{\mathrm{OP}}(\tilde{\pi}) = J(\pi).$$

Moreover, if we show that π^R and π^S are invariant to automorphism, then by Proposition 56, their OP value equals their expected return. Hence, to show that π^R and π^S are optimal, it suffices to show that they are invariant to automorphism, and then compare their expected returns to the expected return of π.

To begin, consider the set of automorphisms in the game, as described in Example 17. Let $g \in \mathrm{Aut}(D)$. Recall that the state permutation is the trivial identity map, and the permutations for actions and observations of both players have to be equal, $g_{A_1} = g_{A_2} = g_{O_1} = g_{O_2}$. Hence, we can represent g as a tuple (g_N, g_A). There are four possible combinations of choices for permutations $g_A, g_N \colon \{1,2\} \to \{1,2\}$. Either permutation can either be the identity map $\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$ or the inversion $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$.

Now consider π^R and π^S. Note that both policies are symmetric in the agents, so $\pi_1^R = \pi_2^R$ and $\pi_1^S = \pi_2^S$. Invariance to automorphism is trivially fulfilled in the first stage, i.e., $\pi_i^R(a_i \mid \emptyset) = \pi^R_{g_N^{-1}i}(g_A^{-1}a_i \mid \emptyset)$ and $\pi_i^S(a_i \mid \emptyset) = \pi^S_{g_N^{-1}i}(g_A^{-1}a_i \mid \emptyset)$ for any $i \in \{1,2\}, a_i \in \{1,2\}$, since both agents randomize uniformly between actions. Now consider the second stage. Note that since automorphisms for the actions and observations of both players have to be identical, an automorphism can never map an action-observation history with $O_{i,1} \neq A_{i,0}$, i.e., such that one's own action and the observed action from the other agent differ, onto one in which they are the same, and vice versa. So we can consider the condition of invariance to automorphism separately for the case in which players achieved coordination in the first stage and for the case where they did not.

In the latter case, players randomize their actions uniformly, so in this case the policy is also trivially invariant. Now consider the former case, i.e., for $i = 1,2$, it is either $\overline{AO}_{i,1} = (1,1)$ or $\overline{AO}_{i,1} = (2,2)$. If both players repeat their action, then for $i \in \{1,2\}$, $a_i = o_i \in \{1,2\}$, and any automorphism $g = (g_N, g_A)$, it is

$$\pi_i^R(a_i \mid a_i, o_i) = 1 = \pi^R_{g_N^{-1}i}(g_A^{-1}a_i \mid g_A^{-1}a_i, g_A^{-1}o_i).$$

In the case where they change their action, for $a_i' \in \{1,2\} \setminus \{a_i\}$, it is $g_A^{-1}a_i' \neq g_A^{-1}a_i$, and thus

$$\pi_i^S(a_i' \mid a_i, o_i) = 1 = \pi^S_{g_N^{-1}i}(g_A^{-1}a_i' \mid g_A^{-1}a_i, g_A^{-1}o_i).$$

In conclusion, this shows that $g^*\pi^S = \pi^S$ and $g^*\pi^R = \pi^R$. Both π^S and π^R have an expected return of

$$J(\pi^S) = J(\pi^R) = \mathbb{E}[R_1 + R_2] = \frac{1}{2}(1+1) + \frac{1}{2}(-1 + \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot(-1)) = \frac{1}{2}$$

i.e., in the first round, they coordinate in half of the cases, in which case they coordinate again, and if they do not coordinate in the first round, they have an equal chance of achieving coordinating or not in the second round.

Now consider π. First, choosing g_N as the inversion and g_A as the identity, it follows that

$$\pi_1 \overset{(!)}{=} \pi_{g_N^{-1}1}(g_A^{-1}\cdot \mid g_A^{-1}\cdot) = \pi_2(\cdot \mid \cdot) = \pi_2,$$

where in (!) we use that π is invariant to automorphism. This shows that π must by symmetric in players, so $\pi_1 = \pi_2$. Moreover, choosing g_N as the identity and g_A as the inversion, it is

$$\pi_i(1 \mid \emptyset) \overset{(!)}{=} \pi_{g_N^{-1}i}(g_A^{-1}1 \mid \emptyset) = \pi_i(2 \mid \emptyset),$$

so under π, too, both agents must choose both actions with equal probability in the first stage.

Turning to the second stage, we assume that π receives a maximal reward of 1 in the second stage if both agents coordinated in the first stage. Now consider the case in which coordination failed in the first step. Note that since π is symmetric in the agents, it is not possible for agents to consistently choose to play either the action of agent 1 or of agent 2 in the second round if those actions did not coincide in the first round. Moreover, due to invariance to action and observation permutations, the probability p that an agent repeats their action from the first round must be the same, no matter whether that action was a 1 or 2, as for $a \neq a' \in \{1,2\}$, it must be $\pi_i(a \mid a, a') = \pi_i(a' \mid a', a)$ for $i = 1, 2$ due to invariance to automorphism. Due to the symmetry of agents, that probability must also be the same for both agents. Hence, we can define

$$p := \pi_i(a \mid a, a') = 1 - \pi_i(a' \mid a, a')$$

for $i = 1, 2$ and $a \neq a' \in \{1, 2\}$. The return in the second stage in this case is then

$$\begin{aligned}
&\mathbb{E}[R_2 \mid A_{2,1} \neq A_{1,1}] \\
&\quad = 1 \cdot \Big(\pi_1(a_{1,1} \mid a_{1,1}, a_{2,1})\pi_2(a_{1,1} \mid a_{2,1}, a_{1,1}) + \pi_1(a_{2,1} \mid a_{1,1}, a_{2,1})\pi_2(a_{2,1} \mid a_{2,1}, a_{1,1})\Big) \\
&\quad - 1 \cdot \Big(\pi_1(a_{1,1} \mid a_{1,1}, a_{2,1})\pi_2(a_{2,1} \mid a_{2,1}, a_{1,1}) + \pi_1(a_{2,1} \mid a_{1,1}, a_{2,1})\pi_2(a_{1,1} \mid a_{2,1}, a_{1,1})\Big) \\
&\qquad = 1 \cdot \Big(p(1-p) + (1-p)p\Big) - 1 \cdot \Big(pp + (1-p)(1-p)\Big) = 0, \qquad (201)
\end{aligned}$$

where $a_{1,1} \neq a_{2,1} \in \{1, 2\}$ are arbitrary. This shows that also

$$J(\pi) \leq \frac{1}{2}(1+1) + \frac{1}{2}(-1+0) = \frac{1}{2}.$$

Thus, it follows that $J(\pi) \leq \frac{1}{2} = J(\pi^R) = J(\pi^S)$, which shows that π^R, π^S are both optimal and thus concludes the proof. □

Now consider the case in which one agent chooses a local policy from π^R and another agent chooses a local policy from π^S. It is clear that this will yield a suboptimal expected return compared with π^R or π^S, as agents will always fail to coordinate in the second round, if they coordinated in the first round.

Lemma 78.

$$J(\pi_1^R, \pi_2^S) = J(\pi_1^S, \pi_2^R) = -\frac{1}{2} < \frac{1}{2}.$$

Proof. Both policies are equal in the first round, in which there is a 50% chance that agents coordinate. If agents do not coordinate in the first round, then they both randomize uniformly in the second stage. If agents do coordinate in the first round, then agent 1 using π_1^R repeats their action from the first round, while agent 2, using π_2^S, will switch to a different action. In this case, the reward is thus always -1. As a result, it is

$$Jwe(\pi_1^R, \pi_2^S) = \frac{1}{2}(1-1) + \frac{1}{2}(-1 + \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1)) = -\frac{1}{2} < \frac{1}{2}.$$

The same argument applies if agent 1 uses π_1^S and agent 2 uses π_2^R. □

E.2. Proof that other-play is suboptimal

Using the two lemmas from the last section, we can now show that if an OP learning algorithm is not concentrated on only one of $[\pi^S]$ or $[\pi^R]$, it is not optimal in the LFC problem for D. Note that we could just choose a particular learning algorithm, for instance, one that randomizes uniformly between π^R and π^S, and show that that algorithm is suboptimal. This would then prove that a suboptimal OP learning algorithm exists. However, we show a more general result. Specifically, take any learning algorithm that learns with positive probability a policy equivalent to π^R in one relabeling of D and, also with positive probability, a policy equivalent to π^L in some potentially different relabeling. Then this learning algorithm is suboptimal in the LFC problem for D.

In the following, let $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D)\}$ be the set of different relabeled Dec-POMDPs of the two-stage lever game D. For any Dec-POMDP $E \in \mathcal{D}$, choose $f_{D,E} \in \mathrm{Iso}(D, E)$ arbitrarily. Below, we define the class of algorithms relevant to our theorem.

Definition 79. Let $\sigma \in \Sigma^{\mathcal{D}}$ be a learning algorithm such that for any $E \in \mathcal{D}$, $E = f^*D$, it is

$$\sigma(E) = \alpha^E \mu^E + (1 - \alpha^E)\mu'^E,$$

where $\alpha_E \in [0, 1]$ and $\mu^E, \mu'^E \in \Delta(\Pi_E)$ are chosen such that

$$\mu^E(f^*_{D,E}[\pi^R]) = 1 = \mu'^E(f^*_{D,E}[\pi^S]).$$

That is, $\sigma(E)$ is a mixture of a distribution μ^E with weight only on the equivalence class of policies corresponding to π^R, and a distribution μ'^E that only puts weight on policies corresponding to π^S. We say that σ learns both $[\pi^R]$ and $[\pi^S]$ if there exist $\tilde{D}, \tilde{D}' \in \mathcal{D}$ such that $\alpha_{\tilde{D}} > 0$ and $\alpha_{\tilde{D}'} < 1$.

Note that Corollary 69 (ii) ensures that this definition does not depend on the chosen isomorphisms $f_{D,E}$.

The above condition is very weak: we only require there to be some relabeled Dec-POMDP on which the learning algorithm chooses an equivalent policy of π^R some of the time, and some potentially different relabeled Dec-POMDP where the algorithm chooses a policy equivalent to π^S some of the time. We now show that a learning algorithm with this property is an OP learning algorithm per Definition 54, but that it is not optimal in the LFC problem for D.

Theorem 80. *In the two-stage lever game, there are two classes of OP-optimal policies, denoted by $[\pi^R]$ and $[\pi^S]$. Any learning algorithm that learns both $[\pi^R]$ and $[\pi^S]$ in the sense of Definition 79 is an OP learning algorithm, but it is not optimal in the LFC problem for that game.*

Proof. First, note that under π^R, π^S, all action-observation histories are reached with positive probability, and by Lemma 77, they are both invariant to automorphism. Hence, by Proposition 65, it is $\Psi(\pi^R) = \pi^R$ and $\Psi(\pi^S) = \pi^S$ (*). This also shows that $[\pi^R]$ and $[\pi^S]$ are distinct. Let σ be a learning algorithm as specified in Definition 79, and define μ^E, μ'^E and α^E for $E \in \mathcal{D}$ as in that definition. Also, let $\tilde{D}, \tilde{D}' \in \mathcal{D}$ such that $\alpha_{\tilde{D}} > 0$ and $\alpha_{\tilde{D}'} < 1$.

We begin by showing that σ is an OP learning algorithm. For any $E \in \mathcal{D}$, it is

$$\begin{aligned}
\mathbb{E}_{\pi\sim\sigma(E)}[J^E_{\mathrm{OP}}(\pi)] &= \alpha_E \mathbb{E}_{\pi\sim\mu^E}[J^E_{\mathrm{OP}}(\pi)] + (1-\alpha_E)\mathbb{E}_{\pi\sim\mu'^E}[J^E_{\mathrm{OP}}(\pi)] && (202)\\
&= \alpha_E \mathbb{E}_{\pi\sim\mu^E}[\mathbb{1}_{f^*_{D,E}[\pi^R]} J^E_{\mathrm{OP}}(\pi))] + (1-\alpha_E)\mathbb{E}_{\pi\sim\mu'^E}[\mathbb{1}_{f^*_{D,E}[\pi^S]} J^E_{\mathrm{OP}}(\pi)] && (203)\\
&= \alpha_E \mathbb{E}_{\pi\sim\mu^E}[\mathbb{1}_{f^*_{D,E}[\pi^R]} J^D_{\mathrm{OP}}(f^*_{E,D}\pi))] \\
&\quad + (1-\alpha_E)\mathbb{E}_{\pi\sim\mu'^E}[\mathbb{1}_{f^*_{D,E}[\pi^S]} J^D_{\mathrm{OP}}(f^*_{E,D}\pi)] && (204)\\
&= \alpha_E \mathbb{E}_{\pi\sim f^*_{E,D}\mu^E}[\mathbb{1}_{f^*_{E,D}(f^*_{D,E}[\pi^R])} J^D_{\mathrm{OP}}(\pi))] \\
&\quad + (1-\alpha_E)\mathbb{E}_{\pi\sim f^*_{E,D}\mu'^E}[\mathbb{1}_{f^*_{E,D}(f^*_{D,E}[\pi^S])} J^D_{\mathrm{OP}}(\pi)] && (205)\\
&= \alpha_E \mathbb{E}_{\pi\sim f^*_{E,D}\mu^E}[\mathbb{1}_{[\pi^R]} J^D_{\mathrm{OP}}(\pi))] + (1-\alpha_E)\mathbb{E}_{\pi\sim f^*_{E,D}\mu'^E}[\mathbb{1}_{[\pi^S]} J^D_{\mathrm{OP}}(\pi)] && (206)
\end{aligned}$$

$$= \alpha_E J^D_{\mathrm{OP}}(\pi^R) + (1-\alpha_E) J^D_{\mathrm{OP}}(\pi^S) \tag{207}$$

$$= \alpha_E \max_{\pi\in\Pi^D} J^D_{\mathrm{OP}}(\pi) + (1-\alpha_E) \max_{\pi\in\Pi^D} J^D_{\mathrm{OP}}(\pi) \tag{208}$$

$$= \max_{\pi\in\Pi^D} J^D_{\mathrm{OP}}(\pi), \tag{209}$$

using Definition 79 in (202) and (203); Corollary 71 in (204); a change of variables for pushforward measures in (205); Lemma 18 and parts (ii) and (iii) of Corollary 69 in (206); the "in particular" part of Theorem 70 in (207); and Lemma 77 in (208). This shows that σ is an OP learning algorithm.

Next, we turn to proving that σ is suboptimal in the LFC problem. To that end, define the auxiliary function

$$X(D_1, D_2, \pi^{(1)}, \pi^{(2)}) := \delta_{D_1,\tilde{D}}\delta_{D_2,\tilde{D}'}\mathbb{1}_{[\pi^R]}(\pi^{(1)})\mathbb{1}_{[\pi^S]}(\pi^{(2)})$$

for any $D_1, D_2 \in \mathcal{D}$ and $\pi^{(1)}, \pi^{(2)} \in \Pi^D$. This function is 1 whenever the first Dec-POMDP is $\tilde{D}$ and the policy $\pi^{(1)}$ is such that its pushforward policy is in $[\pi^R]$, and when the second Dec-POMDP is $\tilde{D}'$ and the pushforward of $\pi^{(2)}$ is in $[\pi^S]$. Otherwise, it is 0.

Recall that by Lemma 18, it is $(f^*_{E,D})^{-1} \in \mathrm{Iso}(D, E)$ for $E \in \mathcal{D}$, and by Corollary 69 (ii), the pushforward of an equivalence class does not depend on the particular chosen isomorphism. Define

$$\beta := \mathbb{E}_{D_1,D_2\sim\mathcal{U}(\mathcal{D})}[\mathbb{E}_{\pi^{(i)}\sim f^*_{\tilde{D},D_i}\sigma(D_i),\, i=1,2}[X(D_1, D_2, \pi^{(1)}, \pi^{(2)})]]$$

and note that $\beta \leq 1$. This number is relevant to us as it is a lower bound on the probability that agents 1 and 2 will have incompatible joint policies in the objective of the LFC problem. By assumption about σ, we know that

$$(f^*_{\tilde{D},D}\sigma(\tilde{D}))([\pi^R]) = \sigma(\tilde{D})((f^*_{\tilde{D},D})^{-1}[\pi^R]) = \alpha_{\tilde{D}} > 0$$

and similarly

$$(f^*_{\tilde{D}',D}\sigma(\tilde{D}'))([\pi^S]) = \sigma(\tilde{D}')((f^*_{\tilde{D}',D})^{-1}[\pi^S]) = 1 - \alpha_{\tilde{D}'} > 0,$$

and it is $\mathcal{U}(\mathcal{D})(\{\tilde{D}\}) = \mathcal{U}(\mathcal{D})(\{\tilde{D}'\}) = \frac{1}{\mathcal{D}}$. Hence, it follows that $0 < \beta \leq 1$ (**). That is, with nonzero probability, using σ leads to incompatible policies.

Our goal is now to prove that $U^D(\sigma) < \frac{1}{2}$. To that end, we also have to show that the value in the cases where $X(D_1, D_2, \pi^{(1)}, \pi^{(2)}) = 0$ is bounded by $\frac{1}{2}$. Note that for any $E \in \mathcal{D}$, it is

$$(f^*_{E,D}\sigma(E))([\pi^R] \cup [\pi^S]) = \sigma(E)(((f^*_{E,D})^{-1}[\pi^R]) \cup ((f^*_{E,D})^{-1}[\pi^S])) = 1,$$

and for $\pi^{(1)}, \pi^{(2)} \in [\pi^R] \cup [\pi^S]$, using the definition of equivalence, it is

$$J^D(\Psi(\pi^{(1)})_1, \Psi(\pi^{(2)})_2) = J^D(\Psi(\pi^C)_1, \Psi(\pi^{C'})_2) \overset{(*)}{=} J^D(\pi_1^C, \pi_2^{C'}) \overset{\text{Lemma 77}}{=} \frac{1}{2} \tag{210}$$

if both policies are from the same class, i.e., $C = C' \in \{R, S\}$, and

$$J^D(\Psi(\pi^{(1)})_1, \Psi(\pi^{(2)})_2) = J^D(\Psi(\pi^C)_1, \Psi(\pi^{C'})_2) \overset{(*)}{=} J^D(\pi_1^C, \pi_2^{C'}) \overset{\text{Lemma 78}}{=} -\frac{1}{2} \tag{211}$$

if they come from different classes, i.e., $C \neq C' \in \{R, S\}$. It follows that for any $D_1, D_2 \in \mathcal{D}$, it is

$$J^D(\Psi(\pi^{(1)})_1, \Psi(\pi^{(2)})_2) \leq \frac{1}{2} \tag{212}$$

almost surely if $\pi^{(1)} \sim f^*_{D_1,D}\sigma(D_1)$ and $\pi^{(2)} \sim f^*_{D_2,D}\sigma(D_2)$.

Using this bound together with (**), it follows that

$$\begin{aligned}
U^D(\sigma) &= U^D(\sigma, \ldots, \sigma) && (213)\\
&= \mathbb{E}_{D_1,D_2\sim\mathcal{U}(\mathcal{D})}[\mathbb{E}_{\pi^{(i)}\sim f^*_{D_i,D}\sigma(D_i),\, i=1,2}[J^D(\Psi_1(\pi^{(1)}), \Psi_2(\pi^{(2)}))]] && (214)\\
&= \mathbb{E}_{D_1,D_2\sim\mathcal{U}(\mathcal{D})}\big[\mathbb{E}_{\pi^{(i)}\sim f^*_{D_i,D}\sigma(D_i),\, i=1,2}\big[J^D(\Psi(\pi^{(1)}_1), \Psi_2(\pi^{(2)}))X(D_1, D_2, \pi^{(1)}, \pi^{(2)})\\
&\quad + J^D(\Psi_1(\pi^{(1)}), \Psi_2(\pi^{(2)}))(1 - X(D_1, D_2, \pi^{(1)}, \pi^{(2)}))\big]\big] && (215)\\
&\le \mathbb{E}_{D_1,D_2\sim\mathcal{U}(\mathcal{D})}\left[\mathbb{E}_{\pi^{(i)}\sim f^*_{D_i,D}\sigma(D_i),\, i=1,2}\left[\left(-\frac{1}{2}\right) X\left(D_1, D_2, \pi^{(1)}, \pi^{(2)}\right)\right.\right.\\
&\quad \left.\left. + \frac{1}{2}\left(1 - X\left(D_1, D_2, \pi^{(1)}, \pi^{(2)}\right)\right)\right]\right] && (216)\\
&= \left(-\frac{1}{2}\right)\beta + \frac{1}{2}(1-\beta) && (217)\\
&\overset{(**)}{<} \frac{1}{2}, && (218)
\end{aligned}$$

where we use Theorem 72 in (214) and Equations 211 and 212 in (216).

To see that this is suboptimal in the LFC problem, consider σ^* defined via $\sigma^*(E) := \delta_{f^*_{D,E}\pi^R}$ for any $E \in \mathcal{D}$. Since

$$f^*_{E,D}\sigma(E) = \delta_{f^*_{D,E}\pi^R} \circ (f^*_{E,D})^{-1} = \delta_{f^*_{E,D}(f^*_{D,E}\pi^R)} = \delta_{\pi^R} \tag{219}$$

for any $E \in \mathcal{D}$, it follows that

$$\begin{aligned}
U^D(\sigma^*) &= U^D(\sigma^*, \ldots, \sigma^*)\\
&\overset{\text{Theorem 72}}{=} \mathbb{E}_{D_1,D_2\sim\mathcal{U}(\mathcal{D})}[\mathbb{E}_{\pi^{(i)}\sim f^*_{D_i,D}\sigma^*(D_i),\, i=1,2}[J^D(\Psi_1(\pi^{(1)}), \Psi_2(\pi^{(2)}))]]\\
&\overset{(219)}{=} J^D(\Psi_1(\pi^R), \Psi_2(\pi^R)) \overset{(210)}{=} \frac{1}{2} \overset{(214)\text{–}(217)}{>} U^D(\sigma), && (220)
\end{aligned}$$

which concludes the proof. □

F. Other-play with tie-breaking

In this section, we formally define OP with tie-breaking as introduced in Section 5, state and prove a rigorous version of Theorem 8, and provide an additional result about random tie-breaking functions. In Appendix F.1, we define OP with tie-breaking and discuss to what degree the formal definition is satisfied by our method. In Appendix F.2, we show that OP with tie-breaking is optimal in the LFC problem and that all principals using OP with tie-breaking is an optimal symmetric Nash equilibrium of any LFC game. In Appendix F.3 we prove that a modification of the tie-breaking function introduced in Section 5 satisfies our definition of a tie-breaking function.

F.1. Definition of other-play with tie-breaking

Recall that OP with tie-breaking was introduced as an extension of OP, to fix the failure of OP in the LFC problem. A tie-breaking function ranks the different OP-optimal equivalence classes of policies in a given problem. For instance, a tie-breaking function could compare the two incompatible policies in the two-stage lever game, π^R and π^S, and choose the policy under which actions are more highly correlated, which is π^R. Another tie-breaking function would be one that samples from an OP learning algorithm and ranks policies in terms of their relative frequencies. If one policy is learned more often than another, that tie-breaking function would be able to distinguish between them. OP with tie-breaking is defined as an algorithm that chooses an OP-optimal policy that maximizes a tie-breaking function.

Note that an obvious alternative approach, making a learning algorithm deterministic by coordinating on a random seed, would not work in the LFC problem. This is because the learned policies have to be compatible even across different, relabeled Dec-POMDPs. A labeling modifies the representation of the data used to train the algorithm, which likely leads to the same effect as a resampling of the random seed. Hence, fixing a random seed fails to ensure that the chosen policies are compatible when principals do not coordinate on labels for the problem. (Note that a learning algorithm that outputs consistent policies on a given problem, but incompatible policies on different, relabeled problems is explicitly included as a non-optimal algorithm in Theorem 80.) Additionally, this would be an unprincipled way to deal with the choice between different maximizers of the OP objective, leaving no possibility, e.g., for an explicit bias towards some policies over others.

Turning to the definitions, recall again from Appendix D that we say that for a Dec-POMDP $D, \pi, \pi' \in \Pi^D$ are equivalent, $\pi \equiv \pi'$, if $\Psi(\pi) = \Psi(\pi')$, where Ψ is the symmetrizer for D that maps a joint policy π to the joint policy $\Psi(\pi)$ corresponding to agents following randomly permuted local policies $(\mathbf{g}_i^*\pi)$ for $\mathbf{g}_i \sim \mathcal{U}(\mathrm{Aut}(D))$. Moreover, we defined $[\pi]$ as the equivalence class of π, and $f^*[\pi] := [f^*\pi]$ for $f \in \mathrm{Iso}(D, E)$.

In the following, define $\Pi_{\mathrm{OP}}^D := \arg\max_{\pi \in \Pi^D} J_{\mathrm{OP}}^D(\pi)$ as the set of OP-optimal policies for any Dec-POMDP D. Let $\mathcal{D}$ be any set of Dec-POMDPs. A tie-breaking function takes in Dec-POMDPs $D \in \mathcal{D}$ and policies $\pi \in \Pi^D$ and outputs values in $[0, 1]$ that can be used to consistently break ties between policies, across different isomorphic Dec-POMDPs.

Definition 81 (Tie-breaking function)**.** Let $\chi\colon \{(D, \pi) \mid D \in \mathcal{D}, \pi \in \Pi^D\} \to [0, 1]$. Then

(a) χ is called a tie-breaking function for $\mathcal{D}$ if

  (i) for any $D \in \mathcal{D}$, χ attains a maximum on the set

$$\{(D, \pi) \mid \pi \in \Pi_{\mathrm{OP}}^D\}.$$

  (ii) for any $D \in \mathcal{D}$ and $\pi, \pi' \in \Pi^D$, it is

$$\chi(D, \pi) = \chi(D, \pi') \Rightarrow \pi \equiv_D \pi'.$$

(b) χ is called invariant to isomorphism if for any $D, E \in \mathcal{D}$, $f \in \mathrm{Iso}(D, E)$, and $\pi \in \Pi^D, \pi' \in \Pi^E$, it is

$$f^*[\pi] = [\pi'] \Rightarrow \chi(D, \pi) = \chi(E, \pi').$$

χ being a tie-breaking function ensures that there is always a unique equivalence class of policies that maximizes the function for a given Dec-POMDP. It being invariant to isomorphism ensures that χ chooses corresponding equivalence classes of policies on different, isomorphic Dec-POMDPs in $\mathcal{D}$.

Using a tie-breaking function that is invariant to isomorphism, we can define OP with tie-breaking.

Definition 82 (Other-play with tie-breaking)**.** Let χ be a tie-breaking function for $\mathcal{D}$ that is invariant to isomorphism. Let $\sigma^\chi \in \Sigma^{\mathcal{D}}$ be a learning algorithm such that for any $D \in \mathcal{D}$, there exists a measurable set $\mathcal{Z} \subseteq \Pi^D_{\mathrm{OP}}$ such that $\sigma^\chi(D)(\mathcal{Z}) = 1$ and $\mathcal{Z} \subseteq \arg\max_{\pi \in \Pi^D_{\mathrm{OP}}} \chi(D, \pi)$. Then we say that σ^χ is an OP with tie-breaking learning algorithm for $\mathcal{D}$.

This definition implies that no matter the problem $D \in \mathcal{D}$, the algorithm always learns the OP-optimal policy that achieves the highest tie-breaking value. Since the tie-breaking function is invariant to isomorphism, this means that policies learned in different training runs and when trained on relabeled Dec-POMDPs are always compatible.

Finally, recall that the practical method introduced in Section 5 consists of sampling $K \in \mathbb{N}$ policies using an OP learning algorithm, applying a tie-breaking function to each policy, and then choosing the one with the highest value. Clearly, if all the OP-optimal equivalence classes of policies in any of the Dec-POMDPs in $\mathcal{D}$ are among the first K learned policies, then this algorithm will satisfy our definition. This appears to be the case for the OP algorithm and toy examples used in our experiments, at least for large enough K, and when we ignore differences between policies that matter little for the agents' expected returns. Moreover, it is easy to see that the algorithm will still always pick equivalent policies if for any two isomorphic Dec-POMDPs $D, E \in \mathcal{D}$, if σ^{OP} learns a policy π with positive probability in D, it also learns an equivalent policy π' in E with positive probability, in the sense that $f^*[\pi] = [\pi']$ for $f \in \mathrm{Iso}(D, E)$. That is, it does not matter if some policies are never learned, if this happens consistently across isomorphic problems. One way for an algorithm σ^{OP} to have this property is by being equivariant (see Appendix C.8).

F.2. Other-play with tie-breaking is an optimal symmetric profile

In the following, fix a Dec-POMDP D and the set of relabeled Dec-POMDPs $\mathcal{D} := \{f^*D \mid f \in \mathrm{Sym}(D\}$, and let χ be a tie-breaking function for $\mathcal{D}$ that is invariant to isomorphism. Let $U(\boldsymbol{\sigma}) := U^D(\boldsymbol{\sigma})$ stand for the payoff in the LFC game for D given the profile of learning algorithms $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$. Recall that the learning algorithm $\sigma \in \Sigma^{\mathcal{D}}$ is optimal in the LFC problem for D if $U(\sigma) \geq U(\sigma')$ for all $\sigma' \in \Sigma^{\mathcal{D}}$, where $U(\sigma) := U(\sigma, \ldots, \sigma)$.

Recall from Appendix C.7 that we call a profile of learning algorithms $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$ symmetric if $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_{g^{-1}i}$ for any principal $i \in \mathcal{N}$ and $g \in \mathrm{Aut}(D)$. $\boldsymbol{\sigma}$ is defined as an optimal symmetric profile if it is symmetric and for any other symmetric profile $\boldsymbol{\sigma}'_1, \ldots, \boldsymbol{\sigma}'_N \in \Sigma^{\mathcal{D}}$, it is $U(\boldsymbol{\sigma}) \geq U(\boldsymbol{\sigma}')$.

Turning to our main theorem, we show that a profile $\sigma^\chi, \ldots, \sigma^\chi$ in which all principals choose OP with tie-breaking is an optimal symmetric profile in the LFC game for D. That is, as long as symmetric principals choose the same learning algorithm, they cannot do better than all choosing OP with tie-breaking. In particular, this implies that σ^χ is optimal in the LFC problem. Note, though, that we prove a stronger statement, as optimality for the LFC problem only requires that $U(\sigma^\chi, \ldots, \sigma^\chi) \geq U(\sigma', \ldots, \sigma')$ for all $\sigma' \in \Sigma^{\mathcal{D}}$, while we show that $U(\sigma^\chi, \ldots, \sigma^\chi) \geq U^D(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N)$ for all symmetric profiles $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$. Afterwards, we will apply Theorem 48 to conclude that all principals using OP with tie-breaking is also a Nash equilibrium, i.e., given that all principals use OP with tie-breaking, no individual principal can do better by switching to a different algorithm.

Theorem 83. *Let σ^χ be an OP with tie-breaking learning algorithm for $\mathcal{D}$, as defined in Definition 82. Then all principals using σ^χ is an optimal symmetric strategy profile in the LFC game for D. In particular, σ^χ is optimal in the LFC problem for D, and it is $U(\sigma^\chi) = \max_{\sigma \in \Sigma^{\mathcal{D}}} U(\sigma) = \max_{\pi \in \Pi^D} J^D_{\mathrm{OP}}(\pi)$.*

Proof. We begin by showing that $U(\sigma^\chi, \ldots, \sigma^\chi) = \max_{\pi \in \Pi^D} J^D_{\mathrm{OP}}(\pi)$. Afterwards, we show that one cannot get a better payoff in the LFC game than that using a symmetric profile of learning algorithms, i.e., that $\sigma^\chi, \dots, \sigma^\chi$ is an optimal symmetric profile. It then follows immediately that also $U(\sigma^\chi) = U(\sigma^\chi, \dots, \sigma^\chi) \geq U(\sigma, \dots, \sigma) = U(\sigma)$ for any $\sigma \in \Sigma^D$, i.e., that σ^χ is optimal in the LFC problem for D.

Recall the definition $\Pi^E_{\text{OP}} := \arg\max_{\pi' \in \Pi^E} J^E_{\text{OP}}(\pi')$ for $E \in \mathcal{D}$. Let $E, F \in \mathcal{D}$ arbitrary and $f, \tilde{f} \in \text{Sym}(D)$ such that $E = f^* D$ and $F = \tilde{f}^* D$. We want to show that σ^χ learns compatible OP-optimal policies in both Dec-POMDPs. To that end, let $\mathcal{Z} \subseteq \Pi^E_{\text{OP}}$ measurable such that $\mathcal{Z} \subseteq \arg\max_{\pi \in \Pi^E} \chi(E, \pi)$ and such that $\sigma(E)(\mathcal{Z}) = 1$. Then since χ is a tie-breaking function, there must exist a policy $\pi' \in \Pi^E_{\text{OP}}$ such that $\pi' \in \mathcal{Z} \subseteq [\pi']$ and thus $\sigma(E)([\pi']) = 1$ and $\pi' \in \arg\max_{\pi \in \Pi^E_{\text{OP}}} \chi(E, \pi)$ (i). Letting $\pi := (f^{-1})^* \pi'$, it follows from Corollary 71 that $\pi \in \Pi^D_{\text{OP}}$ (ii).

Now we want to show that σ^χ learns policies in $[\tilde{f}^* \pi]$ in F. Let $\tilde{\pi} := (\tilde{f} \circ f^{-1})^* \pi'$ and note that $\tilde{f} \circ f^{-1} \in \text{Iso}(E, F)$, so $\tilde{\pi} \in \Pi^F$. Since χ is invariant to isomorphism, it is

$$\chi(F, \hat{\pi}) = \chi(E, (f \circ \tilde{f}^{-1})^* \hat{\pi}) \overset{\text{(i)}}{\leq} \chi(E, \pi'),$$

for any $\hat{\pi} \in \Pi^F$. Here, equality holds for $\hat{\pi} := \tilde{\pi}$, as it is

$$(f \circ \tilde{f}^{-1})^* \tilde{\pi} = (f \circ \tilde{f}^{-1})^* (\tilde{f} \circ f^{-1})^* \pi' = \pi'$$

by Lemma 21. Moreover, due to Corollary 71 it is again $\tilde{\pi} \in \Pi^F_{\text{OP}}$. This shows that $\tilde{\pi} \in \arg\max_{\hat{\pi} \in \Pi^F_{\text{OP}}} \chi(F, \hat{\pi})$, and since χ is a tie-breaking function, it is $\arg\max_{\hat{\pi} \in \Pi^F_{\text{OP}}} \chi(F, \hat{\pi}) \subseteq [\tilde{\pi}]$. Hence, by the definition of σ^χ, it follows that $\sigma^\chi(F)([\tilde{\pi}]) = 1$ (iii).

Now choose any arbitrary isomorphism $f_{F,D} \in \text{Iso}(F, D)$. By Lemma 21, it is $(f^*_{F,D})^{-1} = (f^{-1}_{F,D})^*$. Hence, by the second part of Lemma 69, it is $(f^*_{F,D})^{-1}[\pi] = \tilde{f}^*[\pi]$ (iv). Moreover, using again Lemma 21, it is

$$(f \circ \tilde{f}^{-1})^* \tilde{\pi} = (f \circ \tilde{f}^{-1})^* (\tilde{f} \circ f^{-1})^* \pi' = \pi'. \tag{221}$$

It follows that

$$\begin{aligned}(f^*_{F,D})^{-1}([\pi]) &\overset{\text{(iv)}}{=} \tilde{f}^*[\pi] = \tilde{f}^*[(f^{-1})^* \pi'] \overset{(221)}{=} \tilde{f}^*[(f^{-1})^* (f \circ \tilde{f}^{-1})^* \tilde{\pi}] \\ &\overset{\text{Definition 68}}{=} \tilde{f}^* (f^{-1})^* (f \circ \tilde{f}^{-1})^* [\tilde{\pi}] \overset{\text{Corollary 69 (iii)}}{=} [\tilde{\pi}] \end{aligned} \tag{222}$$

and thus

$$f^*_{F,D} \sigma^\chi(F)([\pi]) = \sigma^\chi(F)((f^*_{F,D})^{-1}([\pi])) \overset{(222)}{=} \sigma^\chi(F)([\tilde{\pi}]) \overset{\text{(iii)}}{=} 1. \tag{223}$$

Since F was arbitrary, we can conclude that the above equation holds for any $F \in \mathcal{D}$ and isomorphism $f_{F,D} \in \text{Iso}(F, D)$.

Using this fact together with the definition of equivalence of policies in Definition 66, as well as the expression for the payoff in the LFC game from Theorem 72, it follows that

$$U(\sigma^\chi, \dots, \sigma^\chi) \overset{\text{Theorem 72}}{=} \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j, D} \boldsymbol{\sigma}_j(D_j),\, j \in \mathcal{N}} \left[J^D \left(\left(\Psi_k(\pi^{(k)}) \right)_{k \in \mathcal{N}} \right) \right] \right] \tag{224}$$

$$\overset{(223)}{=} \mathbb{E}_{D_i \sim U(\mathcal{D}),\, i \in \mathcal{N}} \left[\mathbb{E}_{\pi^{(j)} \sim f^*_{D_j, D} \boldsymbol{\sigma}_j(D_j),\, j \in \mathcal{N}} \left[J^D \left((\Psi_k(\pi))_{k \in \mathcal{N}} \right) \right] \right] \tag{225}$$

$$= J^D(\Psi(\pi)) \tag{226}$$

$$\overset{\text{Theorem 70}}{=} J^D_{\text{OP}}(\pi) \tag{227}$$

$$\overset{\text{(ii)}}{=} \max_{\hat{\pi} \in \Pi^D} J^{\text{OP}}_D(\hat{\pi}), \tag{228}$$

Next, we show that we cannot do better than that with a symmetry-invariant profile of learning algorithms. To that end, let $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N \in \Sigma^{\mathcal{D}}$ arbitrary such that $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_{g^{-1}i}$ for any $g \in \mathrm{Aut}(D)$, $i \in \mathcal{N}$. For $i \in \mathcal{N}$, define the distribution

$$\nu^{(i)} := |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} (f^{-1})^* \boldsymbol{\sigma}_i (f^* D)$$

and let $\mu := \otimes_{i \in \mathcal{N}} \nu_i^{(i)}$. Then it is

$$\begin{aligned}
U(\boldsymbol{\sigma}) &= \mathbb{E}_{\mathbf{f} \sim \mathcal{U}(\mathrm{Sym}(D)^{\mathcal{N}})}[\mathbb{E}_{\pi^{(i)} \sim (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i(\mathbf{f}_i^* D),\, i \in \mathcal{N}}[J^D((\pi_i^{(i)})_{i \in \mathcal{N}})]] && (229)\\
&= \sum_{\mathbf{f} \in \mathrm{Sym}(D)^{\mathcal{N}}} |\mathrm{Sym}(D)|^{-N} \int_{(\pi^{(i)})_{i \in \mathcal{N}} \in (\Pi^D)^{\mathcal{N}}} J^D((\pi_i^{(i)})_{i \in \mathcal{N}}) \mathrm{d} \otimes_{i \in \mathcal{N}} (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i (\mathbf{f}_i^* D) && (230)\\
&= \sum_{\mathbf{f} \in \mathrm{Sym}(D)^{\mathcal{N}}} |\mathrm{Sym}(D)|^{-N} \int_{\pi \in \Pi^D} J^D(\pi) \mathrm{d} \otimes_{i \in \mathcal{N}} (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i (\mathbf{f}_i^* D) \circ \mathrm{proj}_i^{-1} && (231)\\
&= \int_{\pi \in \Pi^D} J^D(\pi) \mathrm{d} \otimes_{i \in \mathcal{N}} \left(|\mathrm{Sym}(D)|^{-1} \sum_{\mathbf{f}_i \in \mathrm{Sym}(D)} (\mathbf{f}_i^{-1})^* \boldsymbol{\sigma}_i (\mathbf{f}_i^* D) \circ \mathrm{proj}_i^{-1} \right) && (232)\\
&= \int_{\pi \in \Pi^D} J^D(\pi) \mathrm{d}\mu && (233)\\
&\overset{(117)}{=} J^D(\mu). && (234)
\end{aligned}$$

Now we want to show that it is $J^D(\mu) \leq \max_{\pi \in \Pi^D} J^D(\pi)$, by proving that μ and thus also the corresponding policy π^μ is invariant to pushforward by automorphism. Let $g \in \mathrm{Aut}(D)$ and let $\mathcal{Z}_i \subseteq \Pi_i^D$ measurable for $i \in \mathcal{N}$. Recall that by Lemmas 13 and 21, g is a bijective self-map on action-observation histories and g^* a bijective self-map on Π^D, and that isomorphisms can be inverted and composed by Lemma 18. In the following, we use the notations $\pi_i \circ g := \pi_i(g \cdot \mid g \cdot)$ and $\mathcal{Z}_i \circ g := \{\pi_i(g \cdot \mid g \cdot) \mid \pi_i \in \mathcal{Z}_i\}$ for $i \in \mathcal{N}$.

By definition, μ has independent local policies. Hence, we can apply Equation 158 from the proof of Lemma 61, which says that

$$(g^*\mu)(\mathcal{Z}) = \prod_{i \in \mathcal{N}^D} \mu_{g^{-1}i}(\mathcal{Z}_i \circ g). \tag{235}$$

Moreover, it is $(g^*)^{-1}(\Pi^D) = \Pi^D$ and by definition, $(g^*\pi)_i = \pi_{g^{-1}i} \circ g^{-1}$, which implies that

$$\begin{aligned}
\mathrm{proj}_{g^{-1}i}^{-1}(\mathcal{Z}_i \circ g) &= \{\pi \mid \pi \in \Pi^D, \pi_{g^{-1}i} \in \mathcal{Z}_i \circ g\} && (236)\\
&= \{\pi \mid \pi \in (g^*)^{-1}(\Pi^D), \pi_{g^{-1}i} \circ g^{-1} \in \mathcal{Z}_i\} && (237)\\
&= \{\pi \mid g^*\pi \in \Pi^D, (g^*\pi)_i \in \mathcal{Z}_i\} && (238)\\
&= \{(g^*)^{-1}\pi \mid \pi \in \Pi^D, \pi_i \in \mathcal{Z}_i\} && (239)\\
&= (g^*)^{-1}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)). && (240)
\end{aligned}$$

Hence, it follows for $i \in \mathcal{N}$ that

$$\nu^{(i)}(\mathrm{proj}_{g^{-1}i}^{-1}(\mathcal{Z}_i \circ g)) = \nu^{(i)}((g^*)^{-1}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i))) \tag{241}$$
$$= |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} (f^{-1})^* \boldsymbol{\sigma}_i(f^* D)((g^*)^{-1}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i))) \tag{242}$$
$$= |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} g^*(f^{-1})^* \boldsymbol{\sigma}_i(f^* D)(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) \tag{243}$$
$$= |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} ((f \circ g^{-1})^{-1})^* \boldsymbol{\sigma}_i(f^* D)(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) \tag{244}$$
$$= |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} ((f \circ g^{-1})^{-1})^* \boldsymbol{\sigma}_i((f \circ g^{-1})^* D)(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) \tag{245}$$
$$= |\mathrm{Sym}(D)|^{-1} \sum_{f \in \mathrm{Sym}(D)} (f^{-1})^* \boldsymbol{\sigma}_i(f^* D)(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) \tag{246}$$
$$= \nu^{(i)}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)). \tag{247}$$

Here, we use in (245) and (246) that $g^{-1} \in \mathrm{Iso}(D, D)$ and thus by Lemma 34, it is $(f \circ g^{-1})^* D = f^* D$ and $\mathrm{Sym}(D) = \mathrm{Sym}(D) \circ g^{-1}$, respectively.

Next, note that by assumption, it is $\nu^{(i)} = \nu^{(g^{-1}i)}$ for $i \in \mathcal{N}$ (v). It follows that

$$\begin{aligned}(g^*\mu)(\mathcal{Z}) &\overset{(235)}{=} \prod_{i \in \mathcal{N}^E} \mu_{g^{-1}i}(\mathcal{Z}_i \circ g) = \prod_{i \in \mathcal{N}^D} \nu^{(g^{-1}i)}(\mathrm{proj}_{g^{-1}i}^{-1}(\mathcal{Z}_i \circ g)) \\ &\overset{(v)}{=} \prod_{i \in \mathcal{N}^D} \nu^{(i)}(\mathrm{proj}_{g^{-1}i}^{-1}(\mathcal{Z}_i \circ g)) \overset{(241)-(247)}{=} \prod_{i \in \mathcal{N}^D} \nu^{(i)}(\mathrm{proj}_i^{-1}(\mathcal{Z}_i)) \\ &= \prod_{i \in \mathcal{N}^D} \mu_i(\mathcal{Z}_i) = \mu(\mathcal{Z}).\end{aligned} \tag{248}$$

Since the sets $\prod_{i \in \mathcal{N}} \mathcal{Z}_i$, $\mathcal{Z}_i \in \mathcal{F}_i$ are a π-system and generate $\mathcal{F}$, this shows that $\mu = g^*\mu$. By Lemma 61, it is thus

$$g^*\pi^\mu = \pi^{g^*\mu} = \pi^\mu \tag{249}$$

for any $g \in \mathrm{Aut}(D)$, where π^μ is the policy corresponding to μ as defined in Section D.2.

Using Proposition 60, it follows that

$$\begin{aligned}U(\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N) &\overset{(229)-(234)}{=} J^D(\mu) \overset{\text{Proposition 60}}{=} J(\pi^\mu) = \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}[J^D(\pi^\mu)] \\ &= \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J^D\left((\mathrm{proj}_i(\pi^\mu))_{i \in \mathcal{N}}\right)\right] \overset{(249)}{=} \mathbb{E}_{\mathbf{g} \sim \mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[J^D\left((\mathrm{proj}_i(\mathbf{g}_i^*\pi^\mu))_{i \in \mathcal{N}}\right)\right] \\ &= J^D_{\mathrm{OP}}(\pi^\mu) \leq \max_{\pi \in \Pi^D} J^D_{\mathrm{OP}}(\pi) \overset{(224)-(228)}{=} U(\sigma^\chi, \ldots, \sigma^\chi).\end{aligned} \tag{250}$$

This concludes the proof. □

Finally, it follows as a corollary of Theorem 48 that the profile $\sigma^\chi, \ldots, \sigma^\chi$ is a Nash equilibrium of the LFC game.

Corollary 84. *All principals using OP with tie-breaking is a Nash equilibrium of the LFC game for D.*

Proof. By Theorem 83, a profile in which all principals use OP with tie-breaking is an optimal symmetric strategy profile in the LFC game for D. Hence, by Theorem 48, it is a Nash equilibrium of the game. □

F.3. Random tie-breaking functions

In this last section, we show that a certain tie-breaking function, based on random hashes of normal forms of histories, is a tie-breaking function that is invariant to isomorphisms, in the sense of Definition 81 above. As we prove that a certain random function is a tie-breaking function with probability 1, one may call this approach a "probabilistic method" (cf. Alon & Spencer, 2016).

In the following, let $\mathcal{D}$ be any set of Dec-POMDPs and $D \in \mathcal{D}$. Recall that in Section 5, we defined a function ι that maps histories to normal forms, where the normal form is a history in which the first occurrence of each state, action, or observation is set to a 0, the second occurrence is set to a 1, and so on, and where, if an element in $\tau \in \mathcal{H}^D$ repeats itself, the number is repeated. For $\iota(\tau) := (s_0, (a_{i,0})_{i\in\mathcal{N}}, r_0, \dots, s_T, (o_{i,T})_{i\in\mathcal{N}}, (a_{i,T})_{i\in\mathcal{N}}, r_T)$, we then define

$$f_N(\iota(\tau)) := (s_0, (a_{f_N^{-1}i,0})_{i\in\mathcal{N}}, r_0, \dots, s_T, (o_{f_N^{-1}i,T})_{i\in\mathcal{N}}, (a_{f_N^{-1}i,T})_{i\in\mathcal{N}}, r_T).$$

for a permutation $f_N \in \mathrm{Bij}(\mathcal{N})$ of $\mathcal{N}$. The tie-breaking function in Section 5 was defined as

$$\tilde{\chi}^{\#}(D,\pi) := \frac{1}{N!}\sum_{f_N\in\mathrm{Bij}(\mathcal{N})} \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\mathbf{g}^*\pi}\left[\#(f_N(\iota(H)))\right]\right], \tag{251}$$

for some neural random neural network $\#$.

To be able to prove a theoretical result, we have to modify this tie-breaking function to make it dependent on the Dec-POMDP that belongs to a normal form $\iota(\tau)$. The above function is simpler to implement, as we do not have to implement a representation of a Dec-POMDP, and it was sufficient for our experimental results. We leave it to future work to determine to what degree the above version works in general.

In order to define the modified function formally, let $(\Omega_{\#}, \mathcal{E}, \mathbb{P}_{\#})$ be some probability space, such that for any $D \in \mathcal{D}$, $f \in \mathrm{Sym}(D)$ and $\tau \in \mathcal{H}^{f^*D}$, there is an independent, identically distributed real-valued random variable $\#(f^*D, \tau)$ on this space. Assume that

$$\mathbb{P}_{\#}(\#(f^*D,\tau) = \lambda) = 0 \quad \forall \lambda \in \mathbb{R}, D \in \mathcal{D}, f \in \mathrm{Sym}(D), \tau \in \mathcal{H}^{f^*D}. \tag{252}$$

For instance, this would be satisfied by uniformly distributed hash values $\#(f^*D,\tau) \sim \mathcal{U}([0,1])$.

Now for a given Dec-POMDP D and history $\tau \in \mathcal{H}^D$, denote $\mathcal{K}(\tau) \subseteq \mathrm{Sym}(D)$ for the set of labelings $f \in \mathrm{Sym}(D)$ such that $f\tau = \iota(\tau)$. It is easy to see that for any history $\tau \in \mathcal{H}^D$, there must exist some labelings $f \in \mathrm{Sym}(D)$ such that $f\tau = \iota(\tau)$, so $\mathcal{K}(\tau)$ is non-empty. Then our tie-breaking function here is defined as

$$\chi^{\#}(D,\pi) := \mathbb{E}_{\Psi^D(\pi)}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right] \tag{253}$$

for $D \in \mathcal{D}, \pi \in \Pi^D$. We have to sample from the symmetrized policy $\Psi(\pi)$ instead of just π, since we need a distribution over histories under which equivalent histories have equal probabilities, and the tie-breaking function needs to be equal for equivalent policies in order to always have a maximizer. The expectation over labelings in $\mathcal{K}(\tau)$ makes the function invariant to isomorphisms. We need to make the hash function depend on the Dec-POMDP, since otherwise, we are unable to prove that it can distinguish all histories that it needs to be able to distinguish.

Remark 85. By Proposition 60, the distribution over histories under $\mathbb{P}_{\Psi(\pi)}$ is the same as the one under $\mathbb{P}_\mu$ where μ is the OP distribution of $\pi \in \Pi^D$. Hence, using the definition of $\mathbb{P}_\mu$ from Equation 118 and the definition of the OP distribution from Equation 172, it follows that

$$\chi^{\#}(D,\pi) = \mathbb{E}_{\Psi^D(\pi)}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right] \tag{254}$$

$$\overset{\text{Proposition 60}}{=} \mathbb{E}_{\mu}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right] \tag{255}$$

$$\overset{(118)}{=} \mathbb{E}_{\tilde{\pi}\sim\mu}\left[\mathbb{E}_{\tilde{\pi}}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right]\right] \tag{256}$$

$$\overset{(172)}{=} \mathbb{E}_{\mathbf{g}\sim\mathcal{U}(\mathrm{Aut}(D)^{\mathcal{N}})}\left[\mathbb{E}_{\mathbf{g}^*\pi}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right]\right] \tag{257}$$

for any $D \in \mathcal{D}, \pi \in \Pi^D$. Given a realization of the hash function $\#$, one can hence compute an estimate of $\chi^{\#}(D,\pi)$ by sampling a Monte Carlo estimate of the expectation in (257). One can easily see that without the dependence on Dec-POMDPs, this formulation is equal to the one in Equation 251.

Now we show that $\chi^{\#}$ is invariant to isomorphism. Note that this statement holds for *any* sample $\omega \in \Omega_{\#}$ and thus for any sample $\chi^{\#}(\cdot)(\omega)$ of the tie-breaking function. Afterwards, we will show that $\chi^{\#}$ is almost surely a tie-breaking function. We first need a small lemma.

Lemma 86. *Let D, E isomorphic Dec-POMDPs with $f \in \mathrm{Iso}(D,E)$ and let $\tau \in \mathcal{H}^D$. Then it is*

$$\{(\tilde{f}^*E, \tilde{f}(f\tau)) \mid \tilde{f} \in \mathcal{K}(f\tau)\} = \{(\hat{f}^*D, \hat{f}\tau) \mid \hat{f} \in \mathcal{K}(\tau)\}.$$

Proof. Note that for any labeling $\tilde{f} \in \mathcal{K}(f\tau)$, $\tilde{f}f\tau = (\tilde{f}\circ f)\tau$ is in a normal form, and $\tilde{f}\circ f \in \mathrm{Sym}(D)$ by Lemma 34. Hence, also $\tilde{f}\circ f \in \mathcal{K}(\tau)$. An analogous argument from considering f^{-1} shows that for any $\hat{f} \in \mathcal{K}(\tau)$, it is $(\hat{f}\circ f^{-1})(f\tau) = \hat{f}\tau$ in a normal form an thus $\hat{f}\circ f^{-1} \in \mathcal{K}(f\tau)$. It follows that $\mathcal{K}(f\tau)\circ f = \mathcal{K}(\tau)$ (i). Moreover, by Lemma 34, it is $\tilde{f}^*E = (\tilde{f}\circ f)^*D$ (ii), and thus

$$\{(\tilde{f}^*E, \tilde{f}(f\tau)) \mid \tilde{f} \in \mathcal{K}(f\tau)\} \overset{\text{(ii)}}{=} \{((\tilde{f}\circ f)^*D, (\tilde{f}\circ f)\tau) \mid \tilde{f} \in \mathcal{K}(f\tau)\} \tag{258}$$

$$\overset{\text{(i)}}{=} \{(\hat{f}^*D, \hat{f}\tau) \mid \hat{f} \in \mathcal{K}(\tau)\}. \tag{259}$$

□

Using this Lemma, we show the first result.

Proposition 87. *$\chi^{\#}$ as defined in Equation 253 is invariant to isomorphism.*

Proof. Let $D, E \in \mathcal{D}$, $f \in \mathrm{Iso}(D,E)$ and let $\pi \in \Pi^D, \pi' \in \Pi^E$ such that $f^*[\pi] = [\pi']$. We need to show that $\chi^{\#}(D,\pi) = \chi^{\#}(E,\pi')$.

Note that $f^*[\pi] = [f^*\pi]$ by definition, and thus it follows from the assumption that $f^*\pi \equiv \pi'$, which means that $\Psi^E(f^*\pi) = \Psi^E(\pi')$ (i) by the definition of $\equiv$.

It follows that

$$\chi^{\#}(E,\pi') = \mathbb{E}_{\Psi^E(\pi')}\left[\mathbb{E}_{\tilde{f}\sim\mathcal{U}(\mathcal{K}(H^E))}\left[\#(\tilde{f}^*E,\tilde{f}H^E)\right]\right] \tag{260}$$

$$\overset{\text{(i)}}{=} \mathbb{E}_{\Psi^E(f^*\pi)}\left[\mathbb{E}_{\tilde{f}\sim\mathcal{U}(\mathcal{K}(H^E))}\left[\#(\tilde{f}^*E,\tilde{f}H^E)\right]\right] \tag{261}$$

$$= \mathbb{E}_{f^*\Psi^D(\pi)}\left[\mathbb{E}_{\tilde{f}\sim\mathcal{U}(\mathcal{K}(H^E))}\left[\#(\tilde{f}^*E,\tilde{f}H^E)\right]\right] \tag{262}$$

$$= \mathbb{E}_{\Psi^D(\pi)}\left[\mathbb{E}_{\tilde{f}\sim\mathcal{U}(\mathcal{K}(fH^D))}\left[\#(\tilde{f}^*E,\tilde{f}(fH^D))\right]\right] \tag{263}$$

$$= \mathbb{E}_{\Psi^D(\pi)}\left[\mathbb{E}_{\hat{f}\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(\hat{f}^*D,\hat{f}H^D)\right]\right] \tag{264}$$

where in (262), we use that isomorphisms and symmetrizer Ψ commute by Corollary 67, in (263), we use Theorem 22, and in (264), we use Lemma 86.

This shows that $\chi^{\#}$ is invariant to isomorphism. □

To show that $\chi^{\#}$ is almost surely a tie-breaking function, we need two technical assumptions. The first, substantial one, is that the set of OP-optimal equivalence classes of policies is finite, to make sure that there always exists a unique maximizer of $\chi^{\#}$. If this was not the case, our method for finding a tie-breaking function could fail. We leave it to future work to investigate whether and under what conditions this may be the case.

The second, less substantial assumption is that, if two policies π,π' are not equivalent, then $\Psi(\pi)$ and $\Psi(\pi')$ also do not lead to the same distribution over histories. Without this assumption, $\Psi(\pi)$ and $\Psi(\pi')$ might differ on action-observation histories that are never reached, such that the policies induce the same distribution over histories but are not equivalent. Choosing policies which differ in that way should not matter for the objective of the LFC problem, but we wanted to avoid dealing with this complication in our proof that OP with tie-breaking is optimal. For this reason, we have stronger requirements for tie-breaking functions here, thus necessitating this technical assumption.[6]

Proposition 88. *Let $\chi^{\#}$ be defined as in Equation 253. Assume that*

(i) for any $D\in\mathcal{D}$, the set $\Pi^D_{\mathrm{OP}}/\equiv$ is finite, where $\Pi^D_{\mathrm{OP}} := \arg\max_{\pi\in\Pi^D} J^D_{\mathrm{OP}}(\pi)$.

(ii) for any $D\in\mathcal{D}$ and policies $\pi,\pi'\in\Pi^D$, if $\mathbb{P}_{\Psi^D(\pi)}(H=\tau)=\mathbb{P}_{\Psi^D(\pi')}(H=\tau)$ for all $\tau\in\mathcal{H}^D$, it follows that $\Psi^D(\pi)=\Psi^D(\pi')$.

Then $\chi^{\#}$ is $\mathbb{P}_{\#}$-almost surely a tie-breaking function for $\mathcal{D}$.

For the proof, we first need a standard lemma in probability theory.

Lemma 89. *Assume $\mathcal{J}$ is a finite set and $X_j, j\in\mathcal{J}$ is a collection of independent real-valued random variables such that for any $j\in\mathcal{J}$, it is $\mathbb{P}(X_j=\lambda)=0$ for any $\lambda\in\mathbb{R}$. Let $\langle\cdot,\cdot\rangle$ be the euclidean scalar product on $\mathbb{R}^{\mathcal{J}}$. Then it is*

$$\mathbb{P}(\langle X,v\rangle=\lambda)=0$$

[6] Note that in Definition 58, each local policy is defined separately, so that $\Psi(\pi)_i(\cdot\mid\tau_{i,t})$ can in principle be defined as something other than the uniform distribution, even if the action-observation history $\tau_{i,t}$ is never reached under the joint policy $\Psi(\pi)$. Nevertheless, if all agents choose local policies from joint policies that lead to the same distribution over histories, the resulting cross-play policy should also induce that distribution, even if there exist opponent distributions where the local policies might differ.

for any $\lambda \in \mathbb{R}$, $v \in \mathbb{R}^{\mathcal{J}} \setminus \{0\}$.

Proof. Let $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{J}} \setminus \{0\}$. Since $v \neq 0$, there exists an index $j \in \mathcal{J}$ such that $v_j \neq 0$. It follows that

$$\mathbb{P}(\langle X, v\rangle = \lambda) = \mathbb{P}_{\#}\left(X_j = \lambda - \frac{1}{v_j}\sum_{j' \neq j} v_{j'} X_{j'}\right) \tag{265}$$

$$= \int_{\mathbb{R}^{\mathcal{J}}} \mathbb{1}_{x_j = \lambda - \frac{1}{v_j}\sum_{j'\neq j} x_{j'} v_{j'}} \mathrm{d}\mathbb{P} \circ X^{-1}(x) \tag{266}$$

$$= \int_{\mathbb{R}^{\mathcal{J}\setminus\{j\}}} \int_{\mathbb{R}} \mathbb{1}_{x_j = \lambda - \frac{1}{v_j}\sum_{j'\neq j} x_{j'} v_{j'})} \mathrm{d}\mathbb{P} \circ X_j^{-1}(x_j) \mathrm{d}\mathbb{P} \circ X_{-j}^{-1}(x_{-j}) \tag{267}$$

$$= \int_{\mathbb{R}^{\mathcal{J}\setminus\{j\}}} \mathbb{P}\left(X_j = \lambda - \frac{1}{v_j}\sum_{j'\neq j} x_{j'} v_{j'}\right) \mathrm{d}\mathbb{P} \circ X_{-j}^{-1}(x_{-j}) \tag{268}$$

$$= \int_{\mathbb{R}^{\mathcal{J}\setminus\{j\}}} 0 \; \mathrm{d}\mathbb{P} \circ X_{-j}^{-1}(x_{-j}) \tag{269}$$

$$= 0, \tag{270}$$

where we use the assumption $\mathbb{P}(X_j = \lambda) = 0$ for any $\lambda \in \mathbb{R}$ in (269). □

Proof of Proposition 88. To check that $\chi^{\#}$ always admits a maximum among the OP-optimal policies, which is part (a), (i) of Definition 81, let $D \in \mathcal{D}$ arbitrary and let $\pi \equiv \pi' \in \Pi^D$ be equivalent policies. By the definition of $\equiv$, it is $\Psi^D(\pi) = \Psi^D(\pi')$ (*). Hence, it is

$$\chi^{\#}(D, \pi) = \mathbb{E}_{\Psi^D(\pi)}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right] \\ \overset{(*)}{=} \mathbb{E}_{\Psi^D(\pi')}\left[\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(H^D))}\left[\#(f^*D, fH^D)\right]\right] = \chi^{\#}(D, \pi'). \tag{271}$$

This shows that $\chi^{\#}(D, \pi') = \chi^{\#}(D, \pi)$ for any $\pi \in \Pi^D_{\mathrm{OP}}$ and $\pi' \in [\pi]$. Hence, we can define the function

$$\tilde{\chi}^{\#}(D, \cdot)\colon \Pi^D_{\mathrm{OP}}/_{\equiv} \to [0, 1], [\pi] \mapsto \chi^{\#}(D, \pi).$$

Then, using assumption (i), $\tilde{\chi}^{\#}(D, \cdot)$ is maximized by some equivalence class $[\pi] \in \Pi^D_{\mathrm{OP}}/_{\equiv}$. By definition of $\tilde{\chi}^{\#}(D, \cdot)$, it follows that for any $\pi' \in \Pi^D_{\mathrm{OP}}$, it is

$$\chi^{\#}(D, \pi') = \tilde{\chi}^{\#}(D, [\pi']) \leq \tilde{\chi}^{\#}(D, [\pi]) = \chi^{\#}(D, \pi),$$

so π maximizes $\chi^{\#}(D, \cdot)$ on the set Π^D_{OP}.

To prove part (a), (ii) of the definition, consider two non-equivalent policies $\pi, \pi' \in \Pi^D$, i.e., assume that $\Psi(\pi) \neq \Psi(\pi')$. We now want to show that $\mathbb{P}_{\#}\left(\chi^{\#}(D, \pi) = \chi^{\#}(D, \pi')\right) = 0$. If this is true, then $\chi^{\#}$ is $\mathbb{P}_{\#}$-almost surely a tie-breaking function.

First, note that actions of automorphisms on histories are group actions (see Section C.4). Thus, we can consider a partition of the set of histories into orbits, $\mathcal{J} := \{\mathrm{Aut}(D)\tau \mid \tau \in \mathcal{H}^D\}$. Moreover, note that since $\Psi^D(\pi), \Psi^D(\pi')$ are both invariant to automorphism by Corollary 67, using Theorem 22, we can conclude that for any $\tau \in \mathcal{H}^D$ and $\tilde{\tau} \in \mathrm{Aut}(D)\tau$, it is $\mathbb{P}_{\Psi(\pi)}(H=\tau) = \mathbb{P}_{\Psi(\pi)}(H=\tilde{\tau})$. The same holds for $\Psi(\pi')$. This step only works since we sample from $\Psi(\pi)$, instead of from the original policy π. As a result, it follows that we can define a vector $v \in \mathbb{R}^{\mathcal{J}}$ such that $v_j := |\mathrm{Aut}(D)\tau|\mathbb{P}_{\Psi(\pi)}(H=\tau)$ for $j \in \mathcal{J}$, where $\tau \in \mathcal{H}^D$ is arbitrary such that $\mathrm{Aut}(D)\tau = j$. Define v' analogously for π'. By assumption (ii), $\Psi(\pi)$ and $\Psi(\pi')$ induce different distributions over histories, so it must also be $v \neq v'$, since the map from orbit-invariant distributions to vectors v is injective.

Second, note that by Lemma 86, it is

$$\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}\left[\#(f^*D, f\tau)\right] = \mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(g\tau))}\left[\#(f^*D, fg\tau)\right]$$

for any $\tau \in \mathcal{H}^D$ and $g \in \mathrm{Aut}(D)$, and thus for any $j \in \mathcal{J}$ we can define the random variable $X_j := \mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}\left[\#(f^*D, f\tau)\right]$, where τ is an arbitrary history such that $j = \mathrm{Aut}(D)\tau$.

We now want to show that $\mathbb{P}_\#(\langle X, v - v'\rangle = 0) = 0$, where $\langle\cdot,\cdot\rangle$ is the euclidean scalar product on $\mathbb{R}^{\mathcal{J}}$. This will then allow us to conclude that $\mathbb{P}_\#\left(\chi^\#(D,\pi) = \chi^\#(D,\pi')\right) = 0$. We already know that $v - v' \neq 0$. To be able to apply Lemma 89 to show this, it remains to prove that for any two $j \neq j' \in \mathcal{J}$, we also have two independent variables $X_j, X_{j'}$, and that $\mathbb{P}_\#(X_j = \lambda) = 0$ for any $\lambda \in \mathbb{R}$.

To prove independence, let $j = \mathrm{Aut}(D)\tau, j' = \mathrm{Aut}(D)\tau'$, and assume towards a contradiction that there exist $f \in \mathcal{K}(\tau), \tilde{f} \in \mathcal{K}(\tau')$ such that $f^*D = \tilde{f}^*D$ and $f\tau = \tilde{f}\tau'$. Then it follows that $\tau = f^{-1}\tilde{f}\tau'$ and thus $j = j'$, which is a contradiction. It follows that the sets of variables $\{\#(f^*D, f\tau) \mid f \in \mathcal{K}(\tau)\}$ and $\{\#(\tilde{f}^*D, \tilde{f}\tau') \mid \tilde{f} \in \mathcal{K}(\tau')\}$ are disjoint. Since all the contained random variables are independent, also $\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}\left[\#(f^*D, f\tau)\right]$ and $\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau'))}\left[\#(f^*D, f\tau')\right]$ are independent random variables.

Next, let $j \in \mathcal{J}$. To prove that $\mathbb{P}_\#(X_j = \lambda) = 0$ for any $\lambda \in \mathbb{R}$, let $\tau \in \mathcal{H}^D$ such that $j = \mathrm{Aut}(D)\tau$ arbitrary and let $\lambda \in \mathbb{R}$. Let $M := \left|\bigcup_{f\in\mathcal{K}(\tau)}\{\#(f^*D, f\tau)\}\right|$ and define $Y_1, \dots, Y_M$ as random variables such that $\{Y_1, \dots, Y_M\} = \bigcup_{f\in\mathcal{K}(\tau)}\{\#(f^*D, f\tau)\}$. Define $w \in [0,1]^M$ via $w_m := \mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}[\delta_{Y_m, \#(f^*D, f\tau)}]$ for $m = 1, \dots, M$. By assumption (252), we have $\mathbb{P}_\#(Y_m = \lambda') = 0$ for any $\lambda' \in \mathbb{R}$ and $Y_m, Y_{m'}$ are by definition independent variables for $m \neq m' \in \{1, \dots, M\}$. We can hence apply Lemma 89 to conclude that

$$\begin{aligned}
\mathbb{P}_\#(X_j = \lambda) &= \mathbb{P}_\#\left(\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}[\#(f^*D, f\tau)] = \lambda\right) && (272)\\
&= \mathbb{P}_\#\left(\sum_{m=1}^{M} Y_m \mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}[\delta_{Y_m, \#(f^*D, f\tau)}] = \lambda\right) && (273)\\
&= \mathbb{P}_\#\left(\sum_{m=1}^{M} Y_m w_m = \lambda\right) && (274)\\
&= \mathbb{P}_\#\left(\langle Y, w\rangle = \lambda\right) && (275)\\
&= 0. && (276)
\end{aligned}$$

Since we have shown above that $v \neq v'$, that $X_j, X_{j'}$ are independent for $j \neq j' \in \mathcal{J}$, and that $\mathbb{P}_\#(X_j = \lambda) = 0$ for any $\lambda \in \mathbb{R}$ and $j \in \mathcal{J}$, we can apply Lemma 89 again to get

$$\mathbb{P}_\#(\langle X, v - v'\rangle = 0) = 0. \tag{277}$$

Finally, for $j \in \mathcal{J}$ let $\tau^{(j)} \in \mathcal{H}^D$ arbitrary such that $j = \mathrm{Aut}(D)\tau^{(j)}$. Then

$$\begin{aligned}\langle X, v\rangle &= \sum_{j\in\mathcal{J}} |\mathrm{Aut}(D)\tau^{(j)}| \mathbb{P}_{\Psi^D(\pi)}(H^D = \tau^{(j)}) X_j \\ &= \sum_{j\in\mathcal{J}} |\mathrm{Aut}(D)\tau^{(j)}| \mathbb{P}_{\Psi^D(\pi)}(H^D = \tau^{(j)}) \mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau^{(j)}))}\left[\#(f^*D, f\tau^{(j)})\right] \\ &= \sum_{\tau\in\mathcal{H}^D} \mathbb{P}_{\Psi^D(\pi)}(H^D = \tau)\mathbb{E}_{f\sim\mathcal{U}(\mathcal{K}(\tau))}\left[\#(f^*D, f\tau)\right] = \chi^\#(D,\pi) \end{aligned} \tag{278}$$

and analogously

$$\langle X, v'\rangle = \chi^\#(D,\pi'). \tag{279}$$

Hence, it is

$$\{\chi^\#(D,\pi) = \chi^\#(D,\pi')\} = \{\langle X, v\rangle = \langle X, v'\rangle\}. \tag{280}$$

It follows that

$$\mathbb{P}_\#(\chi^\#(D,\pi) = \chi^\#(D,\pi')) \overset{(280)}{=} \mathbb{P}_\#(\langle X, v\rangle = \langle X, v'\rangle) = \mathbb{P}_\#(\langle X, v - v'\rangle = 0) \overset{(277)}{=} 0.$$

This concludes the proof. □