Reliably Re-Acting to Partner's Actions with the Social Intrinsic Motivation of Transfer Empowerment

Tessa van der Heiden[1], Herke van Hoof[2], Efstratios Gavves[2] and Christoph Salge[3]

[1]BMW Group
[2]University of Amsterdam
[3] University of Hertfordshire
tessavdheiden@gmail.com

Abstract

We consider multi-agent reinforcement learning (MARL) for cooperative communication and coordination tasks. MARL agents can be brittle because they can overfit their training partners' policies. This overfitting can produce agents that adopt policies that act under the expectation that other agents will act in a certain way rather than react to their actions. Our objective is to bias the learning process towards finding reactive strategies towards other agents' behaviors. Our method, transfer empowerment, measures the potential influence between agents' actions. Results from three simulated cooperation scenarios support our hypothesis that transfer empowerment improves MARL performance. We discuss how transfer empowerment could be a useful principle to guide multi-agent coordination by ensuring reactiveness to one's partner.

Introduction

In this paper we investigate if and how social intrinsic motivation can improve Multi-agent reinforcement learning (MARL). MARL holds considerable promise to help address a variety of cooperative multi-agent problems - both for problem solving and simulation of multi-agent systems. However, one problem with MARL is that agents develop strong policies that are overfitted to their partners' behaviors. Specifically, with centralised training, agents can adopt strategies that expect other agents to act in a certain way rather than reacting to their actions. Such systems are undesirable as they may fail when their partners alter their strategies or have to collaborate with novel partners, either during the learning or deployment phase. Our aim is to avoid this specific lack of robustness and find a guiding principle that makes agents stay reactive to other agents' policy changes.

We want to introduce an additional reward to bias learning towards socially reactive strategies which should fulfil the following constraints: 1) it should, with minimal adaptation, apply to a wide range of problems with different sensor-actuator configurations to preserve the universality of the RL framework, and 2) it should not negatively affect the performance, i.e., once good policies are found, it should not harm exploitation. Fulfilling the above criteria would provide a general-purpose multi-agent learning algorithm for various cooperative tasks - as well as provide insights into general principles that would enable and improve the development of various forms of social interaction.

To address this challenge, we turn towards the idea of using Intrinsic motivation (IM) - a school of computational models (Oudeyer and Kaplan, 2009) that try to capture the essential motivations behind the behavior of (biological) agents - and then use them for behavior generation to obtain plausible and beneficial behavior. The core idea here is to ask if the principles that create single agent behavior can also be used to enhance multi-agent behavior. In this paper specifically, we look at Empowerment, an IM that captures how much an agent is able to affect the world it can itself perceive. Its information-theoretic formulation as the channel capacity between an agent's actions and its own sensors makes it a versatile measure that can be applied to a wide range of models where agent's are defined - satisfying constraint 1. Existing work on coupled empowerment maximisation (Salge and Polani, 2017; Guckelsberger et al., 2018) extends the formalism to a multi-agent setting. In this paper, we focus specifically on the idea of Transfer Empowerment (TE), introduced in those papers, which tries to capture how much one agent can potentially influence the actions of another. We use a slightly modified version of TE, which considers the channel capacity from one agent's actions to another agent's *actions*, rather than to their sensors, as in traditional TE.

Keeping the TE high between two agents means they are in a state were one of them is reliably reacting to the other. Adding this as an additional reward mechanism during training should help to avoid the brittleness of over-fitting we outlined before. We provide here quantitative evidence for our hypothesis that adding transfer empowerment as an additional reward increases upon the performance of state-of-the art MARL methods. Constraint 2 will be evaluated empirically. We also compare this approach to a similar idea of social influence by Jaques et al. (2019). First, we will introduce the concepts of MARL and IM in more detail. We will then define the specific formalism for TE used, and then simulate three increasingly harder, multi-agent, cooperation

scenarios. We will also look at how the better reward was obtained, and discuss the difficult switch from a indexical to an action oriented communication strategy.

Related work

Multi-Agent Reinforcement Learning

There is a large body of research on constructing agents that are robust to their partners. In self-play, for example, agents train against themselves rather than a fixed opponent strategy to prevent developing exploitable strategies (Tesauro, 1994). Population based-training goes one step further by training agents to play against a population of other agents rather than only a copy of itself. For instance, some methods train an ensemble of policies with a variety of collaborators and competitors (Jaderberg et al., 2018; Lowe et al., 2017). By using a whole population rather than only a copy of itself, the agent is forced to deal with a wide variety of potential strategies instead of a single strategy. However, it requires a great deal of engineering because the policy parameters suitable for the previous environment are not necessarily the next stage's best initialization.

Some works combine the minimax framework and MARL to find policies that are robust to opponents with different strategies. Minimax is a concept in game theory that can be applied to find an approach that minimizes the possible loss in a worst-case scenario (Osborne et al., 2004). Li et al. (2019) use it during training to optimize the reward for each agent under the assumption that all other agents act adversarial. We are interested in methods that can deal with perturbations in the training partners' behavior, which differs from dealing with partners with various strategies.

Recent works look at settings in which one RL agent, that is trained separately, must join a group of new agents (Lerer and Peysakhovich, 2018; Tucker et al., 2020; Carroll et al., 2019). For example, Carroll et al. (2019) build a model of the other agents which can be used to learn an approximate best response using RL. Lupu et al. (2021) propose to generate a large number of diverse strategies and then train agents that can adapt to other agents' strategies quickly using meta-learning. A related problem is zero-shot coordination (Hu et al., 2020) in which agents need to cooperate with unseen partners at test time. The focus of our paper is not to perform well with novel partners at test-time or build complex opponent models. Our aim is to train agents together to remain attentive and reactive towards their partners' policies.

Intrinsic Social Motivation

Due to centralized training in MARL, agents might adopt non-reactive strategies that may struggle with other agents' changing behaviors. Social intrinsic motivation can give an additional incentive to find reactive policies towards other agents.

IM in Reinforcement learning (RL) refers to reward functions that allow agents to learn interesting behavior, sometimes in the absence of an environmental reward (Chentanez et al., 2005). Computational models of IM are generally separated into two categories (Baldassarre and Mirolli, 2013), those that focus on curiosity (Burda et al., 2018; Pathak et al., 2017) and exploration (Gregor et al., 2016; Eysenbach et al., 2018), and those that focus on competence and control (Oudeyer and Kaplan, 2009; Karl et al., 2017). The information-theoretic Empowerment formalism (Klyubin et al., 2005) is in the latter category, trying to capture how much an agent is in control of the world it can perceive. Empowerment has produced robust behavior linked to controllability, operationality and self-preservation - in both robots (van der Heiden et al., 2020; Karl et al., 2017; Leu et al., 2013) and simulations (Guckelsberger et al., 2016), with (de Abril and Kanai, 2018) and without (Guckelsberger et al., 2018) reinforcement learning and neural network approximations (Karl et al., 2017).

Empowerment has also been applied to multi-agent simulations, under the term of coupled empowerment maximization (Guckelsberger et al., 2016), in which it was used to produce supportive and antagonistic behavior. Of particular interest is the idea of transfer empowerment - introduced in those two papers - a measure that quantifies concepts such as operational proximity and social influence, and led to behaviours such as collaboration, coordination, and lead-taking (Salge and Polani, 2017).

Similar techniques quantify the interaction between agents for improving coordination between agents. Barton et al. (2018) analyze the degree of dependence between two agents' policies to measure coordination, specifically by using Convergence Cross Mapping (CCM). Strouse et al. (2018) show how agents can share (or hide) intentions by maximizing the mutual information between actions and a categorical goal. One notably relevant work is by Jaques et al. (2019) called social influence, which is the influence of one agent on the policies of other agents, measured by the mutual information between action pairs of distinct agents. Similarly, Mahajan et al. (2019), compute the mutual information between agents' trajectories and a latent variable that captures the joint behavior. Wang et al. (2019) compute the mutual information between the transition dynamics of agents.

In contrast to social influence (SI), transfer empowerment considers the *potential* mutual information or channel capacity. When optimizing for *actual* mutual information, its value is bounded from above by the lowest entropy of both agent's action variables. SI might easily interfere with an exploitation strategy and may need regularization once a good strategy is found. On the other hand, empowerment does not have this limitation and the action sets could have very narrow distributions, while still being reactive.

Model

First, let us define a general model that captures multi-agent scenarios and lets use define transfer empowerment. Let us consider a Dec-POMDP, an extension of the MDP for multi-agent systems, being both decentralized and partially observable (Nair et al., 2003). This means that each of the N agents conditions the choice of its action on its partial observation of the world. It is defined by the following tuple: $\langle S, \boldsymbol{A}, T, O, \boldsymbol{O}, R, N \rangle$. S is the set of states and $\boldsymbol{A} = \times_{i \in [1,..,n]} \boldsymbol{A}^i$ the set of joint actions. At each time step, the state transition function $P(s_{t+1}|s_t, \boldsymbol{a}_t)$ maps the joint action and state to a new state. As the game is partially observable, we have a set of joint local observations, $\boldsymbol{O} = \times_{i \in [1,..,n]} \boldsymbol{O}^i$ and an observation function O. Each agent i selects an action using their local policy $\pi^i(a_t^i|o_t^i)$.

We consider fully cooperative tasks, so agents share a reward $r(s_t, \boldsymbol{a_t})$ which conditions on the joint action and state. The goal is to maximise the expected discounted return $J(\boldsymbol{\pi}) = \mathbb{E}_{\tau \sim \pi}\left[R(\tau)\right] = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$, with discount factor $\gamma \in [0, 1]$ and horizon T. The expectation is taken w.r.t. the joint policy $\boldsymbol{\pi} = [\pi^1, \ldots, \pi^N]$ and trajectory $\tau = (\boldsymbol{o}_0, \boldsymbol{a}_0, \ldots, \boldsymbol{o}_T)$.

Methodology

This section describes an additional heuristic that biases the learning process in obtaining policies that are reactive to other agents' actions. First, we introduce our specific version of transfer empowerment, which rewards the idea of an agent being responsive to adaptations in the other's policy. Then we explain how to train agents in a multi-agent environment.

Transfer Empowerment

Consider two agents, j and k, both taking actions and changing the overall state. Each time agent k acts, the state of agent j is modified, and j's policy indirectly conditions on k's actions. The objective of coordination is that by changing the actions of agent k, agent, j also *reliably* adapts its actions. Here we look at transfer empowerment, namely the *potential* causal influence that one agent has on another. It is defined for pairs of agents by the channel capacity between one agent's action a_t^k and another agent's action a_{t+1}^j at subsequent time steps and conditioned on the current state s_t, which can be computed by maximizing the mutual information $\mathcal{I}$ between those values, with regards to ω^k :

$$\mathcal{E}^{k \to j}(s_t) = \max_{\omega^k} \mathcal{I}\left(A_{t+1}^j, A_t^k \middle| s_t\right). \tag{1}$$

Here, $\omega^k(a_t^k|s_t)$ is the *hypothetical* policy of agent k, that takes an action a_t^k after observing state s_t and influencing a_{t+1}^j at a later time step. Note that the policy $\omega^k(a_t^k|s_t)$ that maximises the mutual information is not necessarily used for action generation, but simply to compute the channel capacity by looking at all potential policies for the one with the highest mutual information $\mathcal{I}$.

Our version of transfer empowerment differs slightly from the one introduced by Salge and Polani (2017), as we consider the potential information flow, or channel capacity, from one agent's actions to another agent's *actions* in subsequent time steps. Salge and Polani (2017) on the other hand, consider the transfer empowerment between one agent's action and another agent's *sensor* state. We make this modification to address to challenges already discussed in those earlier papers.

One issues is that transfer empowerment to another agent's sensory state captures the direct influence on the other agent's environment. This influence, or information flow can take two pathways. The influence can either act directly on the world the other agent perceives, or alternatively, the other agent can perceive the action's of the first agent, and react to them, modifying their own Umwelt. The difference is that in the second case the information flows through the second agent, and requires a degree of attention to, and reactivity to the first agent's actions. In the first case, the second agent can be fully passive and just have its perceived world changed by the first agent. Since we wanted to create a motivation for more reactivity, we used action-to-action TE, because the only way to influence another agent's actions is by having them react to what you do, i.e. have information flow through the other agent.

The other challenge of TE is the necessity to define distinct sensor states for both agents - otherwise TE is identical to self-empowerment. If every agent only has sensor access to a limited part of the world, this is straight forward. But in many simpler models the access to the world is absolute for both agents, or limitations end up being somewhat arbitrary, or design choices that influence the final behaviour. Looking at the actions, rather then the sensor states, offers a principled alternative, as action's of different agent are usually distinct.

Transfer empowerment has ties with, but is different from, social influence (Jaques et al., 2019). Social influence is the mutual information between agents' actions. It is high when both action variables have a particular entropy, e.g., policies taking different actions. However, towards the end of the training, a high entropy policy distribution might be suboptimal. Our method, on the other hand, considers the *potential* and not *actual* information flow, so agents only calculate how they *could* influence and react to each other, rather than carrying out its potential. As such, action sets can have very narrow distributions; as long as the system would still be reactive *if*, those actions change. Therefore it does not interfere with obtaining optimal policies.

Multi-Agent Training

Training with transfer empowerment results in joint policies that are reactive to their partner's actions, because for the value to be high, it requires considering the decisions of others. As such, transfer empowerment rewards a very general idea of coordination that requires paying attention to each other, and reliably reacting to a variation in their actions. While empowerment does not measure how this reaction looks, or even if it is good, combined with the actual reward should lead to the selection of a strategy that both solves the problem while also avoiding the brittleness that comes from not being reactive to the information from other agents' policies. Specifically, we will modify the agents' reward function so that it becomes:

$$\tilde{r}(s_t, \boldsymbol{a_t}) = r(s_t, \boldsymbol{a_t}) + \sum_{s_{t+1}} P(s_{t+1}|s_t, \boldsymbol{a_t}) \sum_{j=1}^{N} \mathcal{E}^{-j \to j}(s_{t+1}), \tag{2}$$

where $-j$ means all agents excluding agent j. To simplify notation, we will use j instead of $-j \to j$ in the superscript. The new RL objective becomes:

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\tau \sim \boldsymbol{\pi}} \left[\sum_{t=0}^{T} \gamma^t \tilde{r}(s_t, \boldsymbol{a_t}) \right].$$

This new return motivates the potential influence of information between agents' actions, thereby stimulating them to act informatively and react reliably.

Efficient Implementation

We now introduce an efficient implementation to estimate empowerment. We use a for agent k's action at time t and a' for agent j's action at time $t+1$. Mutual information is defined as:

$$\begin{aligned} \mathcal{I}(A, A'|s) &= \mathrm{KL}(p(a, a'|s) \| p(a|s)p(a'|s)) \\ &= \sum_a \sum_{a'} p(a, a'|s) \ln \frac{p(a, a'|s)}{p(a|s)p(a'|s)}, \end{aligned}$$

where KL is the KL divergence. We can substitute $p(a, a'|s)$ and cancel out terms:

$$\begin{aligned} &\sum_a \sum_{a'} p(a, a'|s) \ln \frac{p(a, a'|s)}{p(a|s)p(a'|s)} \\ &= \sum_a \sum_{a'} p(a, a'|s) \ln \frac{p(a|a', s)\cancel{p(a'|s)}}{p(a|s)\cancel{p(a'|s)}} \\ &= \sum_a \sum_{a'} p(a, a'|s) \ln \frac{p(a|a', s)}{p(a|s)}. \end{aligned}$$

By choosing a variational approximator $q(a|a', s)$, with the property $\mathrm{KL}(p(a|a', s)||q(a|a', s)) \geq 0$, we obtain a lower bound on the mutual information:

$$\begin{aligned} \mathcal{I}(A, A'|s) &\geq \hat{\mathcal{I}}(A, A'|s) \\ &:= \sum_a \sum_{a'} p(a, a'|s) \ln \frac{q(a|a', s)}{p(a|s)} \\ &= \sum_a \sum_{a'} p(a, a'|s) \left(\ln q(a|a', s) - \ln p(a|s) \right) \\ &= \mathbb{E}_{p(a,a'|s)} \left[\ln q(a|a', s) - \ln p(a|s) \right]. \end{aligned}$$

The gradient of the lower bound can be approximated by Monte-Carlo sampling. Furthermore, the overall training procedure can be implemented efficiently when representing the distributions by neural networks and using gradient ascent. So the gradient computed over S samples:

$$\begin{aligned} \nabla_\theta \hat{\mathcal{I}}_\theta(A, A'|s) &= \nabla_\theta \mathbb{E}_{p(a,a'|s)} \left[\ln q_\theta(a|a', s) - \ln \omega_\theta(a|s) \right] \\ &\approx \frac{1}{S} \sum_{m=1}^{S} \nabla_\theta \left(\ln q_\theta(a_m|a'_m, s) - \ln \omega_\theta(a_m|s) \right), \end{aligned}$$

where we substituted $p(a|s)$ with $\omega_\theta(a|s)$ and $q(a|a', s)$ with $q_\theta(a|a', s)$, to denote functions parametrized by θ.

Partial Observable

The objective in the previous section was to estimate the empowerment value for a particular state s. However, our main goal is to train policies to be reactive towards the actions of their partners. Let a policy for agent j be π^j_χ with parameters χ. As each policy is conditioned on its local observations, the lower bound on mutual information for agent j becomes:

$$\begin{aligned} &\hat{\mathcal{I}}^j_{\theta,\chi}(A, A'|\boldsymbol{o}) = \\ &\quad \mathbb{E}_{\boldsymbol{o'} \sim p_\nu, a^j \sim \pi^j_\chi, a^k \sim \omega^k_\theta} \left[\ln q_\theta(\boldsymbol{a}|\boldsymbol{o}, \boldsymbol{o'}, a'^j) - \ln \boldsymbol{\omega}^{-j}_\theta(\boldsymbol{a}^{-j}|\boldsymbol{o}^{-j}) \right], \end{aligned} \tag{3}$$

where samples are generated by a learned transition model $\boldsymbol{o'} \sim p_\nu(\boldsymbol{o'}|\boldsymbol{o}, \boldsymbol{a})$. The actions are selected by the target policy $a^j \sim \pi^j_\chi(a^j|o^j)$ and behavior policy $a^k \sim \omega^k_\theta(a^k|o^k)$ and the joint action is $\boldsymbol{a} = (a^1, \ldots, a^j, \ldots, a^N)$ where $a^j \sim \pi_\chi$, $a^k \sim \omega_\theta$ and $k \neq j$.

Notice that the actions come from $\boldsymbol{\omega}_\theta$ and π_χ. The former is the joint behavior policy and the latter is the target policy of agent j. The behavioral policy is only used to train agent j's policy with empowerment but will not generate extrinsic environmental rewards.

This training procedure has two interesting properties. First, it estimates a state's empowerment value. This is done by increasing the diversity of agents' actions while ensuring that these are retrievable from agent j's actions. Actions

that affect j's policy, e.g., informative, are chosen more often than those with a lower effect. Second, it trains agent j's policy to be reactive towards the actions of its partners, because we compute the gradient of mutual information w.r.t. χ to directly optimize π_χ. We provide the description of the full algorithm in the Appendix, which also describes how our method applies to settings with more than 2 agents.

Altogether, empowerment prefers states that allow for information flow between agents, altering policies to be more responsive. We will experimentally verify this in the next section.

Experimental Results

We adopt the simulator developed for testing multi-agent reinforcement learning algorithms [1] that allows creating cooperative and competitive environments. Agents have a continuous observation space and a discrete actions space.

Scenarios

We use a cooperative two-dimensional environment consisting of two agents. The (disembodied) speaker agent can, each time step, choose a communication action that is broadcast to the listener. The listener can choose from 5 physical actions, moving up, down, left and right, or doing nothing. The environment contains a series of L, randomly placed, landmarks. Only the speaker has a signal informing it which of the L landmarks is the target. The objective for the randomly placed listener is to reach the target landmark by decoding the speaker's message. The speaker can send a symbol chosen from a set of C distinct symbols. The team reward is the negative squared distance between the listener and the target landmark, which is given out every time step. The game ends after 100 time steps. To perform well the listener has to quickly move onto the landmarks. We developed three tasks in the environment with increasing difficulty. Fig. 2 visualises the tasks.

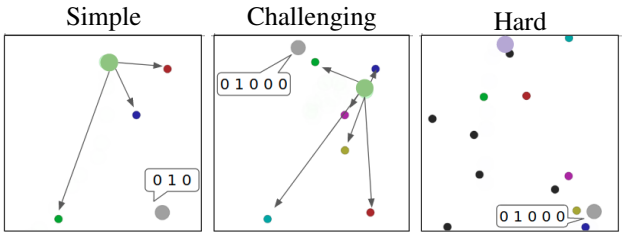

Figure 2: Visualizations of the three tasks. Small dark circles indicate landmarks and obstacles, and the big circles are the listener and speaker. The listener observes the relative distance to the landmarks indicated by the arrows. The speaker's massages are one-hot vectors displayed by the speaker boxes.

Simple The number of symbols $K = |C| = 3$ equals the number of landmarks $L = 3$. The speaker observes the color of the target landmark, while the listener sees a distance vector pointing to each colored landmark.

This scenario could be solved well by an indexical communication strategy, where the speaker simply has to consistently assign a symbol to each landmark color, and then simply relay the information to the speaker, who then has to minimise the distance to the landmark of that color.

Challenging The second task involves more landmarks $L = 6$ than distinct symbols $K = |C| = 5$. The speaker observes the target *position* and the listener's position, while the listener observes the landmarks' positions and the messages sent by the speaker. Here, an *action-oriented strategy*, e.g., indicating movement direction, is likely optimal because the speaker cannot use each symbol for a landmark

[1] https://github.com/openai/multiagent-particle-envs

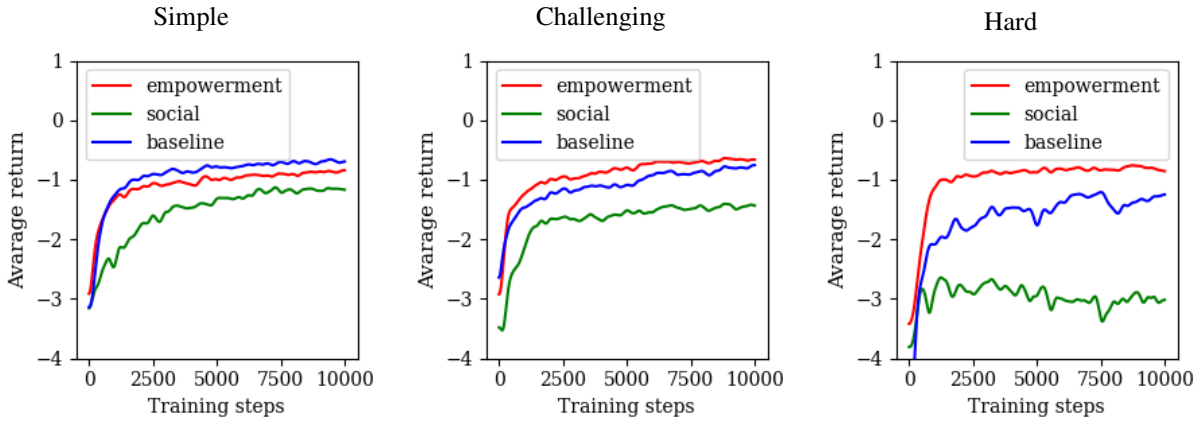

Figure 1: Learning curves for the the three tasks. The rewards are averaged over the steps in an episode to obtain the return. The returns are averaged over three training runs.

uniquely, nor do the landmarks have any identifying features that are easy to community, i.e. they are not colored anymore. Using symbols to direct the listener now requires the speaker to observe and react to the listener's position with an updated signal, and put more cognitive demands on the speaker, who could simply relay its internal signal in the simple scenario.

Hard The last task adds $M = 6$ obstacles, and the reward includes a penalty if the listener hits an obstacle. Furthermore, the landmarks' positions are now unobserved by the listener. These two features increase the difficulty because a higher precision is required. First, the listener has to avoid obstacles, and second, the listener is even more dependent on the speaker because it does not see the landmarks.

Reward The reward function is determined from the position, $\boldsymbol{p} = [p_x, p_y]$, of the listener (agent 1) $\boldsymbol{p}^1$, target $\boldsymbol{p}^g$ and obstacles $\boldsymbol{p}^o$. The state is defined as $s_t = [\boldsymbol{p}_t^1, \boldsymbol{m}_t, \boldsymbol{p}^g, \boldsymbol{p}^{o,1}, \ldots, \boldsymbol{p}^{o,M}, \boldsymbol{p}^{l,1}, \ldots, \boldsymbol{p}^{l,L}]$ for M obstacles, and L landmarks. The reward function is:

$$r(s_t, \boldsymbol{a}_t) = -\|\boldsymbol{p}_{t+1}^1 - \boldsymbol{p}^g\| + \text{penalty} \tag{4}$$

$$\text{penalty} = \begin{cases} -1 & \text{if} \quad \exists j \in [1, ..., M] : \|\boldsymbol{p}_{t+1}^1 - \boldsymbol{p}^{o,j}\| < 0.15 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The observation for the speaker and listener $o_t^0 = [\boldsymbol{p}_t^1, \boldsymbol{p}^g, \boldsymbol{p}^{l,1}, \ldots, \boldsymbol{p}^{l,L}]$ and $o_t^1 = [\boldsymbol{v}_t^1, \boldsymbol{m}_t^0]$, respectively. v is the velocity and m the message, represented by a one-hot vector. The listener's action is a force vector $a^1 = [f_x, f_y]$, while the speaker's action is a message $a^0 = [c^0, \ldots, c^K]$ with vocabulary size K. The listener's position is updated according to the following equation:

$$\begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{v} \\ \dot{\boldsymbol{v}} \end{bmatrix}_{t+1} = \begin{bmatrix} \boldsymbol{p} + \boldsymbol{v}\Delta t \\ \zeta \boldsymbol{v} + \dot{\boldsymbol{v}}\Delta t \\ \frac{\boldsymbol{u}}{\text{mass}} \end{bmatrix}_t \tag{6}$$

with damping coefficient $\zeta = 0.5$ and mass $= 1$. The speaker's messages, will be added to the state at the next time-step: $\boldsymbol{m}_t = \boldsymbol{a}_{t-1}^0$. As is common when working with policies parameterised by neural networks, the actions are one-hot vectors, obtained by Gumbel-Softmax function (Murphy, 2022). For example, the actions of the speaker are converted into

$$\text{one-hot}(\boldsymbol{a^0}) = [\mathbb{I}(a_1^0 = \max(\boldsymbol{a^0})), \ldots, \mathbb{I}(a_K^0 = \max(\boldsymbol{a^0}))].$$

Comparison and Implementation details

We compare our method with MADDPG (Lowe et al., 2017) (baseline) and social influence (Jaques et al., 2019) (social infl.). Our method (empowerment) is built on top of the MADDPG, a centralized actor-critic method. Social influence is a decentralized method. The agents' policies are parameterized by a two-layer ReLU MLP with 64 units per layer. The messages sent between agents are soft approximations to discrete messages, calculated using the Gumbel Softmax estimator. All models are trained for 10k episodes, of which an episode consists of 25 interactions. Source code can be found at https://github.com/tessavdheiden/social_empowerment.

Results and Discussion

Learning Curves

Our two main hypothesis are that adding TE to MARL produces faster adaptation (needs less training steps), and achieves better, overall results. To answer both of the questions, we compare the learning curves, over 10k training steps, averaged over three runs, for both the MARL baseline, and with the addition of TE (empowerment) and Social influence (SI) (social). Figure 1 shows the averaged return after a given number of training steps. A higher score is better, it shows that the listener is closer to the target. Since the listener starts away from the target a score of 0 is impossible, all scores are negative.

The learning speed seems to be comparable between models in difference scenarios, i.e. it takes about the same time for the different algorithm to reach their peak, final performance. Only SI seems to learn slower in both the simple and challenging task. Performances seem to mostly stabilise after some point, so we can also take a closer look at the performances of the trained agents after 10k training steps.

Final Scores

Details of the results are presented in Table 1. It shows the average distance and the percentage of collisions with an obstacle for the final agents, computed for 100 episodes.

The baseline obtains the top performance, with lowest distance of 0.221, in the simple task, requiring a simple, lexical strategy. Here empowerment performs worse with 0.414, with social influence performing even worse with

Table 1: The values show the average distance between the listener and target landmark and the percentage of collisions with obstacles. The results are computed for 100 episodes after training with 10k episodes.

	Simple	Challenging	Hard	
	Average distance	Average distance	Average distance	Obstacle hit
basel.	0.221	0.440	0.520	0.603
SI	0.716	0.949	1.076	0.220
TE	0.414	0.460	0.440	0.266

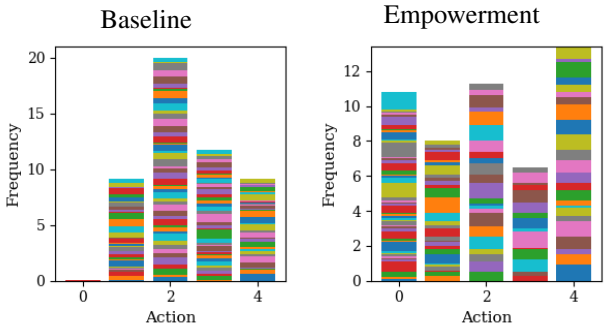

Figure 3: The action distributions for the listener in the hard task for both the baseline and empowerment. The colors indicate each a different episode. 0 is the wait action, 1 - 4 are cardinal accelerations.

0.716. This is an indication of the idea that an added incentive might get in the way of exploiting an easy to find, simple strategy.

In the challenging task, which requires an action-oriented strategy empowerment seems to perform similar to the baseline, while it clearly outperforms the baseline for the hard task, both in terms of average distance, and in terms of obstacles hit. The difference in hit obstacles is particularly large, indicating that TE helps with the higher reactivity required within an episode to navigate around the obstacles. Social influence seems to struggle with both tasks, again likely due to an interference between the added reward and the best exploitation strategy.

In contrast, the better performance of empowerment in the hard challenge, compared to the baseline, must be due to empowerment helping to discover a better overall strategy - as the baseline implementation would be fully capable of producing a strategy identical to the one performance by the TE framework, had it discovered it.

To illuminate this difference, we can take a look at the action distribution for the speaker and listener agents over several episodes, using the agents after 10k training episodes. Fig. 3 shows how often the five available actions were used. Action 0 for the listener is the waiting action - and we see that this one is not used by the baseline. We speculate that it might be difficult for the listener to learn when to use this action, as it is detrimental in most cases. The bias towards reactivity induced by TE might help to keep this rare action as an option - following a symbol by the speaker that might become a "stop" signal.

We can also take three trained listener agents and compare what they will do when we provide them with a fixed speaker signal over several time steps, to figure out what those symbols directed them to do. Fig. 4 shows the trajectories resulting from this, with each color denoting a different forces symbol by the speaker. The baseline has a relatively good separation into cardinal actions, but transfer empowerment

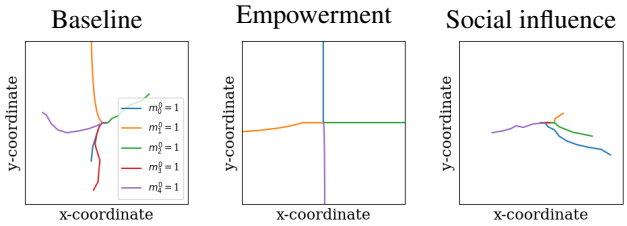

Figure 4: The listener's positions plotted for 10 time steps, given a speaker's message $\boldsymbol{m}^0$, a one-hot vector. The subscript denotes the component in $\boldsymbol{m}^0$ that is equal to a 1.

leads to nearly perfect control by the speaker over the actions of the listener. Note that one signal results in the wait action, leading to no visible trajectory for empowerment.

Conclusions and Future Work

Overall, adding transfer empowerment to MARL seems to improve the overall performance level of cooperative agents - particularly for harder tasks that rely on an action-oriented communication strategy. This seems to indicate that TE helps the learning process to find better solutions to converge on - which remain undiscovered by the baseline MARL with similar training time - while also not getting in the way of exploitation to much. An immediate open question, a direction for future work, is of course the question of generalizability of this approach to different scenarios. Other exciting research directions are scenarios with partners unseen at training time, moving in the direction of one-shot adaptation to partners, and scenarios with competitive, or cooperative-competitive mixed scenarios. Using TE to bias systems towards control, or information hiding to find optimal solutions.

We also showed how an efficient computation of empowerment could be combined with RL for the MARL framework, opening the door for more complex scenarios such as humans interacting with robots. In general, the results in this study are promising for the overall agenda to develop a framework of social intrinsic motivations based on empowerment (or similar measures) to bias an agent towards general social concepts, such as reliable reactivity, or lead-following. The fact that it is based on similar, single-agent intrinsic motivations is also interesting, as it might offer insights into how to transition from single to social agent behavior with only gradual adaptation.

References

Baldassarre, G. and Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Springer.

Barton, S. L., Waytowich, N. R., Zaroukian, E., and Asher, D. E. (2018). Measuring collaborative emergent behav-

ior in multi-agent reinforcement learning. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications*, pages 422–427. Springer.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Chentanez, N., Barto, A. G., and Singh, S. P. (2005). Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pages 1281–1288.

de Abril, I. M. and Kanai, R. (2018). A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv preprint arXiv:1806.06505*.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2018). Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Gregor, K., Rezende, D. J., and Wierstra, D. (2016). Variational intrinsic control. *arXiv preprint arXiv:1611.07507*.

Guckelsberger, C., Salge, C., and Colton, S. (2016). Intrinsically motivated general companion npcs via coupled empowerment maximisation. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE.

Guckelsberger, C., Salge, C., and Togelius, J. (2018). New and surprising ways to be mean. adversarial npcs with coupled empowerment minimisation. *arXiv preprint arXiv:1806.01387*.

Hu, H., Lerer, A., Peysakhovich, A., and Foerster, J. (2020). " other-play" for zero-shot coordination. *arXiv preprint arXiv:2003.02979*.

Jaderberg, M., Czarnecki, W., Dunning, I., Marris, L., Lever, G., Castaneda, A., et al. (2018). Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. arxiv. *arXiv preprint arXiv:1807.01281*.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pages 3040–3049. PMLR.

Karl, M., Soelch, M., Bayer, J., and Van der Smagt, P. (2016). Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv preprint arXiv:1605.06432*.

Karl, M., Soelch, M., Becker-Ehmck, P., Benbouzid, D., van der Smagt, P., and Bayer, J. (2017). Unsupervised real-time control through variational empowerment. *arXiv preprint arXiv:1710.05101*.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer.

Lerer, A. and Peysakhovich, A. (2018). Learning social conventions in markov games. *arXiv preprint arXiv:1806.10071*.

Leu, A., Ristić-Durrant, D., Slavnić, S., Glackin, C., Salge, C., Polani, D., Badii, A., Khan, A., and Raval, R. (2013). Corbys cognitive control architecture for robotic follower. In *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*, pages 394–399. IEEE.

Li, S., Wu, Y., Cui, X., Dong, H., Fang, F., and Russell, S. (2019). Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4213–4220.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pages 6379–6390.

Lupu, A., Cui, B., Hu, H., and Foerster, J. (2021). Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pages 7204–7213. PMLR.

Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. (2019). Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems*, pages 7613–7624.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.

Nair, R., Tambe, M., Yokoo, M., Pynadath, D., and Marsella, S. (2003). Taming decentralized pomdps: Towards efficient policy computation for multiagent settings. In *IJCAI*, volume 3, pages 705–711. Citeseer.

Osborne, M. J. et al. (2004). *An introduction to game theory*, volume 3. Oxford university press New York.

Oudeyer, P.-Y. and Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17.

Salge, C. and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 4:25.

Strouse, D., Kleiman-Weiner, M., Tenenbaum, J., Botvinick, M., and Schwab, D. J. (2018). Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, pages 10249–10259.

Tesauro, G. (1994). Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219.

Tucker, M., Zhou, Y., and Shah, J. (2020). Adversarially guided self-play for adopting social conventions. *arXiv preprint arXiv:2001.05994*.

van der Heiden, T., Weiss, C., Shankar, N. N., Gavves, E., and van Hoof, H. (2020). Social navigation with human empowerment driven reinforcement learning. *arXiv preprint arXiv:2003.08158*.

Wang, T., Wang, J., Wu, Y., and Zhang, C. (2019). Influence-based multi-agent exploration. *arXiv preprint arXiv:1910.05512*.

Appendix

Algorithm 1 explains how we train with empowerment. We omit super- and subscripts denoting time, agent and batch indices whenever clear from the context.

Algorithm 1 Training joint policy $\boldsymbol{\pi}_\theta$ with empowerment

Require: Initialisation of networks $\boldsymbol{\pi}_\chi$, $\boldsymbol{Q}_\psi$, $\boldsymbol{\omega}_\theta$, q_θ and p_ν, and target networks $\boldsymbol{\pi}_\phi$, $\boldsymbol{Q}_\zeta$.
for each episode **do**
  for each time step **do**
    $\boldsymbol{o} = O(s)$, $\boldsymbol{a} \sim \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o})$, $s' \sim f(s, \boldsymbol{a})$
    $\tau = \tau \cup \{(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}', r, \hat{\mathcal{I}}, y)\}$
  end for
  $\mathcal{D} = \mathcal{D} \cup \tau$
  for each agent i **do**
    sample minibatch with S tuples from $\mathcal{D}$
    $\hat{\mathcal{I}}_{\chi^i,\theta^i}(\boldsymbol{o}) = \texttt{computeLowerBound}(\boldsymbol{o}, \boldsymbol{\pi}_\chi, \boldsymbol{\omega}_\theta, q_\theta, p_\nu)$ ▷ Equation 3
    $y = r + \hat{\mathcal{I}}_{\chi^i,\theta^i}(\boldsymbol{o}) + \gamma Q^i_\zeta(\boldsymbol{o}, \boldsymbol{a})$
    $\mathcal{L}(\psi^i) = \frac{1}{S}\sum_j (y - Q^i_\psi(\boldsymbol{o}_j, \boldsymbol{a}_j))^2|_{a^i_j \sim \pi^i_\phi}$
    $\mathcal{L}(\chi^i) = -\frac{1}{S}\sum_j Q^i_\psi(\boldsymbol{o}_j, \boldsymbol{a}_j)|_{a^i_j \sim \pi^i_\chi}$
    $\texttt{updateCritic}(\mathcal{L}(\psi^i), \psi^i)$ ▷ See (Lowe et al., 2017)
    $\texttt{updateActor}(\mathcal{L}(\chi^i), \chi^i)$
    $\texttt{gradientAscent}(\hat{\mathcal{I}}_{\chi^i,\theta^i}, \theta^i, \chi^i)$
    $\texttt{maxLogLikelihood}(\nu, \boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}')$ ▷ See (Karl et al., 2016)
    $\texttt{updateTargets}(\phi^i, \zeta^i, \psi^i, \chi^i)$
  end for
end for

$\boldsymbol{\pi}$	Joint policy with N components $[\pi^1, \ldots, \pi^N]$.
$O(s)$	Deterministic observation function $\boldsymbol{o} = (o^1, \ldots, o^N) = O(s)$.
$\boldsymbol{a}, \boldsymbol{o}$	Joint action and observation $(a^1, \ldots, a^N)$, $(o^1, \ldots, o^N)$.
$\boldsymbol{a}^{-i}, \boldsymbol{o}^{-i}$	Joint action and observation excluding those of agent i.
$J^i(\boldsymbol{\pi})$	Expected return of agent i induced by joint policy $\boldsymbol{\pi}$.
$Q^i(\boldsymbol{o}, \boldsymbol{a})$	Centralised critic of local policy π^i.
$\mathcal{E}^{-i \to i}$	Transfer empowerment from all agents' actions, excluding agent i, towards agent i's action.
$\hat{\mathcal{E}}$	Lower bound on empowerment by employing variational approximation.
$\hat{\mathcal{I}}_\theta$	Lower bound on mutual information computed by θ-parameterized neural networks.

Table 2: Notation