Multi arm bandit optimization to mitigate covid19 risk

R. Aurachman1*
1Telkom University, Bandung, Indonesia
E-mail: rioaurachman@telkomuniversity.ac.id

Abstract. Covid19 pandemic has occurred since 2019. The spread of the virus prevents humans from doing their normal activities. The inability of people to have normal activities also has a bad effect on the economy. The government is trying to start opening public activities in the hope of sustaining economic growth. However, this policy risks doubling the spread of the virus. This research proposes a way to mitigate the risk of the pandemic using multi arm bandit optimization. With this approach, the system is guided to independently look for which sectors can be opened and which sectors should be closed temporarily.

1. Introduction
The Covid19 pandemic has made people reduce their activities outside the home. This slows economic growth and even risks making a country trapped in recession. Some economic sectors are regulated so that they cannot operate within a certain period to reduce the risk of transmission. On the other hand, society needs a better economy. With the limitation of activities and interactions between humans, it will make it difficult for economic conditions to improve. The community is asked to adapt and carry out activities using other means without face-to-face meetings and without close interaction. However, this requires a long learning process. Meanwhile, the impact that occurred due to the economic recession continued and caused many disadvantages.

The government has attempted to allow activity in the community. However, allowing the activity to be carried out runs the risk of causing re-infection to occur. The opening of a school, for example, can cause disease transmission between teachers and students. The reopening of markets for economic activities also creates the risk of disease transmission between traders and buyers. However, there are some places that have a high risk of transmission and there are also some places that have a low risk of transmission. Some places must also continue to operate because the sector is very important for people's daily activities, such as pharmacies, hospitals and shops for purchasing daily necessities. The amount of transmission risk cannot be known with certainty because the virus is not seen directly. The risks of a place can be identified if cases have occurred and reported to government authorities.

The unknown nature of the risk has similar attributes to the nature of the Multi Arm Bandit Optimization problem type. Multi arm bandit optimization is one type of Reinforcement Learning algorithm, which is a subset of Machine Learning. Reinforcement learning solves the problems of stochastic control using a simulation-based approach [1]. Reinforcement learning uses a trial and error approach in a dynamic environment [2]. Through this paper, we propose the application of Multi Arm Bandit Optimization as a decision support system in determining which economic sectors can be opened. One of the main factors in making this decision is the risk of transmission from the economic sector.

The applications related to covid19 are as follows; Observation and analysis of Covid 19 patient data using reinforcement learning [3] and other diseases [4], handling the spread of covid19 based on the decision of which city to lockdown [5][6], handling fake news that arises because of several machine
learning algorithms and reinforcement learning [7]. This research proposes a way to mitigate the risk of the pandemic using multi arm bandit optimization. With this approach, the system is guided to independently look for which sectors can be opened and which sectors should be closed temporarily.

2. Method
As shown in Figure 1, the method used to get this solution is a systems methodology. Problems are studied and analyzed using a systems approach [8]. From the problem formulation and problem situation, a structural approach can be formulated from a suitable mathematical model [9]. Problem formulation can be supported by the application of a rich picture and mind map. The mathematical model is then validated and verified whether it is sufficiently reliable when compared to the real system.

The model that has been obtained is then completed using the completion method that has been studied previously. The method can be a meta heuristic or an exact model. The model is then designed to be integrated and implemented into the decision making system. The decision support system recommends decisions about which sectors of the economy will be allowed to operate. The DSS developed includes three main components which can interact: user interface system, data management system and models management system [10].

![Figure 1. Research Method](image)

3. Result
Multi arm bandit Optimization is a type of Reinforcement Learning algorithm. The reinforcement learning algorithm, in contrast to other types of machine learning algorithms, has a punishment and reward feature. When the algorithm guides the decision to make the right decision, a reward is recorded. If in the process of finding a solution, the algorithm makes a decision to choose a wrong decision, the system accumulates a punishment calculation. With repeated processes, the system will slowly recommend the best decision. The system learns from the impact of reward and punishment from a decision, so that bad decisions will be chosen less often, and good decisions will be chosen more often.

An example of a simple illustration for the application of reinforcement learning is the selection of online advertising. Advertisements displayed online will be responded to by users via clicks (as rewards) and ignoring (as punishments). The system will replace the right ad to be displayed in online media thereby increasing the likelihood of clicks and sales. The system will continue to learn from the rewards and punishments received from website visitors. After several iterations, the featured ad will be selected and the non-superior ad will be eliminated and will not automatically be shown again. The system automatically learns and provides the best recommendations for the best results.

In the decision making for the selection of the economic sector which was opened in the midst of the Ovid 19 pandemic, the alternative decision was not the advertising design, but the economic sector that was opened. The variable that becomes the reward is the number of days during which the economic sector can run without any reports of the transmission of Covid 19. Meanwhile, the variable that becomes punishment is the number of infections that occur from an economic sector. The system is expected to automatically select safe economic sectors. The system is also designed to be able to continuously update the reward and punishment database. The reward and punishment database that occurs on the first day will be different from the reward and punishment database on the other days. On the following day, the reward and punishment data at the end of time is an accumulated reward and punishment since the first day.

This system is integrated in a large area. The updated system does not have to be based on a reward and punishment that comes from the decision-making area. For example, if one day it is proposed to open the economic sector in city A, the reward and punishment database used as a reference can get
information from other cities. Likewise, the results of opening an economic sector in city A will result in rewards and punishments that will be used for decision making in city B.

After the following paragraph, we describe an example of the initial form of multiarm bandit optimization programming using the Python programming language. This programming is adapted from datacamp.com.

```python
n = 10
sectorrisk = [0.5,0.2,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.9]
eps = 0.1

import random
def safetyscore(prob):
safetyscore = 0
for i in range(10):
    if random.random() < prob:
        safetyscore += 1
return safetyscore

import numpy as np
np.random.seed(5)
av = np.array([np.random.randint(0,(n+1)), 0]).reshape(1,2)
def bestSector(a):
    bestSector = 0
    bestMean = 0
    for u in a:
        avg = np.mean(a[np.where(a[:,0] == u[0])][:, 1])
        if bestMean < avg:
            bestMean = avg
            bestSector = u[0]
    return bestSector

import matplotlib.pyplot as plt
"%matplotlib inline"
print(sectorrisk)
```

4. Discussion
The algorithm of the decision making system can be expressed in the form of a flowchart in Figure 2. The flowchart contains several symbols that explain the logic of programming. The algorithm consists of two iterations. The first iteration is to evaluate the entire city or location. The second iteration is to evaluate all economic sectors in the city. When the entire city has been evaluated, the algorithm will direct the system to complete the iteration in that unit of time. The system will perform another analysis on each economic sector in a city. When a sector has been given recommendations, the system will direct the process for making recommendations for the next economic sector. The whole process begins with updating the reward and punishment database. Updates are made based on the information provided by the environment to the system. The information provided is whether there is transmission or not from every sector of the economy. This information will change the level of risk in an economic sector. Changes in risk affect changes in the recommendations that will be given.
It could be that at one time it is recommended that one of the economic sectors be opened. However, after the opening was carried out, there was an infection in the economic sector, so that the database was changed, and recommendations were given to close the related economic sector. The system independently analyzes existing data and makes better recommendations based on historical data. The reward and punishment database does not have to start with a zero value or without an initial preference. Users of the system can provide an initial assessment of which economic sectors are considered to have a high risk of transmission and which economic sectors are considered to have a low risk of transmission. This initial assessment could be a wrong judgment or it could be a correct judgment. The system will update the assessment according to the facts that occur at a later time.
5. Conclusion
Reinforcement learning can be considered as a Knowledge Management System. This system accumulates knowledge and information based on the experiences that occur in the system. The system provides decision recommendations and then the environment provides feedback. This feedback becomes a lesson for the system to be able to make better decisions in the future.

This process runs automatically so as to minimize the bias that humans have in making assessments. Based on the research that has been done, an algorithm and flow have been obtained to support the determination of the economic sectors that can be opened or closed in order to minimize the rate of transmission. The system is designed using the application of reinforcement learning, especially the Multi Arm Bandit Optimization algorithm or it can be seen from another perspective as a multi decision of single arm bandit optimization.

The next stage of this research is to test the system that has been designed on a simulation data. Where possible the system can be tested in real cases while being careful not to recommend wrong decisions to decision makers. Some other developments that can be done are using other Reinforcement Learning algorithms. We can combine reinforcement learning with deep learning [11]

Acknowledgement
We would like to thank Telkom University for providing resources and funds to carry out this research.

References
[1] Bhatnagar S, Prasad H and Prashanth L 2013 Reinforcement learning Lecture Notes in Control and Information Sciences
[2] Kaelbling L P, Littman M L and Moore A W 1996 Reinforcement learning: A survey J. Artif. Intell. Res.
[3] Chaganti S, Grenier P, Balachandran A, Chabin G, Cohen S, Flohr T, Georgescu B, Grbic S, Liu S, Mellot F, Murray N, Nicolaou S, Parker W, Re T, Sanelli P, Sauter A W, Xu Z, Yoo Y, Ziebandt V and Comaniciu D 2020 Automated Quantification of CT Patterns Associated with COVID-19 from Chest CT Radiol. Artif. Intell.
[4] Liu Y, Logan B, Liu N, Xu Z, Tang J and Wang Y 2017 Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017
[5] Khadilkar H, Ganu T and Seetharam D P 2020 Optimising Lockdown Policies for Epidemic Control using Reinforcement Learning Trans. Indian Natl. Acad. Eng.
[6] Khadilkar H, Ganu T and Seetharam D P 2020 Optimising Lockdown Policies for Epidemic Control using Reinforcement Learning: An AI-Driven Control Approach Compatible with Existing Disease and Network Models Trans. Indian Natl. Acad. Eng.
[7] Gradoñ K 2020 Crime in The Time of The Plague: Fake News Pandemic and The Challenges To Law-Enforcement And Intelligence Community Soc. Regist.
[8] Daellenbach H, McNickle D and Dye S 2012 Management Science
[9] Reyniers D and Taha H A 1989 Operations Research: An Introduction (4th Edition) J. Oper. Res. Soc.
[10] Haastrup P, Maniezzo V, Mattarelli M, Rinaldi F M, Mendes I and Paruccini M 1998 A decision support system for urban waste management Eur. J. Oper. Res.
[11] Mousavi S S, Schukat M and Howley E 2018 Deep Reinforcement Learning: An Overview Lecture Notes in Networks and Systems