# Experiments

## SEIR Model and Baseline RL Model

Our study seeks to test designed artifacts using data from Hong Kong in the context of COVID-19, which has rapidly spread since the first case reports in December 2019, becoming a severe global pandemic.

In particular, we utilize the SEIR model to simulate the progression of the COVID-19 pandemic, categorizing individuals into susceptible (S), exposed (E), infected/infectious (I), and removed (R) groups (Djidjou-Demasse et al. 2020; Godio et al. 2020; He et al. 2020). Note that, we integrated the recovered and death cases into the removed category, assuming that individuals who have recovered from the virus have gained immunity and are not susceptible to future infection. Figure 1(a) illustrates the inter-group relations, with , , and denoting the exposure rate, infection rate, and removal rate, respectively. The efficacy of NPIs on disease control is represented by the symbol . Moreover, similar to previous studies using the SEIR model (He et al. 2020; Kompella et al. 2020), random disturbances have been incorporated into the simulation to introduce the complexity of real-world scenarios. In Figure 1(a) Formula 1, represents a 10% probability of increasing or decreasing by a maximum of 10 exposed cases between two sequential statuses, S and E, to simulate accidental infections. Using random numbers with a uniform distribution, the hyper-parameters above have been adjusted (as per Figure 1(a) Formula 2) to simulate fluctuations in transmission and variations in policy implementations.

Meanwhile, we establish the initial RL Model based on previous research (Kompella et al. 2020; Padmanabhan et al. 2021). As shown in Figure 1(b), we have presented a straightforward overview of our baseline RL model. The agent acts as a representative of the policymaker. In order to simulate the environment of the pandemic, we implemented the SEIR model and determined hyper-parameters , , and , through a first-order Euler equation on the pandemic data from January 8 to May 20, 2020, which are verified by prior studies (Djidjou-Demasse et al. 2020; Liu and Rocklöv 2021). We utilized the efficacy of control as the action, restricted by upper and lower bounds of 0.9 and 0, respectively, with a step size of 0.1, resulting in 10 total actions. Apparently, the strictest action results in a reduction of the parameter by 90%. Observations occur at daily intervals with eight attributes, including population numbers within the susceptible (S), exposed (E), infected/infectious (I), and removed (R) populations, the increase in the exposed (E), infected/infectious (I), and removed (R) populations since the last decision round, and the most recent action selected. In terms of reward, the total reward comprises two sub-rewards (as per Figure 1(a) Formula 3, , which means the acceptable number of daily new infected cases): the status of public health reflected by the sigmoid function of the standardized number of new patients exceeding expectations () and the action cost that receives a high value under strict action (). The discount factor, , delineates the relative importance of short-term rewards in comparison to long-term rewards. In our study, we set the value of this parameter to 0.99, and this value has been widely used in prior studies (Chadi and Mousannif 2022; Kompella et al. 2020).

|  |
| --- |
|  |
| **Figure 1.** SEIR Model and Baseline RL Model |

Due to the infinite state space involved in our research, we implemented the deep Q-network (DQN) algorithm, a value-based reinforcement learning method, that leverages an artificial neural network to optimize the action-value function, ensuring that comparable states result in similar action outputs (Mnih et al. 2016; Mnih et al. 2013). We trained our baseline RL model with the pandemic data in Hong Kong from March 27, 2020, featuring , , , and .

## Component of Virus Mutation

As shown in Figure 2, we build an internal SEIR simulation component for Agent. We mark the internal model as , and mark the external model as . While simulating, both two models can receive the same initial population of 4 categories (S, E, I, and R) and the same initial transmission parameters. When artificial mutation happens, can get the changed parameters directly, while will estimate the new parameters with 10 days’ population data of the 4 categories. We provide these settings because virus mutations cannot be detected immediately, and medical testing of virus parameters requires a certain time cost (Kirkeby et al. 2017; Smirnova et al. 2019). At the beginning of the epidemic, due to the unpredictability of the strains, the detection efficiency of pathogens could not be ensured (McIntosh 1974). Our settings can ensure the delay of obtaining disease parameters, which is consistent with the real-world situation.

In order to determine the location of mutation points and predict the reproductive coefficient of mutant strains, we referred to the research of (Perakis et al. 2022) and applied the random martingale process to achieve the CPD algorithm (Shown in Algorithm 1). We set 3 mutations during an episode. The parameters and the date of mutations are shown in Table 1, which are designed according to the existing research on the COVID-19 pandemic (Liu and Rocklöv 2021; Liu and Rocklöv 2022).

|  |
| --- |
|  |
| **Figure 2.** Extended RL Framework and CPD Algorithm’s Performance |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mutations** |  |  |  |  |
| Initial Coefficient | 3.64 | 0.1 | 0.13 | 0 |
| First Mutation | 5.5 | 0.526 | 0.13 | 60 |
| Second Mutation | 7.2 | 0.03 | 0.15 | 120 |
| Third Mutation | 15.2 | 0.025 | 0.2 | 180 |
| **Table 1.** Mutation Data | | | | |

|  |
| --- |
|  |
| **Algorithm 1.** Change Point Detection of Martingale Process |

Figure 3 (a) shows the performance of DQN (marked as Base) and DQNCPD (marked as CPD) in the non-mutation environment. We find that DQNCPD is not as good as DQN in the environment without mutation. In terms of epidemic control, DQN has a significantly better performance in the short term and makes the same performance in the middle and long term. Besides, DQN has a higher total reward in this environment. Figure 3 (b) shows the performance in the mutation environment. DQNCPD achieves excellent control effects in the early and middle term and is in line with the DQN control effects in the long term. And the reward of DQNCPD is slightly higher than DQN. Besides, no matter in which environment, DQNCPD’s action volatility is smaller, which indicates that its learning strategy is more stable. Through these results, we prove the significance of considering virus mutation, which may perform better in real situations with kinds of COVID-19 strains.

This result shows that considering virus mutations will help the system perform better. It also confirms that it is necessary for us to consider the gap in the general system design and narrow the gap, taking it as one of the design perspectives.

|  |
| --- |
|  |
| **Figure 3.** Performance of DQN and DQNCPD |

## Component of Multi-rewards

The values of we used are 10, 5, 2, 1, 0.5, 0.2 and 0.1. And when decreases to 1 or below, the system will take due to the large proportion of action costs. Therefore, the paper only shows the results when is greater than or equal to 1.

In Figure 4, the red line represents the DQNCPD, and the blue line represents DQN, the same As shown in both the non-mutation and mutation environments, we can see that with the increase of , the 2 systems obviously take significantly more stringent actions, and instead, both 2 systems will take the action when drop to 1. In addition, the larger the , the better the performance of epidemic control, which is reflected in the control of daily new infected.

We also make a comparison of the 2 models under the same weights. In the non-mutation environment, DQNCPD performs slightly better in controlling the daily new infected and the total reward when , instead, it performs significantly worse when and . However, in the mutation environment, the performance of the 2 models is almost the same when . From the change of the average of actions taken by the 2 models, the fluctuation of the actions taken by DQNCPD is significantly smaller than that of DQN, which is more reasonable in the real situation and can make sense.

|  |
| --- |
|  |
| **Figure 4.** Performance with Various Weights in Non-Mutation/Mutation Environment |

## Component of Guided Rewards

We summarize the previous research on RL-based policy support systems in epidemic control and conclude 2 kinds of AI immoral behaviors, threshold fluctuation (**TF**) and algorithm cheating (**AC**). We find that the 2 immoral behaviors are related to reward function design. We record the reward settings in 4 research with immoral behaviors and summarize them in Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Immoral Behavior** | **Paper** | **Reward Settings** | **reference** |
| **Threshold Fluctuation (TF)** | 1 |  | (Arango and Pelov 2020) |
| 2 |  | (Chadi and Mousannif 2022) |
| 3 |  | (Ohi et al. 2020) |
| 4 |  | (Padmanabhan et al. 2021) |
| **Algorithm Cheating (AC)** | 5 | Same as 4 | (Padmanabhan et al. 2021) |
| **Table 1.** Reward Function in Selected Research with Immoral Behaviors | | | |

In order to better prove the real existence of **TF** and **AC**, we simulate the reward function design in these researches and verify our summary of reward design errors. The reward we designed for TF is shown in Formula 4 and for AC in Formula 5.

The results of reward function for TF are shown in Figure 5 (a), which clearly shows an obvious phenomenon of threshold fluctuation. We also observed significant changes in the selected action under all weights. The results of reward function for AC are shown in Figure 5 (b), which demonstrate the existence of AI cheating behavior. The DQN gains positive rewards through the “short loosening, long tightening” policy, which leads to a significantly worse control effect of the COVID-19 epidemic. The number of daily new infections is much higher than in the original reward setting, which contradicts the original intention.

|  |
| --- |
|  |
| **Figure 5.** Performance of DQN and DQNCPD in Mutation Environment |

## Component of Comparison and Interpretation

The four policies we use as benchmarks are designed as follows.

1. **Double thresholds policy (DTP)**: DTP refers to taking the strictest action when the population of I exceeds the first threshold , taking the action when the population of I is less than the second threshold , and taking the moderate action when the population between the 2 thresholds.
2. **Complete intervention policy (CIP)**: CIP refers to taking the strictest action at each step, whose control effect is 0.9.
3. **Random intervention policy (RIP)**: RIP refers to taking arbitrary action at each step.
4. **Non-intervention policy (NIP)**: NIP refers to making the epidemic develop free and taking action at each step.

# Appendix

## Base RL-based Policy Support System

**Agent:** The agent represents the policymaker. Our goal is to train it to be an intelligent policymaker with the ability to select the optimal NPIs strategy under different environmental states.

**Environment:** We use the COVID-19 data from Hong Kong as the data of the SEIR model. The parameters , , and , which are competed through first-order Euler equation on the epidemic data from January 8 to May 20, 2020, and verified with prior studies (Djidjou-Demasse et al. 2020; Liu and Rocklöv 2021).

**Observation:** We set 1 day as the interval between each decision point. In order to match the reality, we assume that the virus parameters cannot be observed by the agent. The observation contains 8 features, including the number of the population of the S, E, I, and R, the increasing number of the E, I, and R since the last decision point, and the last selected action.

**Action:** We use control effects instead of specific NPIs as the action directly, set the upper limit to 0.9, the lower limit to 0, and give a graduation value of 0.1, which means we have 10 actions in total. With the strictest action, the parameter will be reduced by 90%.

**Reward:** The reward function is the key to guiding the direction of RL. In this study, we subdivide the reward into two sub-rewards. The first one represents the effect of public health and is reflected in the number of daily new cases. The second one represents the action cost, which will get a high value under strict action. The total reward is completed by Formula 1. And we set .

**Discount Factor**: The discount factor , defines how valuable short-term rewards are compared to the long-term reward. In this case, we set this parameter as 0.99.

According to the Bellman equation, the Q-function defines the expected future discounted reward for taking action in state and then following policy thereafter, as shown in Formula 2 (Watkins 1989).

Given the infinite state space in our case, we select the deep Q-network (DQN) algorithm. DQN is a value-based RL method and uses an artificial neuron network to optimize the action-value function so that similar states can get similar output actions (Mnih et al. 2016; Mnih et al. 2013).

# References

Arango, M., and Pelov, L. 2020. "Covid-19 Pandemic Cyclic Lockdown Optimization Using Reinforcement Learning." arXiv.

Chadi, M.-A., and Mousannif, H. 2022. "A Reinforcement Learning Based Decision Support Tool for Epidemic Control: Validation Study for Covid-19," *Applied Artificial Intelligence* (36:1), p. 2031821.

Djidjou-Demasse, R., Michalakis, Y., Choisy, M., Sofonea, M. T., and Alizon, S. 2020. "Optimal Covid-19 Epidemic Control until Vaccine Deployment," *medRxiv*), pp. 1-21.

Godio, A., Pace, F., and Vergnano, A. 2020. "Seir Modeling of the Italian Epidemic of Sars-Cov-2 Using Computational Swarm Intelligence," *Int J Environ Res Public Health* (17:10), pp. 1-19.

He, S., Peng, Y., and Sun, K. 2020. "Seir Modeling of the Covid-19 and Its Dynamics," *Nonlinear Dynamics* (101:3), pp. 1667-1680.

Kirkeby, C., Halasa, T., Gussmann, M., Toft, N., and Græsbøll, K. 2017. "Methods for Estimating Disease Transmission Rates: Evaluating the Precision of Poisson Regression and Two Novel Methods," *Scientific Reports* (7:9496), pp. 1-11.

Kompella, V., Capobianco, R., Jong, S., Browne, J., Fox, S. J., Meyers, L. A., Wurman, P. R., and Stone, P. 2020. "Reinforcement Learning for Optimization of Covid-19 Mitigation Policies." arXiv.

Liu, Y., and Rocklöv, J. 2021. "The Reproductive Number of the Delta Variant of Sars-Cov-2 Is Far Higher Compared to the Ancestral Sars-Cov-2 Virus," *Journal of Travel Medicine* (28:7), pp. 1-3.

Liu, Y., and Rocklöv, J. 2022. "The Effective Reproductive Number of the Omicron Variant of Sars-Cov-2 Is Several Times Relative to Delta," *Journal of Travel Medicine* (29:3), pp. 1-4.

McIntosh, K. 1974. "Coronaviruses: A Comparative Review," *Current Topics in Microbiology and Immunology / Ergebnisse der Mikrobiologie und Immunitätsforschung,* W. Arber, R. Haas, W. Henle, P.H. Hofschneider, N.K. Jerne, P. Koldovský, H. Koprowski, O. Maaløe, R. Rott, H.G. Schweiger, M. Sela, L. Syruček, P.K. Vogt and E. Wecker (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 85-129.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. 2016. "Asynchronous Methods for Deep Reinforcement Learning," *International conference on machine learning*: PMLR, pp. 1928-1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. J. a. p. a. 2013. "Playing Atari with Deep Reinforcement Learning." arXiv.

Ohi, A. Q., Mridha, M. F., Monowar, M. M., and Hamid, M. A. 2020. "Exploring Optimal Control of Epidemic Spread Using Reinforcement Learning," *Scientific Reports* (10:1), p. 22106.

Padmanabhan, R., Meskin, N., Khattab, T., Shraim, M., and Al-Hitmi, M. 2021. "Reinforcement Learning-Based Decision Support System for Covid-19," *Biomedical Signal Processing and Control* (68), p. 102676.

Perakis, G., Singhvi, D., Skali Lami, O., and Thayaparan, L. 2022. "Covid-19: A Multiwave Sir-Based Model for Learning Waves," *Production and Operations Management* (0:0), pp. 1-19.

Smirnova, A., deCamp, L., and Chowell, G. 2019. "Forecasting Epidemics through Nonparametric Estimation of Time-Dependent Transmission Rates Using the Seir Model," *Bulletin of Mathematical Biology* (81:11), pp. 4343-4365.

Watkins, C. J. C. H. 1989. "Learning from Delayed Rewards," *Robotics & Autonomous Systems*.