Collaboration Promotes Group Resilience in Multi-Agent RL

Ilai Shraga, Guy Azran, Matthias Gerstgrasser, Ofir Abu, Jeffrey S. Rosenschein, Sarah Keren

Keywords: Multi-Agent Reinforcement Learning, Group Resilience, Collaboration, Deep Reinforcement Learning.

Summary

To effectively operate in various dynamic scenarios, RL agents must be resilient to unexpected changes in their environment. Previous work on this form of resilience has focused on single-agent settings. In this work, we introduce and formalize a multi-agent variant of resilience, which we term *group resilience*. We further hypothesize that collaboration with other agents is key to achieving group resilience; collaborating agents adapt better to environmental perturbations in multi-agent reinforcement learning (MARL) settings. We test our hypothesis empirically by evaluating different collaboration protocols and examining their effect on group resilience. Our experiments show that all the examined collaborative approaches achieve higher group resilience than their non-collaborative counterparts.

Contribution(s)

1. We introduce a novel definition of group resilience and formalize this notion, which corresponds to the group's ability to adapt to unexpected changes.
   Context: Prior work primarily focused on resilience in single-agent settings or adversarial MARL scenarios without a unified resilience measure.
2. We demonstrate that collaboration promotes group resilience, providing empirical evidence across multiple MARL benchmarks.
   Context: Since prior work has mostly explored single-agent or adversarial settings, collaboration and its effects have not been considered.

Collaboration Promotes Group Resilience in Multi-Agent RL

Ilai Shraga[1], Guy Azran[1], Matthias Gerstgrasser[2], Ofir Abu[3], Jeffrey S. Rosenschein[3], Sarah Keren[1]
ilaishraga@campus.technion.ac.il, guy.azran@campus.technion.ac.il, matthias@seas.harvard.edu, ofirabu@cs.huji.ac.il, jeff@cs.huji.ac.il, sarahk@technion.ac.il

[1]**Taub Faculty of Computer Science, Technion - Israel Institute of Technology.**
[2]**School of Engineering and Applied Sciences, Harvard University.**
[3]**School of Computer Science and Engineering, The Hebrew University of Jerusalem.**

Abstract

To effectively operate in various dynamic scenarios, RL agents must be resilient to unexpected changes in their environment. Previous work on this form of resilience has focused on single-agent settings. In this work, we introduce and formalize a multi-agent variant of resilience, which we term *group resilience*. We further hypothesize that collaboration with other agents is key to achieving group resilience; collaborating agents adapt better to environmental perturbations in multi-agent reinforcement learning (MARL) settings. We test our hypothesis empirically by evaluating different collaboration protocols and examining their effect on group resilience. Our experiments show that all the examined collaborative approaches achieve higher group resilience than their non-collaborative counterparts.

1 Introduction

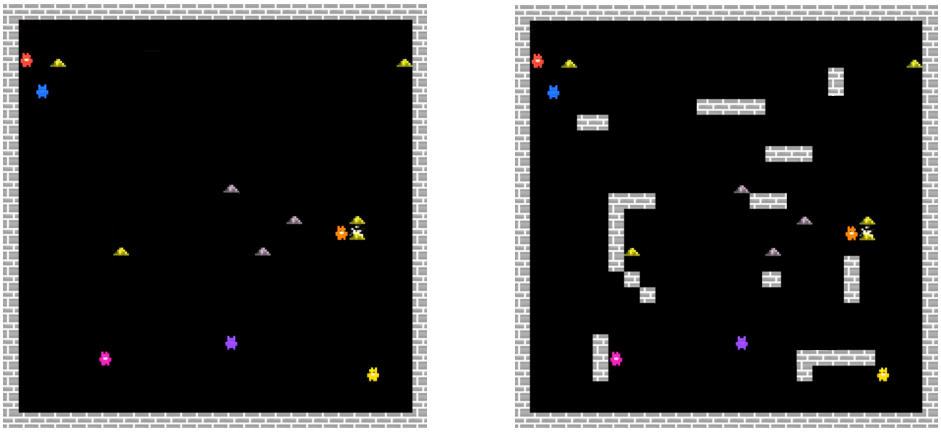
Reinforcement Learning (RL) agents are typically required to operate in dynamic environments and must develop an ability to quickly adapt to change. Promoting this ability is hard, even in single-agent settings (Padakandla, 2022). When the RL agent operates alone, it needs to adapt its behavior to the changing and possibly partially observable environment. For a group, this is even more challenging. In addition to the dynamic nature of the environment, agents need to deal with high variance caused by the changing behavior of other agents. Nevertheless, the ability of autonomous agents, individually or as a group, to adapt to environmental changes is highly desirable in real-world settings where dynamic environments are the norm. Therefore, if a group is to reliably pursue its objective, it should be able to handle unexpected environmental perturbations.

Many recent Multi-Agent RL (MARL) studies show that collaboration boosts agents' performance (Christianos et al., 2020; Foerster et al., 2016b; Honhaga & Szabo, 2024; Jaques et al., 2019b; Lowe et al., 2020; Qian et al., 2019; Xu et al., 2012). We aim to understand how collaboration promotes the *group resilience*, i.e., how collaboration affects the ability of a group to rapidly regain a significant portion of their prior performance after unexpected perturbations occur in their environment. We therefore investigate whether agents that collaborate in different ways indeed reclaim a larger share of their former performance after a disturbance occurs.

Contrary to investigations of *transfer learning* (Liang & Li, 2020; Zhu et al., 2023) or *curriculum learning* (Portelas et al., 2020), we do not assume a stationary target domain in which agents will be deployed, nor do we have a dedicated training phase to prepare agents for various environments.

Moreover, unlike typical generalisation scenarios, which evaluate final performance on previously unseen tasks, our resilience metric explicitly captures how group performance recovers following random but bounded perturbations. We focus on multi-agent reinforcement learning (MARL) settings, for which previous work demonstrates how collaboration enables groups to operate efficiently in complex but stationary environments (Christianos et al., 2020; Foerster et al., 2016b; Jaques et al., 2019b). We offer empirical evidence that collaboration promotes resilience in non-stationary environments, facilitating it via existing (Jaques et al., 2019b) and custom communication protocols.

The literature offers a range of definitions for resilience in both single and multi-agent settings (Pattanaik et al., 2017; Phan et al., 2020; Vinitsky et al., 2020; Zhang et al., 2017), primarily focusing on resilience in the face of adversarial behavior (Saulnier et al., 2017; Phan et al., 2020), and on algorithms for training resilient agents that define resilience in a model-specific way. However, these frameworks do not define resilience in a unified, measurable way. We concentrate on non-adversarial settings, emphasizing the agents' ability to enhance resilience through collaboration. We define and measure group resilience based on agents' performance under unexpected and random environmental perturbations of bounded magnitude. Our goal is to provide a unified, general measure of resilience that is based on a user-specified distance metric between environments and that is adaptable to various settings. The perturbations we model represent unexpected changes in the real world, reflecting practical scenarios that agents might encounter.



(a) (b)

Figure 1: An illustration of the coop-mining domain. There are 5 miner agents (colorful creatures). Two types of ores (resources) appear randomly on the map: iron (grey mounds) and gold (yellow mounds). 1a shows a clean mine with walls only on the boundaries. 1b shows a perturbation of the environment; perturbations are newly introduced non-traversable walls (solid lines).

Example 1 *Figure 1a depicts a multi-agent variation of the coop-mining domain (Leibo et al., 2021). All miners are employed by the same mining company, which aims to maximize the group's total revenue. Miners may find two types of ores, iron and gold, anywhere in the mine. A single miner can extract one unit of iron alone, which sells for $100. Mining gold requires two miners to strike the ore multiple times in unison, but it sells for $800. As the miners excavate, unstable terrain or mild earthquakes may cause cave-ins that create unexpected blockages in random locations within the mine (see Figure 1b). At this point, the miners may benefit from sharing information about these blockages to enhance the group's resilience, i.e., their ability to adapt to the perturbations. We aim to show that agents trained to communicate prior to perturbations are more resilient to unexpected disruptions and can more quickly adapt to the changing mine layout.*

Our contributions are threefold. First, we suggest a **new unified measure** called *group resilience*, which corresponds to the group's ability to adapt to unexpected changes. This formulation covers a wide range of real-world multi-agent problems. Second, we design and implement a MARL frame-

work for testing multi-agent resilience in the face of environment perturbations. Finally, we provide empirical evidence supporting our hypothesis that collaboration promotes resilience in MARL.

2 Background

Reinforcement learning **(RL)** is a learning paradigm in which agents learn by observing the world, acting within it, and receiving rewards (positive or negative) for achieving certain states or state transitions. RL problems commonly model the world as a *Markov decision process* (MDP) (Bellman, 1957) $M = \langle S, A, R, P, \gamma \rangle$ where S is a set of possible states, A is a set of agent actions, $P : S \times A \times S \to [0, 1]$ is the state transition function, $R : S \times A \times S \to \mathbb{R}$ is the reward function, and γ is the reward discount factor. The objective is to find a policy π^* such that $\pi^* \in \arg\max_\pi \mathbb{E}[J(\pi)]$, where: $J(\pi) = \mathbb{E}_{s_t,\, s_{t+1} \sim P;\, a_t \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t R\big(s_t, a_t, s_{t+1}\big)\right]$ is the expected return of policy π.

Multi-Agent Reinforcement Learning **(MARL)** extends RL to settings with multiple agents. We model the environment as a *Markov Game* (MG) (Littman, 1994), in which each agent selects an individual action, receives a reward, and the transition to the next state depends on the joint action of all agents. We denote by $\mathcal{U}_\pi(M)$ the cumulative discounted return obtained by a joint policy π in environment M. Since we focus on settings in which the environment might change, we define the environment-specific optimum as

$$\mathcal{U}^{\max}(M) = \sup_\pi \mathcal{U}_\pi(M)$$

We follow Jordan & Trott (2020) in using the *normalised utility* $\hat{\mathcal{U}}(M) = \mathcal{U}(M)/\mathcal{U}^{\max}(M) \in [0, 1]$, so scores remain comparable even when perturbations affect rewards or dynamics.

3 Measuring Group Resilience

We aim to promote the ability of a group to adapt to random perturbations in their environment. We refer to this ability as *group resilience* and formally define it below. Our definition explicitly quantifies the group's adaptive capacity, i.e., the extent of performance recovery following unexpected perturbations. We will then explore our hypothesis on promoting group resilience by facilitating collaboration.

To measure group resilience, we take inspiration from the field of multi-agent robotics. Specifically, the work of Saulnier et al. (2017) aims to produce a control policy that allows a team of mobile robots to achieve desired performance in the presence of faults and attacks on individual group members. Accordingly, a group of robots achieves *resilient consensus* if the cooperative robots' performance remains in some desired range, even in the presence of a bounded number of non-cooperative robots.

Similarly, we define *group resilience* as the ability of agents to maintain at least a fixed fraction of their original performance when the environment undergoes an unexpected perturbation bounded in magnitude. Unlike resilience to adversarial agents, our definition explicitly addresses changes in the environment itself, implicitly accounting for the corresponding changes in agents' observations and experiences. Formally, our definition relies on (i) a distance measure $\delta(M, M')$ quantifying the magnitude of change between an original MDP M and a perturbed MDP M', and (ii) the utility measure that quantifies group performance. Given these measures, we require that any perturbed environment within a bounded distance K from the original environment results in decreased performance by at most a constant factor C_K[1].

We note that a variety of subtly different formal definitions can satisfy this intuitive requirement. We provide here only those definitions directly relevant to our experiments. Specifically, the following definitions assume that a designer aims to guarantee resilience over a subset of perturbed environments within a specified distance from a reference MDP M (see Figure 2).

[1] Note the similarity to the classical ϵ-δ continuity definition.

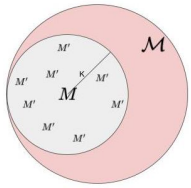


Figure 2: Resilience considers performance differences for perturbed environments M' within some distance from a reference environment M.

An important issue to consider is that once an environment is perturbed, the optimal utility that can be achieved might change. Thus, in principle, it might be beneficial to consider the performance level of a team with regard to the normalized utility $\hat{\mathcal{U}}$ in each environment as follows.

Definition 1 (Relative-to-optimum C_K-resilience) *Given a class of MDPs $\mathcal{M}$, a reference MDP $M \in \mathcal{M}$, and a bound $K \in \mathbb{R}$, we say a group of agents is C_K-*resilient *over $\mathcal{M}$ relative to optimum if*

$$\forall M' \in \mathcal{M} : \delta(M, M') \leq K \Longrightarrow \hat{\mathcal{U}}(M') \geq C_K \cdot \hat{\mathcal{U}}(M)$$

Setting $C_K \in [0, 1]$ ensures performance degradation remains bounded.

In some settings, it may be impossible to know or approximate the optimal utility. Nevertheless, if the goal is to compare the resilience of multiple groups, it is sufficient to compare their utilities in the original and perturbed environments without normalization.

Definition 2 (Relative-to-origin C_K-resilience) *Given a class of MDPs $\mathcal{M}$, a reference MDP $M \in \mathcal{M}$, and a bound $K \in \mathbb{R}$, we say a group of agents is C_K-*resilient *over $\mathcal{M}$ relative to origin M if*

$$\forall M' \in \mathcal{M} : \delta(M, M') \leq K \Longrightarrow \mathcal{U}(M') \geq C_K \cdot \mathcal{U}(M)$$

The above definition is especially relevant to our aim of comparing the performance of groups that undergo the same perturbation but adopt different collaboration approaches.

While resilience over all M' within distance K provides a strong guarantee, it can be impractical. Thus, given a probability distribution Ψ over $\mathcal{M}$, we define resilience-in-expectation as follows:

Definition 3 (Relative-to-origin C_K-resilience in expectation) *Given an MDP M, a distribution Ψ over a class of MDPs, and a bound $K \in \mathbb{R}$, a group of agents is C_K-*resilient in expectation *over Ψ relative to origin M if*

$$\mathbb{E}_{M' \sim \Psi | \delta(M, M') \leq K} \left[\mathcal{U}(M') \right] \geq C_K \cdot \mathcal{U}(M)$$

This definition guarantees expected group performance remains above a threshold when averaging over perturbed environments within distance K from the reference environment.

Polynomially many samples from Ψ suffice to approximate this expectation with high probability. Formally, if we assume that $\mathcal{U}(M')$ is a random variable with finite variance σ^2 drawn i.i.d. from Ψ under the condition $\delta(M, M') \leq K$, then by Chebyshev's inequality, the expected normalized utility can be approximated within distance ϵ of the true expectation with probability at least δ using $\frac{\sigma^2}{\epsilon^2(1-\delta)}$ samples.

Note that Definitions 2 and 3 compare agents' performance under perturbation directly against their performance in the unperturbed environment, without explicitly considering baseline performance. Hence, a non-optimal baseline policy (e.g., no-op actions) could misleadingly indicate high resilience. Depending on analysis objectives, our measures may thus require complementary baseline normalization.

3.1 Perturbations

In this work, we are interested in settings in which we have an initial environment and a set of *perturbations* that can occur. A perturbation $\phi : \mathcal{M} \mapsto \mathcal{M}$ is a function transforming a source MDP into a modified MDP. An *atomic perturbation* is a perturbation that changes only one of the basic elements of the original MDP. Given an MDP $M = \langle S, A, R, P, \gamma \rangle$ and perturbation ϕ, we denote the resulting MDP after applying ϕ by $M^\phi = \langle S^\phi, A^\phi, R^\phi, P^\phi, \gamma^\phi \rangle$.

Among the variety of perturbations that may occur, we focus here on three types of atomic perturbations. *Transition function perturbations* modify the distribution over the next states for a single state-action pair. *Reward function perturbations* modify the reward of a single state-action pair. *Initial state perturbations* change the initial state of the MDP.

Definition 4 (Transition Function Perturbation) *A perturbation ϕ is a* transition function perturbation *if for every MDP $M = \langle S, A, R, P, \gamma \rangle$, M^ϕ is identical to M except that for a single action-state pair $s \in S$ and $a \in A$, $\mathbb{P}_s^a[S] \neq \mathbb{P}_s^{a,\phi}[S]$.*

Definition 5 (Reward Function Perturbation) *A perturbation ϕ is a* reward function perturbation *if for every MDP $M = \langle S, A, R, P, \gamma \rangle$, M^ϕ is identical to M except that for a single action-state pair $s \in S$ and $a \in A$, $r_s^a \neq r_s^{a,\phi}$.*

Definition 6 (Initial State Perturbation) *A perturbation ϕ is an* initial state perturbation *if for every MDP $M = \langle S, s_0, A, R, P, \gamma \rangle$, M^ϕ is identical to M except that $s_0 \neq s_0^\phi$.*

Example 1 (continued) *In our Coop-Mining domain, a path blockage can be modeled as an initial state perturbation or as a transition function perturbation that stops the agent from transitioning to the adjacent cell. This is represented by changing P to express a probability distribution $\mathbb{P}_s^a[S]$ that is set to be $\mathbb{P}_s^a(s') = 0$ when (s, a, s') represents crossing the blocked area, and $\mathbb{P}_s^a(s) = 1$, which represents staying in the same place. A change in an ore's (gold) location can be represented by two atomic perturbations: one that replaces the reward for mining at the original location with a negative reward, and one that adds a positive reward for mining at the new location.*

A straightforward way to measure the distance δ between two MDPs is to count the minimal number of atomic perturbations that transition the original MDP into the transformed one. While this metric has some limitations, it is sufficient for our experimental settings. Future work should consider more sophisticated distance measures between MDPs. For instance, one can define the overall distance based on a state-level distance $d(s, s')$, instantiated using established measures from the literature, such as Kantorovich (Wasserstein) distances Dobrushin (1970); Song et al. (2016), bisimulation-based distances Ferns et al. (2004), or transfer-oriented distance metrics Bou Ammar et al. (2014).

4 Facilitating Collaboration via Communication

Equipped with a measure for group resilience, we now focus on maximizing the resilience of a group of RL agents by facilitating collaboration. Recent MARL work suggests various approaches for promoting collaboration (Jaques et al., 2019b; Mahajan et al., 2019; Rashid et al., 2018). In this work, we focus on communication (Christianos et al., 2020; Foerster et al., 2016b).

To support collaboration, communication protocols produce messages that encode information valuable to other agents' learning. Prior literature in MARL explicitly defines effective collaboration as relying upon the exchange of meaningful information, allowing agents to coordinate their actions and adapt jointly to dynamic environmental changes (Foerster et al., 2016a; Jaques et al., 2019a). We examine different communication protocols based on broadcasting observations that are least aligned with their previous experiences. Misalignment corresponds to the agents' familiarity with the environment, which may decrease due to perturbations. By communicating misaligned transitions, agents increase the other agents' familiarity with the environment.

We present two definitions of misalignment. The first is taken from (Gerstgrasser et al., 2023), and measures misalignment using the Temporal Difference (TD) error of a given observation. Formally, let $e_t = \langle s_t, a_t, s_{t+1}, r_t \rangle$ be the experience at time t, that is, transitioning to state s_{t+1} after taking action a_t in s_t and receiving reward r_t. Let π_p represent the policy of agent p, and let Q^{π_p} represent the Q-function of policy π_p, i.e., the expected value of taking action a in state s and following π thereafter. Given an experience, the *TD-error* is:

$$\mathrm{TD}(e_t) = \left| r_t + \gamma \max_a Q^{\pi_p}(s_{t+1}, a) - Q^{\pi_p}(s_t,\, a_t) \right| \tag{1}$$

This definition is inspired by Prioritized Experience Replay (Schaul et al., 2016, PER), according to which a deep Q-Network (DQN) agent (Mnih et al., 2015) maintains a buffer of past transitions and prioritizes them in a way that expedites training.

A second measure of *misalignment* of a transition considers the difference between the observed reward r_t and the expected reward $\hat{r_t}$. The misalignment for agent p at s_t after taking action a_t, denoted by $J^p_{s_t,a_t}$, is defined as

$$J^p_{s_t,a_t} \;=\; \frac{\left| r_t \;-\; \hat{r_t} \right|}{r_t} \tag{2}$$

where $\hat{r_t} \approx Q^{\pi_p}(s_t, a_t) - Q^{\pi_p}(s_{t+1}, \pi_p(s_{t+1}))$.

Using these two measures, we created different communication protocols in which misalignment is used to decide which messages to broadcast to other agents, described below:

1. **No Communication** – Agents do not share information (used as a baseline).
2. **Mandatory Broadcast** – Each agent p broadcasts at state s_t its most misaligned experiences, i.e., transitions with the highest $J^p_{s_t,a_t}$. A message consists of the misaligned transitions τ. Messages are received by all other agents and inserted into their replay buffers. The number of transitions broadcast at each time step is bounded by the channel bandwidth parameter m_l.
3. **Emergent Communication** – Each agent p broadcasts a discrete communication symbol m^p_t among a given set of symbols, at each state s_t. Individual messages of all agents are concatenated into a single vector $m_t = [m^1_t \dots m^N_t]$, which is included as an additional observation signal that all agents receive at the next time step $(t+1)$. We distinguish between two subcases: Self-Centric and Global-Centric. In **Emergent Self-Centric Communication**, each agent p uses counterfactual reasoning and chooses a symbol m^p_t that would have minimized *its own* misalignment at time step $t-1$. Formally, the loss function for π_m is:

$$L_t \;=\; \left| \arg\min_m J^{p_m}_{s_{t-1},a_{t-1}} \;-\; J^p_{s_{t-1},a_{t-1}} \right|$$

   where $J^{p_m}_{s_{t-1},a_{t-1}}$ is the misalignment at $(t-1)$ had it received message m. In **Emergent Global-Centric Communication**, agents observe the misaligned observations of all other agents at each time step. Each agent p is rewarded for choosing a symbol m^p_t that would have minimized the *total* misalignment of the group at $(t-1)$. Agents maintain a model that predicts the global (average) misalignment of the other agents given an observation and messages. Formally:

$$L_t \;=\; \left| \arg\min_m \hat{J}^{P_m}_{s_{t-1},a_{t-1}} \;-\; J^P_{s_{t-1},a_{t-1}} \right|$$

   where $\hat{J}^{P_m}_{s_{t-1},a_{t-1}}$ is the counterfactual predicted average misalignment of the other agents, having received m, and $J^P_{s_{t-1},a_{t-1}}$ is the actual average misalignment of the other agents.
4. **suPER – Selectively Sharing Experiences** (Gerstgrasser et al., 2023): Each agent p broadcasts at state s_t its experiences with the highest TD-errors (Equation 1). These transitions are inserted into the replay buffers of receiving agents. The number of broadcast transitions per time step is bounded by m_l. suPER leverages PER insights so that not all experiences are equally relevant for learning, and thus it supports decentralized training with minimal communication overhead, compatible with standard DQN (Mnih et al., 2015) variants.

5 Empirical Evaluation

Our empirical evaluation aims to assess the effect collaboration has on group resilience. Specifically, we measure and compare the utility of groups of agents in randomly perturbed environments (various types of atomic perturbations), where each group implements a different communication protocol and learning approach[2]. We conducted two sets of experiments. The first uses our custom communication protocols in environments requiring relatively simple forms of collaboration. The second uses SOTA communication in an abstraction of Experiment 1. Experiments ran on a `x86_64` CPU running Ubuntu 20.04.6.

Experiment 1: Simple Collaboration We experiment with three multi-agent RL environments. **Cleanup** (Vinitsky et al., 2019): Seven agents must balance harvesting apples (individual rewards) and cleaning a river (enables regrowth but prevents harvesting). Agents can fine each other, and each observes a raw image of its surroundings. **Harvest** (Vinitsky et al., 2019): Similar to Cleanup, but apple regrowth depends on proximity rather than a river, requiring coordinated harvesting to avoid depletion. **Multi-Taxi** (Azran et al., 2024): Taxis transport passengers in a configurable grid world with perturbations. Observations are symbolic state vectors. Rewards include high positive for drop-offs, small negative for time steps, and large negative for collisions. Grid sizes range from 5×5 to 8×8 with 2–3 taxis and passengers.

We trained individual neural networks for every RL agent using the distributed Asynchronous Advantage Actor-Critic (A3C) algorithm (Mnih et al., 2016). For the Multi-Taxi domain, agents were implemented using a Deep Q-Network (Mnih et al., 2015). We evaluate a spectrum of communication protocols within our framework. The **No Communication** protocol involves agents operating independently without exchanging information. In the **Social Influence** protocol, as suggested by (Jaques et al., 2019b), agents broadcast messages aimed at maximizing their impact on the immediate behaviors of other agents. The **Mandatory Communication** protocol requires agents to share their top m_l most misaligned transitions, as detailed in Section 4. In the **Emergent Self-Centric Communication** protocol, agents broadcast a discrete symbol at each step that would have minimized their previous step's misalignment. In the **Emergent Global-Centric Communication** protocol, agents monitor the current misalignment levels of their peers and broadcast symbols that aim to minimize the group's overall misalignment.

In the Cleanup and Multi-taxi domains, we used two types of perturbations. The first is a transition function perturbation, randomly adding non-traversable obstacles (e.g., walls) to the map. The second is an initial state perturbation, randomly changing the initial configuration (e.g., changing the river location in Cleanup, or initial taxi/passenger locations in Multi-taxi). In Harvest and Multi-taxi, we used a reward function perturbation, randomly reallocating rewards/resources (e.g., eliminating passengers or apples). We measure the perturbation's magnitude using the state-distance approach of (Song et al., 2016) described in Section 3. We experiment with bounds $K \in \{50, 150, 200\}$. For each initial environment M and bound K, we uniformly sample from possible perturbed M' such that $\delta(M, M') \leq K$, applying random atomic perturbations until the desired magnitude is reached.

To measure the effect of perturbations on group performance, we calculate the average utility throughout training before and after perturbation. We repeat each experiment 8 times with different random seeds. The process described above is used to generate perturbations for each seed.

To measure resilience, we use Definition 3 with a uniform distribution Ψ. We let $C_K = \frac{\text{avg}\left(\mathcal{U}(M')\right)}{\mathcal{U}(M)}$, where M' is drawn from Ψ within distance K.

Experiment 2: Cooperative Mining Here, we examine the **Coop-Mining** domain, designed to test various social interactions. There are five agents, and, as described in Example 1, the domain incentivizes coordinating to gather resources, balancing reliable low-reward iron versus high-yield gold that requires cooperation.

[2]Our code base and full results are available at https://github.com/CLAIR-LAB-TECHNION/CollaborationForResilience

We trained agents similarly to Experiment 1, but using Deep Q-Networks (DQNs) (Mnih et al., 2015). We employ the **suPER** advanced communication protocol with the hyperparameters in the original work, whereby agents transmit their highest TD-error experiences into others' replay buffers, leveraging PER ideas for decentralized training with minimal overhead.

We apply transition function perturbations of varying magnitudes, measuring distance similarly to Experiment 1. Resilience and utility are calculated likewise.

5.1 Results

Table 1 shows results for Experiment 1, reporting mean C_K-resilience with perturbations of varying magnitudes, alongside $\mathcal{U}$ in the non-perturbed environment (std. dev. in parentheses). Figure 5 compares group resilience in the Cleanup domain across different communication protocols, grouped by perturbation intensity, while Figure 3 shows the utility throughout training for $K = 200$. Figures 6 and 4 similarly present results for Experiment 2 (Coop-Mining).

Table 1: Average (and standard deviation) C_K-resilience for Cleanup, Harvest, and Multi-Taxi.

	Cleanup				**Harvest**				**Multi-Taxi**			
	$C_{K=50}$	$C_{K=150}$	$C_{K=200}$	$\mathcal{U}$	$C_{K=50}$	$C_{K=150}$	$C_{K=200}$	$\mathcal{U}$	$C_{K=50}$	$C_{K=150}$	$C_{K=200}$	$\mathcal{U}$
No communication	0.62 (0.25)	0.21 (0.14)	0.06 (0.05)	3.56 (1.32)	0.64 (0.15)	0.38 (0.11)	0.25 (0.12)	128.18 (94.84)	0.67 (0.14)	0.35 (0.17)	0.27 (0.17)	141.35 (40.59)
Social Influence	**0.78** (0.17)	0.40 (0.17)	0.25 (0.13)	7.49 (2.59)	0.77 (0.13)	0.50 (0.13)	0.36 (0.11)	132.68 (100.51)	0.69 (0.10)	**0.51** (0.17)	**0.38** (0.13)	149.45 (45.18)
Mandatory Communication	0.69 (0.19)	0.40 (0.23)	0.21 (0.17)	4.47 (1.59)	0.72 (0.14)	0.48 (0.11)	**0.38** (0.13)	169.02 (105.65)	0.70 (0.11)	0.48 (0.13)	0.35 (0.13)	**221.25** (51.63)
Emergent Global-Centric	0.77 (0.16)	**0.43** (0.14)	**0.27** (0.08)	**11.41** (2.05)	**0.81** (0.11)	0.48 (0.13)	0.36 (0.12)	**186.50** (101.67)	**0.74** (0.10)	0.45 (0.17)	0.34 (0.14)	197.75 (67.33)
Emergent Self-Centric	0.64 (0.21)	0.33 (0.17)	0.15 (0.14)	4.81 (0.87)	0.74 (0.14)	**0.52** (0.13)	0.33 (0.12)	131.68 (94.76)	0.67 (0.11)	0.49 (0.14)	**0.38** (0.15)	140.15 (40.42)

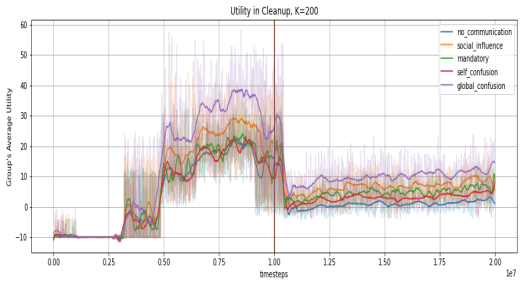


Figure 3: Average utility for $K = 200$ in the Cleanup environment domain before and after a perturbation occurs.

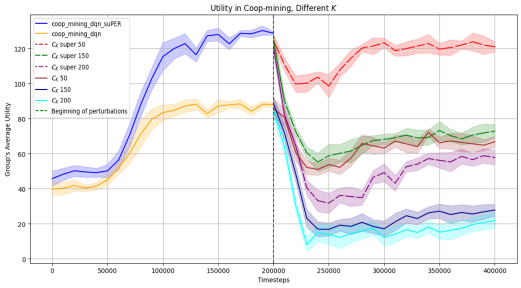


Figure 4: Average utility for different perturbation magnitudes in the Coop-mining environment domain before and after a perturbation occurs.

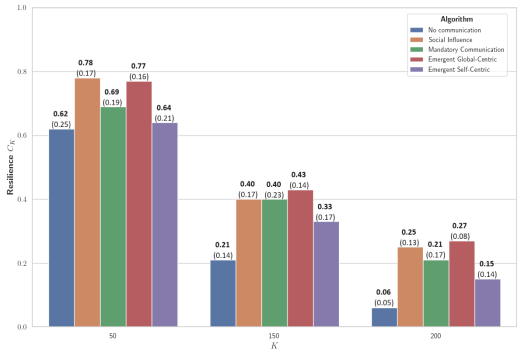


Figure 5: Cleanup environment resilience.

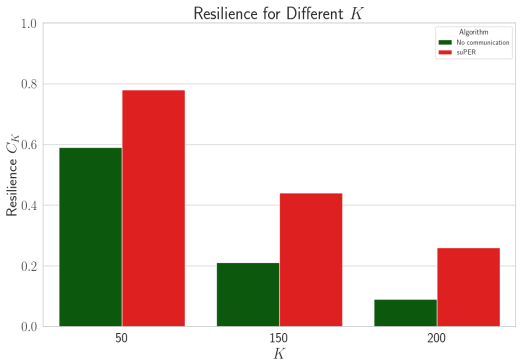


Figure 6: Coop-mining environment resilience.

In both experiments, we observe that all collaborative approaches achieve higher resilience than the no-communication approach, supporting our main hypothesis. The effect is more pronounced for larger-magnitude perturbations (e.g., a small 3% increase in resilience with $K = 50$ in Cleanup versus a 180% increase with $K = 200$). We also observe that the global-centric approach generally outperforms the self-centric approach (higher or similar resilience, and higher initial performance). This further reinforces that collaboration promotes resilience in that agents can recover after a perturbation a larger fraction of their previous performance, even if they are self-interested.

6 Conclusion

We suggest that collaboration promotes resilience: we hypothesized that agents who learn to collaborate will adapt more quickly to changes in their environment. In support of this agenda, we introduced a novel formulation for group resilience. To the best of our knowledge, this is the first measurement of group resilience that is relevant to MARL settings. In addition, we presented an empirical evaluation of various MARL settings and communication protocols that show that collaboration via communication can significantly increase resilience to changing environments.

While we examined our approach in MARL settings with homogeneous agents that collaborate via communication, we intend to examine additional methods for collaboration in settings with heterogeneous groups of agents as a next step. Additionally, we intend to explore resilience in real-world domains, including multi-robot settings.

It is noteworthy that the recent global pandemic perturbed many aspects of the environments in which we operate. In such cases, people used to certain kinds of collaboration before the pandemic may have found it easier to adjust to the unfamiliar constraints that were imposed. We believe our results reflect a quite specific benefit that automated agents can derive from collaborating with one another. We do note that many usual caveats on AI research apply, especially concerning tasks that might not be of societal benefit. We leave this for future work, noting potential solutions in existing research on differential privacy and federated learning.

References

Guy Azran, Mohamad H. Danesh, Stefano V. Albrecht, and Sarah Keren. Contextual Pre-planning on Reward Machine Abstractions for Enhanced Transfer in Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10953–10961, March 2024. ISSN 2374-3468. DOI: 10.1609/aaai.v38i10.28970.

Richard Bellman. A Markovian Decision Process. *Indiana University Mathematics Journal*, 6(4): 679–684, 1957. ISSN 0022-2518. DOI: 10.1512/iumj.1957.6.56038.

Haitham Bou Ammar, Eric Eaton, Matthew E. Taylor, Decebal C. Mocanu, Kurt Driessens, Gerhard Weiss, and Karl Tuyls. An automated measure of MDP similarity for transfer in reinforcement learning. In *Machine Learning for Interactive Systems Workshop (WS-14-07), 28th AAAI Conference on Artificial Intelligence*, pp. 31–37, Québec City, QC, Canada, 2014. AAAI Press.

Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10707–10717. Curran Associates, Inc., 2020.

R. L. Dobrushin. Prescribing a System of Random Variables by Conditional Distributions. *Theory of Probability & Its Applications*, 15(3):458–486, January 1970. ISSN 0040-585X, 1095-7219. DOI: 10.1137/1115049.

Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 162–169, Banff, AB, Canada, 2004. AUAI Press.

Jakob Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2137–2145, 2016a.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning, May 2016b. URL http://arxiv.org/abs/1605.06676. arXiv:1605.06676 [cs].

Matthias Gerstgrasser, Tom Danino, and Sarah Keren. Selectively Sharing Experiences Improves Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36: 59543–59565, December 2023.

Ishan Honhaga and Claudia Szabo. A simulation and experimentation architecture for resilient cooperative multiagent reinforcement learning models operating in contested and dynamic environments. *SIMULATION*, pp. 00375497241232432, 2024.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 3040–3049. PMLR, 2019a.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning, June 2019b. URL http://arxiv.org/abs/1810.08647. arXiv:1810.08647 [cs, stat].

Spencer Jordan and Alex Trott. Normalization of returns across multiple environments. In *Proceedings of Machine Learning Research*, pp. 1–12, 2020. To appear; fill in exact volume & pages if available.

Joel Z. Leibo, Edgar Dué nez Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*. PMLR, 2021. DOI: 10.48550/arXiv.2107.06857. URL https://doi.org/10.48550/arXiv.2107.06857.

Yongyuan Liang and Bangwei Li. Parallel Knowledge Transfer in Multi-Agent Reinforcement Learning, March 2020. URL http://arxiv.org/abs/2003.13085. arXiv:2003.13085 [cs].

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In William W. Cohen and Haym Hirsh (eds.), *Machine Learning Proceedings 1994*, pp. 157–163. Morgan Kaufmann, San Francisco (CA), January 1994. ISBN 978-1-55860-335-6. DOI: 10.1016/B978-1-55860-335-6.50027-1.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments, March 2020. URL http://arxiv.org/abs/1706.02275. arXiv:1706.02275 [cs].

Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: Multi-Agent Variational Exploration. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 1476-4687. DOI: 10.1038/nature14236.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous Methods for Deep Reinforcement Learning, June 2016. URL http://arxiv.org/abs/1602.01783. arXiv:1602.01783 [cs].

Sindhu Padakandla. A Survey of Reinforcement Learning Algorithms for Dynamically Varying Environments. *ACM Computing Surveys*, 54(6):1–25, July 2022. ISSN 0360-0300, 1557-7341. DOI: 10.1145/3459991. URL http://arxiv.org/abs/2005.10619. arXiv:2005.10619 [cs, stat].

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust Deep Reinforcement Learning with Adversarial Attacks, December 2017. URL http://arxiv.org/abs/1712.03632. arXiv:1712.03632 [cs].

Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, and Claudia Linnhoff-Popien. Learning and Testing Resilience in Cooperative Multi-Agent Systems. *New Zealand*, 2020.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic Curriculum Learning For Deep RL: A Short Survey, May 2020. URL http://arxiv.org/abs/2003.04664. arXiv:2003.04664 [cs, stat].

Yichen Qian, Jun Wu, Rui Wang, Fusheng Zhu, and Wei Zhang. Survey on Reinforcement Learning Applications in Communication Networks. *Journal of Communications and Information Networks*, 4(2):30–39, June 2019. ISSN 2509-3312. DOI: 10.23919/JCIN.2019.8917870.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning, June 2018. URL http://arxiv.org/abs/1803.11485. arXiv:1803.11485 [cs, stat].

Kelsey Saulnier, David Saldana, Amanda Prorok, George J. Pappas, and Vijay Kumar. Resilient Flocking for Mobile Robot Teams. *IEEE Robotics and Automation Letters*, 2(2):1039–1046, April 2017. ISSN 2377-3766, 2377-3774. DOI: 10.1109/LRA.2017.2655142. URL http://ieeexplore.ieee.org/document/7822915/.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized Experience Replay, February 2016. URL http://arxiv.org/abs/1511.05952. arXiv:1511.05952 [cs].

Jinhua Song, Yang Gao, Hao Wang, and Bo An. Measuring the Distance Between Finite Markov Decision Processes. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pp. 468–476, Richland, SC, May 2016. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-4239-1.

Eugene Vinitsky, Natasha Jaques, Joel Leibo, Antonio Castenada, and Edward Hughes. An open source implementation of sequential social dilemma games. https://github.com/eugenevinitsky/sequential_social_dilemma_games/issues/182, 2019. GitHub repository.

Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust Reinforcement Learning using Adversarial Populations, September 2020. URL http://arxiv.org/abs/2008.01825. arXiv:2008.01825 [cs, stat].

Cheng-Zhong Xu, Jia Rao, and Xiangping Bu. URL: A unified reinforcement learning approach for autonomic cloud management. *Journal of Parallel and Distributed Computing*, 72(2):95–105, February 2012. ISSN 07437315. DOI: 10.1016/j.jpdc.2011.10.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0743731511001924.

Tan Zhang, Wenjun Zhang, and Madan Gupta. Resilient Robots: Concept, Review, and Future Directions. *Robotics*, 6(4):22, September 2017. ISSN 2218-6581. DOI: 10.3390/robotics6040022.

Zhuangdi Zhu, Kaixiang Lin, Anil K. Jain, and Jiayu Zhou. Transfer Learning in Deep Reinforcement Learning: A Survey, July 2023. URL http://arxiv.org/abs/2009.07888. arXiv:2009.07888 [cs, stat].