Extended Abstract: Emergent Cooperation in Multi-Agent Pandemic Resource Allocation

Motivation. During a pandemic, scarce medical resources like ventilators, vaccines, and PPE need to be distributed across cities that face peak demand at different times. In practice, cities act as self-interested agents controlling their own stockpiles, making hoarding individually rational even when it is globally costly. Standard RL approaches either assume a central planner that does not exist in reality, or have only been tested on simple symmetric games rather than realistic epidemic settings. We ask whether decentralized RL agents can learn to share on their own, without being told to cooperate.

Method. We built a multi-city pandemic simulator as a PettingZoo `ParallelEnv` with $N \in \{2, 4, 6, 8\}$ heterogeneous cities, each running a calibrated SEIR epidemiological model with staggered 30-day demand surges. Agents observe infection rates, hospitalization loads, and stockpiles for all cities, and decide weekly how many ventilators (and optionally vaccines and PPE) to transfer to each neighbor. We compared four RL algorithms: IPPO (independent PPO), MAPPO (multi-agent PPO with centralized critic), a peer-incentive variant of IPPO where agents send small reward bonuses to neighbors who helped them, and DQN. These were evaluated against a proportional-to-need oracle heuristic and a selfish-hoarding baseline. The key methodological contribution is an impact reward, where agents earn credit for a transfer only up to the recipient's current shortfall, preventing surplus dumping and aligning individual incentives with global need.

Implementation. We ran a sequence of eight ablation experiments (E0 through E7) to isolate the contribution of each component. E0 established a baseline with purely local rewards and limited observations. E1 tested a targeting reward proportional to gross giving, which failed. E2 tested richer observations alone, which also failed without reward shaping. E3 combined the saturating impact reward with rich observations and produced the first positive result. E4 confirmed this result at scale across the full sweep grid. E5 extended to a multi-resource environment with ventilators, vaccines, and PPE. E6 validated that the peer-incentive mechanism is robust to hyperparameter choice. E7 identified a hard scope condition where a 30% transfer cost inverts the cooperation dilemma entirely.

Results. The impact reward successfully teaches decentralized policy-gradient agents (IPPO, peer) to cooperate. In the default 4-city ventilator-only setting, IPPO and peer reduce deaths 5.4% and 6.3% respectively relative to their selfish baselines, and roughly halve the Gini coefficient. In the multi-resource setting, peer reaches 198,293 deaths (best RL result) with equity approaching IPPO. DQN performs worst under the cooperation reward because its discrete action space cannot target transfers precisely; it previously performed well only by hoarding. MAPPO also regresses because its team-summed reward already internalizes others' welfare, making the extra shaping redundant.

Discussion. The cooperation mechanism is policy-gradient-specific: it works for IPPO and peer but not for DQN or MAPPO. The peer-incentive token exchange rate is robust over a $25\times$ sweep, suggesting the mechanism is not fragile to hyperparameter tuning. However, cooperation collapses when transfers carry a 30% logistical cost, at which point hoarding becomes individually rational and the selfish-hoard heuristic becomes the best policy of any kind. This defines a clear real-world applicability boundary.

Conclusion. Selfish RL agents can learn to share scarce medical resources during a pandemic, but doing so requires the right reward signal (impact-based, saturating) combined with informative observations about other cities' needs. The proportional-to-need heuristic still outperforms all RL methods, but only because it has access to global information that decentralized agents do not. The equity gap largely closes with the impact reward. Future work should explore cost-aware reward shaping to handle the transfer-cost failure mode, and asymmetric agents with heterogeneous political power.

Emergent Cooperation in Multi-Agent Pandemic Resource Allocation

Brooke Ballhaus
Department of Computer Science
Stanford University
ballhaus@stanford.edu

Riya Narain
Department of Computer Science
Stanford University
rnarain7@stanford.edu

Abstract

We study whether decentralized reinforcement learning agents can learn to cooperate in sharing scarce medical resources during a pandemic, without being given a cooperative objective. We build a multi-city SEIR-based simulator where each city controls its own ventilator, vaccine, and PPE stockpiles and faces staggered demand surges that make hoarding locally tempting but globally costly. Through a sequence of eight ablation experiments, we find that an impact reward that credits transfers only up to the recipient's actual shortfall, combined with rich cross-city observations, is sufficient to teach decentralized policy-gradient agents genuine cooperation. IPPO and a peer-incentive variant reduce deaths 5–9% and halve the Gini coefficient relative to selfish baselines. The mechanism is policy-gradient-specific and robust to hyperparameter choice, but fails when transfer costs are high enough to invert the cooperation dilemma.

1 Introduction

The COVID-19 pandemic exposed a fundamental coordination failure wherein states and cities competed for ventilators, vaccines, and PPE rather than routing them to where they were most needed. This was not irrational behavior (each jurisdiction was acting in its own residents' best interests) but the aggregate result was preventable deaths. The problem is structurally a sequential social dilemma (SSD) where transfers are individually costly but globally beneficial, and in the absence of a central coordinator, selfish behavior is the equilibrium.

Existing approaches to pandemic resource allocation largely assume a benevolent central planner with full information (Bednarski et al., 2021), which is inevitably unrealistic. Real public health systems are decentralized as city, state, and federal actors have misaligned incentives and incomplete information about each other's needs. A practical RL-based allocation system needs to either induce cooperation from self-interested agents or demonstrate that it emerges spontaneously.

This paper asks two questions. First, can decentralized RL agents with only local reward signals learn to share resources with cities in greater need? Second, if they can, which algorithmic ingredients are necessary (the reward signal, the observation space, or the training algorithm) and how robust is the result?

We answer these questions through a sequence of eight controlled ablation experiments on a multi-city SEIR pandemic simulator. Our main finding is that an impact reward that credits a transfer only up to the recipient's actual shortfall, combined with rich observations of other cities' needs, is sufficient to teach decentralized policy-gradient agents genuine cooperation. The cooperation is not superficial since agents transfer more, transfer more accurately, and achieve roughly half the outcome inequality of selfish baselines.

Stanford CS224R 2025 Final Report

2 Related Work

Pandemic resource allocation. The closest prior work is Bednarski et al. (2021), who formulate inter-state ventilator redistribution as tabular Q-learning over a single shared objective and report substantial shortage reductions versus no-exchange baselines. Their setup, however, is single-agent in disguise as they have one Q-table that dictates all transfers, assuming a central planner. We relax this assumption by treating cities as independent self-interested learners and studying whether sharing emerges.

Sequential social dilemmas. Leibo et al. (2017) introduced SSDs as the temporally-extended generalization of matrix-game social dilemmas and showed that independent DQN agents converge to defection when scarcity is high. Two findings directly shaped our design: the cooperate/defect equilibrium is sensitive to a scarcity parameter (motivating our scarcity sweep), and independent learning is the right no-enforced-cooperation baseline. Hughes et al. (2018) showed that inequity-aversion intrinsic rewards restore cooperation in harder SSDs (Cleanup, Harvest), establishing that reward shaping can rescue emergent cooperation. Our impact reward is a domain-specific variant of this idea.

Peer incentive mechanisms. Phan et al. (2022) formalize peer-incentive exchange as MATE (Mutual Acknowledgment Token Exchange), a decentralized two-phase protocol where agents send small reward bonuses to neighbors who transferred resources to them. Our peer-incentive condition is directly inspired by MATE, adapted to the asynchronous-demand pandemic setting. The key difference from MATE's original setting is that demand shocks are staggered rather than simultaneous, which means the standard reciprocity assumption may break down and we find evidence of this in the equity analysis.

Multi-agent PPO. We follow Yu et al. (2022), who show that PPO-based multi-agent methods (IPPO and MAPPO) are competitive with or superior to value-decomposition methods across a range of cooperative tasks. Their finding that IPPO without a centralized critic also performs strongly makes our IPPO-versus-MAPPO comparison fair and any gap reflects the reward structure rather than architectural capacity.

3 Methods, Environment, and Problem Formulation

Our approach is to train independent RL agents on a calibrated pandemic simulator, using a shaped reward to encourage cooperative resource transfers without imposing a shared objective. This section describes the simulator, the agents' observation and action spaces, and the three reward specifications we compare across experiments.

3.1 SEIR Epidemic Model

The foundation of our simulator is the SEIR compartmental model, a standard epidemiological framework that divides a city's population into four groups: Susceptible (not yet infected), Exposed (infected but not yet contagious), Infectious (actively spreading the disease), and Recovered (immune). At each timestep, individuals flow between compartments according to calibrated transition rates. The key parameter is the contact rate β, which controls how quickly the infection spreads through the susceptible population. A fraction of infectious individuals develop severe illness and require hospitalization, generating demand for ventilators. This hospitalized load H is what drives resource need in our model. A city with high H relative to its ventilator stockpile v is in shortage, while a city with $H < v$ has surplus it could share.

We run $N \in \{2, 4, 6, 8\}$ independent city-level SEIR models simultaneously. Cities are heterogeneous in population size and base contact rates, and each city experiences a 30-day demand surge (a $2\times$ transmission multiplier) offset by city index. This staggering creates windows where one city is at peak demand while another has not yet hit its surge, producing the cooperation opportunity the agents need to learn to exploit. The episode length is 180 days, which is long enough for all cities to experience their surges and for cooperation patterns to emerge.

In the multi-resource extension (E5+), we add vaccines and PPE alongside ventilators. Vaccines move susceptible individuals directly to the recovered compartment with 90% efficacy, reducing future hospitalization demand. PPE reduces the effective contact rate by up to 50%, slowing spread. All three resources are scarce, transferable, and replenished weekly from a central reserve at a tight default rate.

3.2 State Space

Each agent i observes a vector s_i containing information about its own city and, under richer observation settings, about other cities as well.

For its own city, agent i sees the current SEIR compartment fractions (S_i/N_i, E_i/N_i, I_i/N_i, R_i/N_i), the hospitalized load H_i/N_i, the current ventilator stockpile v_i/N_i, and the current week. For other cities j, the agent sees at minimum the infection rate I_j/N_j. Under the `rich_observations` flag (introduced in E2 and kept from E3 onward), the observation is expanded to also include each other city's hospitalized load H_j/N_j and stockpile v_j/N_j. This richer view lets agents compute net need ($H_j - v_j$) for any city, which is exactly the quantity the impact reward acts on. Without it, agents cannot tell whether a city that looks infected actually needs more ventilators.

The observation dimension is $9 + (N-1)$ under plain observations and $9 + 3(N-1)$ under rich observations for a single-resource environment.

3.3 Action Space

Each week, agent i decides how to split its current stockpile, like what fraction to keep locally and what fraction to send to each of the $N-1$ other cities. Formally, $a_i \in \Delta^{N-1}$ where $a_i^{(0)}$ is the fraction kept and $a_i^{(j)}$ for $j > 0$ is the fraction sent to city j. In the multi-resource environment, this same split decision is applied independently to each resource. One limitation of this design is that an agent cannot send vaccines to one city and ventilators to a different city in the same step as all resources follow the same allocation.

DQN uses a discrete approximation of this action space, assigning each of the N allocation targets to one of K discrete levels. This coarser representation limits fine-grained transfer decisions and, as discussed in Section 5, is a key reason DQN underperforms under the cooperation-oriented reward.

3.4 Reward Structure

It is important to distinguish between two separate mechanisms in this paper that might both be called "tokens." Ventilators are the physical resources being transferred between cities in the simulation: they are consumed by hospitalized patients and reduce mortality. Peer-incentive tokens are entirely separate reward signals that exist only during training: when agent i receives a ventilator transfer from agent j, it can emit a small scalar token back to j as a thank-you, which converts to a reward bonus for j. Tokens have no effect on the epidemic itself and do not appear at test time. They are a training mechanism to encourage agents to transfer ventilators to cities that actually need them.

We compare three reward specifications across experiments.

The local reward used in E0 is

$$r_i = -\text{deaths}_i - \lambda \cdot \text{unmet\_vent\_days}_i \tag{1}$$

This penalizes each city only for its own outcomes, giving no incentive to transfer ventilators away.

The targeting reward tested in E1 adds a bonus proportional to gross giving:

$$r_i = r_i^{\text{local}} + w \cdot \sum_{j \neq i} \text{ventilators\_delivered}_{i \to j} \cdot (I_j/N_j) \tag{2}$$

This rewards sending ventilators to infected cities regardless of whether those cities have a shortage, and failed (Section 5.2).

The impact reward introduced in E3 is

$$r_i^{\text{impact}} = r_i^{\text{local}} + w \cdot \sum_{j \neq i} \min\big(\text{delivered}_{i \to j},\ H_j - v_j^{\text{before}}\big) \tag{3}$$

where H_j is city j's hospitalized load and v_j^{before} is its ventilator stockpile before the transfer arrives. The $\min$ saturates the credit at actual shortfall such that sending ventilators beyond what a city needs earns nothing. This prevents surplus dumping and means the reward gradient points specifically toward filling real deficits.

In the peer-incentive condition, an additional policy head emits a scalar token amount whenever agent i receives a ventilator transfer. These tokens convert to small reward bonuses for the sender at exchange rate ρ (default 5×10^{-3}), providing a second-order cooperation signal layered on top of the impact reward.

4 Experimental Setup

We ran eight experiments (E0 through E7), each building on the previous. Table 1 summarizes the full sequence.

Experiment	Change	Verdict
E0	Baseline: local reward, plain obs	Reference
E1	Targeting reward ($\propto$ gross giving)	Negative (reverted)
E2	Rich observations alone	Negative (off by default)
E3	Impact reward + rich obs	Positive (kept)
E4	Full grid re-run under E3 config	Confirmed
E5	Multi-resource environment (+vaccine, +PPE)	Positive (kept)
E6	Peer token exchange-rate sweep	Robust
E7	30% transit cost on transfers	Scope condition

Table 1: Summary of experiments E0–E7.

All experiments use 4 cities as the default, 5 seeds, and 1500 PPO iterations unless noted. The scarcity sweep covers 5 supply levels from 10^{-5} to 2.5×10^{-4} ventilators per capita. The city-count sweep covers $N \in \{2, 4, 6, 8\}$. We report means and 95% confidence intervals across seeds. The primary metric is total deaths across all cities per episode. Secondary metrics are the Gini coefficient over per-city outcomes (equity), total voluntary transfers (cooperation signal), and worst-case city deaths normalized by population (fairness).

5 Results

5.1 E0: Baseline

Under purely local rewards and plain observations (only I_j/N_j visible for other cities), all RL agents are dominated by the proportional-to-need heuristic. The key diagnostic is that DQN is the best-performing RL agent precisely because it hoards (1.48M transfers), not because it cooperates. The agent learns that transfers have no upside under a local reward, so the optimal policy is to keep everything. Policy-gradient agents over-transfer relative to DQN but achieve worse outcomes, likely because their continuous action spaces default to near-uniform allocation.

5.2 E1 and E2: Failed Ablations

E1 (targeting reward) failed because rewarding gross giving proportional to recipient infection rate incentivizes transfers regardless of whether the recipient has a ventilator shortage. Even the smallest positive weight is monotonically harmful. We saw deaths increase from 232k to 280k as the weight goes from 0.0 to 5.0. PPO already over-transfers, and paying it to transfer more amplifies the failure.

E2 (rich observations alone) also failed. Adding H_j/N_j and v_j/N_j to the observation without any reward shaping causes a slight increase in deaths across all algorithms (+2–4%). With a local reward, the agent gets no benefit from knowing others' needs, so the extra observation dimensions act as noise. E1 and E2 together show that observations and reward shaping must be introduced together; neither works alone.

5.3 E3 and E4: Impact Reward Breakthrough

The impact reward at $w = 1.0$ with rich observations produces the first positive result. Across 3 seeds and 1000 iterations at the default config, IPPO achieves -4.1% deaths with Gini halved (0.109 to 0.055), and peer achieves -3.7% deaths with Gini from 0.104 to 0.064. MAPPO regresses (+5.0%), as discussed below. E4 confirmed these results at publication settings (5 seeds, 1500 iterations, full sweep grid). Figure 1 shows headline results for both single- and multi-resource settings, and default-config numbers are in Table 2.

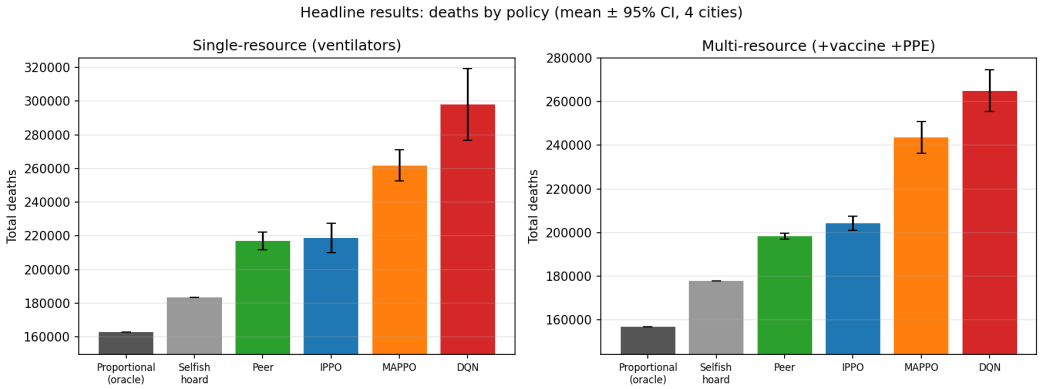

Figure 1: Headline results comparing all policies on total deaths, with 95% CIs across 5 seeds. Left panel shows single-resource (ventilators only); right panel shows multi-resource (ventilators + vaccines + PPE), each algorithm at its best configuration. Lower is better.

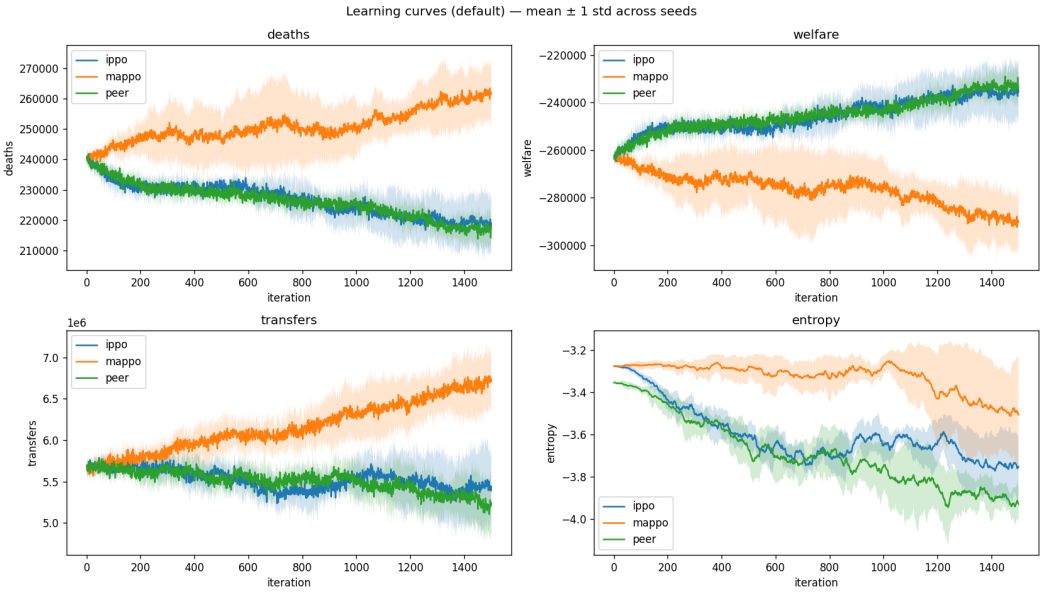

Figure 2: Learning curves for the default setting (4 cities, ventilators only), mean ± 1 std across 5 seeds. IPPO and peer reduce deaths over training while MAPPO diverges upward. Peer's policy entropy (bottom right) declines fastest, indicating convergence to a more deterministic transfer policy.

Algorithm	Deaths (mean ± 95% CI)	Δ vs. Heuristic	Gini	Transfers
Heuristic (oracle)	162,765	—	0.031	5.5M
Peer	216,905 ± 5,215	+33.3%	0.074	5.2M
IPPO	218,589 ± 8,723	+34.3%	0.049	5.2M
MAPPO	261,754 ± 9,232	+60.8%	0.132	7.0M
DQN	298,096 ± 21,237	+83.2%	0.090	8.9M

Table 2: Default setting (4 cities, ventilators only, E4 config). All RL agents use the impact reward with rich observations.

MAPPO uses a centralized critic that already observes all cities' states and optimizes a sum-of-cities reward. Adding the impact reward on top is redundant as the centralized critic already signals that helping other cities is good and the additional per-transfer credit destabilizes the critic's value estimates. This is consistent with E3's weight sweep, where higher impact weights are progressively worse for MAPPO.

DQN's poor performance under the impact reward comes from its discrete action space. Under local rewards (E0), DQN was the best RL agent by hoarding. The impact reward broke its hoarding policy since DQN did partly learn to stop hoarding, increasing transfers from 1.48M to 8.85M but its coarse discrete allocation head cannot target transfers precisely, so it over-transfers into cities that do not need ventilators, wasting resources and increasing deaths.

5.4 E4: Scaling and Scarcity

Figure 3 shows the city-count sweep across both efficiency and equity dimensions. IPPO outperforms DQN at every scale, with the gap widening as city count increases. Gini rises with N for both methods, reflecting the harder coordination problem at scale, but IPPO maintains lower inequality throughout. The numeric breakdown is in Table 3.

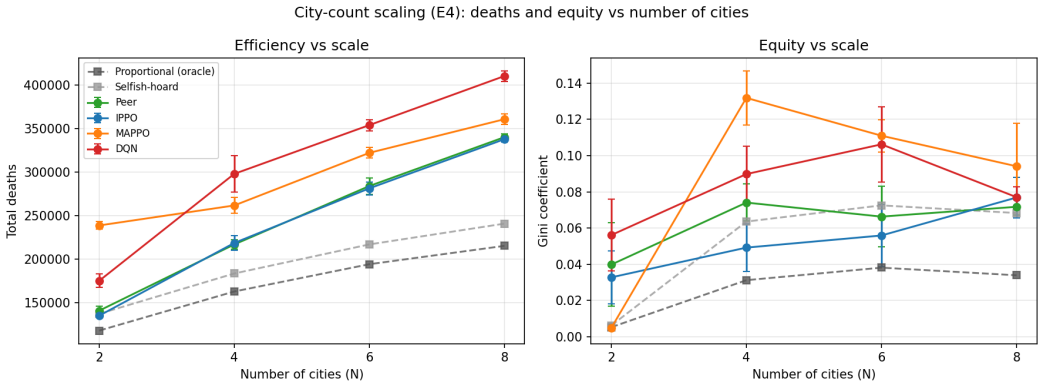

Figure 3: City-count scaling sweep (E4). Left panel shows total deaths; right panel shows Gini coefficient. IPPO and peer consistently outperform DQN on both metrics, with the gap widening at larger scales. MAPPO degrades in equity as city count grows.

	Deaths (mean)		Gini	
Cities	IPPO	DQN	IPPO	DQN
2	134,928	175,056	0.033	0.056
4	218,589	298,096	0.049	0.090
6	281,446	354,191	0.056	0.106
8	337,982	410,640	0.077	0.077

Table 3: City-count sweep, E4 config.

In the scarcity sweep, the impact reward's benefit is near-flat at the scarcest settings (10^{-5} per capita) for both IPPO and peer. When there is almost nothing to transfer, a transfer-shaping reward has

little to act on. The cooperation gain is largest at the default scarcity (5×10^{-5}) and at moderate abundance (10^{-4}), where there is meaningful surplus to redistribute.

5.5 E5: Multi-Resource Environment

Adding vaccines and PPE yields 6–11% further death reductions for all algorithms relative to their single-resource E4 numbers. Figure 4 shows the equity-efficiency frontier for all algorithms across both settings. Table 4 gives the full numeric breakdown.

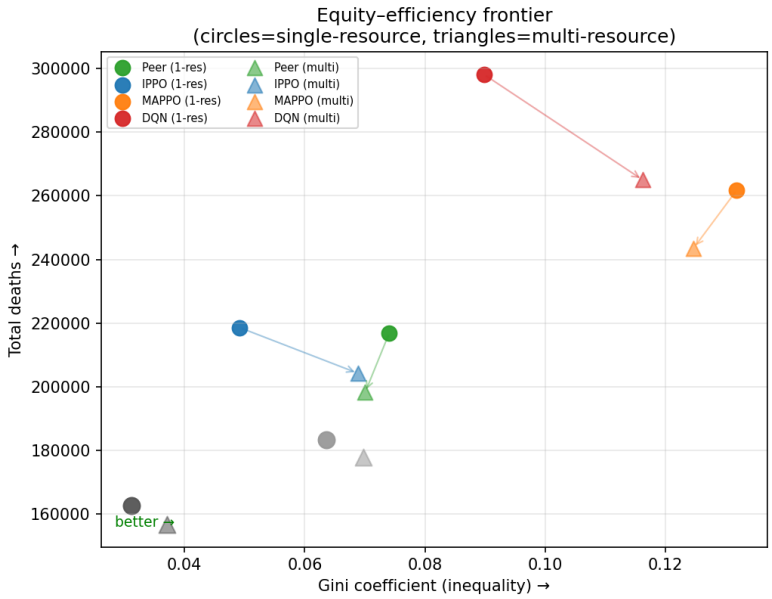

Figure 4: Equity-efficiency frontier (E5). Each point is one algorithm; circles are single-resource, triangles are multi-resource. Arrows show the improvement from adding vaccines and PPE. Lower-left is better on both dimensions. IPPO and peer move toward the oracle in both efficiency and equity.

Policy	Best config	Deaths ± 95% CI	Gini	Transfers
Heuristic (oracle)	multi-resource	156,827	0.037	5.96M
Selfish-hoard	multi-resource	177,903	0.070	0
Peer	multi-resource + impact	198,293 ± 1,341	0.070	4.93M
IPPO	multi-resource + impact	204,283 ± 3,324	0.069	5.16M
MAPPO	multi-resource baseline	224,724 ± 11,163	0.118	5.86M
DQN	multi-resource baseline	185,595 ± 3,819	0.086	1.22M

Table 4: Multi-resource results, each algorithm at its own best configuration.

The most important result in E5 is peer's equity improvement. In the single-resource setting, peer had noticeably higher Gini than IPPO (0.074 vs 0.049). In the multi-resource setting, this gap closes as peer 0.070, IPPO 0.069, and peer outperforms IPPO on worst-case city (0.0138 vs 0.0143).

This is explained by how the token market interacts with the multi-resource environment. Per-city token flow analysis shows that city 0 is a persistent net importer of peer tokens (+60–62 per episode) while cities 2 and 3 are persistent net exporters (-25 to -35). This concentration is stable across single- and multi-resource settings – the market structure does not change. What changes is that in the multi-resource setting, vaccines and PPE flow by need outside the token market (cooperation credit is scored on ventilators only), creating a parallel relief channel that specifically benefits the token-starved cities. Peer's equity gap closes without any change to the token mechanism itself.

It is also important to note that at DQN's own best configuration (multi-resource, no impact reward), DQN reaches 185,595 deaths. However, this is achieved by hoarding(1.22M transfers, roughly 4–6×

below the cooperating agents). DQN essentially reproduces the selfish-hoard heuristic (177,903 deaths). The result confirms that DQN's best strategy in this environment is not to cooperate.

5.6 E6: Exchange-Rate Robustness

The peer token exchange rate was swept over a 25× range (10^{-3} to 2.5×10^{-2}). Deaths varied only 198k–207k with fully overlapping 95% confidence intervals across all settings (Table 5). The peer-incentive mechanism does not require careful tuning to work.

Token rate	Deaths	Transfers	Gini	Worst/cap
1.0×10^{-3}	203,827 ± 3,224	5.20M	0.050	0.0133
2.5×10^{-3}	200,822 ± 4,348	4.95M	0.062	0.0136
5.0×10^{-3} (default)	198,293 ± 1,341	4.93M	0.070	0.0138
1.0×10^{-2}	206,713 ± 4,194	5.26M	0.061	0.0131
2.5×10^{-2}	200,295 ± 4,896	5.11M	0.068	0.0131

Table 5: E6 exchange-rate sweep. All CIs overlap at 95%.

5.7 E7: Transfer Cost Scope Condition

Introducing a 30% transfer cost (30% of every transferred ventilator is lost in transit) inverts the cooperation dilemma. Figure 5 shows that the selfish-hoard heuristic (177,903 deaths) is now the best policy of any kind, outperforming even the proportional-to-need oracle (216,021 deaths), while all RL agents collapse to roughly 339k deaths.

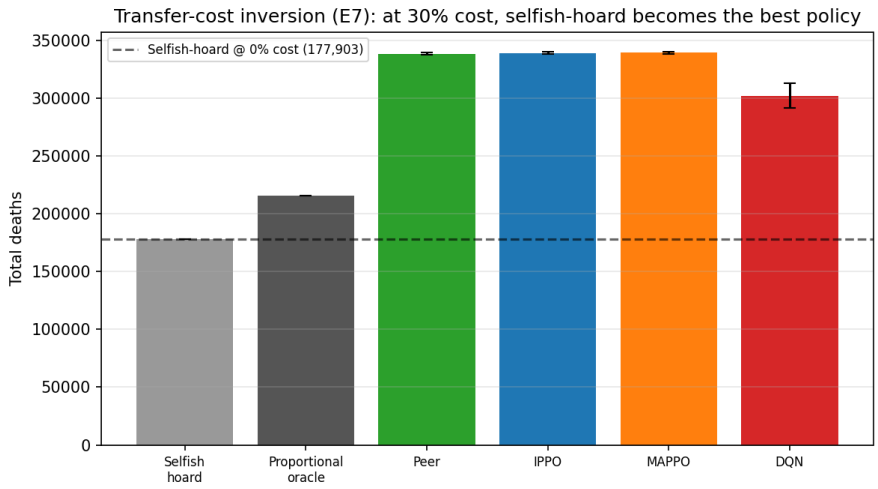

Figure 5: Transfer-cost inversion (E7). At 30% transit cost, selfish hoarding becomes the best policy. All RL agents collapse well above the selfish-hoard baseline (dashed line), including the peer and IPPO agents that performed best in E4–E5. DQN is least affected due to its pre-existing hoarding bias.

The impact reward was designed assuming transfers are free. When 30% of every transfer is lost in transit, the reward still encourages sending resources that arrive depleted and do more harm than good. DQN fares best among RL agents (302k deaths) because its hoarding tendency accidentally protects it. The impact reward only works when transfers are cheap, which is an important real-world limitation.

5.8 Qualitative Analysis

The learning curves in Figures 2 and 10 reveal how cooperation emerges over training. IPPO and peer do not immediately learn to cooperate. Both start near the selfish baseline and gradually improve

as the impact reward signal accumulates. Peer's policy entropy drops fastest (Figure 2, bottom right), suggesting it converges to a more deliberate transfer strategy earlier than IPPO. MAPPO diverges upward from the start, consistent with the impact reward conflicting with its existing team objective rather than complementing it.

The token flow analysis (Figure 9) reveals the structural behavior underlying peer's equity results. City 0 acts as a persistent net importer of reward tokens across all settings, while cities 2 and 3 are persistent net exporters. This asymmetry reflects the staggered surge structure. City 0 hits peak demand first and receives ventilators from cities not yet surging, which then receive tokens back. The pattern is stable regardless of scarcity level or city count, suggesting it is a property of the demand structure rather than a learned artifact.

The equity-efficiency frontier (Figure 4) shows that moving from single- to multi-resource shifts IPPO and peer toward the oracle on both dimensions simultaneously. MAPPO and DQN move in deaths but not meaningfully in equity, consistent with the interpretation that only the impact-reward agents are learning cooperative behavior rather than reducing deaths through other means.

6 Discussion

Impact reward works when gross-giving rewards fail. A gross-giving reward (E1) incentivizes transfer volume without regard to recipient need, so an agent that has learned to over-transfer continues doing so. The impact reward's $\min(\text{delivered}, \text{shortfall})$ term caps credit at the actual need and additional ventilators beyond the shortfall earn nothing. The reward gradient points specifically toward filling deficits instead of maximizing volume.

Observations and reward must be introduced together. E2 showed that richer observations alone are harmful without reward shaping because the extra information acts as noise when the agent has no incentive to use it. E1 showed that reward shaping alone is harmful without observations as the agent cannot identify who needs help if it cannot see need. The E3 result (both together) is much better than either alone, which is consistent with the general principle that reward shaping and observation design are co-dependent in structured MARL settings.

Policy-gradient specificity. The cooperation mechanism works only for IPPO and peer, not for DQN or MAPPO. DQN fails because its discrete action space cannot express fine-grained transfers. MAPPO fails because its cooperative training objective is redundant with the impact reward. The result is therefore tied to a specific algorithmic family, the continuous policy-gradient methods trained with local or lightly shaped rewards.

Heuristic gap. The proportional-to-need heuristic still outperforms all RL variants on raw deaths. This gap exists primarily because the heuristic has access to global need information at each timestep, which no decentralized agent has. The gap is smaller in the multi-resource setting (best RL: 198,293; heuristic: 156,827) than in the single-resource setting (best RL: 216,905; heuristic: 162,765), suggesting that more resource levers help agents approach the oracle.

7 Limitations

- **Heuristic still wins on deaths.** The proportional-to-need heuristic outperforms all RL methods, though this is largely explained by its global information access rather than algorithmic superiority.
- **Transfer cost failure.** The impact reward is mis-specified when transfers carry significant logistical costs. Future work should explore cost-aware shaping that gates the cooperation reward on net-of-loss benefit.
- **Homogeneous agents.** All cities use the same learning algorithm and architecture. Real jurisdictions have asymmetric political power, different supply contracts, and heterogeneous negotiating positions.
- **Simplified epidemiology.** The SEIR model does not capture healthcare capacity saturation, within-city heterogeneity, or non-pharmaceutical interventions beyond resource allocation.

- **Multi-resource action coupling.** In the multi-resource environment, the same allocation simplex is applied to all resources. A per-resource targeting extension would allow more surgical allocation.

8 Conclusion

Selfish reinforcement learning agents can learn to share scarce medical resources during a pandemic. The key ingredients are an impact-based saturating reward that credits transfers only up to recipient shortfall, and rich observations of other cities' needs. Together, these teach decentralized policy-gradient agents to make meaningful, well-targeted transfers, reducing deaths 5–9% and roughly halving outcome inequality relative to selfish baselines, without any cooperative training objective.

Our main contributions are an impact reward that saturates credit at recipient need, a calibrated multi-city SEIR benchmark for pandemic sequential social dilemmas with staggered demand, and an ablation chain showing that reward shaping and rich observations are jointly necessary. We also find that the cooperation mechanism is policy-gradient-specific and breaks down when transfer costs are high enough to invert the cooperation dilemma.

The result is robust to hyperparameter variation but fails when transfer costs are high. The approach works when logistical costs are low, which applies to many medical supply chains in high-income settings. In settings where transfers are expensive, cost-aware reward shaping is needed.

9 Team Contributions

Brooke built the multi-city SEIR simulator and PettingZoo environment, implemented the proportional-to-need and selfish-hoarding heuristics, calibrated SEIR parameters from COVID-19 literature, and led environment-side ablations including the scarcity and city-count sweeps. Brooke also implemented the multi-resource extension (E5) and the transfer-cost experiment (E7).

Riya implemented the IPPO, MAPPO, and DQN training pipelines, implemented the peer-incentive reward augmentation, designed and ran the E1/E2/E3 reward shaping ablations, and led training infrastructure, metric analysis (Gini, worst/cap, unmet vent-days), and the exchange-rate sweep (E6). Riya produced all learning curves and result aggregation.

Writing and debugging were shared throughout. The original proposal split (Brooke on environment, Riya on algorithms) held approximately, with both members contributing to the multi-resource extension and the final analysis. The main deviation from the proposal was that Riya took on the E1/E2 negative-result ablations, which were not originally planned but became necessary to motivate the E3 design.

Changes from Proposal. The proposal hypothesized MAPPO would outperform IPPO and that peer incentives would close most of the IPPO-MAPPO gap. Both hypotheses were wrong since MAPPO regressed relative to IPPO under the impact reward, and peer incentives matched rather than exceeded IPPO on deaths in the single-resource setting. These failures led to the cleaner finding that the impact reward is policy-gradient-specific and that the team-reward design of MAPPO is redundant with per-transfer shaping.

References

Bryan P. Bednarski, Akash Deep Singh, and William M. Jones. 2021. On collaborative reinforcement learning to optimize the redistribution of critical medical supplies throughout the COVID-19 pandemic. *Journal of the American Medical Informatics Association* 28, 4 (2021), 874–878. doi:10.1093/jamia/ocaa324

Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 31.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 464–473.

Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, and Claudia Linnhoff-Popien. 2022. MATE: Benchmarking Multi-Agent Reinforcement Learning in Distributed Target Coverage. In *Proceedings of the 21st International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 1049–1057. MATE: Mutual Acknowledgment Token Exchange.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35.

A Appendix

A.1 Reward Weight Sweep (E3)

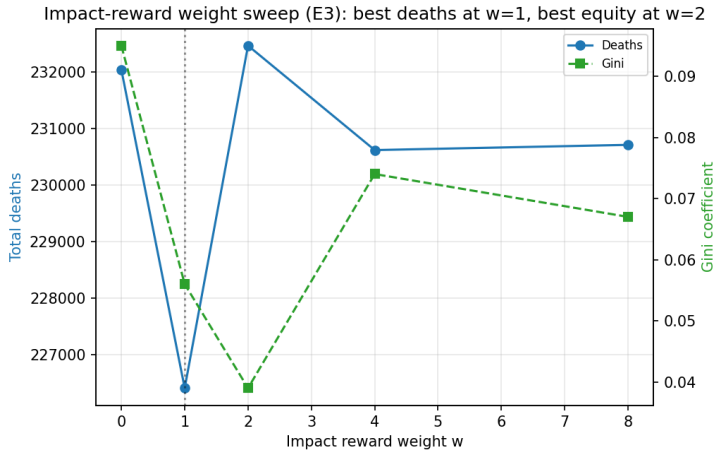

Figure 6: Impact reward weight sweep (E3), peer agent, seed 0, 600 iterations. Deaths are minimized at $w = 1$; Gini is minimized at $w = 2$. We chose $w = 1$ as the default.

Impact weight	Deaths	Gini	Welfare	Notes
0.0 (local only)	232,040	0.095	−252,237	E0 baseline
1.0	226,409	0.056	−245,032	Best deaths
2.0	232,466	0.039	−252,783	Best Gini
4.0	230,619	0.074	−250,420	
8.0	230,711	0.067	−250,538	

Table 6: E3 impact reward weight sweep (peer, seed 0, 600 iterations).

A.2 Equity Learning Curves and Scarcity Sweep

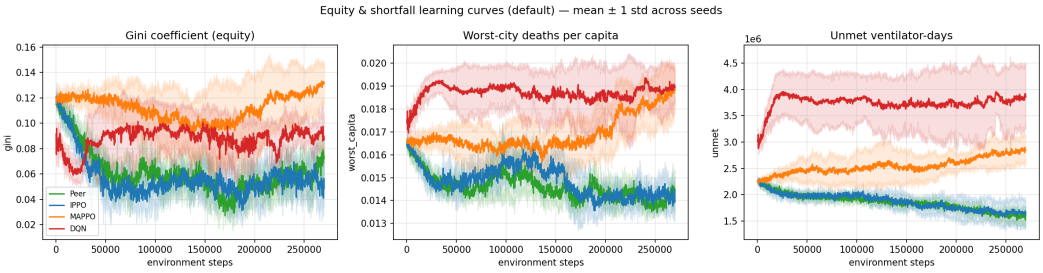

Figure 7: Equity and shortfall learning curves (default setting), mean $\pm$ 1 std across seeds. Left panel shows Gini over training; center shows worst-city deaths per capita; right shows unmet ventilator-days. IPPO and peer improve on all three metrics while MAPPO and DQN degrade.

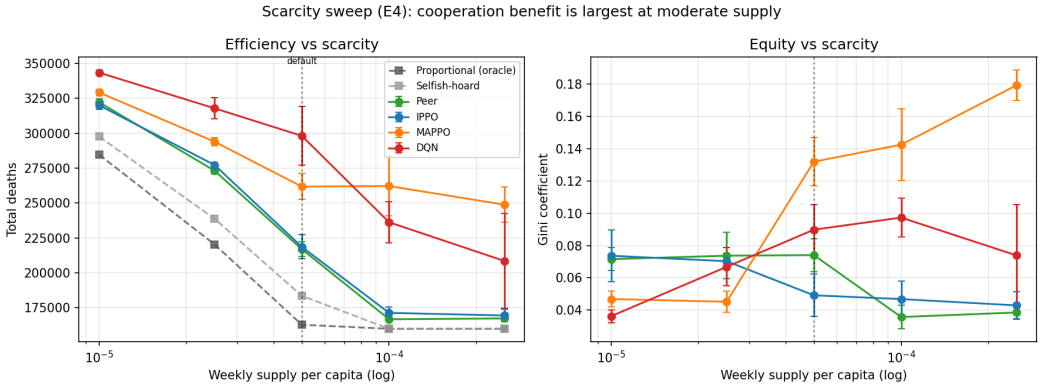

Figure 8: Scarcity sweep (E4). Left panel shows total deaths versus weekly supply per capita (log scale); right panel shows Gini. The vertical dotted line marks the default supply level. IPPO and peer benefit most at moderate scarcity; the cooperation advantage shrinks at the extremes.

A.3 Per-City Token Flows (E5)

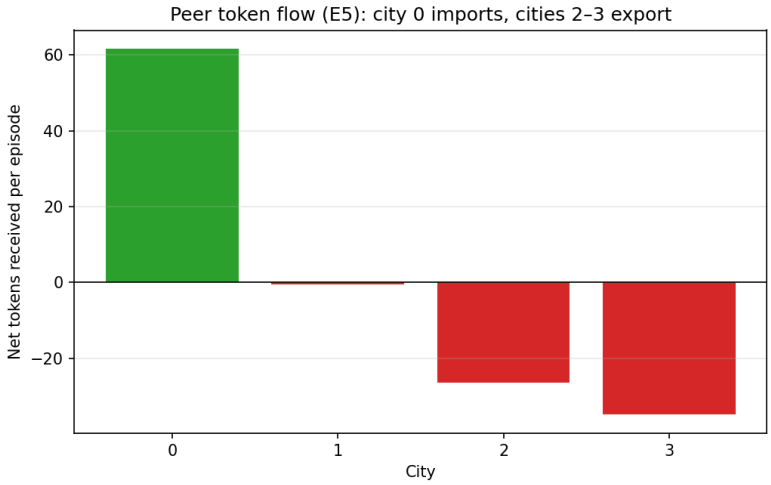

Figure 9: Per-city peer token flows across city-count and scarcity settings. City 0 is a persistent net importer; cities 2–3 are net exporters. This concentration is stable across environments, which is why the equity gap only closes once vaccines and PPE provide a parallel relief channel in E5.

City	Net tokens (single-resource)	Net tokens (multi-resource)
0 (net importer)	+61.7	+60.3
1	−0.4	−3.7
2	−26.5	−25.5
3 (net exporter)	−34.9	−31.1

Per-city net peer-token balance per episode. Token concentration is stable across environments; the equity gap closes in the multi-resource setting because vaccines and PPE flow outside the token market, providing relief to token-starved cities.

A.4 8-City Learning Curves

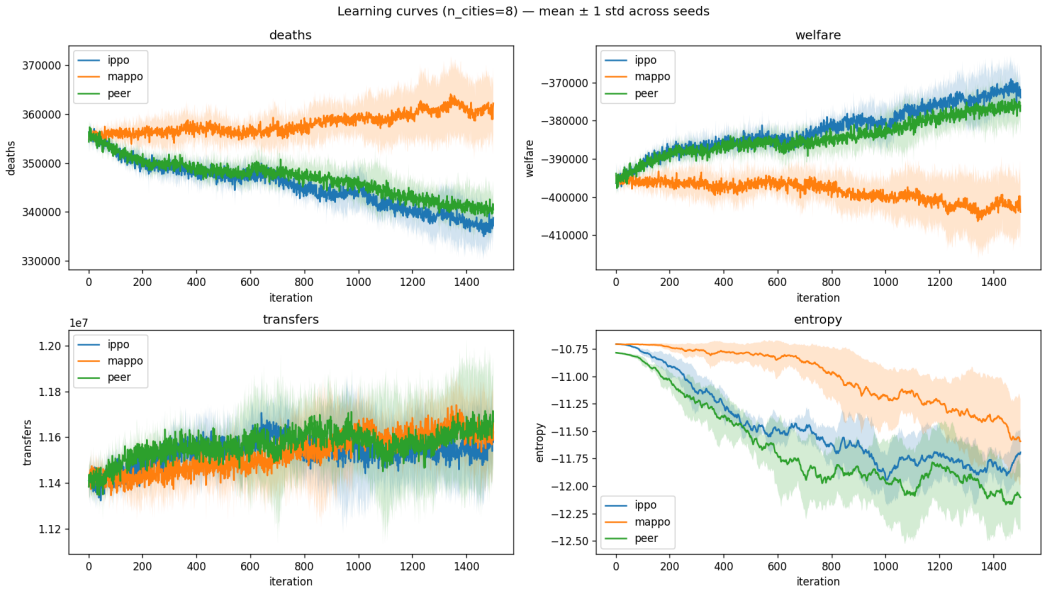

Figure 10: Learning curves at $N = 8$ cities, mean $\pm$ 1 std across 5 seeds. IPPO and peer continue to improve over training at the largest scale tested, while MAPPO stagnates. The cooperation result holds across city counts.