ARTIFICIAL INTELLIGENCE THEORY and APPLICATIONS

ISSN: 2757-9778 || **ISBN :** 978-605-69730-2-4

More information available at **aita.bakircay.edu.tr**

Deep Reinforcement Learning for Simulation-Based Determination of COVID-19 Pandemic Mitigation Policies

Mahmut L. ÖZBİLEN *, Emre EĞRİBOZ, Ruşen HALEPMOLLASI, İsmail BİLGEN, Mehmet HAKLIDIR

TÜBİTAK, Informatics and Information Security Research Center, Information Technologies Institute, Turkey

* Corresponding Author: TÜBİTAK, *Informatics and Information Security Research Center, TURKEY*
Tel. : +90 262 675 30 00 E-Mail : mahmut.ozbilen@tubitak.gov.tr

Publication Information

Keywords :
- Reinforcement learning;
- Deep learning;
- COVID-19;
- Pandemic.

Category : Research article

Received : 14.07.2021
Accepted : 25.09.2021

© 2021 Izmir Bakircay University.
All rights reserved.

A B S T R A C T

A global health emergency has been declared by WHO at the beginning of 2020 based on the increasing number of cases in the COVID-19 epidemic. Governments around the world have taken unprecedented measures. However, there is no guarantee that the measures taken are best to mitigate the effect of pandemic. We investigate the impact of government policies regarding interventions on deaths related to the COVID-19 and mitigation of the economic decline. In a simulation environment, we use Reinforcement Learning (RL) to explore the optimal policies to prevent COVID-19 outbreak. We use a specific simulator called PandemicSimulator which has detailed abilities to simulate spread of disease and people interactions at different locations. The simulator is utilized to train RL agents to take mitigation policies with minimum economic damage of the pandemic without exceeding the hospital capacity. We use Deep Q Networks to train the RL agent. We compare the performance of the our agent's policy with the policy applied by the United Kingdom in terms of critical patients, deaths and economic damage. Results show that policies improved by the RL agent can help decision makers in the pandemic mitigation policies.

1. Introduction

Coronavirus disease 2019 (COVID-19) appeared in Wuhan, China in the last months of 2019 and has turned into a pandemic with extensive spreads in some countries such as Italy, Iran (World Health Organization, 2020). Moreover, COVID-19 has spread to millions of persons worldwide and the public health threat, which it impacts, and is the most severe seen in a respiratory virus the influenza pandemic (H1N1 virus) that emerged nearly a century ago (Taubenberger et al., 2006). Many people have had to change their lifestyle due to this rapidly transmitted virus. Also, countries have used different approaches to test, cleaning, social distancing, and quarantine measures to balance the spread of the virus and its impact on the economy. Governments have endeavored to keep deaths from COVID-19, the highest priority for individuals, and the intensity of healthcare as low as possible while they have also endeavored to keep the

inevitable economic decline caused by the restrictions imposed to prevent the spread of the virus as low as possible (Elgin et al., 2020; Anderson et al. 2020).

Artificial Intelligence (AI) based solutions, in particular, deep learning and machine learning are increasingly used in several domains in healthcare such as identifying disease (Jaiswal et al., 2019), detecting bias in methods used to diagnose diseases (Dervisoglu et al., 2021), detecting of COVID19 from symptomatic information (Najar, 2021), etc. Also, there are many opportunities in various medical domains where promising AI techniques, particularly Reinforcement learning (RL) based solutions can be applied. RL becomes prominent in exploring the optimal policies in various problems such as treatment of a patient, the tracking of the spread of the epidemic and mitigation of its effect (Ozbilen et al., 2021). Therefore, in healthcare informatics, generating solutions based on RL approaches has considerably attracted the attention of researchers in recent years (Yu et al., 2019; Hu et al., 1994; Schaefer et al., 2005).

In the comment focus study, Gottesman et al. have presented guidelines for RL on decisions about patient treatment. According to the authors, in healthcare, RL models need to access all information that affects decision-making and be trained in an adequate sample size to establish convergence between the policies learned and the actual policies and to provide effective results (Gottesman et al., 2019).

Li et al. have proposed a hybrid retrieval-generation RL method, namely HRGR-Agent, for medical image report generation, which was trained to integrate clinicians' knowledge with artificial neural networks. HRGR-Agent that is guided through RL method based on sentence level and word level rewards uses a hierarchical decision-making process. They have emphasized that although the findings in medical reports are mostly normal, the detection of rare and diverse abnormal findings is extremely important and can be detected with alone neither a generation approach nor a template approach. Therefore, they have trained the retrieval module together with the generation module by using RL framework in HRGR-Agent. For a sentence obtained from a medical image, the retrieval policy module decides either usage of a template sentence from a template database or generation of a new sentence (Li et al., 2018).

Moreover, Awasthi et al. have offered a novel pipeline joining ACKTR/DQL model with Contextual Bandits model in a feed-forward way to get an optimal distribution of vaccine. They have used RL models to suggest better actions. Contextual Bandits method has been used to enable daily basis modifications that may need to be implemented in the real-world scenario (Awasthi et al., 2020). Liu et al., in its yet immature study (Liu et al., 2020), have proposed a multi-agent Q learning under the microscopic epidemic model to predict the spread of the disease based on the decisions of individuals where every agent can choose its activity level. Optimal decisions obtained by minimizing agents' cost functions are used to estimate the spread of the epidemic. They have filled the gap between the relationship between agent decisions and the spread of the disease. In this way, they attempt to reveal the difference in results of the same policies such as "lock down", "stay-at-home" across different regions by accounting the human behavioral factor (Liu et al., 2020).

Another challenge in pandemic diseases such as COVID-19 is the shortages of medical equipment. This shortage requires systematic redistribution of necessary equipment throughout the need in the hospitals. Bednarski et al. propose a solution that incorporates both deep learning and reinforcement learning to alleviate the challenges of redistribution of ventilators throughout the COVID-19 pandemic. Their solution consists of three stages: 1. data processing, 2. demand inference, 3. collaborative exchange. In stage 2, they use an inference method like LSTM or RNN to predict the future demand. In stage 3, depending on the predicted future demand, they use RL to decide a policy such as, "no exchange", "maximum need first", "minimum need first", "random order". They use two RL algorithms: value iteration and Q-learning. They

created a simulation environment with data from the Institute of Health Metrics, Centers for Disease Control and Prevention, and Census Bureau to evaluate their system (Bednarski et al., 2020).

Martin-Calvo et al. have used real-world mobility and census data of the Boston area to build a co-location network at three different layers (community, households and schools). Also, they used a data-driven SEIR (susceptible, exposed, infected, recovered) model (Arregui et al., 2018) to test six different social distancing strategies, namely (i) school closures, (ii) self-distancing and teleworking, (iii) self-distancing, teleworking, and school closure, (iv) restaurants, nightlife, and closures of cultural venues, (v) non-essential workplace closures, and (vi) total confinement (Martin-Calvo et al., 2020). In another study, to explain the transmission dynamics of SARS-CoV-2 in Boston, authors have also built a detailed agent-based model that explores the strategies regarding the lifting of social-distancing interventions in combination with testing and isolation of cases and tracing and quarantine of exposed contacts. According to their results, following the reduction of the epidemic through strict restriction and the ceasing of unimportant activities, a proactive testing policy, the contacts tracing and their household quarantine led to the gradual reopening of economic activities. Meanwhile, low COVID-19 incidence in the population and a manageable impact on the healthcare system can also be achieved (Aleta et al., 2020).

In this paper, we investigate to optimize mitigation policies aimed at minimizing the economic impact and the risk of COVID-19 by balancing between economy and public health goals. For this purpose, we study RL algorithms that run in a controlled pandemic simulator and learning to decide mitigation policies. Then we compared the performance of the RL algorithms with the real world policies that are applied in the United Kingdom at the beginning of the pandemic.

The rest of the paper is organized as follows: Preliminaries are mentioned in Section 2 and the method of the study is presented in Section 3. After obtaining United Kingdom data is mentioned in Section 4, the experiments are presented in Section 5. Finally, in Section 6, the conclusion is discussed.

2. Preliminaries

Reinforcement learning is a machine learning approach that aims to ensure that the agent taking actions in an environment reaches the highest cumulative reward. Environment that the agent interacts with is modeled with Markov Decision Process (MDP) that contains the tuple of elements (S, A, P, R) where:

- *S:* set of states
- *A:* set of actions available in the current state
- *P:* is the function $P_a(s,s')$ that is probability of getting state *s'* for the action *a* in the state *s*
- *R:* is the reward of the action in the current state

Interaction of an agent with the environment is as follows: with each act of the agent, its state changes, and each state is fed back to the agent by the environment with a positive or negative reward. We have a framework on which RL agents are trained to develop a policy that maximizes cumulative reward by defining the problem as an MDP (Figure 1).

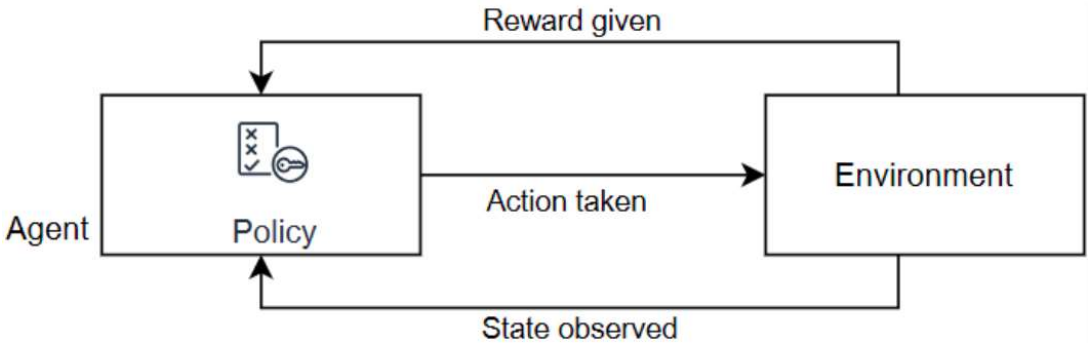

Figure 1. Reinforcement Learning Setting

There are two main approaches to develop such a policy, i.e., value based learning and policy based learning. In value-based learning, the agent assigns states a value upon given reward and previous incidents. In policy-based methods, the aim is to directly develop a policy that maximizes cumulative reward.

3. Methodology

3.1 PandemicSimulator

PandemicSimulator is an agent-based pandemic simulator that can be used to analyze and improve mitigation policies for the economic impact of pandemics without exceeding hospital capacity. It has the features of defining a community and location, customizing human interactions at locations, and modeling government policies. The simulator was developed for the purpose of training RL algorithms to improve pandemic mitigation policies with the joint contributions of artificial intelligence researchers and epidemiologists to simulate pandemics. Furthermore, it was implemented with the motivation of dealing with the COVID-19 outbreak (Kompella et al., 2020).

The simulator has many more unknown parameters, such as the rate at which the virus spreads from interactions in different environments, the effect on the rate of spread of different measures such as wearing a mask or washing hands. However, it is possible to calibrate these variables with the data obtained since the beginning of the pandemic. The more accurate and healthy the data calibration process has, the better the information that the model can extract from historical data. In the original work, the calibration of the simulator was carried out with historical data from Sweden using the assumption of the linear relationship between mean of the spread rate distribution and the number of deaths. In this way, the decisions to be made to mitigate the pandemic policies will be much accurate (Kompella et al., 2020).

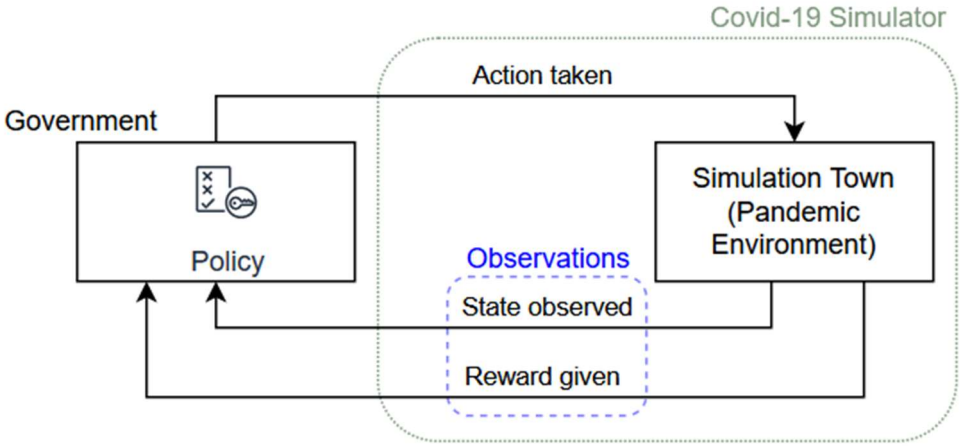

Figure 2. Reinforcement Learning Setting in PandemicSimulator

In this study, environment is the area simulated in the PandemicSimulator, the agent is the government that decides policies, actions are the restrictions, and reward is calculated with infections and economic damage, which is illustrated in Figure 2.

At the end of a simulated day, the results of the tests performed to diagnose positive COVID-19 and the hospitalization status for the patients are monitored by PandemicSimulator and fed to RL agent as states. A state observation includes the following numbers:

- critical patients
- dead
- infected persons
- not infected persons
- recovered persons

Actions of the RL agents are the restrictions bundled in stages that will be used in policies. Restrictions are:

- **Stay-at-home** is grouped into two sets. Zero (0), covers "no measure" and recommendation of not leaving houses, and one (1) comprises actions requiring staying at houses with different levels of exceptions, e.g. leaving allowed only once a week or only essential daily needs.

- **Face covering** to be 0 means there is no policy for wearing a mask or it is only recommended and of 1 means it is required in all shared/public spaces outside the home.

- **Social distance** is defined as a value between 0 and 1, where 0 can be considered as no social distance restriction, and 1 as situations in which any close distance is prohibited. Unlike the other restriction criterions, it was estimated based on the resident's personal communication.

- **Limiting gatherings** defines the restrictions on gatherings of people by considering low and high risks. The numbers next to LOW and HIGH tag indicate the maximum number of people who can come together according to defined risk.

- **Locked locations** refers to the closing of locations such as school, hair salon, office, retail store, and bar.

As illustrated in Table 1, there are four stages for which Stage 0 is the least strict and Stage 4 is the strictest.

Table 1. Restrictions in PandemicSimulator by Stages

Stages	Stay home if sick	Wear facial masks	Social distancing	Limiting gatherings	Locked locations
Stage 0	False	False	None	None	None
Stage 1	True	False	None	Low risk: 50, High risk: 25	None
Stage 2	True	True	0.3	Low risk: 25, High risk: 10	School, Hair Salon
Stage 3	True	True	0.5	Low risk: 0, High risk: 0	School, Hair Salon, Bar
Stage 4	True	True	0.7	Low risk: 0, High risk: 0	School, Hair Salon, Office, Retail Store, Bar

There are two main components of the reward: ratio of critical patients to the hospital capacity and economic damage. Economic damage is calculated by the ratio to the strictest stage, which is four in our case. Reward of the agent is calculated as follows:

$$r = a\ max(\frac{n^c - c^{max}}{c^{max}}, 0) + b\frac{stage^p}{max_j\ stage_j^p} \quad (1)$$

where r is the reward calculated, n^c is number of person in critical condition, c^{max} is the maximum capacity of hospital, a and b, which are -0.4 and -0.1, respectively, are variables that determine the importance of the component. In the formula, constant p, which is 1.5, denotes the exponent of the stage, and max_j returns maximum $stage_j$ in the stages array. We set the default values for *a, b, and p* as in the PandemicSimulator which are identified experimentally in (Kompella et al., 2020). Note that a and b must be negative to produce negative reward when critical number of patients exceeds the hospital capacity and stricter regulations are applied.

3.2 RL for optimization of policies for COVID-19 pandemic mitigation

Mitigating COVID-19 pandemic is a decision-making problem where governments decide policies according to the course of the pandemic. To find the optimal policies, we used RL that is highly suitable for this problem because it is basically a decision-maker. In this study, we used Deep Q Networks (DQN), a deep reinforcement learning technique, to train reinforcement learning agent.

3.2.1 Deep Q Networks

Classic Q-Learning algorithm stores values of state-action pairs in a table. This approach is useful for low dimensional environments but limited in environments where the number of state-action pair is too large or state-actions are continuous. Therefore, memory size would be the main problem for large state-action environments. For example, in video games, states are the screen images that are represented with pixels so the size of state dimension would be too high. To tackle this problem, Mnih et al. proposed DQN model that approximates the values of the actions using neural networks instead of storing all the state-action pairs (Figure 3) (Mnih et al., 2013).

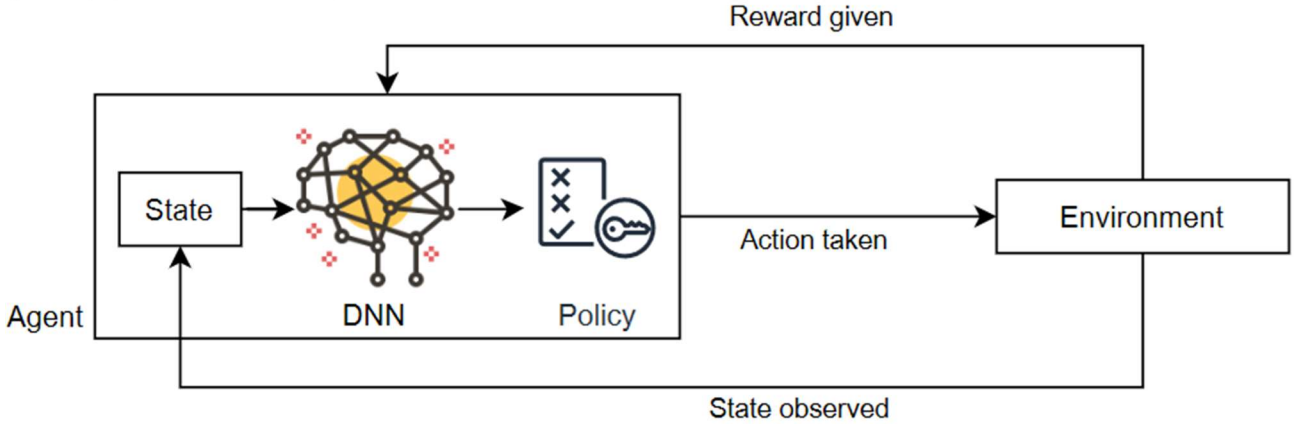

Figure 3. Deep Reinforcement Learning Setting

Neural network models approximate the values of the state s for each action a and update the network to minimize loss to real value of the state. Loss is calculated by the difference of target value y_i and output of neural network approximator $Q(s, a)$. The target value is not a label like in classic machine learning problems but it is derived from Bellman equation (Eq. 2) (Mnih et al., 2013).

$$y_i = r + \Upsilon max_a Q(s, a) \quad (2)$$

In the proposed DQN architecture, inputs are the states that are daily testing results, and output of the network is the value of each action for the given state. We used multilayer perceptron with two hidden layers as a value approximator instead of convolutional neural networks because input size is low so using convolutional neural networks would be too complex for this task.

4. Case Study: United Kingdom

To benchmark our model findings, we determined a policy according to United Kingdom approaches to mitigate the effects of the pandemic. United Kingdom that announces regulations to manage the pandemic imposes the restrictions in stages as illustrated in Table 2. We utilized the policies applied by governments in the time domain from a publicly available source (Roser et al., 2020). We retrieved the policy information of United Kingdom for the reference time interval and define benchmarking strategies. In addition, we referenced personal communication of residents and news sources.

Table 2. Restrictions in United Kingdom by Stages

Stages	Stay home if sick	Wear facial masks	Social distancing	Limiting gatherings	Locked locations
Stage 0	False	False	None	None	None
Stage 1	False	False	None	Low risk: 0, High risk: 0	School, Hair Salon, Office, Retail Store, Bar
Stage 2	False	False	0.5	Low risk: 0, High risk: 0	School, Hair Salon, Office, Retail Store, Bar
Stage 3	True	False	0.7	Low risk: 0, High risk: 0	School, Hair Salon, Office, Retail Store, Bar

5. Experiments

The experiments are set to compare RL agents and the United Kingdom's policy against the COVID-19 pandemic. We limited our PandemicSimulator simulation environment to a small town of 1000 people and a hospital with a capacity of 10 people to conduct the experiment. When the simulation starts, one person is infected and no action is taken until at least five people are infected for the virus to spread sufficiently. While the simulation is running, the epidemic spreads from person to person in the environment as in reality. The applied stages of United Kingdom during the course of the pandemic include the various restrictions, such as no household mixing indoor, the closing of business, shops, and schools, mentioned, in detail, Section 4. Hyperparameters such as the contamination rate or the protection of the facial mask were used with reference to the calibration done by the original study from historical data.

We benchmarked RL agent’s policy that is trained with DQN with the policy of United Kingdom. We performed the experiments 30 times, starting with random seeds at each trial. We compared the two policies from different criteria. Main benchmarking criterion is to control the critical patient number under the hospital limit (should be less than 10 in our case). To compare the economic damage, we can use cumulative reward since it is calculated with negative economic impact along with the critical patient number. Figure 4 and Figure 5 illustrate results comparing the average performance of the RL agent and United Kingdom policy in 30 trials. Results show that, the policy, learned by RL agent, is better in terms of both critical patients and cumulative reward.

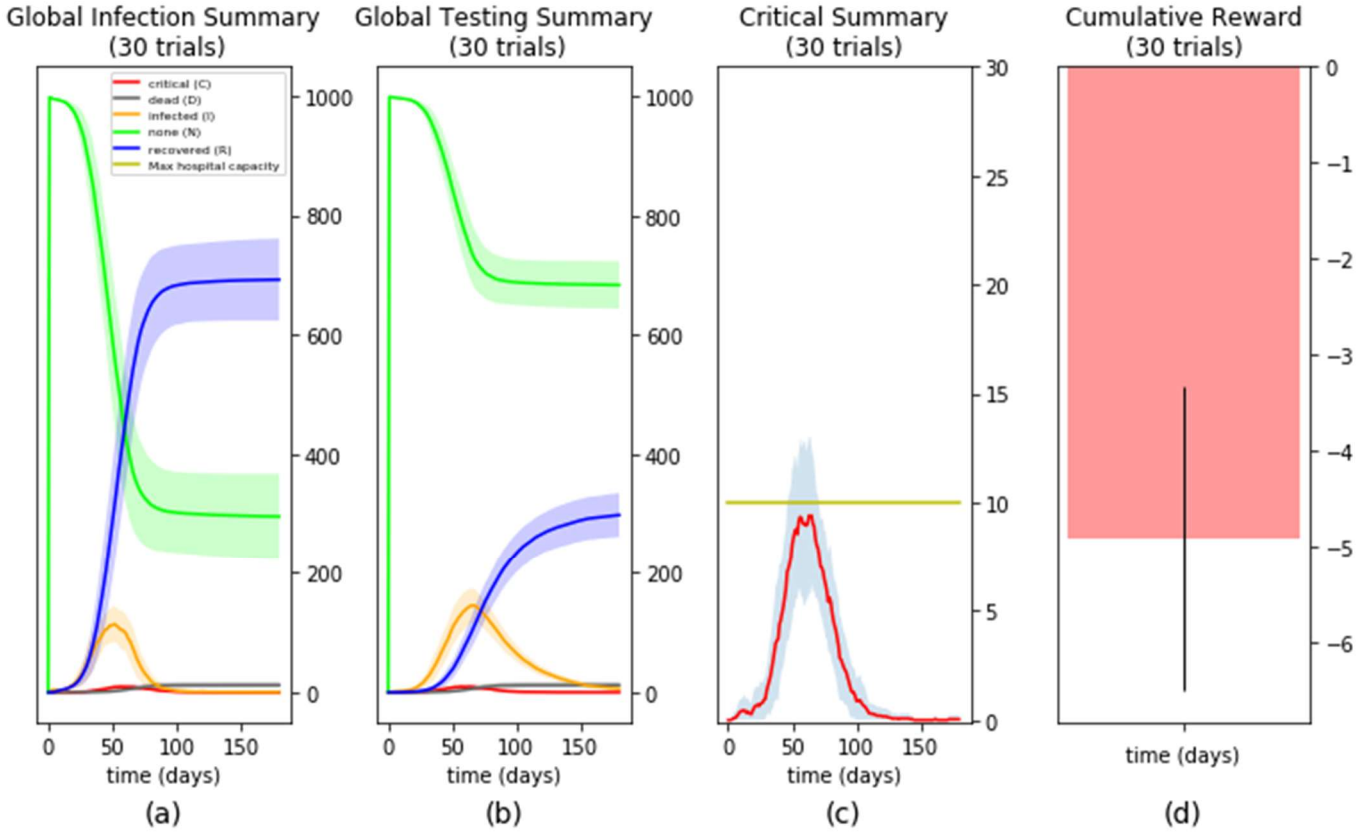

Figure 4. Results of RL agent's policy in 30 trials

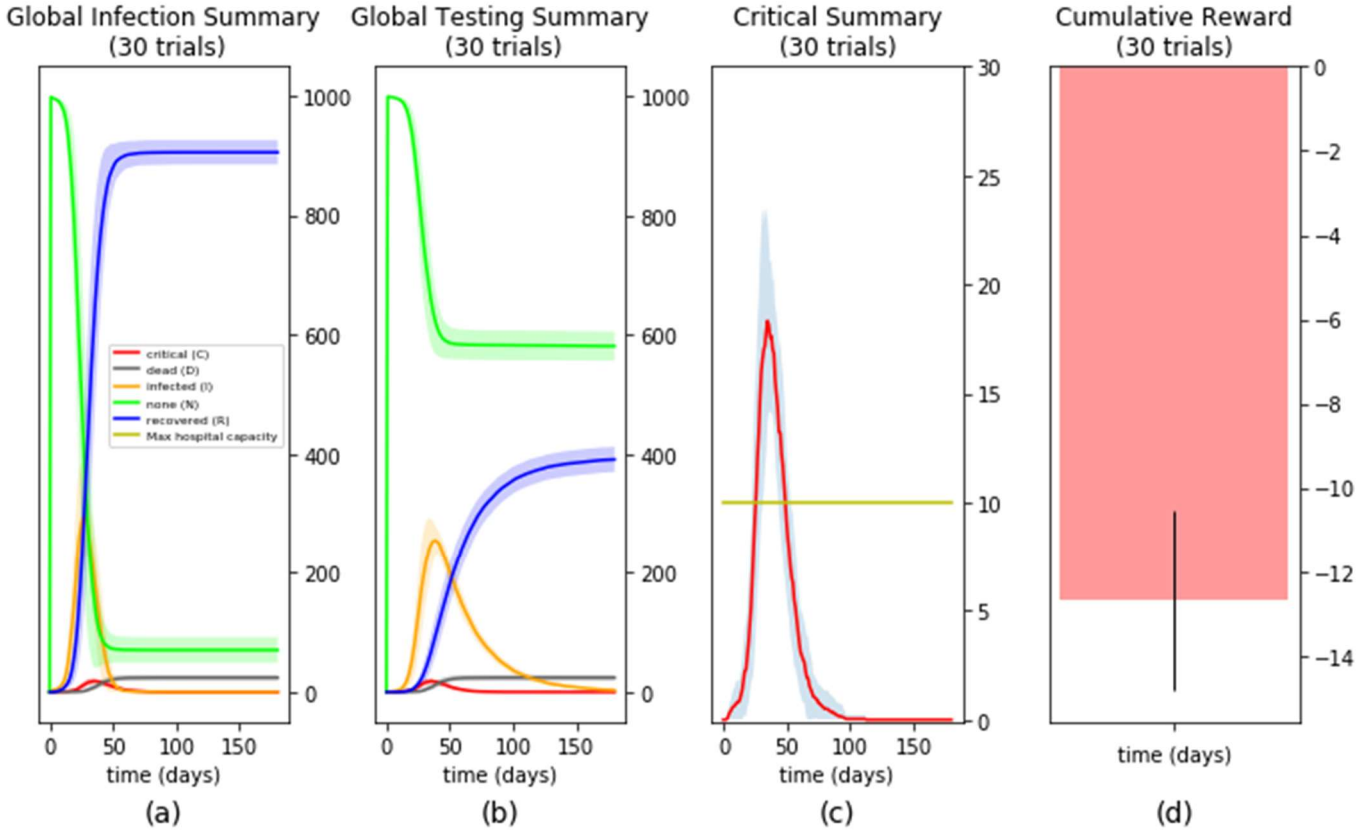

Figure 5. Results of United Kingdom's policy in 30 trials

6. Conclusion

COVID-19 emerged in last months of 2019 and affected millions of people worldwide. In a pandemic, economic difficulties and the number of critical hospitalizations are dependent on prevention policies. Hence, it is very important to observe the situation and to decide prevention policies based on this observation. Eventually, governments had to take measures to prevent spread of virus whereas avoiding economic downturn caused by policies against the spread. Since the consequences of these restrictions are not confident, the mitigation policies are always open for improvement in aspects of controlling the spread of the virus, reducing the intensity of healthcare, and preventing economic downturn.
We thank the reviewer for sharing this insightful remark. We used the hospital capacity parameter of the Simulator which is determined, in general. Unfortunately, we didn't perform an optimization procedure on those parameters with respect to the UK actual data. We agree with this valuable comment that suggests a better comparison in terms of fairness. In this context, we included the colored part below to the paragraph in the Conclusion Section as a future work suggestion.

We used RL algorithms to explore optimum mitigation policies that will minimize both the spread of the virus and its economic damage. Our benchmarking results show that RL framework can be utilized to achieve better policies than the United Kingdom in a controlled simulation environment and can be advisory to the authorities. Possible future works can be exploring off-policy algorithms and adding missing pandemic factors such as contact tracing and vaccinating to the simulator. In addition, a more fair and transparent comparison result can be achieved by adjusting the simulator parameters based on the population and hospital capacity data of the government of interest.

Acknowledgement

We would like to thank to Samira Balayoglu, to whom we referred as a personal communication regarding UK pandemic policies in this study, for her valuable contributions.

References

Aleta, A., Martin-Corral, D., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini Jr, I. M., Merler, S., et al. (2020). Modelling the impact of testing, contact tracing and house- hold quarantine on second waves of covid-19. Nature Human Behaviour, 4 (9), 964–971.

Awasthi, R., Guliani, K. K., Bhatt, A., Gill, M. S., Nagori, A., Kumaraguru, P., & Sethi, T. (2020). Vacsim: Learning effective strategies for covid19 vaccine distribution using reinforcement learning. arXiv preprint arXiv:2009.06602.

Arregui, S., Aleta, A., Sanz, J., & Moreno, Y. (2018). Projecting social contact matrices to different demographic structures. PLoS computational biology, 14 (12), e1006638.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., & Hollingsworth, T.D. (2020) How will country-based mitigation measures influence thecourse of the covid-19 epidemic? *The lancet*, *395* (10228), 931–934.

Bednarski, B. P., Singh, A. D., & Jones, W. M. (2020). On collaborative re- inforcement learning to optimize the redistribution of critical medical supplies throughout the covid-19 pandemic. *Journal of the American Medical Informatics Association*.

Dervisoglu, H., Bilgen, I, Halepmollasi, R., Can, B., Haklidir, M., (2021) Unfairness of Deep Learning Methods Arising Gender Bias in Covid-19 Diagnosis of Medical Images. *Artificial Intelligence Theory and Application,* 2, (Special Issue), 81-94.

Elgin, C., Basbug, G., & Yalaman, A. (2020). Economic policy responses to a pandemic: Developing the covid-19 economic stimulus index. *Covid Economics*, *1* (3), 40–53.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi- Velez, F., & Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine*, *25* (1), 16–18.

Hu, C., Lovejoy, W. S., & Shafer, S. L. (1994). Comparison of some control strategies for three-compartment pk/pd models. *Journal of Pharma- cokinetics and Biopharmaceutics*, *22* (6), 525–550.

Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., & Rodrigues,J. J. (2019). Identifying pneumonia in chest x-rays: A deep learning approach. Measurement, 145, 511–518.

Kompella*, V., Capobianco*, R., Jong, S., Browne, J., Fox, S., Meyers, L., Wurman, P., & Stone, P. (2020). Reinforcement learning for optimization of covid-19 mitigation policies.

Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. *arXiv preprint arXiv:1805.08298*.

Liu, C. (2020). A microscopic epidemic model and pandemic prediction using multi-agent reinforcement learning. *arXiv preprint arXiv:2004.12959*.

Martin-Calvo, D., Aleta, A., Pentland, A., Moreno, Y., & Moro, E. (2020) Effectiveness of social distancing strategies for protecting a community from a pandemic with a data driven contact network based on census and real-world mobility data. *Complex. Dig.*

Max Roser, E. O.-O., Hannah Ritchie, & Hasell, J. (2020). Coronavirus pan- demic (covid-19) [https://ourworldindata.org/coronavirus]. *Our World in Data*.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *CoRR*, *abs/1312.5602*. http://arxiv.org/abs/1312.5602

Najar, O., (2021) Hypothetical Framework For Early Detection of Covid19 From Symptomatic Information by Using Deep Learning. *Artificial Intelligence Theory and Application,* 2, (Special Issue), 115-121.

Organization, W. H. et al. (2020). Coronavirus disease 2019 (covid-19): Situation report, 82.

Ozbilen, M., Egriboz, E., Halepmollasi, R., Bilgen, I, Haklidir, M., (2021) A Deep Reinforcement Learning Approach to Explore Optimal Policies for Covid-19 Pandemic Mitigation: Preliminary Analysis. *Artificial Intelligence Theory and Application,* 2, (Special Issue).

Schaefer, A. J., Bailey, M. D., Shechter, S. M., & Roberts, M. S. (2005). Modeling medical treatment using markov decision processes. *Operations research and health care* (pp. 593–612). Springer.

Taubenberger, J. K., & Morens, D. M. (2006). 1918 influenza: The mother of all pandemics. *Revista Biomedica*, *17* (1), 69–79.

Yu, C., Liu, J., & Nemati, S. (2019). Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*.