Hierarchical Reinforcement Learning for Scarce Medical Resource Allocation with Imperfect Information

Qianyue Hao[1], Fengli Xu[1]*, Lin Chen[2], Pan Hui[2], Yong Li[1]
[1]Beijing National Research Center for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing, China.
[2]Dept. of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China.
liyong07@tsinghua.edu.cn

ABSTRACT

Facing the outbreak of COVID-19, shortage in medical resources becomes increasingly outstanding. Therefore, efficient strategies for medical resource allocation are urgently called for. Reinforcement learning (RL) is powerful for decision making, but three key challenges exist in solving this problem via RL: (1) complex situation and countless choices for decision making in the real world; (2) only imperfect information are available due to the latency of pandemic spreading; (3) limitations on conducting experiments in real world since we cannot set pandemic outbreaks arbitrarily. In this paper, we propose a hierarchical reinforcement learning method with a corresponding training algorithm. We design a decomposed action space to deal with the countless choices to ensure efficient and real time strategies. We also design a recurrent neural network based framework to utilize the imperfect information obtained from the environment. We build a pandemic spreading simulator based on real world data, serving as the experimental platform. We conduct extensive experiments and the results show that our method outperforms all the baselines, which reduces infections and deaths by 14.25% on average.

CCS CONCEPTS

• Information systems → Data mining; **• Computing methodologies → Reinforcement learning**; **• Applied computing → Life and medical sciences**.

KEYWORDS

Hierarchical reinforcement learning; medical resource allocation; imperfect information; COVID-19 pandemic.

ACM Reference Format:
Qianyue Hao[1], Fengli Xu[1]*, Lin Chen[2], Pan Hui[2], Yong Li[1] . 2021. Hierarchical Reinforcement Learning for Scarce Medical Resource Allocation with Imperfect Information. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467181

1 INTRODUCTION

In face of a suddenly outbreak pandemic, such as COVID-19, the globe tends to suffer from severe shortage of medical resources [22]. Key equipment, including ventilators, is needed to care for critically ill patients while personal protective equipment (PPE) are necessary for both medical staffs and the public. Under such circumstance, efficient strategies for allocating the scarce medical resources are of great importance to maximize the protection of the public health and to minimize the damage caused by the pandemic. In our work, we consider two critical kinds of medical resources, i.e., hospital beds and surgical masks.

The problem of medical resource allocation has long been studied by public health researchers. Strategies based on individual characteristics such as age, occupation [4, 20] or symptom seriousness [6] have been proposed and adopted. Other guidelines including saving the most lives or saving the most life years [4, 20] clarify the ultimate objective, but it is nontrivial to turn such high-level guidances into practical strategies. On another aspect, machine learning for pandemic interventions have also long been studied, including works on efficient lockdown strategies [14, 19], etc. However, methods that focus on the problem of medical resource allocation with AI, especially with reinforcement learning, stay almost unexplored.

Despite the importance and necessity of efficient strategies on scarce medical resource allocation, there exist several key challenges in solving such problems via reinforcement learning. (1) Medical resource allocation in real world scenarios means countless choices for decision making, i.e., an overlarge action space, leading to difficulties in designing and training the RL agent. (2) Due to the latency of pandemic spreading, we cannot obtain a precise description of the overall pandemic spreading situation and only imperfect information is available [25], which adds to difficulties in decision making. (3) Since we cannot set pandemic outbreaks in the real world arbitrarily, it is impossible to train and test the RL agent during real world pandemic spreading processes, leading to limitations on conducting experiments.

In this paper, we propose a hierarchical reinforcement learning method to solve this problem in view of the above challenges. To tackle the difficulty of the overlarge action space, we design a decomposed action space for the RL agent, which is able to generate efficient strategies in complex situations on the real world scale. Meanwhile, we design an according training algorithm, ensuring an efficient training process. In order to solve the difficulty caused by imperfect information, we design a framework based on gated recurrent units to utilize the imperfect information. By rebuilding intact information of the pandemic spreading situation from the imperfect one obtained from the environment, it provides a basis

∗ Corresponding author.

KDD '21, August 14–18, 2021, Virtual Event, Singapore
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467181

for the RL agent to make decisions, which contributes to the performance greatly. We also build a pandemic spreading simulator based on the disease model and real world data, which turns out to simulate the real world situations precisely and serves well as our experimental platform. We conduct extensive experiments and the results clearly exhibit the outstanding performance of our method.

Our main contributions can be summarized as follows:

- We propose a hierarchical reinforcement learning method with a decomposed action space, as well as a corresponding training algorithm. Due to our optimized design, the training process achieves high efficacy while efficient and real time strategies are available on the real world scale.
- We propose an imperfect information utilization framework, and thus achieve good performance under imperfect information situation, which is close to the real world.
- We design a pandemic spreading simulator based on real world data and disease model. We justify its accuracy comparing the real world situation, based on which we use the simulator as the experimental platform. We conduct extensive experiments in two cities under various settings, where the results indicate that our method outperforms all the baselines by 14.25% on average.

2 PRELIMINARIES

2.1 Disease Modeling

In this paper, we focus on the widely spreading COVID-19 pandemic, which brings about public health crises. We use a disease model with 8 states to describe the development of such disease in extension of the typical SEIR model [15]. S is used to denote susceptible while E is used for exposed. The state infected (I) is divided into 4 states for better description of the disease. I_u refers to those who are infected but not tested, able to move around without any constraints. I_t refers to those who are infected, tested and diagnosed with COVID-19, whose movements are constrained inside one region. Since the widely applied nucleic acid testing for COVID-19 infection cannot reach 100% accuracy, we specially design the state I_a to denote those who are infected but turn out to be false negative in nucleic acid testing. And finally I_h refers to those who are adopted to the hospital. The state R (removed) in SEIR is divided into recovered (R) and dead (D), to better describe the death caused by such pandemic.

All possible state transitions are described in Figure 1 with the respective transition probabilities. The parameter β is the average number of new E cases caused by one existing I individual per unit time, through infected-susceptible ($I-S$) contacts. Thus it measures the infectiousness of the pandemic. For a exposed-susceptible ($E-S$) contact, the infectiousness is weighted by r_E. The parameter ϵ describes the probability for an S individual to become I per unit time. The parameters α and μ describe the $I \longrightarrow D$, $I \longrightarrow R$ probabilities per unit time.

2.2 Medical Resources

We consider two major kinds of critical medical resources in fighting against COVID-19. The first kind is hospital beds (staffed with medical workers and equipped with medical facilities), which are necessary for treating the infected people and saving the lives of those who are critically ill [4]. The second kind is surgical masks, which play an indispensable role in protecting the healthy people from infection. We study the pandemic spreading in a city with a total population of N, divided into M regions according to main road networks and the population size in region i is denoted as $N^{(i)}, i = 1, 2, ...M$. We adopt the following indices to describe the attributes of medical resources in the city:

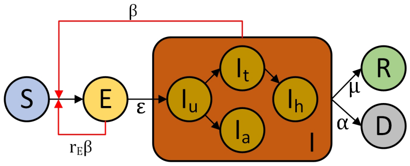


Figure 1: All possible state transitions (disease model).

- **Total Bed Number** N_B: The total number of hospital beds that the medical system of the city can provide.
- **Available Bed Number** $N_B(t)$: The number of available hospital beds at time t.
- **Total Mask Number** N_M: We assume that each person consumes one surgical mask per day and N_M is the number of daily supply of surgical masks in the city.
- **Remaining Mask Number** $N_M(t)$: The number of remaining surgical masks at time t.
- **Mask Efficacy Factor** γ: Considering surgical masks cannot filter the virus absolutely, γ equals to the percentage of virus the surgical masks can filter efficiently.

The raw infectious rate of the disease is denoted as β_0. Medical research has shown that infectious rate reduces almost linearly as the product of mask coverage percentage (denoted as $\pi(t)$ at time t) and mask efficacy factor goes up [3]. Thus the exact infectious rate in region $i, i = 1, 2, ...M$ at time t, denoted as $\beta^{(i)}(t)$, is:

$$\beta^{(i)}(t) = \beta_0(1 - \pi(t)\gamma). \quad (1)$$

2.3 Problem Formulation

We consider the situation when medical resources are in severe shortage, i.e., time t when $N_M(t)$ is less than the number of people need masks and $N_B(t)$ is less than the number of patients waiting for hospital adoptions. We need to find out efficient and real time strategies to allocate the currently available medical resources among the M regions according to the pandemic spreading situation, minimizing the damage caused by the pandemic.

The damage caused by the pandemic after a spreading process lasting T is measured quantitatively by the following four metrics:

- **Infected Cases**: Total number of people who have ever been infected (recovered, dead or still under infection) until T.
- **Dead Cases**: Total number of deaths until T.
- **Max Daily Infection**: The maximum number of newly infected people in 24 hours during the spreading process.
- **Max Daily Death**: The maximum number of newly died people in 24 hours occurred during the spreading process.

Since the former two measure how serious the pandemic is while the latter two measure how fast the pandemic is spreading, to minimize the damage caused by the pandemic is to minimize the above metrics, obviously. Therefore, the formal definition of the research problem is given as follow:

Problem 1 (Medical Resources Allocation). *At time t, in the city with M regions and population size $N^{(i)}, i = 1, 2, ...M$, given available beds number $N_B(t)$, remaining masks number $N_M(t)$ and the current pandemic spreading situation, with the goal of minimizing the damage caused by the pandemic, find out the number of hospital beds and surgical masks allocated to each region $\{N_B^{(i)}(t)\}$ and $\{N_M^{(i)}(t)\}$, ensuring $\sum_i N_B^{(i)}(t) \leq N_B(t)$, $\sum_i N_M^{(i)}(t) \leq N_M(t)$.*

Using the expressions above, we can calculate the exact infectious rate in region i at time t as follow:

$$\beta^{(i)}(t) = \beta_0(1 - \pi(t)\gamma) = \beta_0(1 - \frac{N_M^{(i)}(t)}{N^{(i)}}\gamma). \tag{2}$$

3 METHOD

3.1 System Overview

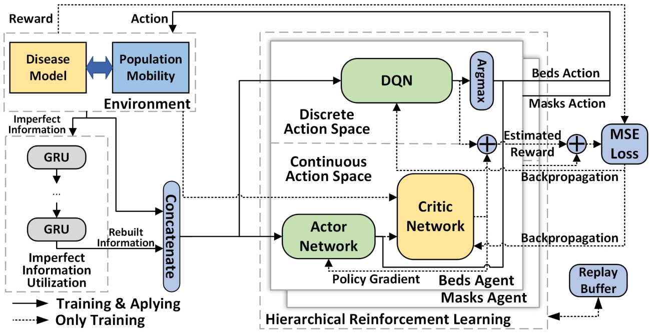


Figure 2: Overview of the hierarchical reinforcement learning framework with imperfect information.

To obtain real time strategy for medical resource allocation, we propose a hierarchical reinforcement learning system with imperfect information utilization design and the structure overview is shown in Figure 2. The system mainly consists of two parts:

- **Imperfect Information Utilization**: Rebuilding intact information from the imperfect one from the environment.
- **Hierarchical Reinforcement Learning**: Finding out the real time strategies according to rebuilt intact information.

Detailed information about these two parts can be found in the following two sections, and a specially designed training algorithm is shown in Section 3.4. Our method is built based on PyTorch and the source code is available at https://github.com/KYHKL-Q/Hierarchical-RL.

3.2 Imperfect Information Utilization

Firstly, we give a definition of pandemic spreading information:

Definition 1 (Pandemic Spreading Information). *In a city with M regions, the pandemic spreading information of state-X, $X \in \{S, E, I_u, I_t, I_a, I_h, R, D\}$ is a vector with the length of M, denoted as $\theta^{[X]} \in \mathbb{N}^M$, (e.g., $\theta^{[I_u]}$). Element i, $\theta_i^{[X]}, i = 1, 2, ...M$, is the number of people in region i who are in state-X.*

In the real world, it is almost impossible for us to obtain the pandemic spreading information of all the 8 states in the disease model. In other words, we are usually in the situation of imperfect information. According to the actual situation, we can obtain the pandemic spreading information of state-I_t and I_h for sure. The number of deaths (D) caused by the pandemic is also accessible. However, the pandemic spreading information of state-S, E and I_u are not accessible due to the latency in pandemic spreading. Mentioning that the pandemic spreading information of state-R is also difficult to get because some people suffer from mild symptoms and recover at home without being tested or adopted to the hospital.

In the real world pandemic spreading process, people who are not infected, i.e., in state-S, stay the majority, while those who are infected but turn out to be false negative in nucleic acid tests only account for a very small percentage. Thus, the pandemic spreading information of state-E, I_u, I_t, I_h, R and D has a wider range of variation and can better reflect the pandemic spreading situation, playing a more important role in aiding the RL decision making. Therefore, the aim of imperfect information utilization is to infer the unknown pandemic spreading information of state-E, I_u and R according to the obtained ones of state-I_t, I_h and D.

We design a gated recurrent unit (GRU) [2] based recurrent neural network (RNN) for information rebuilding, utilizing its strong ability in time sequence modeling. The detailed structure is shown in Figure 3. The GRU-RNN framework consists of three branches, corresponding to the information of three states to be rebuilt. Each branch contains layers consists of 1024 GRUs using Tanh as the activation function. We design a 1024×M (M is the number of regions in the studied city) full connection layer at the end of each branch, projecting the hidden states in the RNN to the final outputs. Since the inputs and outputs are number of people in certain disease processing states, whose value is relatively large considering the mathematical features of Tanh, we design 1:1000 rescale layers right after the inputs and before the outputs, adjusting the values to a proper range. We apply backward propagation through time (BPTT) [27] and Adam [13] optimizer for training, receiving a good convergence.

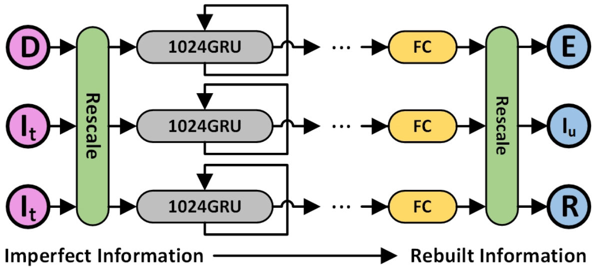


Figure 3: Detailed structure of the GRU-RNN based imperfect information utilization framework.

3.3 Hierarchical Reinforcement Learning

3.3.1 *Action Space Decomposition.* A great challenge in solving the problem of medical resource allocation with typical RL methods is the overlarge action space faced in the real world situation. There are usually hundreds of regions in a city and various kinds of medical resources to be allocated (the number of kinds of resources is denoted as K). Therefore, the output strategy is supposed to be a matrix with the shape of $M \times K$, which lays in a rather high dimensional space, leading to a very large action space. Also, there exists constraint between the elements: the total available amount of each kind of medical resources is fixed, which adds to the difficulty for the RL agent to make proper decisions.

To solve this challenge, we design a decomposed action space for decision making, rather than giving the matrix described above directly. For each kind of medical resource, we firstly give a significance rank of all the M regions. We design several ranking principles, including ranking according to the number of infections, population density, strength of population flows in the regions as well as their combinations. The first step of decision making is to choose a ranking principle, which lays in a discrete action space, and to give the significance rank. The second step is to give a satisfaction factor, a float number between 0 and 1, denoted as f. In the process of medical resource allocation, we consider the regions by the order of the given significance rank. Using U_i to denote the amount of a certain kind of medical resource needed by region i, we allocate $f \times U_i$ of such kind of medical resource to it and turn to the next region, i.e., only satisfy f of its demand. When there is no more medical resource, the allocation process comes to an end. Therefore, the second part of decision making is to give the float number and thus it lays in a continuous action space.

As is described above, we decompose the decision making process into two steps and decompose the overlarge action space into a discrete one and a continuous one, which are relatively small, making it possible for the RL agent to make efficient decisions. Besides, there are two things worth mentioning:

- Both discrete and continuous action are independent among different kinds of medical resources, i.e., different kinds of medical resources can be allocated according to different significance rank and satisfaction factor.
- Both discrete and continuous action for each kind of medical resource vary over time, i.e., real time strategies.

These two things make sure that the actions have enough freedom and thus maintain countless possibilities, making it possible to achieve the aim of minimizing the damage caused by the pandemic.

3.3.2 *Reward Function.*

We use a specially designed reward function to measure the seriousness of the pandemic spreading situation in the RL training. We use s_i to denote the pandemic spreading situation after i time steps. $X[s_i], X \in \{S, E, I_u, I_t, I_a, I_h, R, D\}$ refers to the total number of people in all regions who are in state-X in the situation s_i. Then we define single step reward for time step i as $\mathcal{G}_i(s_{i-1}, s_i)$, as follows,

$$\mathcal{G}_i(s_{i-1}, s_i) = C + \log\frac{1}{2}\Big(\frac{1}{\max(E[s_i] - E[s_{i-1}], 0) + 1} + \frac{1}{\max(I_{ut}[s_i] + I_t[s_i] - I_{ut}[s_{i-1}] - I_t[s_{i-1}], 0) + 1}\Big), \quad (3)$$

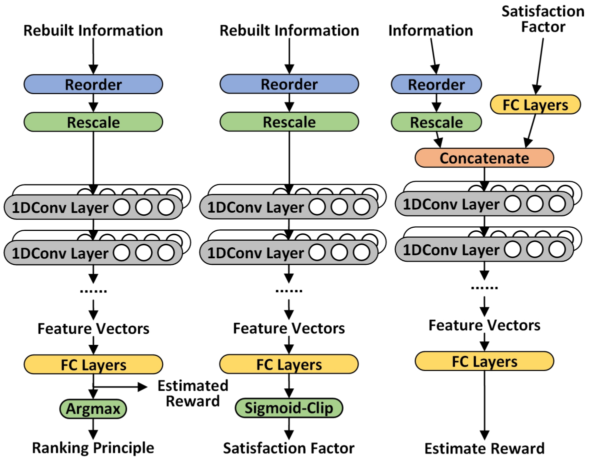
where C is a constant for bias. We further define long term reward of the action taken in time step i as $\mathcal{F}_i$, as follows,

$$\mathcal{F}_i = \mathcal{G}(s_{i-1}, s_i) + \kappa\mathcal{F}_{i+1} = \sum_{n=i}^{\infty} \kappa^{n-i}\mathcal{G}(s_{n-1}, s_n). \quad (4)$$

$\kappa \in (0, 1)$ is the factor weighting the future profit. In problems with finite time steps, the sum operation with infinite items can be cut off at a certain point.

3.3.3 *Network Structures.*

We apply one deep Q-Network (DQN) for each kind of medical resource in the discrete action space, while one pair of Actor-Critic networks for each kind of medical resource in the continuous action space. The detailed structures of these networks are shown in Figure 4, the shape of tensors varies with the number of regions in the studied city.



(a) Discrete DQN (b) Continuous actor network (c) Continuous critic network

Figure 4: Detailed structure of the neural networks designed in the hierarchical reinforcement learning.

All the networks use Leaky-ReLU as the activation function. The DQN takes in the rebuilt information of the pandemic spreading situation and gives the estimated long term reward for each ranking principles while the principle with the highest reward is selected to be in the real time strategy. The actor network in the continuous action space takes in the rebuilt information and gives the best satisfaction factor. To make sure the satisfaction factor is restricted between 0 and 1, we set a Sigmoid-Clip layer at the end of this network as follows,

$$F(x) = \max(\delta, \frac{1}{1 + e^{-x}}), \quad (5)$$

where δ is a manual lower bound of the satisfaction factor that can be set approximately equals to the total amount of medical resource divided by the population size in the city. The critic network in the continuous action space only acts in the training process. It takes in the information of the environment along with the satisfaction factor given by the actor network and outputs the estimated long term reward of the given satisfaction factor. Therefore, it supervises the training of the actor network through policy gradient.

Specially, we design reorder layers right after the input of pandemic spreading information, which exchanges the elements in the input vectors. Role of such layers is to maximize the relational degree among the adjacent elements and thus benefit the convolutional feature extractions. Relational degree is measured by the strength of population flow between the corresponding regions (defined in Section 4.1). We denote the M elements in the input vector as $V_i, i = 1, 2, ...M$ and use $C(i, j), i{\neq}j$ for the relational degree between the V_i and V_j. Thus the problem can be formulated as:

Problem 2 (Input Element Reordering). *Given the raw sequence $V_1, V_2, ...V_M$ and $C(i, j)$, find a reordered sequence $V_{r^1}, V_{r^2}, ...V_{r^M}$ where $r^j{\neq}r^k, j{\neq}k$, maximizing $\sum_{i=1}^{M-1} C(r^i, r^{i+1})$.*

It can be proved that the problem is equivalent to the traveling salesman problem (TSP), where there is no polynomial time algorithm to solve it precisely.

Proof. Considering a complete graph $G = (V, E)$ where the M regions corresponding to its M nodes. The length of edge connecting node i and j is $-(C(i, j) - L)$, where $L = \max C(i, j) + \epsilon, \forall i, j; \epsilon > 0$. There is an extra node O, connecting all the other nodes with the edge length L. Thus solving the input element reordering problem is to solve the TSP problem in G. □

We apply a stochastic method by searching 100 million possible reordered sequences ($M!$ possibilities in total) and selecting the best one. Although not precise, we obtain a solution that is good enough to satisfy the need of convolutional feature extractions.

3.4 Training Algorithm

Existing studies mainly focus on algorithms for training an RL agent either in a discrete action space or a continuous one. However, in our decomposed action space design, an algorithm to train an RL agent taking actions in both continuous and discrete action spaces simultaneously is needed, which stays unexplored. We propose a new algorithm by modifying the widely applied DQN training [18] and DDPG algorithm [17]. We replicate each network to yield a corresponding target network, and use soft replacement between the pair of networks during training. We also apply experience replay during training. It is worth mentioning that the algorithm can be used to train an RL agent with arbitrary D discrete actions and arbitrary C continuous actions. In our case, when considering two kinds of medical resources, we have $C = D = 2$.

We use $\mathcal{D}_i(s, a|\theta^{\mathcal{D}_i})$ to denote the DQN with weights $\theta^{\mathcal{D}_i}$, use $C_j(s|\theta^{C_j})$ to denote the continuous actor networks with weights θ^{C_j} and use $Q_j(s, a|\theta^{Q_j})$ to denote the continuous critic networks with weights θ^{Q_j}, where $i = 1, 2, ...D$, $j = 1, 2, ...C$. $\mathcal{D}'_i(s, a|\theta^{\mathcal{D}'_i})$, $C'_j(s|\theta^{C'_j})$ and $Q'_j(s, a|\theta^{Q'_j})$ refer to the corresponding target networks. We use $\{d_i^{(t)}\}$ to denote the discrete actions at time t, while use $\{c_j^{(t)}\}$ to denote the continuous ones. $s^{(t)}$ refers to the pandemic spreading information at time t and $r^{(t)}$ refers to the long term reward for actions at time t.

When updating the weights using a mini-batch with N samples from the experience replay buffer, denoted as

$$(s^{(n)}, \{d_i^{(n)}\}, \{c_j^{(n)}\}, r^{(n)}, s^{(n+1)}), n = 1, 2, ...N.$$

We calculate the loss L as follows:

$$y^{(n)} = r^{(n)} + \kappa(\sum_i \max_{d'} \mathcal{D}'_i(s^{(n+1)}, d'|\theta^{\mathcal{D}'_i}) + \sum_j Q'_j(s^{(n+1)}, C'_j(s^{(n+1)}|\theta^{C'_j})|\theta^{Q'_j})), \tag{6}$$

$$L = \frac{1}{N}\sum_n (y^{(n)} - \sum_i \mathcal{D}_i(s^{(n)}, d_i^{(n)}|\theta^{\mathcal{D}_i}) - \sum_j Q_j(s^{(n)}, c_j^{(n)}|\theta^{Q_j}))^2. \tag{7}$$

Algorithm 1 shows the outline of our training algorithm.

Algorithm 1 Hierarchical Reinforcement Learning Training

```
Input: Training episode $M$; reward function $\mathcal{F}(s, s')$.
Output: Trained discrete and continuous models.
Initialize $\mathcal{D}_i(s, a|\theta^{\mathcal{D}_i})$, $C_j(s|\theta^{C_j})$ and $Q_j(s, a|\theta^{Q_j})$ randomly.
Initialize $\mathcal{D}'_i(s, a|\theta^{\mathcal{D}'_i})$, $C'_j(s|\theta^{C'_j})$, $Q'_j(s, a|\theta^{Q'_j})$ by replicating from the origin networks.
Initialize experience reply buffer $\mathcal{R}$.
for $episode = 1, 2, ...M$ do
    Initialize a random process $\mathcal{N}$ for action exploration.
    Obtain initial observation of the environment $s^{(1)}$.
    for $t = 1, 2, ...T$ do
        Obtain $d_i^{(t)} = \max_{d_i} \mathcal{D}_i(s^{(t)}, d_i|\theta^{\mathcal{D}_i})$ with $\epsilon$-greedy and $c_j^{(t)} = C_j(s^{(t)}|\theta^{C_j}) + \mathcal{N}^{(t)}$.
        Execute $\{d_i^{(t)}\}$, $\{c_j^{(t)}\}$, obtain new state $s^{(t+1)}$ and reward $r^{(t)} = \mathcal{F}(s^{(t)}, s^{(t+1)})$.
        Store transition $(s^{(t)}, \{d_i^{(t)}\}, \{c_j^{(t)}\}, r^{(t)}, s^{(t+1)})$ into $\mathcal{R}$.
        Randomly sample $N$ transitions from $\mathcal{R}$, calculate $L$.
        Update $\theta^{\mathcal{D}_i}$, $\theta^{Q_j}$ minimizing $L$, update $\theta^{C_j}$ with gradient:
            $-\nabla_{\theta^{C_j}}(\frac{1}{N}\sum_n \sum_j Q_j(s^n, C_j(s^n|\theta^{C_j})|\theta^{Q_j}))$.
        Update target networks with soft replacement factor $\tau$.
    end for
end for
return $\mathcal{D}_i(s|\theta^{\mathcal{D}_i}), i = 1, 2, ...D$; $C_j(s|\theta^{C_j}), j = 1, 2, ...C$.
```

4 EXPERIMENTS

4.1 Pandemic Spreading Simulation

Since we cannot set pandemic outbreaks in the real word arbitrarily, training an RL agent to test our method under the real world pandemic spreading is not practical. Thus we design a pandemic spreading simulator, which consists of disease modeling and population mobility modeling. We apply the 8-state disease model of COVID-19 mentioned in Section 2.1 while using the parameters describing the features of COVID-19 from the medical researches [7, 8, 10, 16, 28] to guarantee a realistic disease modeling.

As mentioned in Section 3.1, we divide the city into M regions according to road networks, which match the real world functional areas in the city well. The total population size of the city (denoted as N) is accessible on the official website of the government. Population mobility modeling is based on real world mobility data, obtained in cooperation with the Internet services provider, by collecting GPS coordinates of users accessing location-based services. Considering privacy protection, we eliminate the detailed information of users and keep the following tracing:

- The number of users traveling from region i to j during time step t, denoted as $m_{ij}(t)$.

- The number of users in region i at time step t, denoted as $n_i(t)$.

We collect data in two cities, i.e., Beijing and Shanghai, with the detailed information shown in Table 1. Note that in Beijing, we only consider the downtown area and use Beijing referring to it for short later.

Table 1: Details of the datasets.

City	Population (N)	Regions (M)	Record Duration
Downtown Beijing	11 Million	673	21 days
Shanghai	24 Million	657	7 days

We calculate the average of $n_i(t)$ along the time dimension and obtain the average number of users in region i (denoted as $n^{(i)}$). Then we estimate the population size in region i (denoted as $N^{(i)}$) by comparing N and $n^{(i)}$, for the Internet services users have a similar spatial distribution as the whole population. Therefore, population density in region i is also available, calculating the area of the region. We can estimate the population mobility strength, i.e., the possibility for people region i to travel to region j at time t, which is denoted as $P_{ij}(t)$, by the ratio of $m_{ij}(t)$ and $n_i(t)$. Note that $P_{ij}(t)$ varies with time due to difference on population mobility between morning and evening, workdays and weekends.

The simulation process is shown in Algorithm 2, where the length of each time step is 30 minutes, to provide high temporal resolution results.

Algorithm 2 Pandemic Spreading Simulation

```
Input: Simulation duration $T$; information of the target city, i.e.,
    $N[i], P_{ij}(t),\ i, j = 1, 2, ..., M$; disease model (state transition
    possibility) and initially infected situation.
Output: Simulation results.
Initialize $\mathcal{R}^{(0)}$, records of number of people in each state each
    region, according to $N[i]$, $M$ and initially infected situation.
for $t = 1, 2, ...T$ do
    Obtain $\Delta\mathcal{R}$, the number of people experience state transition
        based on $\mathcal{R}^{(t-1)}$ according to state transition possibilities.
    Generate $\mathcal{R}^{(t)} \leftarrow \mathcal{R}^{(t-1)} + \Delta\mathcal{R}$.
    Obtain $\Delta\mathcal{R}'$, the number of people travel across regions based
        on $\mathcal{R}^{(t)}$ according to $P_{ij}(t)$.
    Update $\mathcal{R}^{(t)} \leftarrow \mathcal{R}^{(t)} + \Delta\mathcal{R}'$.
end for
return $\mathcal{R}^{(0)}, \mathcal{R}^{(1)}, ...\mathcal{R}^{(T)}$.
```

We evaluate the precision of our implemented simulator, making sure it performs well as the experimental platform. We perform simulations lasting 2 weeks in downtown Beijing, and then compare the results with the official reports[1] during the local pandemic outbreak in Xinfadi, Beijing, June 2020. We perform 10 simulations considering the randomness in simulations and average number of infected cases in simulations are plotted along with those in the real world in Figure 5. The results show a coefficient of determination $R^2 = 0.984$ between simulations and the real world situation. It demonstrates that the simulator can simulate the real world situation precisely, which can support the evaluations for our hierarchical reinforcement learning method.

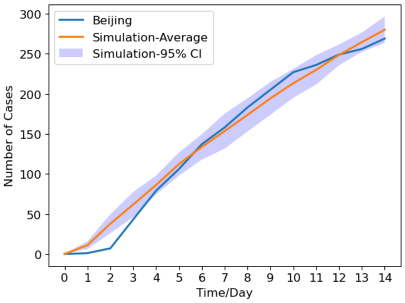


Figure 5: Comparison between simulations and the real world situation in Beijing.

4.2 Experimental Settings

In order to fully examine the performance of our method, we design three experiments with various initially infected situations in the two cities. The settings are shown in Table 2, I_0 initially infected cases are averagely distributed to K initial regions in each experiment. We simulate the pandemic spreading processes for four months (120 days), during which we set the former two months to be a freely-spreading period, and responsive actions are taken in the later two months. We adjust the real time strategy every 24 hours, and set the total number of hospital beds equipped to treat the COVID-19 patients (N_B) in both cities to be 10000. We set the number of daily supply of surgical masks in both cities (N_M) to be slightly less than 10% of the population size, which outline a situation of scarce medical resources. We fix the mask efficacy factor γ to be 0.9, a typical value for the surgical masks [3].

Table 2: Experimental settings.

Experiment	1	2	3
City	Beijing	Beijing	Shanghai
Initial cases (I_0)	80	80	100
Initial regions (K)	4	40	5
Hospital beds (N_B)	10000	10000	10000
Surgical masks (N_M)	1 Million	1 Million	2.2 Million

Contrasts between experiment 1 and 2 show the differences between aggregated and scattered outbreak while contrasts between experiment 1 and 3 cover the situation in completely different cities.

4.3 Baselines

We compare our method with several baselines, including expert solutions in public health and RL based methods. The expert solutions include:

- **No Medical**: The situation without allocating medical resources, serve as a blank control group.
- **Expert Seriousness** [1]: Allocating the medical resources according to the order of seriousness of pandemic spreading. Regions with more infected cases have absolute priority.

[1] National Health Commission, http://www.nhc.gov.cn/

Table 3: Experiment results in situations with perfect information.

Experiment	Metric	No medical	Expert seriousness	Expert population	DQN greedy	DQN expert	**Our method**
1	Infection (σ)/%	43.99 (0.12)	19.92 (0.18)	27.23 (0.29)	26.68 (0.32)	17.24 (0.22)	**15.71 (0.13)**
	Death (σ)/‰	27.11 (0.11)	10.88 (0.12)	14.51 (0.16)	14.30 (0.21)	9.45 (0.12)	**8.62 (0.09)**
	Max daily infection (σ)/‰	19.17 (0.03)	11.65 (0.07)	16.05 (0.13)	15.64 (0.11)	10.26 (0.13)	**9.37 (0.08)**
	Max daily death (σ)/10^{-4}	16.76 (0.14)	7.32 (0.12)	10.33 (0.14)	10.05 (0.15)	6.22 (0.11)	**5.63 (0.06)**
2	Infection (σ)/%	37.40 (0.21)	9.85 (0.17)	14.68 (0.34)	13.52 (0.31)	8.27 (0.19)	**7.53 (0.16)**
	Death (σ)/‰	21.63 (0.17)	4.97 (0.10)	7.11 (0.18)	6.59 (0.16)	4.28 (0.10)	**3.90 (0.09)**
	Max daily infection (σ)/‰	18.77 (0.04)	6.52 (0.12)	10.23 (0.21)	9.43 (0.21)	5.33 (0.12)	**4.79 (0.11)**
	Max daily death (σ)/10^{-4}	14.35 (0.12)	3.50 (0.09)	5.41 (0.15)	4.96 (0.13)	2.92 (0.10)	**2.59 (0.09)**
3	Infection (σ)/%	29.21 (0.14)	11.59 (0.13)	16.33 (0.27)	16.40 (0.19)	11.23 (0.12)	**9.62 (0.11)**
	Death (σ)/‰	15.73 (0.07)	6.02 (0.08)	8.07 (0.14)	8.15 (0.11)	5.86 (0.05)	**5.07 (0.06)**
	Max daily infection (σ)/‰	17.09 (0.04)	7.62 (0.07)	11.27 (0.16)	11.26 (0.10)	7.36 (0.08)	**6.29 (0.07)**
	Max daily death (σ)/10^{-4}	11.26 (0.05)	4.29 (0.08)	6.18 (0.11)	6.23 (0.11)	4.12 (0.04)	**3.52 (0.04)**

Table 4: Experiment results in situations with imperfect information.

Experiment	Metric	No medical	Expert population	DQN greedy	DQN expert	**Our method (imperfect Info.)**
1	Infection (σ)/%	43.99 (0.12)	27.23 (0.29)	26.53 (0.21)	18.29 (0.24)	**15.88 (0.21)**
	Death (σ)/‰	27.11 (0.11)	14.51 (0.16)	14.23 (0.14)	10.05 (0.14)	**8.73 (0.11)**
	Max daily infection (σ)/‰	19.17 (0.03)	16.05 (0.13)	15.52 (0.06)	10.78 (0.11)	**9.44 (0.10)**
	Max daily death (σ)/10^{-4}	16.76 (0.14)	10.33 (0.14)	9.99 (0.12)	6.66 (0.15)	**5.77 (0.09)**
2	Infection (σ)/%	37.40 (0.21)	14.68 (0.34)	13.99 (0.25)	8.73 (0.22)	**7.40 (0.16)**
	Death (σ)/‰	21.63 (0.17)	7.11 (0.18)	6.82 (0.14)	4.46 (0.11)	**3.85 (0.08)**
	Max daily infection (σ)/‰	18.77 (0.04)	10.23 (0.21)	9.69 (0.15)	5.71 (0.15)	**4.70 (0.11)**
	Max daily death (σ)/10^{-4}	14.35 (0.12)	5.41 (0.15)	5.17 (0.13)	3.07 (0.10)	**2.56 (0.06)**
3	Infection (σ)/%	29.21 (0.14)	16.33 (0.27)	16.41 (0.18)	11.18 (0.19)	**9.62 (0.08)**
	Death (σ)/‰	15.73 (0.07)	8.07 (0.14)	8.07 (0.09)	5.84 (0.11)	**5.06 (0.05)**
	Max daily infection (σ)/‰	17.09 (0.04)	11.27 (0.16)	11.46 (0.11)	7.30 (0.12)	**6.29 (0.04)**
	Max daily death (σ)/10^{-4}	11.26 (0.05)	6.18 (0.11)	6.22 (0.10)	4.12 (0.09)	**3.52 (0.03)**

- **Expert Population** [9]: Allocating the medical resources according to population density. Regions with higher population densities have absolute priority.

The RL methods and the combination of RL and expert solution are as follows:

- **DQN Greedy**: A non-hierarchical reinforcement learning method that applies one DQN to select the significance ranking principle for each kind of medical resource. Regions with higher rankings have absolute priority.
- **DQN Expert** [5]: A combination of DQN and expert solution. Public health research shows that if mask coverage rate reaches 86% approximately, the pandemic of COVID-19 can be controlled efficiently. Thus a DQN is used to select ranking principle for surgical masks while the satisfaction factor is fixed to 0.86. The action on hospital beds stays identical to DQN greedy.

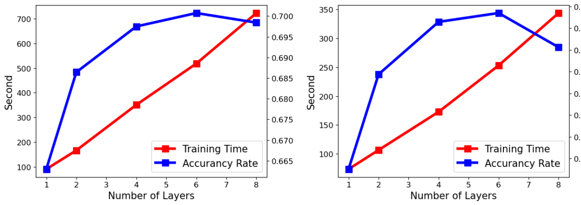
These baselines cover both authoritative expert solutions in public health studies and typical RL solutions. Also, by combining RL and expert solution, the last baseline outperforms the former ones and turns out to be the strongest one.

4.4 Main Results and Analysis

We evaluate the trained models based on the simulator. We repeat all simulations for 10 times considering the randomness. Firstly, we consider perfect information situations, where our method takes the precise pandemic spreading information of all 8 states directly from the environment without information rebuilding and so do the baselines. The results measured by the four metrics defined in Section 2.3 (the lower, the better) are shown in Table 3, where we divide values of metrics by the population size and standard deviations in 10 simulations are shown in the brackets. It turns out that our method outperforms all baseline methods in all 3 experiments.

Then we study the imperfect information situation, where we train the DQN based baselines only with the information of the 3 states (I_t, I_h and D) and expert seriousness fails to work due to lack of the information of infected cases. Our method works through information rebuilding, the effect of number of RNN layers in the imperfect information utilization framework is shown in Figure 6. Accuracy rate is the percentage of output values which have an relative error less than 30%. We notice that in both Beijing and Shanghai situation, there is no obvious increase on accuracy or even decrease due to overfitting using more than 4 layers, while training time goes up greatly as more layers are hired. Therefore, we use 4 RNN layers, which are enough for information rebuilding, considering both efficacy and accuracy.

The final results measured by the same metrics are shown in Table 4. Note that the results of no medical and expert population



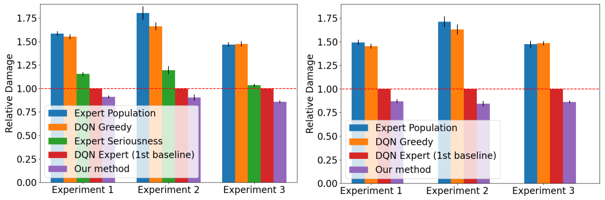
(a) Beijing

(b) Shanghai

Figure 6: Effect of layer number in information rebuilding.

stay unchanged for actions in these two baselines have no relation with the pandemic spreading information. It also turns out that our method has an outstanding performance.

Figure 7 shows the comparison between our method and baselines. We normalize the performance of the strongest baseline (DQN Expert) to 1.0 and calculate relative damage caused by the pandemic through an average of all the 4 metrics.



(a) Perfect Information

(b) Imperfect Information

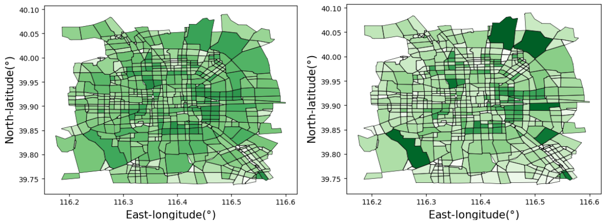
Figure 7: Performance comparison in both perfect and imperfect information situations.

All the results prove that our method can give more efficient real time medical resource allocation strategies which decrease the damage caused by the pandemic greatly. Quantitatively, the average performance gain compared with the strongest baseline is 11.00% under perfect information situation while 14.25% under imperfect one. The higher performance gain under imperfect information proves that the design of imperfect information utilization brings an extra benefit under imperfect information, which is closer to the real world situation.

We take hospital beds allocation in experiment 1 as an example and strategies after applying our method for 10 and 30 days are visualized on the map of Beijing in Figure 8, where darker color indicates more medical resources. It shows that the strategy varies during the pandemic spreading process and some key regions are allocated with more resources, i.e., efficient and real time strategies.

4.5 Convergency and Efficiency in Training

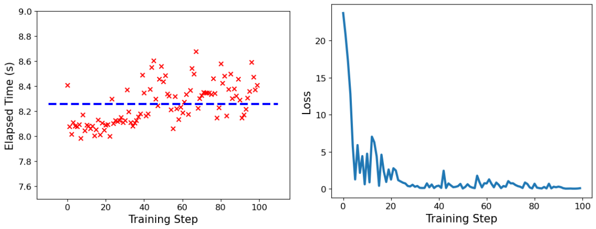
We train RL agents on the datasets of Beijing and Shanghai, and the training process in Beijing is shown in Figure 9. Due to the design of decomposed action space, the decision making process is greatly simplified and thus lightweight networks can satisfy our needs. Combining with our training algorithm, we can conduct the training process efficiently. It only consumes 8.26s for each step on average and less than 15 minutes for the whole 100 steps on a laptop without high performance GPUs. Training loss defined in Section 3.4 decreases quickly during the training process, indicating a good convergency.



(a) 10 days

(b) 30 days

Figure 8: Strategies visualization in experiment 1 where darker color indicates more medical resources.



(a) Elapsed Time

(b) Loss

Figure 9: Elapsed time and Loss in the model training.

5 RELATED WORK

We focus on solving the problem of medical resource allocation with reinforcement learning, which closely relates to the studies on medical resource allocation. We also introduce widely related applications of reinforcement learning in the field of public health.

5.1 Medical Resource Allocation

It has long been a thorny issue about how to allocate scarce medical resources in face of public health disasters. Generally, medical resources include healthcare workers, hospital beds, medicine and medical equipment, e.g., CT/MRI scanners and ventilators [4]. Specifically, in the context of COVID-19, surgical masks are also considered to be critical [3]. Our work focuses on the allocation of hospital beds and surgical masks, which can also be easily extended to any resource type.

High-level guidelines on allocating scarce medical resources among different population groups, e.g., stratified by ages, occupations [4, 20] or seriousness [6], have been framed and widely accepted, in the hope of maximizing social benefits out of limited resources [4, 23]. These guidelines serve as a starting point, but need detailed designs to facilitate real world implementations.

It is largely unexplored how to conduct real time medical resource allocation on the urban scale. In contrast, our method offers adaptive and real time strategies on medical resource allocation among urban regions, considering both real world population mobility and pandemic spreading situations.

5.2 Reinforcement Learning and Public Health

Reinforcement learning is a powerful method for decision making [26], which has long been used in the field of public health. In the context of COVID-19, RL algorithm can make predictions on the pandemic spreading situation [12]. Researchers have proposed various RL solutions for lockdown strategies, trying to intervene the pandemic spreading while to minimize the negative effect on social economy [14, 19]. A reinforced algorithm is also proposed for redistribution of ventilators among states [?]. However, efficient and real time strategies for medical resource allocation within cities stays largely unexplored and our work focus on this problem.

Besides, RL methods perform well in medical image analysis [24], drug design [21] and medical robots control [11], etc. They are making contributions to protect the public health and there exists vast unexplored space in other health related applications of RL.

6 CONCLUSIONS

In this paper, we study the problem of scarce medical resources allocation with imperfect information. We propose a hierarchical reinforcement learning method with imperfect information utilization framework. We conduct extensive experiments under various settings using a pandemic spreading simulator built based on real world data. The results demonstrate that our method outperforms the baseline methods all the time, especially in situations with imperfect information.

There are several future directions for our work. Firstly, we currently only allocate the medical source on the level of urban regions. In the future, we plan to extend our method to allocate medical resource on fine-grained level, i.e., to individuals, by considering personal characteristics such as age and occupation. Secondly, our current method only focus on minimizing the number of infections and deaths. In the future, we expect to take more factors into consideration, such as maximizing the equality among different social groups.

ACKNOWLEDGMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000, the National Natural Science Foundation of China under U1936217, 61971267, 61972223, 61941117, 61861136003, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

REFERENCES

[1] Hui Cao and Simin Huang. 2012. Principles of Scarce Medical Resource Allocation in Natural Disaster Relief: A Simulation Approach. *Medical Decision Making* 32, 3 (2012), 470–476.

[2] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR* abs/1406.1078 (2014).

[3] Steffen E. Eikenberry, Marina Mancuso, Enahoro Iboi, Tin Phan, Keenan Eikenberry, Yang Kuang, Eric Kostelich, and Abba B. Gumel. 2020. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infectious Disease Modelling* 5 (2020), 293 – 308.

[4] Ezekiel J. Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P. Phillips. 2020. Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *New England Journal of Medicine* 382, 21 (2020), 2049–2055.

[5] João A.M. Gondim. 2021. Preventing epidemics by wearing masks: An application to COVID-19. *Chaos, Solitons Fractals* 143 (2021), 110599.

[6] Simmy Grover, Alastair McClelland, and Adrian Furnham. 2020. Preferences for scarce medical resource allocation: Differences between experts and the general public and implications for the COVID-19 pandemic. *British journal of health psychology* 25, 4 (2020), 889–901.

[7] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David SC Hui, et al. 2020. Clinical characteristics of coronavirus disease 2019 in China. *New England journal of medicine* 382, 18 (2020), 1708–1720.

[8] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine* 26, 5 (2020), 672–675.

[9] Hiromasa Horiguchi Seizan Tanabe Manabu Akahane Toshio Ogawa Soichi Koike Tomoaki Imamura Hideo Yasunaga, Hiroaki Miyata. 2011. Population density, call-response interval, and survival of out-of-hospital cardiac arrest. *International Journal of Health Geographics* 10, 26 (2011).

[10] Chaolin Huang, Y Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395, 10223 (2020), 497–506.

[11] Brenton Keller, Mark Draelos, Kevin Zhou, Ruobing Qian, Anthony N Kuo, George Konidaris, Kris Hauser, and Joseph A Izatt. 2020. Optical coherence tomography-guided robotic ophthalmic microsurgery via reinforcement learning from demonstration. *IEEE Transactions on Robotics* 36, 4 (2020), 1207–1218.

[12] Soheyl Khalilpourazari and Hossein Hashemi Doulabi. 2021. Designing a hybrid reinforcement learning based algorithm with application in prediction of the COVID-19 pandemic in Quebec. *Annals of Operations Research* (2021), 1–45.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Gloria Hyunjung Kwak, Lowell Ling, and Pan Hui. 2021. Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. *Plos one* 16, 5 (2021), e0251550.

[15] Michael Y. Li and James S. Muldowney. 1995. Global stability for the SEIR model in epidemiology. *Mathematical Biosciences* 125, 2 (1995), 155 – 164.

[16] Q Li, X Guan, P Wu, X Wang, L Zhou, Y Tong, R Ren, Leung Ksm, Lau Ehy, J Y Wong, et al. 2020. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine* 382, 13 (2020), 1199–1207.

[17] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[18] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[19] Abu Quwsar Ohi, MF Mridha, Muhammad Mostafa Monowar, and Md Abdul Hamid. 2020. Exploring optimal control of epidemic spread using reinforcement learning. *Scientific reports* 10, 1 (2020), 1–19.

[20] Govind Persad, Alan Wertheimer, and Ezekiel J Emanuel. 2009. Principles for allocation of scarce medical interventions. *The Lancet* 373, 9661 (2009), 423–431.

[21] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. 2018. Deep reinforcement learning for de novo drug design. *Science advances* 4, 7 (2018), eaap7885.

[22] Megan L Ranney, Valerie Griffeth, and Ashish K Jha. 2020. Critical supply shortages—the need for ventilators and personal protective equipment during the Covid-19 pandemic. *New England Journal of Medicine* 382, 18 (2020), e41.

[23] Sara J Rosenbaum et al. 2011. Ethical considerations for decision making regarding allocation of mechanical ventilators during a severe influenza pandemic or other public health emergency. (2011).

[24] Farhang Sahba, Hamid R Tizhoosh, and Magdy MA Salama. 2006. A reinforcement learning framework for medical image segmentation. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings.* IEEE, 511–517.

[25] Julian Sheather. 2006. Ethics in the face of uncertainty: preparing for pandemic flu. *Clinical Ethics* 1, 4 (2006), 224–227.

[26] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[27] P. J. Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[28] Joseph T Wu, Kathy Leung, and Gabriel M Leung. 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* 395, 10225 (2020), 689 – 697.