Calculus of Consent via MARL: Legitimating the Collaborative Governance Supplying Public Goods

Yang Hu
Tsinghua University
Beijing, China
`huy18@mails.tsinghua.edu.cn`

Zhui Zhu
Tsinghua University
Beijing, China
`zhuz18@mails.tsinghua.edu.cn`

Sirui Song
McGill University
Montréal, Canada
`sirui.song@mail.mcgill.ca`

Xue Liu
McGill University
Montréal, Canada
`xueliu@cs.mcgill.ca`

Yang Yu
Tsinghua University
Beijing, China
`yangyu1@tsinghua.edu.cn`

Abstract

Public policies that supply public goods, especially those involve collaboration by limiting individual liberty, always give rise to controversies over governance legitimacy. Multi-Agent Reinforcement Learning (MARL) methods are appropriate for supporting the legitimacy of the public policies that supply public goods at the cost of individual interests. Among these policies, the inter-regional collaborative pandemic control is a prominent example, which has become much more important for an increasingly inter-connected world facing a global pandemic like COVID-19. Different patterns of collaborative strategies have been observed among different systems of regions, yet it lacks an analytical process to reason for the legitimacy of those strategies. In this paper, we use the inter-regional collaboration for pandemic control as an example to demonstrate the necessity of MARL in reasoning, and thereby legitimizing policies enforcing such inter-regional collaboration. Experimental results in an exemplary environment show that our MARL approach is able to demonstrate the effectiveness and necessity of restrictions on individual liberty for collaborative supply of public goods. Different optimal policies are learned by our MARL agents under different collaboration levels, which change in an interpretable pattern of collaboration that helps to balance the losses suffered by regions of different types, and consequently promotes the overall welfare. Meanwhile, policies learned with higher collaboration levels yield higher global rewards, which illustrates the benefit of, and thus provides a novel justification for the legitimacy of, promoting inter-regional collaboration. Therefore, our method shows the capability of MARL in computationally modeling and supporting the theory of calculus of consent, developed by Nobel Prize winner J. M. Buchanan.

1 Introduction

As brilliantly revealed in J. M. Buchanan's celebrated book *The Calculus of Consent* [1], collective actions are always formed by (probably conflicting) individual actions, during which process some individuals should rationally alienate or waive their rights of liberty based on the calculus of consent. This outstanding theory of public choice applies to the supply of public goods, which generally requires the collaboration among multiple parties that could restrict the liberty of some parties for the sake of public interest. Thus, the controversy between the extent of liberty restriction and the

necessary level of collaboration is naturally raised for debate, and governors will face the problem of legitimacy whenever they fail to explain the necessity and intensity of collaborative policies.

It should be noticed that the making of optimal policies and their legitimacy explanations involve delicate quantified calculation of individual and public interests, or more specifically in the scope of public goods supply, the balance of costs and benefits. The calculation is of such high computational complexity that is beyond the capability of human instincts, which impairs the legitimacy and credibility of policy-makers, and thus restricts them from fully functioning to pursue the greatest good for the public. Fortunately, the rapidly developing computational tools for decision making, such as optimal control and reinforcement learning, have enabled us to find optimal solutions for complicated problems, so they can effectively serve as convincing justifications for policy-makers.

There is no doubt that the most recent large-scale debate over the supply of public goods is about COVID-19 pandemic control policies. In the highly globalized and interconnected world today, pandemic outbreaks have proven to be more devastating to human society than any time in history — they not only cause millions of deaths around the world, but also jeopardize the global value chain (GVC) [2] and give rise to billions of dollars of economic losses. Economists have realized that the control of infectious diseases falls into the category of global public goods (GPGs) that cannot be efficiently supplied without intensive inter-regional collaboration [3]. And here naturally comes the problem of legitimacy: how can a central government convince its local governments, or an international association convince its member countries, that enforcing strict lockdown policies is an appropriate and necessary measure to take? Why should the latter be willing to carry out those policies that might harm their own economic interests?

Computational approaches are playing an increasingly important role in search of optimal pandemic control policies. For example, Tsay et al. proposed an optimization-based decision-making framework to calculate optimal pandemic control policies [4]. Artificial intelligence tools such as Multi-Agent Reinforcement Learning (MARL) algorithms have also be employed to allow more complex models, balancing health and economic costs [5, 6]. However, all of these studies focus on centralized control policies only, so that they do not take into consideration how multiple independent regions should and would collaborate to fight against the pandemic. Therefore, we feel motivated to invent new computational models that helps to analyze the underlying patterns of inter-regional collaboration, which also helps to find and legitimize optimal pandemic control policies.

The structure of the multi-region collaborative pandemic control problem bears several challenges, such as intensive coupling among regions brought by pandemic spread, the exponential explosion nature of pandemic spread, and the instability of observed environment due to inderdependency of regional policies. To address these inherent challenges, we design an MARL model called Intelligent Region Collaboration (IRC), where each region is represented by a learning agent, and each agent's action is to block a proportion of traffic from any other regions into it, which abstracts the mobility restriction policy in reality. Regions are assigned different *pandemic-tolerance levels* and *lockdown-tolerance levels* to capture the major factors that influence policy-making. To reflect different collaboration levels, each agent is assumed to receive both an individual reward and a global reward, and the trade-off weight between these two terms, known as the *reward mixing ratio*, quantifies collaboration levels. The MARL model is trained with a specifically designed actor-critic algorithm that "decouples" the highly unstable environment faced by each agent into a stable local environment learned by the agent itself, and a stable global environment learned globally.

The IRC method proposed in this paper gives hint for effective tools to find optimal pandemic control policies under different inter-regional collaboration levels. These policies not only help policy-makers in the real world, but also provide a perspective to explore the influence of different tolerance levels and collaboration levels on agents' behavioral patterns and the final outcomes. For this purpose, we construct an exemplary environment with agents of different tolerance levels, and observe the policy learned by our model under different collaboration levels. The learned policies are evaluated by multiple metrics, compared with two heuristic baseline policies. Experimental results show that different collaboration patterns appear under different collaboration levels in an interpretable way. It is observed that, when a region only cares about its own interest, its optimal pandemic control policy is almost completely based on its own tolerance levels, and thus the amount of blocked traffic is highly imbalanced among different regional types; on the other hand, as global interests become dominant in a region's reward composition, it is more likely for a region to sacrifice a small portion of its local reward by blocking extra amount of traffic into it, in exchange for lower

lockdown penalties of more vulnerable regions. The collaboration is also characterized by more evenly distributed mobility among different types of regions.

The contribution of this paper is that we design an MARL model to find optimal region-collaboration pandemic control policies, which contains an adjustable parameter that quantifies the collaboration level among regions, or more specifically, the proportion of global reward in each agent's mixed reward. In this way, we obtain a new perspective to analyze the collaboration behavior of regions with different types under different collaboration levels, and by examining the patterns in the learned optimal policies we can better understand the underlying logic of inter-regional collaboration, and thus provide new justifications for the legitimacy of collaborative governance supplying public goods.

2 Related Works

In this section, we discuss the research that is the most pertinent to the work in this paper.

2.1 Multi-Agent Reinforcement Learning

Reinforcement learning (RL) has long been an active research field for many decision-making problems [7, 8]. Despite its success in single-agent settings, Multi-Agent Reinforcement Learning (MARL) that deals with multiple autonomous agents is much more complicated, and has seen a rapid development in the past decade [9, 10]. State-of-the-art MARL algorithms are able to show human-level performance in complex multi-agent games such as DOTA [11] and StarCraft II [12].

In the typical cooperative MARL setting, the environment responds to the aggregate action of all agents with a single global reward, and agents take joint efforts to maximize total global reward. Therefore, a natural idea is to learn the decomposition of the global reward first, and then apply value-based RL algorithms for each agent, which is adopted by popular algorithms like VDN [13], QMIX [14], QTRAN [15], etc. Another approach is to adapt policy gradient methods for multi-agent settings, among which the most successful algorithms are MADDPG [16], COMA [17] and MAAC [18]. There is also much research investigating the cooperative behaviour of MARL agents in various environments [19, 20].

The multi-agent setting in this paper is different from the typical cooperative setting in most existing literature. Our pandemic control task is special in that agents are allowed to retrieve their own local rewards in response to their actions, and thus the value decomposition network is unnecessary. Meanwhile, collaboration level is abstracted as an adjustable hyperparameter, so that collaborative behavior can be controlled rather than only observed.

2.2 Pandemic Control

In epidemiological studies, the Susceptible-Infected-Recovered (SIR) model is a widely used mathematical model for pandemic transmission, on which many research on pandemic control strategies are based [21]. It is also modified for more realistic modelling: SEIR model added an Exposed state to make the model more generalizable [22], while DURLECA proposes the SIHR model to capture the challenges posed by asymptomatic infections [5].

Based on these pandemic transmission models, researchers have discussed how to find the optimal pandemic control strategy in different settings using traditional optimization approaches [23–25]. More recently, the application of RL methods is proposed, where the actions of all regions are determined by one agent in a centralized manner [5]. Compared with these existing studies, our IRC method is the first to apply MARL methods for optimal pandemic control policies, and to investigate the behaviour patterns of different regions under different collaboration levels.

3 Modeling and Method

3.1 Environment Modeling

The environment model focuses on the traffic (or *mobility*) and pandemic transmission among different regions. Therefore, each region can be simply regarded as a single node within a fully connected graph, and all within-node details are excluded from our model.

Mobility. The amount of mobility demand $M_{d,i,j}^t$ between each pair of regions i, j at each time step is pre-determined. Each region regulates its pandemic control policy by determining the proportion of incoming mobility demand that can be fulfilled (note that traffic is controlled by destination region rather than the departure region). Therefore, region j is allowed to determine $p_{i,j}^t \in [0, 1]$ for all $i \neq j$, and fulfills only that proportion of the total mobility demand. A more rigorous formulation of mobility demand and actions can be found in Appendix A.1.

It is reasonable to abstract pandemic control policy as incoming mobility restrictions. Mobility between regions reflects the quantity of economic activities, and mobility restrictions will lead to losses of economic interests; mobility of people and commodities can be easily controlled by region governments via border and public transportation controls; mobility is also closely related to pandemic control as unrecognized infections will spread the virus to the destination. Therefore, in our model, agents are expected to maximize the fulfillment of mobility demands, but also wisely restrict mobility to reduce the spread of pandemic simultaneously.

Pandemic transmission. The pandemic transmission model is based on the Susceptible-Infected-Hospitalized-Recovered (SIHR) model proposed in [5]. The *pandemic state* of region i at time step t is $E_i^t = (S_i^t, I_i^t, H_i^t, R_i^t)$, which consists of the number of susceptible, infected, hospitalized, and recovered population within that region. Agents are able to observe the number of hopitalization and recovery, but cannot distinguish between susceptible (healthy) people and infected people that are both asymptomatic. Therefore, the *visible pandemic state* of region i at time step t is $E_{v,i}^t = (S_i^t + I_i^t, H_i^t, R_i^t)$. The detailed model for pandemic transmission is deferred to Appendix A.2.

3.2 Problem Formulation

The target of our multi-region collaborative pandemic control problem is two-fold — agents are expected to minimize the spread of infection, and fulfill the maximum proportion of mobility demand at the same time. Therefore, the *local reward* R_i of region i is determined by two factors: how much mobility demand into this region is fulfilled, and how serious the pandemic state is in this region. More rigorously, each region i receive a local reward R_i for its pandemic situation at each time step, which is a sum of the *pandemic-spread cost* $C_{p,i}$ and the *mobility-control cost* $C_{m,i}$

$$R_i = -C_{p,i}^t - C_{m,i}^t, \tag{1}$$

The *global reward* R_g (the common interest of all regions) is the sum of all local rewards

$$R_g = \sum_{i=1}^n R_i = \sum_{i=1}^n \left(-C_{p,i}^t - C_{m,i}^t \right). \tag{2}$$

Here costs $C_{p,i}^t$ and $C_{m,i}^t$ grow exponentially with regard to the number of hospitalized patients (H_i^t) and an accumulated amount of blocked mobility, respectively. Two characteristic hyperparameters, namely *pandemic-tolerance level* $H_{0,i}$ and *lockdown-tolerance level* $L_{0,i}$, represent the exponential growth rate mentioned above. The pair $(H_{0,i}, L_{0,i})$ is also referred to as *region types* since they sufficiently reflect a region's resilience for pandemic outbreak and lockdown — the higher the hyperparameter, the more resilient a region is. Detailed definitions of these costs and tolerance levels are deferred to Appendix A.3.

Although the two tolerance levels seem highly abstract and artificial, they are meaningful in practice. We would like to emphasize that finding effective pandemic control measures involves both epidemiological and socioeconomic considerations. On the epidemiological side, research has revealed the effect of human mobility on the COVID-19 pandemic [26, 27], pointing out the fact that respiratory infectious diseases like the COVID-19 can hardly be controlled without strict international restrictions on the mobility of population and commodities. However, in the socioeconomic perspective, strict lockdown policies will probably damage local economy [28], leading to unfavorable consequences like unemployment, supply shortage or even depression. Therefore, policy-makers face an inherent trade-off between pandemic control and economic well-being — countries and cities have different supply of medical resources, and those with more medical resources tend to be more resilient to a pandemic outbreak; meanwhile, regions whose economy rely heavily on travellers and imported commodities will be more reluctant to carry out complete lockdown policy, as it may hurt the economy even more severely than the pandemic outbreak. Therefore, the simple abstraction is sufficient for demonstrating conflicting factors of interest, as the equilibrium is largely determined

by the *pandemic-tolerance level* (whether it has abundant medical resources to endure a long-term pandemic outburst) and the *lockdown-tolerance level* (whether its economic structure can survive a long-term lockdown) as we have defined above.

Now we can follow the multi-region collaborative pandemic control problem. The objective of each agent is a weighed sum of local and global rewards, namely

$$p_i^t = \arg\max_p \left\{ R_i(p; M_{d,i}^t, E_i^t) + \alpha R_g(p; M_d^t, E^t) \right\}. \tag{3}$$

where we introduce a *reward mixing ratio* α to represent the *collaboration level*; i.e., the proportion that the global reward accounts for in the mixed reward $\tilde{R}_i = R_i + \alpha R_g$.

Indeed, regions are not equally willing to collaborate at the cost of their own benefits — regions in a collaborative system are expected to enforce strict lockdown policies to guarantee global well-being, even if they will damage the local economy; regions in a self-centered system, however, will more likely carry out loose restrictions to save the local economy, at the cost of wider-spread diseases and global utility losses which they do not really care. As a result, different behavior patterns will be observed in systems of regions that vary greatly in the the willingness to collaborate, and thus a legitimate policy in one system might not be so in another — strict lockdown policies can be easily enforced in unitary governments, but harder in federal governments; vaccines can be efficiently distributed within a country, but significantly less smoothly within an international association. Differences in the willingness to collaborate root in the difference in social capitals, which plays a decisive role in the organization of collaborative supply of public goods. Such "willingness" should be emphasized in the analysis of inter-regional pandemic control policies.

Differences in collaboration levels may root from different sources — it could be the natural consequence of a coherent and unified political structure (e.g., as in China) that always acts in pursuit of greatest global utility; or it could be due to the political relationship and economic inter-dependence (e.g., as in the EU), so that the high externality of pandemic control prompts the regions to care more about the global utility.
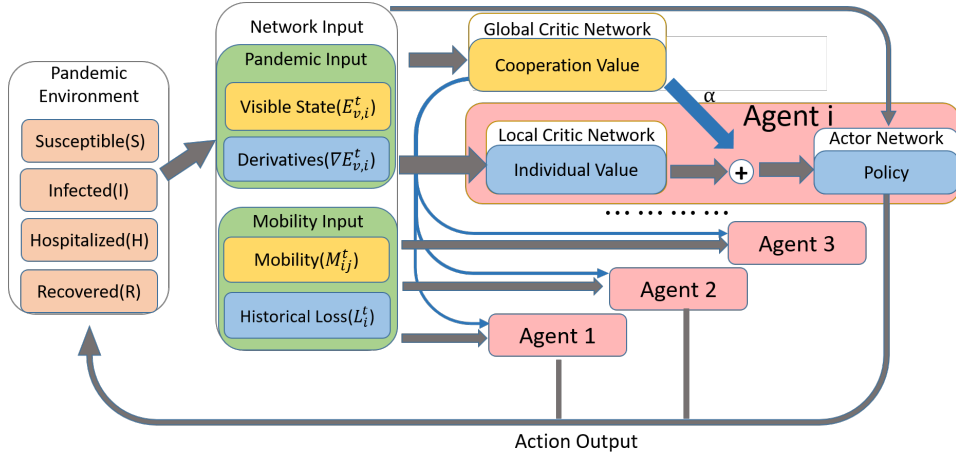
3.3 Intelligent Region Collaboration (IRC)

Figure 1: Schematic diagram of Intelligent Region Collaboration (IRC) model.

To solve the multi-region collaborative pandemic control problem, we design an MARL method called Intelligent Region Collaboration (IRC), which implements a multi-agent actor-critic RL algorithm enhanced by GNN (graph neural networks). Each region is modeled as an RL agent. For agent i at time step t, it first acquires an observation consisting of the visible pandemic state E_v^t and mobility demands $(M_d^t, M_d^{t+1}, M_d^{t+2}, M_d^{t+3})$ for a period of 4 time steps from the environment; then based on this information, the agent determines its action $p_i^{t:(t+3)}$ for the next 4 time steps. Note that control actions are carried out every several time steps, which makes practical sense since changes in mobility demands and control policies can only be applied periodically in reality.

Details of our RL framework are presented below:

- **State.** The state includes the visible pandemic state E_v^t and its temporal derivative $\nabla_t E_v^t$.

- **Action.** The action of each RL agent i is a column vector $p_i^t = (p_{1,i}^t, \cdots, p_{n,i}^t)^\top$ (n is the number of regions), where $p_{j,i}^t$ is the fulfilled proportion of mobility demand $M_{d,j,i}^t$.

- **Reward.** Agent i receives a *mixed reward* $\tilde{R}_i = R_i + \alpha R_g$ as defined previously.

- **Learning Algorithm.** The IRC method is based on the GNN-enhanced Deep Deterministic Policy Gradient (DDPG) method used in DURLECA [5]. The GNN is used to estimate the future pandemic status from the observations, which can utilize the underlying graph structure. Each agent is trained by the DDPG algorithm introduced in [29].

To deal with the challenge of mutually coupled and volatile environment observed by each agent that obstructs efficient training, critics in our IRC model are specially designed to utilize physical decomposablity of local and global rewards in our setting. Therefore, instead of employing a single critic for each agent and introducing an extra value-decomposition network, as in standard MAD-DPG agents [16], we employ an individual *local critic* for each agent and an additional shared *global critic* for all agents to learn local and global rewards separately. More specifically, each DDPG agent in our IRC method is equipped with an actor and a local critic, and has access to a shared global critic. The local critic is customized for each agent, which learns the Q-function associated with the local reward R_i; the global critic is shared by all agents, which learns the Q-function associated with the global reward R_g. The actor generates actions that maximizes the weighed sum of the local critic output and the global critic output. In this way, the highly unstable environment faced by each agent is decoupled into a stable local environment learned by the agent itself and a stable global environment learned by all the agents from a global view.

4 Experiment and Evaluations

In this section, we present the experimental results in a representative setting, in order to reveal the patterns of collaboration behavior among regions with different types under different collaboration levels, based on the optimal policies learned by our IRC method. We first describe the setup and evaluation metrics used in the experiments, and then present the results and our interpretations.

4.1 Experimental Settings

The experiment focuses on the early "outbreak" stage of pandemic transmission, where most regions contain few infected people, but one specific *source region* (labelled as region #0) is suffering from an already out-of-control pandemic outbreak ($R_0 > 1$), and infections will spread from the source region to other regions. This stage is critical to successful pandemic control. Since we care more about preventing the vast pandemic spread among all the regions than controlling infections in the source region (which is configured to be impossible, as $R_0 > 1$ in region #0), we specially exclude the pandemic penalty from region #0's local reward.

Environment and region types. With the main purpose of revealing the influence of tolerance levels and collaboration level on the outcome, the artificial setting contains only 5 regions. Each region has an initial population of 10,000,000 people, a daily traffic of 5,000 people for each travelling route, and 2,000 initial infections in source region #0. Apart from the source region, other regions (#1 through #4) are designed to demonstrate different strategies learned by agents of different region types. Here we assume the tolerance levels are set to be either "high" (H_+ or L_+) or "low" (H_- or L_-), where $H_+ > H_-$ and $L_+ > L_-$. It is evident that 4 non-source regions are sufficient to cover all possible region types, namely $(H_+, L_+), (H_+, L_-), (H_-, L_+), (H_-, L_-)$, respectively.

We point out that region types are abstract hyperparameters specific to our pandemic model, so that it is their relative relationship rather than absolute values that matter in our illustrative experiment, which justifies the two-level setting. For the sake of implementation, we assign the specific values as $H_+ = 0.003$, $H_- = 0.001$, $L_+ = 72$ and $L_- = 24$. The source region #0 is special in that its pandemic situation is not concerned, so we assign to it a lockdown-tolerance level $L_{0,0} = 0.05$.

Collaboration levels. To investigate agents' behavioral patterns under different collaboration levels, our IRC models are trained with 3 different reward mixing ratios $\alpha \in \{0.01, 0.40, 10.0\}$. For each

configuration, the training procedure is repeated for 10 times with different random seeds, and the model that yields the highest reward is selected for further evaluation and analysis.

4.2 Evaluation Protocols

Baselines. Our MARL models are compared against the following two baselines (or conceptually "expert policies"), which are abstracted from common and most easily implementable practices.

- *Fixed policy.* Each region allows a fixed proportion p_{fix} of all incoming traffic for entrance. The proportion is global in the sense that it works for all regions and throughout the pandemic period. This abstracts typical control strategies like permanently restricting international flights.

- *Threshold policy.* This is a more flexible strategy that dynamically "respond to" the changing pandemic situation, where lockdown is enforced when there are many observed (i.e., hospitalized) patients, indicating that the pandemic situation is worsening, and the cumulative mobility loss is still acceptable, indicating that the lockdown will not hurt economy too seriously. More specifically, the policy sets each action $p_{i,j}^t$ by the rule

$$p_{i,j}^t = \begin{cases} 0 & H_i^t > H_{\text{th}} \text{ and } L_i^t < L_{\text{th}} \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

Metrics. Several metrics reflecting the effectiveness of pandemic control and the extent of collaboration are used to quantitatively evaluate the performance of our models against the baselines:

- *Mean global reward* $\overline{R}_g$. This is a direct revelation of the performance that the model is trained to optimize. Here we only need the mean of global reward, since the local rewards of all agents sum to the global reward.

- *The mean* $\overline{H}$ *and maximum* H_{max} *number of total hospitalized patients.* These two metrics represent the pressure on local medical services brought by the pandemic control policy, which reflect the feasibility of the pandemic control policy in medical sense.

- *Mean action* $\overline{p}$. Mean action is defined as the ratio of total actual mobility over total mobility demand into a region. It represents the strictness of inter-regional lockdown enforced by a region, and acts as an indicator for the economic loss induced by the pandemic control policy.

- *Type-wise analysis.* To analyze the collaboration behavior of agents with different regional types, the mean hospitalization $\overline{H}$ and mean mobility $\overline{p}$ within regions of the same type are also calculated, and presented in the form of 2-by-2 matrices and/or radar plots.

4.3 Experimental Results

The results of baseline polices and our IRC models are shown in Table 1. It is clear that, regardless of the mixing ratio α, our IRC models outperform both baselines by receiving larger global rewards, higher mobility and lower mean hospitalization rate. Therefore, our IRC models achieve better performance at balancing the pandemic cost and lockdown cost of each region, and at balancing the demand of different regions.

Table 1: The performance of different models.

model	parameter	$\overline{R}_{\mathbf{g}}$	$\overline{H}$	$H_{\mathbf{max}}$	$\overline{p}$
Fixed	$p_{\text{fix}} = 0.5$	64.16	88.29	276.68	0.5000
Threshold	$H_{\text{th}} = 1, L_{\text{th}} = 168$	88.82	92.65	202.17	0.8198
IRC	$\alpha = 0.01$	**103.89**	59.46	373.25	0.7753
IRC	$\alpha = 0.40$	**106.97**	31.29	327.90	0.7090
IRC	$\alpha = 10.0$	**108.13**	27.21	232.73	0.6909

Despite the simplicity of the setting, this environment clearly illustrates different behavior patterns of agents of different regional types. For agents trained under different mixing ratios, the type-wise analysis of mean actions and mean hospitalization rates are shown in Figure 2 and 3. It is observed that, when agents make decisions based largely on local rewards (i.e., in the case of $\alpha = 0.01$),
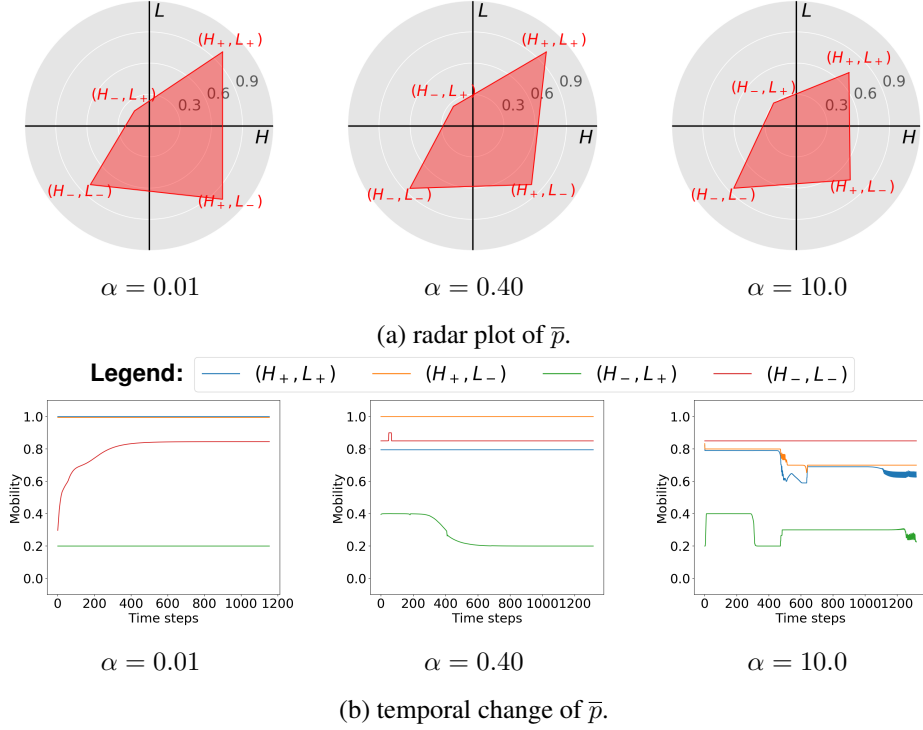
(a) radar plot of $\bar{p}$.

Legend: —— (H_+, L_+) —— (H_+, L_-) —— (H_-, L_+) —— (H_-, L_-)

(b) temporal change of $\bar{p}$.

Figure 2: Type-wise analysis of mean action $\bar{p}$ for different mixing ratios.

they tend to select actions that best serve their own interests — the (H_-, L_+)-region is extremely vulnerable to outbursts of infections, so its optimal policy is to enforce strict lockdown policies; the (H_+, L_-)-region cannot afford long-term lockdowns, so it prefers to largely open up inward traffic; the (H_-, L_-)-region suffers from both a pandemic outbreak and a large cumulative mobility loss, so it strictly blocks all inward traffic initially, but is forced to open up later to avoid cumulative lockdown penalty. This is clearly shown in the radar plot (Figure 2(a)), where the colored area (representing mobility) leans towards (H_+, L_-)-direction, indicating that regions that can afford more patients but not strict lockdowns are more likely to open up.

The above behavioral patterns are similar to what we have observed in the very early stage of COVID-19 outbreak, where regional governments have not established unified collaboration protocols. The (H_-, L_+)-region corresponds to those regions in reality that have limited medical resources and relatively self-sufficient economies, so that they are willing to avoid local panic for medical resources at the cost of economic losses. The (H_+, L_-)-region represents those regions with export-oriented economies that rely heavily on inter-regional population and commodity mobility, so they will not voluntarily enforce lockdown policies, but would rather maintain economic vitality even at the cost of increasing number of infections. The (H_-, L_-)-region, on the other hand, is a representative of under-developed regions that are the most vulnerable to a pandemic outbreak — without help from other regions, they will soon fall into the dilemma that neither open-up nor lockdown is favorable.

However, as the global reward gains more weight in agents' optimization objectives, their behavior gradually changes. The agents with higher lockdown-tolerance level (L_+) tend to share the lockdown burden with those agents who have lower lockdown-tolerance level (L_-), as they enforce stricter lockdown from the source region. Such changes will not be observed if each agent focuses on its own interest only, since for those H_+-regions that voluntarily share the burden, slightly larger infection numbers do not hurt them as badly as the extra lockdown costs; however, if they choose not to cooperate, the most vulnerable (H_-, L_-)-region has to block more traffic from H_+-regions to avoid imported infections, which vastly increases the lockdown cost of the (H_-, L_-)-region or even leads to oscillations in action. Therefore, enforcing extra lockdown by those H_+-regions against the source city helps more vulnerable cities to suffer less from pandemic outbreaks and/or cumulative mobility losses, which is indeed a kind of altruistic collaboration that promotes overall welfare and

realizes better inter-regional pandemic control. The trend of changing behavior is illustrated in Figure 3, and is also clear from the radar plot (Figure 2(a)), as the colored area becomes more balanced among different types of regions.

The influence of the collaborative behavior of our IRC agents can also be reflected by the mean hospitalization matrix, as illustrated in Figure 3. It is observed that, as the mixing ratio α increases, the number of hospitalized people in the (H_-, L_-)-region gradually decreases even when its mobility increases at the same time. Meanwhile, the majority of pandemic penalties are still undertaken by H_+-regions (since they still allow most inward mobility), which have more medical resources and are thus more resilient to such mild increases in infections. Therefore, our IRC agents indeed learn to balance the demand of regions with different regional types, and such balanced collaboration behavior is helpful to promote overall welfare (indicated by increased $\overline{R}_g$ and lower $\overline{H}$).
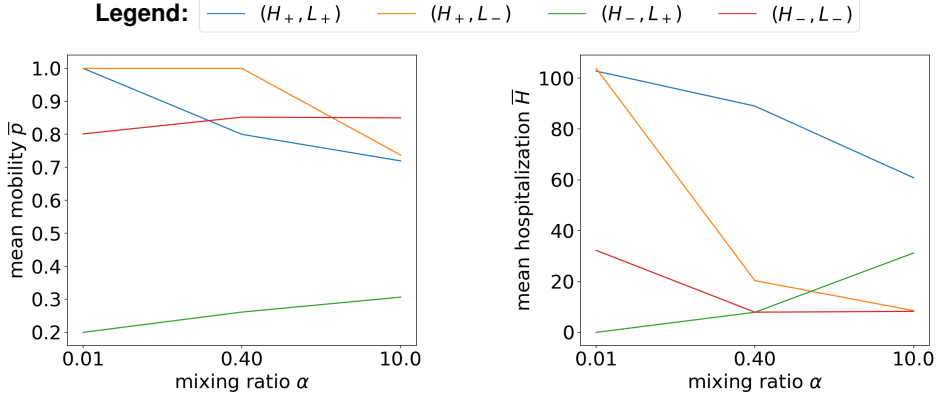

Figure 3: Type-wise trend of change for $\overline{p}$ and $\overline{H}$ when mixing ratio increases.

As a short summary, our IRC models successfully learn different policies for different mixing ratios, and the learned behavior shows an interpretable changing pattern of collaboration behavior — regions with abundant medical resources and high resilience for pandemic outbreak will voluntarily enforce stricter lockdown policies to help those vulnerable regions at higher collaboration levels. It is also worth mentioning that, as the mixing ratio α increases, our IRC models are more likely to find policies that yield higher global rewards, which demonstrates the benefit of collaboration, and confirms the mixing ratio to be a good adjustable parameter representing the collaboration level.

5 Conclusions and Future Work

In this paper, we analyze collaborative patterns of multi-region pandemic control behavior under different collaboration levels, using the optimal policies found by our specially-designed MARL method. We introduce a reward mixing ratio to abstract the inter-regional collaboration level that can be adjusted to reveal agents' different behavior difference under different collaboration levels. Experimental results in an exemplary environment show that, as the global reward accounts for a larger portion of each agent's objective, the actions and rewards of regions of different types become more balanced, and agents tend to learn policies that not merely promote their own interests, but also collaborate to help reduce the loss of other regions. One major take-away message here is that higher collaboration level will lead to better and more balanced pandemic control policies, in that regions will voluntarily help to coordinate the demands of different regions, and regions more resilient to lockdown tend to sacrifice a mild portion of its own utility to protect more vulnerable regions.

Our framework provides a novel computational perspective for understanding and promoting collaboration for the supply of public goods. As a representative type of them, pandemic control cannot be supplied with a few regions or through a few isolated policies, while different attitudes towards inter-regional collaboration enables different pandemic control policies, and will possibly lead to different outcomes. This reminds us that, to better cope with global catastrophes like COVID-19 and more efficiently provide public goods to promote global welfare, international collaboration should be encouraged and all related parties should be called for to actively assume international responsibilities, so that the strong will protect the weak by sharing the losses. The legitimacy of

such collaborative policies that restrict individual liberty can be justified by the optimality guarantee of the computational methods. We believe this framework can be generalized to analyze the supply of other public goods that involves collaboration.

The model of this paper is limited in that it regards inter-regional mobility control as the only form of action. However, regions' within-regional pandemic control strategies will also induce epidemiological and economic costs, which also potentially influences inter-region collaboration. For example, a destination region may be more willing to accept inward mobility if the source region enforces better control within itself. Therefore, a more comprehensive modelling is left as future work.

References

[1] James M. Buchanan and Tullock Gordon. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. Ann Arbor: University of Michigan Press, 1965.

[2] Fukunari Kimura, Shandre Mugan Thangavelu, Dionisius Narjoko, and Christopher Findlay. Pandemic (covid-19) policy, regional cooperation and the emerging global production networks. *Asian Economic Journal*, 34(1):3–27, 2020.

[3] Gordon Brown and Daniel Susskind. International cooperation during the COVID-19 pandemic. *Oxford Review of Economic Policy*, 36(Supplement-1):S64–S76, 09 2020.

[4] C. Tsay, F. Lejarza, M. A. Stadtherr, and M. Baldea. Modeling, state estimation, and optimal control for the us covid-19 outbreak. *Scientific reports*, 10(1):1–12, 2020.

[5] S. Song, Z. Zong, Y. Li, X. Liu, and Y. Yu. Reinforced epidemic control: Saving both lives and economy. *arXiv preprint arXiv:2008.01257*, 2020.

[6] H. Khadilkar, T. Ganu, and D. P. Seetharam. Optimising lockdown policies for epidemic control using reinforcement learning. *Transactions of the Indian National Academy of Engineering*, 5(2):129–132, 2020.

[7] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[8] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[9] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[10] L. Buşoniu, R. Babuška, and B. de Schutter. Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221, 2010.

[11] OpenAI. Openai five. `https://blog.openai.com/openai-five/`, 2018.

[12] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

[13] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087, 2018.

[14] T. Rashid, M. Samvelyan, C. S. Witt, G. Farquhar, J. Foerster, and S. Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4292–4301, 2018.

[15] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5887–5896, 2019.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatchr. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.

[17] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] S. Iqbal and F. Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 2961–2970, 2019.

[19] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4): e0172395, 2017.

[20] J. K. Gupta, M. Egorov, and M. Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.

[21] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[22] M. Y. Li and J. S. Muldowney. Global stability for the seir model in epidemiology. *Mathematical biosciences*, 125(2):155–164, 1995.

[23] C. C. McCluskey. Complete global stability for an sir epidemic model with delaydistributed or discrete. *Nonlinear Analysis: Real World Applications*, 11(1):55–59, 2010.

[24] P. Ogren and C. F. Martin. Optimal vaccination strategies for the control of epidemics in highly mobile populations. In *Proceedings of the 39th IEEE Conference on Decision and Control (Cat. No. 00CH37187)*, volume 2, pages 1782–1787. IEEE, 2000.

[25] L. Sattenspiel and K. Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical biosciences*, 128(1-2):71–91, 1995.

[26] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. Du Plessis, N. R. Faria, R. Li, W. P. Hanage, et al. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497, 2020.

[27] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner. Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254, 2020.

[28] Suborna Barua. Understanding coronanomics: The economic implications of the coronavirus (covid-19) pandemic. MPRA Paper 99693, University Library of Munich, Germany, April 2020. URL `https://ideas.repec.org/p/pra/mprapa/99693.html`.

[29] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

A Environmental Modelling

A.1 Mobility

At each time step t, the mobility demands of all n regions are represented by a matrix $M_{\mathrm{d}}^t \in \mathbb{R}^{n \times n}$, where $M_{\mathrm{d},i,j}^t$ denotes the mobility demand from region i to region j.

All actions at time step t are arranged as a matrix $p^t \in [0,1]^{n \times n}$, where $p_{i,j}^t$ denotes the proportion of mobility demand that can be fulfilled from region i to region j, which is decided by the destination region j. Therefore, region j is allowed to determine the jth column $p_j \in [0,1]^n$ in matrix p.

Eventually, the actual allowed mobility M_{a}^t at time step t is calculated by $M_{\mathrm{a},i,j}^t = M_{\mathrm{d},i,j}^t p_{i,j}^t$.

A.2 Pandemic Transmission

The pandemic transmission model is based on the Susceptible-Infected-Hospitalized-Recovered (SIHR) model proposed in [5]. The *pandemic state* of region i at time step t is $E_i^t = (S_i^t, I_i^t, H_i^t, R_i^t)$, which consists of the number of susceptible, infected, hospitalized, and recovered population within that region, respectively. Agents are able to observe the number of hospitalization and recovery, but cannot distinguish between susceptible (healthy) people and infected people that are both asymptomatic. Therefore, the *visible pandemic state* of region i at time step t is $E_{\mathrm{v},i}^t = (S_i^t + I_i^t, H_i^t, R_i^t)$.

At each step t, pandemic transmission is calculated in two phases: mobility happens first, and then the pandemic spreads within each region.

At the mobility happening stage, people move between regions according to actual allowed mobility M_{a}^t, calculated in the previous section. After the mobility happening stage, the intermediate state becomes $\hat{E}_i^t = (\hat{S}_i^t, \hat{I}_i^t, \hat{R}_i^t)$, which is given by

$$E_{\mathrm{s},i}^t = E_i^t - \sum_j \frac{M_{\mathrm{a},i,j}^t}{N_i^t} E_i^t, \quad E_{\mathrm{m},i}^t = \sum_j \frac{M_{\mathrm{a},j,i}^t}{N_j^t} E_j^t, \quad \hat{E}_i^t = E_{\mathrm{s},i}^t + E_{\mathrm{m},i}^t. \tag{5}$$

Here N_i^t is the population of region i at time step p, $E_{\mathrm{m},i}^t$ stands for the pandemic state of the moving people, and $E_{\mathrm{s},i}^t$ stands for the pandemic state of the staying people.

At the pandemic spreading stage, the eventual pandemic state is calculated by

$$
\begin{aligned}
S_i^{t+1} &= \hat{S}_i^t - \frac{\beta_{\mathrm{s},i}^t S_{\mathrm{s},i}^t I_{\mathrm{s},i}^t}{N_{\mathrm{s},i}^t} - \frac{\beta_{\mathrm{m},i}^t S_{\mathrm{m},i}^t I_{\mathrm{m},i}^t}{N_{\mathrm{m},i}^t}, \\
I_i^{t+1} &= \hat{I}_i^t + \frac{\beta_{\mathrm{s},i}^t S_{\mathrm{s},i}^t I_{\mathrm{s},i}^t}{N_{\mathrm{s},i}^t} + \frac{\beta_{\mathrm{m},i}^t S_{\mathrm{m},i}^t I_{\mathrm{m},i}^t}{N_{\mathrm{m},i}^t} - \gamma \hat{I}_i^t, \\
H_i^{t+1} &= H_i^t + \gamma_i^t \hat{I}_i^t - \theta_i^t H_i^t, \\
R_i^{t+1} &= \hat{R}_i^t + \theta_i^t H_i^t,
\end{aligned}
\tag{6}
$$

where β_i^t is the transmission rate of the susceptible people, γ_i^t is the hospitalization rate of the infected people, and θ_i^t is the recovery rate of the hospitalized people.

A.3 Tolerance Levels

The local reward of an agent's action on that region is a combination of the pandemic-spread cost and the mobility-control cost. The pandemic-spread cost $C_{\mathrm{p},i}$ for agent i is given by

$$C_{\mathrm{p},i} = k_{\mathrm{h}} \exp\left(\frac{H_i}{H_{0,i}}\right), \tag{7}$$

where k_{h} is the hyperparameter that abstracts the cost induced by the first observed unit of infection within the region, while $H_{0,i}$ is a hyperparameter called *pandemic-tolerance level* that determines the exponentially increasing rate of the cost as the number of hospitalized patients increases. The

mobility-control cost $C_{m,i}$ of region i is given by

$$M_{d,i}^t = \sum_j M_{d,j,i}^t , \quad M_{a,i}^t = \sum_j M_{a,j,i}^t ,$$

$$L_i^t = \sum_{\tau=0}^{t-1} \lambda^{t-\tau} \frac{M_{d,i}^{t-\tau} - M_{a,i}^{t-\tau}}{\bar{M}_{d,i}} , \tag{8}$$

$$C_{m,i}^t = \exp\left(\frac{L_i^t}{L_{0,i}}\right) \frac{M_{d,i}^{t-\tau} - M_{a,i}^{t-\tau}}{\bar{M}_{d,i}} .$$

Here $M_{d,i}^t$ is the total mobility demand into region i, while $M_{a,i}^t$ is the actual amount of mobility that is allowed to enter region i at time step t; $L_{0,i}^t$ is the cumulative penalty caused by continuous mobility restriction of the same region, where λ is the temporal discount factor of historical penalties; L_0 is the hyperparameter called *lockdown-tolerance level* that determines the exponentially increasing rate of the cost as the allowed amount of mobility decreases.

B Detailed Experimental Results

Detailed experimental results are displayed below in Table 2.

Table 2: Type-wise behavior patterns for agents of different types.

(a) Type-wise comprison table of mean action $\bar{p}$ for different mixing ratios.

$\bar{p}$	L_+	L_-
H_+	1.000	1.000
H_-	0.200	0.801

$\alpha = 0.01$

$\bar{p}$	L_+	L_-
H_+	0.800	1.000
H_-	0.261	0.852

$\alpha = 0.40$

$\bar{p}$	L_+	L_-
H_+	0.720	0.737
H_-	0.307	0.850

$\alpha = 10.0$

(b) Type-wise comprison table of mean hospitalization $\overline{H}$ for different mixing ratios.

$\overline{H}$	L_+	L_-
H_+	102.76	103.84
H_-	0.00	32.25

$\alpha = 0.01$

$\overline{H}$	L_+	L_-
H_+	89.00	20.35
H_-	7.89	7.93

$\alpha = 0.40$

$\overline{H}$	L_+	L_-
H_+	60.80	8.54
H_-	31.22	8.26

$\alpha = 10.0$