

SIMULATION-BASED REINFORCEMENT LEARNING FOR SOCIAL DISTANCING

by

F. Ege Hoşgüngör

213150

A Dissertation

Submitted to the Department of Engineering & Informatics

University of Sussex

In Partial Fulfilment of the Requirements

For the Degree of Master of Science

June 2020

TABLE OF CONTENTS

[1 ACKNOWLEDGEMENT 2](#_Toc48663792)

[2 ABSTRACT 3](#_Toc48663793)

[3 INTRODUCTION 3](#_Toc48663794)

[4 RELATED WORK 5](#_Toc48663795)

[5 METHODOLOGY 7](#_Toc48663796)

[5.1 SOCIAL DISTANCING 7](#_Toc48663797)

[5.2 THEORETICAL FRAMEWORK 7](#_Toc48663798)

[5.3 RESEARCH DESIGN 7](#_Toc48663799)

[5.4 POLICY OPTIMIZATION 7](#_Toc48663800)

[6 RESULTS 7](#_Toc48663801)

[7 DISCUSSION AND CONCLUSION 7](#_Toc48663802)

[8 REFERENCES 7](#_Toc48663803)

[9 Appendix 8](#_Toc48663804)

# ACKNOWLEDGEMENT

The report may be freely copied and distributed provided the source is acknowledged.

Throughout the writing of this dissertation, I have received a great deal of support and assistance. I would first like to thank my supervisor, Prof. T. Nowotny, whose expertise was invaluable in the formulating of the research topic and methodology in particular. In addition, I would like to thank my parents for their wise counsel and hospitality during these rough times. Specially thanks to my friend Utku Demir for helping me out with Cloud Computing when my laptop broke down. You are always there for me. Finally, there are my friends, who were of great support in deliberating over our problems and findings, as well as providing happy distraction to rest my mind outside of my research.

# ABSTRACT

Through agent-based modelling, and standard reinforcement learning algorithms at scale, we found AI agents can give insights about ongoing epidemics by simulating the disease. An epidemic simulation has been created with a physics engine and analyzed its results with SIR graphs. We found clear evidence of the relation between social distancing and getting infected rates. We further provide evidence that multi agent cooperation may scale better with increasing environment complexity and lead to a behavior that closer to far more human behavior. Ya da maskenin takılma yüzdesi ne kadar etkiliyor bunu yazıcaz.

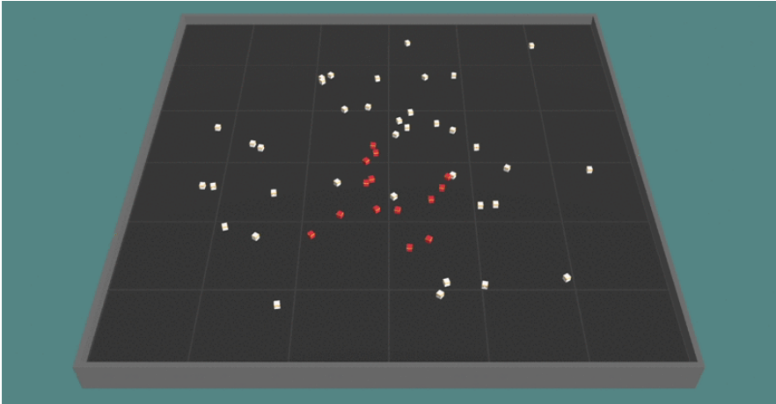
# KEYWORDS

Agent-based-modelling, reinforcement learning, cooperative multi-agent environment, epidemic simulation, social distancing, covid-19

<https://www.scribbr.co.uk/thesis-dissertation/abstract/>

# INTRODUCTION

Ever since the outbreak of Severe Acute Respiratory Syndrome as known as Covid-19 came out, life has changed drastically. Many researchers dedicated themselves to fight against the spread of this fatal virus and minimize the loss. Artificial Intelligence researchers are focusing their expertise knowledge to develop mathematical models for analyzing this epidemic disease.[1] An epidemic disease requires quick decisions to be made about interventions that could reduce or contain the disease spread. Decision-makers need to be agile about their strategy since they race with the time. Every second that is wasted doubles the damage to humanity. In contrast, in order to decide confidently which strategy will work, decision makers need to analyze many scenarios and variables since if they give a wrong decision, that can also cause harm. This is an optimization problem which researchers cope by creating their own data and utilizing Reinforcement Learning to develop optimal strategies.[2]

Reinforcement Learning is an area of Machine Learning where an agent learns the best behavior by interacting with the environment. Creating these complex environments and artificial intelligent agents that solve complex human relevant tasks has been life-long challenge for RL researchers.[3] Environment which is a crucial component of RL, describes the task that the agent attempts to solve. Agent and environments cannot be considered separately, and it is only a design choice for the researcher to determine where the environment starts and the agent ends. Most of the time environment is defined as anything that agent cannot have a direct control on it. In this article we introduce an environment which can simulate the epidemic spread with physics engine of Unity. With the help of Reinforcement Learning, agents are trained to learn social distancing by their own.

*Figure 1:The Environment and agents. The square shape court limits the area and the cubes are the agents. White ones represent healthy and red ones represent infectious agents. The simulation shows how one infected agent starts to spread the disease in each time step.*

Social distancing is one of the most effective precaution that individuals can apply to their daily life. It is a successful strategy to prevent the infectious disease from spreading. It has many forms but at its core, the aim is to keep people apart enough from each other by putting physical distance and/or confining them to their homes. In this article, we took the idea of social distancing and simplified it to physical interaction. In other words, we assume social distancing is just having a physical distance between individuals. An infection mechanism starts when individuals are closer to each other than the threshold value. Being closer means individual has higher chance to get infected. In following chapter, this mechanism will be explained in detail.

To analyze possible outcomes, we used SIR graphs which is a widely common mathematical model that provides insights about the infectious disease outbreak. The model divides individuals to three categories. Susceptible, Infectious and Recovered. Recovered ones can be also called as Removed since they don’t have any affect to simulation. We also used the same number of categories in our simulation. Although it has been proven that the individual has short-term immunity after recovering from the covid-19, the character of the virus is still ambiguous. To avoid ambiguity, we assume when individuals recovered from the disease, they will immune to it forever and after recovery, they don’t have any effect to the curve.

For our deep reinforcement learning task, we used Proximal Policy Optimization Algorithm (PPO). This algorithm is designed by OpenAI. [3] and it is used in numerous different tasks from robotics to atari games. On a collection of benchmark tasks, PPO outperformed other online policy gradient methods and had a better balance between sample complexity and simplicity. OpenAI defines their algorithm with three features. Easy code, sample efficiency and ease of tune.

We introduce a new cooperative multi-agent physics-based reinforcement learning environment for control of epidemic spread. Through only a health status-based sparse reward function, agents learn many human-relevant skills to protect themselves from epidemic outbreak including social distancing and self-isolation when they get sick. For example, agents learned how to maintain a balance between collecting reward boxes and not risking getting infected. We find that for several tests, curriculum learning can visibly change the results of the training. Setting the task step by step harder helped agents to learn better and to converge the loss function closer to the global minimum. In addition to that we showed pretrained agents learned the task faster than agents which is trained from scratch. Moreover, we observe signs of collaboration and simple communication between agents even though they don’t get direct reward from their actions. For instance, in one of the sessions, infected agents learned to gather up in a location where they avoid infecting others without knowing each other’s health status. This behavior demonstrates that they found a way to tell that they are infected or not to other agents.

The main contributions of this work are: 1) clear evidence that social distancing is mathematically a correct way to flatten the SIR curve. 2) A demonstration of how agent-based strategies and advances in computing can be leveraged to determine the optimal policy in an epidemic outbreak for a particular environment without expert human guidance.[4] 3) Evidence that self-isolation buralara bir şeyler gelmeli ve acil gelmeli



*Figure 2: Different states of the agents. Through the simulation, agent’s status of health changes. To represent the change, we used 4 different colors. a) White Bots indicates that the agent is not controlled by a brain. It only has simple hard-coded actions such as directly going targeted locations or bouncing from the walls. This represent individuals in a community who are not acting logically. b) Blue Bots are agents with a brain which controls them. c)Red Bots indicates*

*whether with brain or not the bot is infected. d)Purple indicates that agent is not infectious anymore. In SIR models’ purple agents call as Recovered-Removed. Please see* [[*https://github.com/Hsgngr/Pandemic\_Simulation*](https://github.com/Hsgngr/Pandemic_Simulation)](https://openai.com/blog/emergent-tool-use)*for example videos.*

# RELATED WORK

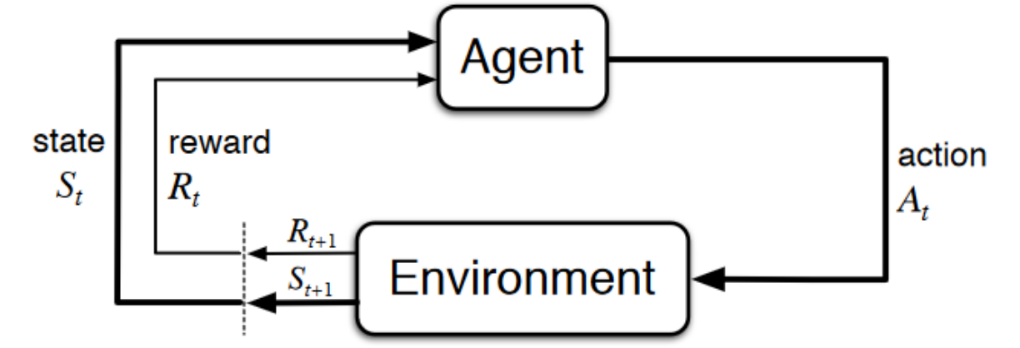
The history of agent-based modelling (ABM) can be traced back to the Simula programming language, which is developed in the mid-1960s and widely used as the first framework for automating step-by-step agent simulations.[5] ABM use simple rules which can result in different sort of complex behavior. These models consist of interacting rule-based agents to create real-world-like complexity. Back in the days, the rules were strictly defined by researchers as hard-coded and therefore it was hard to generalize and get interesting results. [6] Over time, an extensive literature has developed on creating simulations and a series of studies has indicated that there are two major approach in developing them.[7] At the one end is what we call the “ brute force” method which is basically includes designing every piece of the simulation. This method works much faster and it doesn’t need any AI training since every action is pre-decided. However, finding an optimal behavior of how the simulation works requires redoing large chunks of the model or even starting over, depending on the significance of the change. In addition to that, in some cases the wanted behavior is impossible to code. For example, in a real-world autonomous driving task, researchers create their own synthetic data with a single virtual camera which gets RGB images from the environment.[8]. They used augmented data in training since coding images by hand was simply impractical.

*Figure 3:Two-dimensional agent-based epidemic simulation visualization. Each square represents an individually programmable agent. Color-coding allows easy visual tracking of agents with different health status. (Microbial Threats to Health: Emergence, Detection, and Response 2003)*

The other end of this spectrum is an extremely flexible simulation in a way that researchers define rules as minimal as possible. This creates a suitable environment for AI.[7] In this kind of simulations, at first agents don’t have any assumptions about their environment. By trial and error, they develop an internal model which represents how they understand the system by observing the surrounding and collecting data. Letting the agent to create its own strategy rather than coding one, creates an opportunity to generate more comprehensive and complex behavior. Additionally, agent-based simulations generally have many parameters which requires to be tuned. By using brute force, it would take much longer to explore every possible scenario that can happen. On the other hand, utilizing sub-sampling and creating scenarios most likely to occur, AI can deliver dramatic speed increases for large-scale ABM simulations.

Mathematics and statistics have been crucial for analyzing infectious disease and control them since 1766 when Bernoulli published his evaluation about life expectancies and death rates.[9] Early work is formulated deterministic differential equations models for the transmission called SIR by Kermack and McKendrick[10]. Epidemiology is a science which interested investigating all the factors that determine the presence or absence of diseases and disorders. Earlier, the biggest obstacle in front of the epidemiology was not being an experimental science. Since the experiments were not practical nor ethical study populations were having limitations even though the discipline concerns itself with large populations of ill humans. Embracement of new powerful computational technologies to analyze, model, and simulate the dynamics of infectious disease has accelerated research in the field of epidemiology in 90’s.[11] Such simulations served as dry lab where new interventions could be designed, evaluated, and optimized on outbreaks, with many advantages for real-world epidemic prevention and control efforts.[6] The development of this new science lead a new interdisciplinary, collaborative area which consist of epidemiologists and other computationally oriented academic disciplines. In his paper, Hofmeyr gets inspiration from biological immune system to design a better computer security in the form of a network intrusion detection system called “ARTIS”. [12] He asserts immune system is a highly complex system and precisely tuned to detect and eliminate any infection disease. To create such a system, he uses a computer simulation in order to imitate immune system and apply for computer security. More recent work of modelling and simulating an epidemic spread classified in two categories: host and spread.[13] While host models investigate the effect of disease on individuals, spread models focuses on predicting how disease spread among group of people. In this paper we created a spread class simulation where we investigate how an infectious disease passes from one to another instead of how it effects the host individual which will be discussed in following chapter.

Sutton explains RL with following example: An Infant explores itself and its environment without any explicit teacher, but it does have a direct sensorimotor connection to its surroundings. By practicing same actions, it produces a wealth of information about cause and effect, about the consequences of actions, and about what to do in order to achieve goals. [8] Humans learn by interacting with their environment and try to make inference from their experiences. Whether this is learning how to ride a bike or hold a conversation, the process works the same. We have an awareness about our environment via our observations and we constantly try optimizing ourselves by predicting our action’s consequences. That’s how any human learn from interaction and it is a foundational idea underlying nearly all theories of learning and intelligence.[14] The problems which is solved by interacting them, can be defined as Reinforcement Learning tasks. These tasks are essentially closed-loop problems in a way that taken actions shape future inputs. At each step, agent comes to a state with a reward of . With its current policy, agent gives a decision and takes an action which gets the agent one step further in a state called . Agent doesn’t get any external support about what to do to solve the problem, but instead it learns from its own actions and consequences which may result with getting a reward. By taking the difference between actual and expected rewards it tries to optimize its policy and takes another action. Thus, the process consists of trial-and-error search.



*Figure 4: Simplified diagram of RL closed-loop process representation (Sutton, Introduction to Reinforcement Learning)*

Although Reinforcement Learning have been widely studied in the literature there has not been many studies about epidemic spread control with RL. In epidemic outbreaks, particularly this method of machine learning is beneficial since researchers are not limited with real data. In addition, a brute force approach to this problem is computationally intractable and inefficient, since calculating every states -even not at all useful ones- require large amount of computation power[15] In lieu of identifying optimal policies , other computational AI-related methods have been used. For instance, Big data is used for estimating the severity of seasonal influenza.[16]. Shrimp disease occurrence prediction has been made with neural networks and logistic regression[17] In parallel, a network-based contact-tracing model has been developed to learn about outbreak propagation in STD’s.[18] Even genetics algorithms is used for finding optimal vaccination strategies in influenza.[19] In [2], Yanez has created a baseline for how to design a reinforcement learning environment to represent the problem of epidemics and finding optimal interventions. In her study, agents represent the decision-makers such as governments, health institutions and the task is finding the optimal intervention strategy in 3 categories: preventive inverventions, treatment of disease and reduce-transmission interventions. The state includes infection rates, reproducibility etc. besides whether they are susceptible, infected or recovered. The action set includes mask-wearing, social distancing, contact tracing, closing schools, lockdown etc. The reward is given related to the death or infection spread rates depending on selection. We abstracted the idea of having epidemic simulation to physical form and changed the representation of the agents to people. Instead of intervention strategies which governments can take, we have investigated precautions which can be applied by individuals. The action set was also inspired from Yanez,[2] but rather than abstractly representing mask-wearing or social distancing, our task was physically show that agents can create these actions by their own. A recent study in 2018, [20] approach the problem of selecting optimal strategies for influenza as k-bandit problem, which is a common problem in reinforcement learning.[14][21]

In the context of Multi-agent, there have been many researches in the literature. Many reinforcement learning tasks involve the participation of more than one single agent which fall into the area of multi-agent reinforcement learning (MARL),.[22]. Recent years have witnessed astonishing advances in MARL such as OpenAI’s Hide and Seek game and Dota 2 AI OpenAI Five[3][23] These developments became possible due to the development of deep neural networks (DNNs). In multi-agent RL tasks, agents operate in a common environment each of which aims to optimize its own cumulative reward by interacting with the environment and other agents.[24] Due to the interaction between agents, the tasks complexity and according to that needed computation power can be increased exponentially. On the other hand, under the favor of the complexity it provides, multi-agents systems reserve a place in many areas from social science to finance.[25][26] Multi-agent algorithms can be divided into 3 categories; cooperative, competitive, and the combination of this two depending on the task which agents solve. In the cooperative settings, agents collaborate with each other while try to optimize the common long-term cumulative return. On the contrary, in competitive multiagent tasks, the cumulative reward of agents sums up to zero. The combination of these two also called as “Ecosystem” which is multiple interacting agents with independent reward signals. This kind of environments can be thought of as an environment with full of animals where some of them will collaborate and some of them will compete. Apart from interaction between agents in MARL, the basic framework of multi-agents differs from single agent settings in terms of stationarity. Agents improve their policies concurrently which creates self autocurricula [3] and the environment faced by agents become non-stationary in MARL.[27] Eliminating the stationary environment settings is also a choice of design. Although RL methods optimized themselves in stationary environments better than non-stationary ones they are restrictive and usually overfit the task. As researchers our goal was to get as close as possible to a real-world scenarios since almost every one of the real-life applications are non-stationary.[27] Furthermore in [28], authors discuss that MARL systems suffer from the curse of dimensionality also known as combinatorial nature of MARL. One way to overcome this difficulty in multiagent scenarios is the use of search parallelization which is possible with neural networks. In upcoming paragraphs we will discuss the advantage of using deep neural networks with approximation algorithms such as Proximal Policy Algorithm (PPO) in deep reinforcement learning (DRL).

Cooperation between agents fits the definition of surviving in an epidemic outbreak therefore we focused on developing a cooperative environment. Agents creates strategies even though they don’t share the same observations. The collaboration comes from having the common goal. Cooperative also called utilitarian agents can be seen as competitive too if they accidentally learn an aggressive policy work and stuck with it. Therefore, researcher needs to design carefully the dynamics of the environment. There are two types of learning strategy in cooperative multi agent RL. *Team learning* and *concurrent learning*. Team learning is a common and easy way to design agents where they learn same set of behavior.[29] The advantage of team learning approach is that it can use single agent machine-learning techniques which sidesteps the complexity of co-adaptation of several learners as it is the case in concurrent learning.

Team learning may be divided into two categories. *Homogeneous* and *heterogeneous* team learning.[29] In homogeneous team learning all agents have the same goals and actions. The only differences among them are their sensory observations and their current states which differentiate their decisions. They learn single agent behavior which is known by every agent in the environment. On the other hand, heterogenous team agents can develop more complex behaviors with different roles in team. Heterogeneous teams have larger actions space but generally converge to a better solution through agent specialization.[29] Choosing among these approaches depends on the reinforcement task and some problems do not require agent specialization. In our task we believed individuals do not need specialization and therefore we implemented a homogeneous team learning multi agents to represent the individuals in equal conditions. As an advantage of utilizing same brain, the search space is remarkably reduced during the training. Surprisingly in our case, as author mentions in [29] homogeneous agents learned to act heterogeneously do to development of sub-behavior that differs based on agent’s health status. This behavior is discussed in results chapter.

Single-brain olmalarından da bahset

* Cooperative Multi-Agent Literature
* Sparse-Reward Functions
* PPO History
* Curriculum Learning
* Simulation Environment OpenAI – Unity Benchmark MDRL a survey and critique of multiagent drl. Sayfa 29 Diğer enginelerden bahset
* Social Distancing – Hayvanlarda da var, psikolojide mantıklı
* Niye flocking gibi algoritmalar kullanmıyorum. – Aslında burada yine agent-based modellingde AI’ın önemini vurgulamış olucam.

Social Distancting hakkında bir şeyler -hayvanlarda da görülüyor. Ödül ceza sistemine göre RL.

Flocking- gibi başka distance ayarlama algoritmaları

Flocking

Early work explored

Was further explored

More recent work attempted

In the context of Multi agent

Multi agent (cooperation) -Food Collection

Epidemic Simulation -SIR Model RL in Covid

To the best of our knowledge, there has been no research recently done on physics-based epidemic simulation with reinforcement learning. Authors in [2] offer approaches about how to design RL environments for epidemic spread however it is neither agent-based nor physics-based simulation. Authors in [6],[11] use ABM simulations without utilizing the artificial intelligence.

# METHODOLOGY

* SOCIAL DISTANCING

Social distancing’in mekanizmasını anlat. Proximity mechanism

* THEORETICAL FRAMEWORK

The theoretical framework is the structure that can hold or support a theory of a research study. The theoretical framework introduces and describes the theory which explains why the research problem under study exists.

* RESEARCH DESIGN

RL can be roughly divided into Model-free and Model-based methods. In Model-based methods, researcher define a cost function to

Environment for Epidemic Control

Önce single agent çalıştırdım. Single agent çalışmanın avantajlarını anlat. Diğer agentlar işin içine girmiyor çok en azından task daha kolaylaşıyor. Böylelikle ilk başta setting’I daha sağlıklı kurmamı sağladı.

Multi-agentları işin içine katarak nasıl non-stationary hale getirdiğini anlat. [27]

RewardCube ekledim.

* POLICY OPTIMIZATION

Agent policies are optimized using Proximal Policy Optimization (PPO) and Soft-Actor Critic (SAC). Both algorithms are compared, and PPO is selected for this task. The training is performed using Unity engine and open source Unity ML-Agents Toolkit. The agents trained in single-agent environments which is located next to each other in scene but do not have any interaction between them. At execution time, each agent act by using only their own observations and at optimization time, we use all agents’ observations to update our policy. So even though 8 different environments are used during training there was only neural network as an output. In other words, agents share the same policy parameters but act and observe independently as each of them were in different states.

# RESULTS

# DISCUSSION AND CONCLUSION

We have demonstrated that an epidemic simulation with a simple infection mechanism, multi-agent cooperative environment and standard reinforcement learning algorithms at scale can induce agents to learn complex strategies and human-like behaviors. We observed many strategies from social distancing to self-quarantine that agents developed suggesting that it is possible to flatten the SIR curve by taking individual precautions in an epidemic outbreak.

Our results with epidemic simulation should be viewed as a proof of concept showing a agent-based simulation with reinforcement learning can be use to assist decision makers during the epidemic.

# REFERENCES

[1] N. S. Punn, S. K. Sonbhadra, and S. Agarwal, “COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms,” *medRxiv*, p. 2020.04.08.20057679, 2020.

[2] A. Yañez, C. Hayes, and F. Glavin, “Towards the control of epidemic spread: Designing reinforcement learning environments,” *CEUR Workshop Proc.*, vol. 2563, pp. 188–199, 2019.

[3] B. Baker *et al.*, “Emergent Tool Use From Multi-Agent Autocurricula,” 2019.

[4] O. Bent, S. L. Remy, S. Roberts, and A. Walcott-Bryant, “Novel Exploration Techniques (NETs) for Malaria Policy Interventions,” *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 7735–7740, Dec. 2017.

[5] B. C. Perley, “Programming Language Maintenance,” in *Defying Maliseet Language Death*, UNP - Nebraska, 2017, pp. 63–84.

[6] E. C. Mark S. Smolinski, Margaret A. Hamburg, and Joshua Lederberg, on E. M. T. to H. in the 21st C. B. on Global, and H. I. of Medicine, *Microbial Threats to Health*. Washington, D.C.: National Academies Press, 2003.

[7] B. Zeigler, A. Muzy, and L. Yilmaz, “Artificial Intelligence in Modeling and Simulation,” in *Encyclopedia of Complexity and Systems Science*, New York, NY: Springer New York, 2009, pp. 344–368.

[8] B. Osiński *et al.*, “Simulation-based reinforcement learning for real-world autonomous driving,” 2019.

[9] Dietz and H. JAP, “Bernoulli was ahead of modern epidemiology,” *Nature*, 2000.

[10] Robert E. Serfling, “Summary for Policymakers,” in *Climate Change 2013 - The Physical Science Basis*, vol. 53, no. 9, Intergovernmental Panel on Climate Change, Ed. Cambridge: Cambridge University Press, 2019, pp. 1–30.

[11] J. S. Koopman, “Emerging objectives and methods in epidemiology.,” *Am. J. Public Health*, vol. 86, no. 5, pp. 630–632, May 1996.

[12] S. A. Hofmeyr and S. Forrest, “Architecture for an artificial immune system.,” *Evol. Comput.*, vol. 8, no. 4, pp. 443–473, 2000.

[13] M. Shatnawi, S. Lazarova-Molnar, and N. Zaki, “Modeling and simulation of epidemic spread: Recent advances,” *2013 9th Int. Conf. Innov. Inf. Technol. IIT 2013*, no. March, pp. 118–123, 2013.

[14] S. Richard and A. G. Barto, “An introduction to reinforcement learning,” *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions*. pp. 63–80, 2011.

[15] W. J. M. Probert *et al.*, “Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 374, no. 1776, p. 20180277, Jul. 2019.

[16] L. Simonsen, J. R. Gog, D. Olson, and C. Viboud, “Infectious disease surveillance in the big data era: Towards faster and locally relevant systems,” *J. Infect. Dis.*, vol. 214, no. Suppl 4, pp. S380–S385, 2016.

[17] P. Leung and L. T. Tran, “Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression,” *Aquaculture*, vol. 187, no. 1–2, pp. 35–49, Jul. 2000.

[18] K. T. D. Eames and M. J. Keeling, “Contact tracing and disease control,” *Proc. R. Soc. B Biol. Sci.*, vol. 270, no. 1533, pp. 2565–2571, 2003.

[19] R. Patel, I. M. Longini, and M. E. Halloran, “Finding optimal vaccination strategies for pandemic influenza using genetic algorithms,” *J. Theor. Biol.*, vol. 234, no. 2, pp. 201–212, 2005.

[20] P. Libin *et al.*, *Machine Learning and Knowledge Discovery in Databases*, vol. 11053, no. June. Cham: Springer International Publishing, 2019.

[21] T. Alamo, D. G. Reina, and P. Millán, “Data-Driven Methods to Monitor, Model, Forecast and Control Covid-19 Pandemic: Leveraging Data Science, Epidemiology and Control Theory,” pp. 1–65, Jun. 2020.

[22] K. Zhang, Z. Yang, and T. Başar, “Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms,” pp. 1–72, Nov. 2019.

[23] OpenAI *et al.*, “Dota 2 with Large Scale Deep Reinforcement Learning,” Dec. 2019.

[24] L. Busoniu, R. Babuska, and B. De Schutter, “A Comprehensive Survey of Multiagent Reinforcement Learning,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[25] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, “Multi-agent reinforcement learning in sequential social dilemmas,” *Proc. Int. Jt. Conf. Auton. Agents Multiagent Syst. AAMAS*, vol. 1, pp. 464–473, 2017.

[26] J. W. Lee, J. Park, J. O, J. Lee, and E. Hong, “A multiagent approach to Q-learning for daily stock trading,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*, vol. 37, no. 6, pp. 864–877, 2007.

[27] S. Padakandla, P. K. J, and S. Bhatnagar, “Reinforcement Learning in Non-Stationary Environments,” *Appl. Intell.*, May 2019.

[28] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, “A Survey and Critique of Multiagent Deep Reinforcement Learning,” *Auton. Agent. Multi. Agent. Syst.*, vol. 33, no. 6, pp. 750–797, Oct. 2018.

[29] L. Panait and S. Luke, “Cooperative Multi-Agent Learning: The State of the Art,” *Auton. Agent. Multi. Agent. Syst.*, vol. 11, no. 3, pp. 387–434, Nov. 2005.

# Appendix