Annals of Operations Research (2023) 328:245–277
https://doi.org/10.1007/s10479-022-04926-7

ORIGINAL RESEARCH

A simulation-deep reinforcement learning (SiRL) approach for epidemic control optimization

Sabah Bushaj[1] · Xuecheng Yin[2] · Arjeta Beqiri[1] · Donald Andrews[3] · İ. Esra Büyüktahtakın[4]

Accepted: 28 July 2022 / Published online: 26 September 2022

Abstract

In this paper, we address the controversies of epidemic control planning by developing a novel Simulation-Deep Reinforcement Learning (SiRL) model. COVID-19 reminded constituents over the world that government decision-making could change their lives. During the COVID-19 pandemic, governments were concerned with reducing fatalities as the virus spread but at the same time also maintaining a flowing economy. In this paper, we address epidemic decision-making regarding the interventions necessary given of the epidemic based on the purpose of the decision-maker. Further, we intend to compare different vaccination strategies, such as age-based and random vaccination, to shine a light on who should get priority in the vaccination process. To address these issues, we propose a simulation-deep reinforcement learning (DRL) framework. This framework is composed of an agent-based simulation model and a governor DRL agent that can enforce interventions in the agent-based simulation environment. Computational results show that our DRL agent can learn effective strategies and suggest optimal actions given a specific epidemic situation based on a multi-objective reward structure. We compare our DRL agent's decisions to government interventions at different periods of time during the COVID-19 pandemic. Our results suggest that more could have been done to control the epidemic. In addition, if a random vaccination strategy that allows super-spreaders to get vaccinated early were used, infections would have been reduced by 32% at the expense of 4% more deaths. We also show that a behavioral change of fully quarantining 10% of the risky individuals and using a random vaccination strategy leads to a reduction of the death toll by 14% and 27% compared to the age-based vaccination strategy that was implemented and the New Jersey reported data, respectively. We have also demonstrated the flexibility of our approach to be applied to other locations by validating and applying our model to the COVID-19 case in the state of Kansas.

✉ İ. Esra Büyüktahtakın
esratoy@vt.edu

[1] Department of Management Information Systems and Analytics, School of Business and Economics, SUNY Plattsburgh, Plattsburgh, NY, USA

[2] Yale School of Public Health, New Haven, CT, USA

[3] Trinity College Dublin, School of Natural Sciences, Dublin, Ireland

[4] Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA

Keywords COVID-19 · Epidemic control planning · Deep reinforcement learning · Agent-based simulation · Simulation-deep reinforcement learning · Vaccination strategy

1 Introduction

Coronavirus disease of 2019 (COVID-19) very quickly paralyzed the world as we know it. After starting as a local epidemic, in a short time, it reached across the world and was declared a pandemic by World Health Organization (WHO) on March 11, 2020. As of June 29, 2021, 33,640,572 individuals have been infected, and 604,115 have died from the COVID-19. No government had it easy to come up with regulations and interventions. Apart from the loss of human lives, COVID-19 also caused an economic recession in the world economy. The global stock market has experienced the worst crash since 1987 (Jones et al. 2021), and the International Labor Organization estimated a loss of 400 million full-time jobs across the world (McKeever 2020). COVID-19's economic impact is also felt in agriculture (Poudel et al. 2020), manufacturing (Tareq et al. 2021), arts and sports (Grix et al. 2021), and tourism (Sigala 2020). Even the United States (US) was hit hard by the disruption of supply chains (Nikolopoulos et al. 2021), change of lifestyle (Giuntella et al. 2021), and limited resources (Galanakis et al. 2021). This disruption was accompanied by a huge job loss (Bell and Blanchflower 2020). The effects of COVID-19 on human health and economies led to many controversies. Former US President Trump declared a national emergency and accepted that his administration "played it down" to not cause panic in public after the WHO declared the pandemic. On March 16, the Trump government puts in place the first interventions stopping gatherings of more than ten people and canceling non-essential trips for the next 15 days. In fear of an economic failure, this is the closest thing to a nationwide shutdown that the US implemented. What is worse, a tentative reopening provoked a harsher spread of infections, kindling the debates regarding how the government handled the pandemic (Ashraf 2020).

At the end of a dark year, hope arose when the first vaccines were introduced. On December 11, 2020, the US Food and Drug Administration (FDA) authorized the Pfizer-BioNTech vaccine for emergency use, and just a week later, on December 18, the Moderna vaccine was authorized as well. The government started vaccination using age and comorbidity-based strategies aiming to protect individuals more prone to the critical effects of a COVID-19 infection. Another discussion arose as to whether a strategy where the super-spreaders, individuals contributing the most to the spread of the virus, were targeted first (or at least were allowed to get vaccinated) could have higher benefits. Moghadas et al. (2021) study a vaccination strategy with a delayed second dose due to the limited vaccine supply. Their experiments show that a delay of 9 weeks for the second dose of the Moderna vaccine could avert at least an additional 17.3 infections per 10,000 population and reduce deaths by 0.34 individuals per 10,000 population compared to the four-week interval between the two doses. Gupta and Morain (2021) investigate different prioritization approaches and assess the likeliness of those approaches to reduce morbidity and mortality.

In this study, our goal is to develop an approach incorporating two components: a decision-maker and an evaluation mechanism of the decisions taken. An agent-based simulation is a quite suitable approach to mimic epidemic spread and population movements and quantify interventions (Shamil et al. 2021; De Mooij et al. 2021). Hence, we use an agent-based model as an evaluation of the interventions' impact on the population's health status. Kerr et al. (2021) provide an agent-based model, namely Covasim, which is an online simulation involving multiple characteristics of individuals in a population and different contact layers for each individual where interventions can be applied at any particular point and between any group of agents. Such bottom-up modeling of human interaction in different environments transcends other simulation models in terms of representing stochastic situations. Considering the high usability of, and advancements in reinforcement learning (RL) (see, e.g., Bushaj and Büyüktahtakın 2021; Delarue et al. 2020; Kong et al. 2018) and the computational limitations of mathematical optimization models to deal with large populations, we employ a deep reinforcement learning (DRL) agent as a governing decision-maker that can intervene in the simulation and apply available measures to change the course of an epidemic. DRL models have been useful in many decision-making environments across different fields such as healthcare (Zhou et al. 2021), policy-making (Lin et al. 2020), autonomous systems (Chen and Chan 2021), and logistics (Joe and Lau 2020). DRL models lack the conciseness of a typical mathematical program but are useful in learning and quickly providing policies that aim to improve a certain objective. To do this, we propose a Simulation-Deep Reinforcement Learning (SiRL) framework where an agent-based simulation model is integrated inside a reinforcement learning environment. We set out to construct a model that can capture the details of human individualism and present them as aggregate information to a general decision-maker. Thus, we believe this approach is a natural representation of the relationship between a government and its constituents. It combines the stochasticity of individual decisions and the data aggregation to a system governor that can generate a policy for the greater good of the whole system.

The structure of the rest of the paper includes the related work in Sect. 2, followed by the details on the agent-based model (ABM) simulation and deep reinforcement learning (DRL) environment in Sects. 3 and 4, respectively. Further, in Sect. 5, we integrate the ABM and DRL approaches into a SiRL framework. Finally, we show our experiments and results in Sect. 6 and conclude the paper in Sect. 7.

2 Related work

In many real-world problems, it is difficult to obtain necessary data and reproduce complex situations. Hence, simulation has been a useful methodology to express the environment with all its variables and dynamics. The choice of modeling is highly dependent on the type of the problem, the complexity of the problem, and the decision-makers' requirements.

Agent-based simulation has emerged and matured over the last 20 years, expanding both its realm of applications and its sophistication as technology and computing have improved. Drawbacks of agent-based modeling are the lack of easy-to-use software and implementation and the amount of time needed to come up with a structured and detailed environment. System dynamics (SD) are powerful in designing systems that can illuminate behavior and provide policies. For example, former SD studies have modeled relevant biological and behavioral mechanisms as well as critical feedback processes to make empirical estimates of the COVID-19 progression (Rahmandad et al. 2021; Ghaffarzadegan and Rahmandad 2020). Rather than rely on the SD approaches that model system-level disease dynamics, our main motivation to use an agent-based simulation lies in its capability to capture individual-level disease dynamics and the stochasticity in human contact networks and behavior. Because ABM is a bottom-up approach, we are focused on providing agents (individuals) full action power, which models our system according to the individual effect in the environment. Agent-based simulation can be utilized to predict epidemic trends and dynamics (Müller et al. 2021; Kieu et al. 2020), evaluate containment strategies and intervention decisions (Kerr et al.

2021; Shamil et al. 2021; Hinch et al. 2020; Alzu'bi et al. 2021), and mitigate risks of reopening (D'Orazio et al. 2020; Li et al. 2021). Agent-based modeling can be quite helpful in representing real-world interactions of populations and offer a decision-maker the chance to intervene and evaluate the outcomes of each decision. Epstein (2009) suggests that ABM is perfectly suitable for modeling the dynamics of an epidemic across a population. In the context of COVID-19, agent-based modeling has the potential to assist public health officials in responding to outbreaks with an appropriate level of intervention while minimizing the economic impact of those restrictions.

Among recent ABM simulations, the Covasim (COVID-19 Agent-based Simulation) developed by Kerr et al. (2021) models the dynamics of COVID-19 spread in a population by considering demographics based on age, different transmission characteristics among contact layers, and specific viral properties of the disease itself. Kerr et al. (2021) model human contact from different environments in a very effective form, capturing system dynamics and the uncertainty associated with them. Covasim has been very effective in simulating disease spread and comparing the simulation with other offered non-pharmaceutical interventions, such as social distancing, reducing contacts, testing, contact tracing, and quarantining. Li et al. (2021) extend the Covasim by also implementing a vaccination strategy and performing simulations according to Operation Warp Speed (an intervention proposed by the former Trump administration) to facilitate and accelerate the development, manufacturing, and distribution of vaccines and diagnostics and the plan of one million vaccines per day, proposed by the Biden administration. During the current pandemic, different countries have tried to implement measures to deal with the epidemic despite having scarce medical resources, all while aiming to lower the spread of COVID-19 and minimize the economic and human costs of the epidemic. Most countries have tried to keep a balance and optimize decision-making based on their available resources.

Typically, epidemiological methods can be categorized into compartments, e.g., Susceptible (S), Infected (I), and Recovered (R) (Kermack and McKendrick 1927) and agent-based (Epstein 2009; Kerr et al. 2021). Compartmental models tend to be faster, while agent-based models are slower and more complex. The advantage of ABM is that it can build a realistic model by capturing the stochasticity in the system by expressing the relationship between individual agents.

In compartmental models, individuals progress through the compartments which are distinguished based on the population's health status. These models are often run with ordinary differential equations and are useful in predicting how disease spreads, estimating effective reproductive number, and investigating how different interventions affect the epidemic spread. Giordano et al. (2020) present an extended compartmental model, SIDARTHE, which discriminates between infected individuals, thus presenting a realistic view of the diagnosed infections, their severity, and non-diagnosed individuals. Higazy (2020) models the COVID-19 pandemic using a fractional-order model of SIDARTHE and predicts the evolution of the pandemic to understand the impact of possible plans that can reduce the diffusion with different values of the fractional order. Such models do a very good job in modeling infectious diseases but are generally deterministic. An individual in the SIDARTHE model is defined by the compartment they belong to and the probability of moving to other compartments. These probabilities are defined by the disease severity and progression. Specifically, the infected individuals are moved to the *Diagnosed* compartment with a probability ε. The agent-based models (e.g., Covasim) have a similar underlying compartmental structure, but in addition to that, an individual is represented as a more complicated and well-represented entity. Individuals have contact networks (household, school, workplace), existing conditions, specified viral loads, and a certain age. This allows for individualism, meaning that two individuals in the same compartment do not necessarily have the same characteristics. Due to this heterogeneous and individual-based representation of the epidemic dynamics, adopting an agent-based model can better express the stochastic nature of individual decisions.

In addition to simulation studies that approximate the dynamics of an epidemic, mathematical optimization has often been used for decision-making to control epidemic outbreaks. In an epidemic situation, proper resource allocation contributes to better public health outcomes as well as to a healthier economy. Different mathematical programming methodologies are presented to tackle the resource allocation challenges in a pandemic, such as mixed-integer programming (Büyüktahtakın et al. 2018), multi-stage stochastic programs (Bushaj et al. 2022; Yin and Büyüktahtakin 2021; Bushaj et al. 2020; Yin and Büyüktahtakın 2022; Kıbış et al. 2020), stochastic programs (Tanner et al. 2008; Mehrotra et al. 2020), and approximate dynamic programming (Coşgun and Büyüktahtakın 2018). Dasaklis et al. (2012) critically review the roles of logistics operations and their management in epidemic control and identify possible literature gaps. They claim that the issue of epidemic control in the supply chain literature is fragmented. Most of the available frameworks have very little correlation to the real world scenarios, and the applicability of the modeling approaches is limited. Queiroz et al. (2020) prepare a detailed review on the impacts of epidemic outbreaks in supply chains and present a series of open research questions to frame a research agenda for scholars and practitioners. In addition, they identify multiple suitable approaches to support supply chain responsiveness, adaptation, and sustainability. Among others, they claim that a combination of simulation theories with dynamic capabilities could make up for complex scenarios to cope with resource scarcity and sequential decisions throughout the pandemic.

Büyüktahtakın et al. (2018) propose a mixed-integer programming formulation that integrates epidemic dynamics into a logistics model to project the disease growth while minimizing the total number of infections and fatalities from the Ebola outbreak in West Africa. They provide insights regarding intervention timing and intensity for each region in Guinea, Liberia, and Sierra Leone. Yin and Büyüktahtakin (2021) present a multi-stage stochastic programming compartmental model to tackle the uncertain disease progression and resource allocation in an infectious outbreak. They introduce equity constraints in their model and apply them to the Ebola disease spread in West Africa. Yin et al. (2023) present a risk-averse multi-stage stochastic epidemics-ventilator-logistics compartmental model addressing the resource allocation changes of COVID-19. The authors modify the lower and upper bounds of Büyüktahtakın (2022) to region-based bounds to tackle problem complexity. Their results show that short-term migration significantly influences the disease transmission. Ventilator allocation depends on multiple factors, including initial infections, ICU capacity, the population of a geographic location, and the availability of the ventilators.

Optimization models that oversee the impact of all possible interventions and budget allocation scenarios on the growth of the disease simultaneously (see, e.g., Büyüktahtakın et al. 2018; Yin and Büyüktahtakin 2021; Bushaj et al. 2020, 2022; Yin and Büyüktahtakın 2022; Kıbış and Büyüktahtakın 2019) are powerful tools to model epidemic logistics and optimize decision strategies for resource allocation. Such operations research (OR) approaches focus on modeling disease dynamics on a large-scale population over multiple regions and time periods.

However, optimization models in combination with agent-based simulations can be extremely difficult to solve. When we focus on a specific population and heterogeneity among disease compartments such as age-specific transmission rates, agent-based models could capture individual-level interactions and detailed disease dynamics better than mathematical programming models. However, in that case, agent-based models should be supported by a powerful optimization tool.

Deep Reinforcement Learning (DRL) has lately been very attractive for evaluating optimal policies based on a given situation. In the last decade, Reinforcement Learning (RL) has shifted from the use of tabular formats of actions and states (Watkins and Dayan 1992; Hasselt 2010) to the usage of Deep Neural Networks (DNN) due to their immense benefits. The use of DNN in RL has led to advances, such as Deep Q-Learning (Schaul et al. 2015), Double Deep Q-Learning (Van Hasselt et al. 2016), and Actor-Critic Methods (Mnih et al. 2016; Wu et al. 2017). DRL has proven its strength in various applications such as games (Mnih et al. 2013; Silver et al. 2018), combinatorial optimization (Bushaj and Büyüktahtakın 2021; Delarue et al. 2020), and healthcare (Mahmud et al. 2018).

Due to the devastating COVID-19 pandemic, recent studies have already used DRL to help in different applications related to COVID-19 (Kompella et al. 2020; Wan et al. 2020; Bednarski et al. 2020). Kompella et al. (2020) aim to use RL to optimize decisions during the pandemic in a way that minimizes the economic impact and keeps hospitals at a normal capacity. Bednarski et al. (2020) investigate the use of deep learning models to provide near-optimal distribution of healthcare equipment to better deal with public health crises similar to COVID-19. Awasthi et al. (2020) tackle the problem of distributing a limited vaccine supply by using a sequential decision strategy based on RL. They propose VacSIM which formulates sequential decision-making into a Contextual Bandits approach to optimize the distribution of the COVID-19 vaccine. They claim that up to 9,039 additional lives could be saved when evaluating their policy against a naive distribution policy. Ohi et al. (2020) implement a DRL agent based on a short-term memory DDQN to learn an optimal policy for maintaining a balance between mitigating epidemic spread and economic cost. Khalilpourazari and Doulabi (2021b) use reinforcement learning as a facilitator to solve a compartmental model (SIDARTHE) in a reasonable time. Khalilpourazari and Doulabi (2021a) design a hybrid reinforcement learning approach that combines the benefits of machine learning and evolutionary computation. They claim their approach exploits the solution space very intelligently, accelerating the algorithm and enabling them to resolve complicated large-scale problems. These studies present an interesting use of RL in enhancing state-of-the-art methodologies. However, they differ from our study because, in their case, RL is used as an aid, while in our study RL is the main decision-maker and has full acting power to change the disease dynamics.

In essence, Simulation Optimization (SO) is the optimization of an objective subject, so some constraints and system dynamics are updated using a simulation. Gillisa et al. (2021) propose a simulation-optimization framework that combines an age-based SEIR compartmental simulation model and a genetic algorithm to discover good strategies and optimize intervention strategies. They extract insights from the COVID-19 pandemic to aid policymakers in making closure, protection, and travel decisions by minimizing the total number of infections under a limited budget. Their results highlight that social distancing and wearing masks are of the highest importance, while closures and travel restrictions are more flexible policy restrictions. Onal et al. (2021) extend the simulation-optimization framework of Onal et al. (2020) to search and treat invasive species under a limited budget and completely random dispersal (Büyüktahtakın and Haight 2018). The simulation is responsible for representing the growth of the invader spatially for up to 25 years, and then the optimization model finds an optimal response such that it minimizes the economic damage caused by the invader.

Simulation-optimization studies using agent-based modeling can be very effective but often suffer from the dimensionality curse. Specifically, applying those models to a large population not only becomes more challenging to simulate, but might also become impossible to optimize. To overcome this challenge, recent studies have used RL-based techniques to utilize the information obtained from the agent-based simulation. Several studies present distinct frameworks combining agent-based models with DRL tools to explore decision-making options. Ohi et al. (2020) implement a simulation model which serves as a virtual environment for training a DRL agent to make non-pharmaceutical decisions based on a specific situations within the epidemic. They demonstrate how agents select possible available actions to reduce the spread of the disease while still considering the economic factors. They present different lockdown strategies that the DRL agent undertakes to halt the propagation of the disease. Kompella et al. (2020) present a pandemic simulator that models the epidemic spread, including the interactions between individuals in a community, testing with false positive/negative rates, imperfect public adherence to social distancing measures, and contact tracing. They then use an RL-based methodology to optimize mitigation policies within the pandemic simulator.

Inspired by these achievements of DRL, our goal is to develop a self-sufficient framework that fully represents the relationship between the evolution of a disease in a population with individual-level interactions and the government's intervention actions to control an outbreak. We propose a Simulation-Deep Reinforcement Learning (SiRL) approach to epidemic disease modeling and decision-making where the simulation is agent-based and optimization is handled by a DRL agent based on environment compartmental data.

The Covasim methodology of Kerr et al. (2021) has been successfully used to represent the realism of the COVID-19 pandemic and make future predictions. Hence, we extend the open-source simulation to better fit with the simulation strategy inside our SiRL framework. Covasim is a stochastic agent-based simulator developed by researchers from the Institute for Disease Modeling, Global Health Division, Bill & Melinda Gates Foundation in the U.S, Burnet Institute in Austria, and Big Data Institute at the University of Oxford, United Kingdom, to analyze COVID-19. The simulations provide projections regarding the numbers of infections and peak hospital demand and help to explore the potential impact of different interventions, including social distancing, school closures, testing, contact tracing, and quarantining.

Covasim is used extensively in the literature for spatio-temporal simulation (Gharakhanlou and Hooshangi 2020) to derive strategies for non-pharmaceutical interventions (Contreras et al. 2021), and to evaluate reopening strategies (Bilinski et al. 2021).

2.1 Key contributions

This paper provides the following contributions in terms of the simulation and reinforcement learning models and their integration as well as insights into decision making to control the COVID-19 epidemic.

2.1.1 Simulation

In Covasim, all the details for each intervention are defined at the start of the simulation. We extend the Covasim simulation to be flexible towards incorporating interventions in real-time and over multiple time periods. We modify Covasim to incorporate an online intervention at a current time step by feeding the Covasim model with an action from the DRL agent, who represents the decision-maker, at a preset frequency and enforcing the intervention internally based on the details defined. This way, Covasim becomes more flexible, and new interventions can be enforced up to a defined time period. The preset frequency serves as a simulation step size. This step size is set based on a manager's decision-making schedule. If a manager wants to intervene daily, the step size is set as 1, but we can set the step size to any number.

Another extension to the Covasim model is the incorporation of vaccination strategies. In addition to Covasim's disease progression mechanism, we add vaccination strategies that can be used for any two-shot or single-shot vaccine. Currently, we introduce only vaccines approved under the Emergency Use Authorization (EUA) of the FDA, but an extension to other single-shot or two-shot vaccines can easily be done. An individual can be exposed to other infected individuals at any point during the vaccination process. Depending on the state at which an individual is, we calculate the likeliness of getting infected based on the type of vaccine and the number of shots they had received. In addition to the age and comorbidity-based vaccination strategy, we also develop a random vaccination strategy where no priorities are set. In the random vaccination strategy, an individual from each group has the same chance of being selected for vaccination, given that they belong to either susceptible or recovered compartments.

2.1.2 Reinforcement learning

To our knowledge, this study is the first multi-reward DRL approach that is integrated with a very detailed agent-based simulation to guide the government with sequential intervention methods to curb an epidemic outbreak. Typically, reinforcement learning models build upon an internal environment that represents the dynamics of the systems. Upon that, states are defined, and actions are used as triggers to switch from one state to another. We develop a reinforcement learning agent that is fed by an external system, an agent-based simulation model. The RL agent is guided by a multi-objective reward function that incorporates a perception of the economy in the population and different compartmental statistics regarding healthy, infected, and dead individuals. The multi-objective reward function is tailored to allow the decision-maker to emphasize one problem over another based on a trade-off they are willing to accept. The RL agent is responsible for learning how to intervene in the agent-based simulation after looking at the state (information given from the simulation) and then translating that into an optimal action. Then, the RL action is transformed into an intervention on the agent-based simulation.

2.1.3 Insights into decision making

Our simulation-deep reinforcement learning approach demonstrates that more could have been done by the Trump government to tackle the disease spread when it started to proliferate. Even if the introduced measures had been implemented in a timely manner, then the tentative reopening in mid-April of 2020 would have proven successful. Further, our experiments implementing an age-based vaccination strategy, same as the one employed by the state of New Jersey, show that a reopening was possible with only vaccination and mandatory masking by mid-February 2021. Our random vaccination strategy suggests that due to the super-spreaders getting vaccinated early, a faster reopening was possible at the beginning of February 2021.

Our multi-objective reward function experiments show that giving more attention to maximizing the number of healthy individuals and to an effective vaccination process has the highest impact on reducing the epidemic spread. In addition, putting more weight on saving the economy may render interventions useless and explode into an uncontrollable wave of widespread infection, causing more hospitalizations and a higher death toll.

Comparing vaccination strategies, we investigate the applied age-based vaccination strategy versus a random vaccination strategy where everyone has an equal chance of getting vaccinated. Our results show that using a random vaccination strategy that allows superspreaders to get vaccinated reduces the number of infected individuals by 32%. In addition, when total infections decrease, the number of hospitalizations and critical cases decreases as well.

Accounting for behavioral change of individuals at risk can play a significant role in the effectiveness of random and age-based vaccination strategies. We show that fully quarantining 10% of the risky individuals and using a random vaccination strategy reduces the death toll by 14% with respect to age-based vaccination and 27% compared to the NJ reported data.

3 Simulation environment

To simulate the Covid-19 pandemic in the population of around 9 million people who reside in New Jersey, we enhance the Covasim model developed by Kerr et al. (2021) (Version 2.1.2, 2021-03-31) and adapt it to our needs and purpose. Kerr et al. (2021) propose an open-source ABM developed to project epidemic trends and explore intervention scenarios. The Covasim ABM has many useful features, such as age-structured agents, and transmission networks with different social layers such as households, schools, workplaces, and communities. Covasim further includes intrahost viral dynamics with viral-load-based transmissibility. Covasim also supports a wide range of already built interventions such as physical distance, protective equipment, testing, and quarantine, as well as the capability to extend and make custom interventions.

Kerr et al. (2021) also implement a process of calibration calculating the loss using a normalized absolute error. They formulate an equation to find parameters that minimize the function that measures the difference between the observed data and the model predictions. In their calibration module, most of the parameters are fixed based on the values available from the literature, and the only parameter allowed to vary is β, which is the probability of virus transmission when a susceptible individual comes in contact with an infectious individual.

We extend the agent-based simulation of Covasim to the one in Fig. 1. Here, the susceptible compartment includes all healthy individuals. Once a healthy individual is exposed (Exposed), they get infected but not yet contagious. The yellow shading shows the states at which an individual is infectious and transmits the disease. As the incubation days are over, an individual either has no symptoms (Asymptomatic) and is recovered (Recovered), or symptoms start to manifest (Symptomatic). An individual might experience mild symptoms (Mild) and then transition to the recovered compartment. If the symptoms become severe, then there is still a chance that the individual will recover, but medical attention such as hospitalization (Hospital) might be needed. If symptoms become critical (Intensive Care Unit [ICU]), then the individual still has a slim chance of recovering, but if not, the individual will be transitioned to the death compartment (Dead). Vaccinated 1 and Vaccinated 2 (enclosed in the blue dotted rectangle) represent the individuals who get the first and second shot of a two-shot vaccination, respectively. In our model, susceptible individuals are eligible to be vaccinated for the first dose. Asymptomatic cases will automatically transition to Recovered after some time. Other symptomatic individuals might worsen and eventually die, but those individuals might recover with some probability as well. After eight months, the antibodies of recovered individuals cannot protect them anymore. Thus, they will transition to susceptible again (Dan et al. 2021).

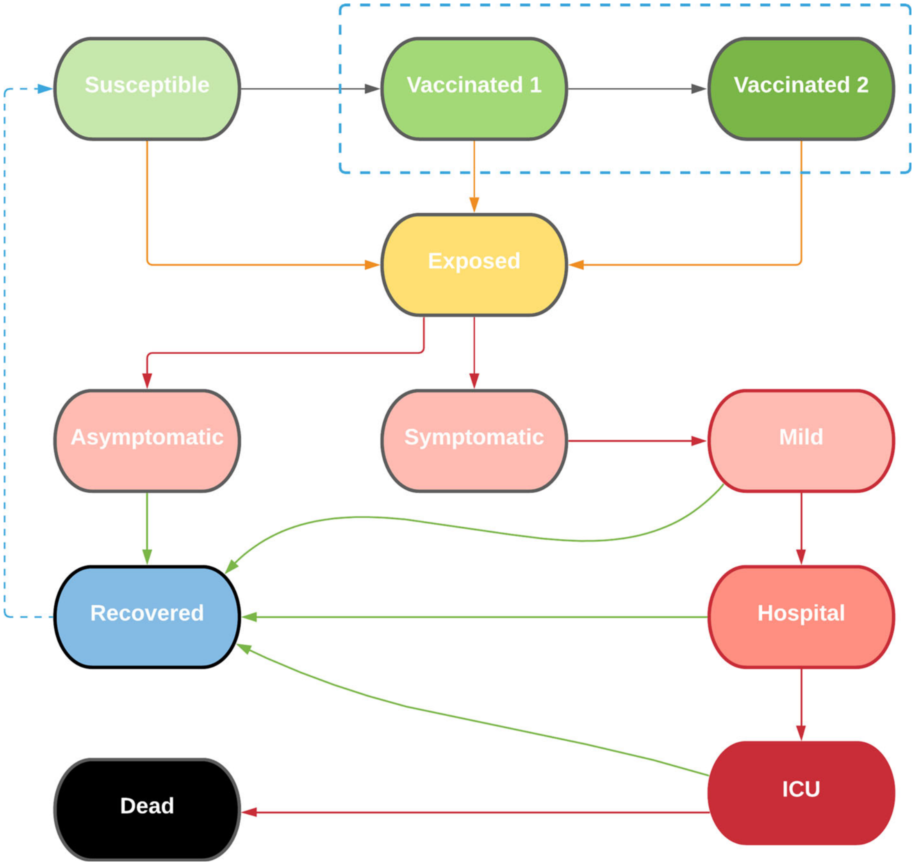


Fig. 1 Covasim disease progression, compartments, and final outcomes in the extended model. The blue dashed box and arrow show the new aspects of the simulation model extended from that of Kerr et al. (2021)

In their study, Kerr et al. (2021) portray Covasim as a simulation tool with intervention strategies for the whole simulation predefined at the start of it. And using the multi-simulation feature, they can compare how each intervention affects the disease spread. In our study, we are interested in defining the best intervention strategy for a certain situation of the pandemic periodically over multiple time stages. Hence, we extend Covasim to be more flexible where an action can be defined at any point in time, and a new intervention will be enforced up to a defined time period. Furthermore, we extend the Covasim interventions by implementing two additional vaccine interventions to perform single-shot or double-shot vaccines. While one of the vaccination strategies considers vaccination with equal probability for each individual, namely the random vaccination model (RVM), the other considers people with comorbidities and older in age with a higher likelihood of getting vaccinated than young and healthy individuals, called the age-based vaccination model (AVM). With this, we aim to provide insights into the discussion regarding the priority of vaccinating critical individuals or super-spreaders.

4 DRL environment

Counting on simulation to describe the compartmental situation of the population, we need to model a DRL environment where the current state of the simulation is represented. We define a state in simulation as the population statistics (percentage of the population in each compartment) in each disease compartment on a particular day during the pandemic. This information is also used to express the state of the RL environment. A state is composed of the following information: the ratio of the susceptible population (S), the ratio of the population who received only the first shot of the vaccine (V_1), the ratio of the population who received both shots of the vaccine (V_2), the ratio of total infections (I), the ratio of hospitalized cases (H), the ratio of individuals in an ICU (C), the ratio of the recovered individuals (R), and the ratio of the dead individuals (D) over all the population.

4.1 Episode and states

4.1.1 Episode

We define an episode as the full cycle of simulation and DRL agent intervention decisions-making. Before starting the framework, we define the step size and the full length of the simulation. For example, assuming that we want to simulate for a year and our step size is a month, at the beginning of each month, the DRL agent would enforce interventions in the simulation. Then, the simulation is run for a month based on the intervention given by the DRL agent. The episode starts with the first intervention of the first month and ends after the simulation for the last month of the year.

4.1.2 States

In our RL environment, we formulate our state as a one-dimensional array containing information for the current compartmental situation of the epidemic and denote it as $\theta := [E_t, S, I, H, C, D, R, V_1, V_2]$, where E_t is the economic index at time t. A state represents the proportion of the population in each disease compartment defined on Fig. 1. A state is generated after one simulation run.

4.2 Multi-objective reward function

At first, due to the fast spread of the COVID-19 pandemic, many governments were faced with tough choices. COVID-19 started taking lives daily, but most governments were slow to enforce closures since they feared economic collapse (Rocha 2020). In such situations, a government or a decision-maker needs a tool to do a sensitivity analysis and find trade-offs between different objectives, such as reducing the overall disease spread, keeping the economy performing, and protecting people's health or decreasing the death toll. Particularly, during the COVID-19 pandemic, a full closure would threaten the economy, while no interventions would result in more infections, deaths, and side economic costs related to degradation of the quality of human life or loss of lives, workforce reduction, and hospital expenses. We formulate a multi-objective reward function to offer the decision-maker the option of shifting between a strategy to keep the economy flowing to another where they would like to reduce the total death toll. Hence, economic stability and well-being are two dimensions that make dealing with an epidemic a more difficult challenge. Without considering the epidemic's impact on the economy, which we call the economic index, decision-making would not be complete. Hence, we quantify the contribution of each individual to the economy. Specifically, the health condition of a person defines the level of their contribution to the economy. Quarantining and work, school, and business closures come at a high cost. We assume that every healthy individual contributes to the economy with a value of 1. In our validation, this contribution is given from susceptible, recovered, and vaccinated individuals. Individuals who get infected will not be able to fully contribute to the economy. Depending on the severity of the infection, it might also become a cost to the economy. Finally, the deaths of infected people result in the worst economic loss because the economic contribution of an individual is completely lost.

4.2.1 Economic index

To quantify the economic situation of a particular day during the pandemic, we formulate the economic contribution at time t, E_t, as follows:

$$E_t = S + V_1 + V_2 + R - \alpha \times I - \beta \times H - \gamma \times C - D, \tag{1}$$

where α, β, and γ can also serve as tuning parameters of the economic index at a time t. Here, we assume that healthy people would function normally in the economy while infected (I), hospitalized (H), people in ICUs (C), and dead (D) individuals would mean a loss economically.

4.2.2 Multi-objective reward function

Using the formulation of E_t above, we maximize the following multi-objective reward function:

$$R(\theta) = \lambda \times E_t - \mu \times I - \rho \times D + \pi \times (S + V_1 + V_2) \tag{2}$$

where state $\theta := [E_t, S, I, H, C, D, R, V_1, V_2]$. Tuning parameters λ, μ, π and ρ are determined based on which part of the objective we want to emphasize more.

4.3 Actions or intervention measures

4.3.1 Actions

For the learning of our DRL agent, we investigate different possible actions that are realistic but not necessarily exclusive. In practice, we can combine different non-pharmaceutical interventions and vaccines with social distancing measures. In total, we define nine possible actions that our agent can choose from. At the start of the pandemic, we will only include six of these actions as vaccines might not be available at that point. Once the vaccines are available, all nine actions can be applied. The various actions considered are defined below:

0. *Do Nothing*: We allow the agent to not enforce any restriction on the population.
1. *Testing, Contact Tracing, and Quarantine*: This action performs tests and traces contacts of positive tests and quarantines them. Usually, these actions go together as the traced contacts are notified and they either get tested too, or they are ordered to remain in quarantine.

2. *Close Schools and Non-Essential Workplaces*: Governments might decide to close schools and non-essential workplaces and limit gatherings up to a certain number to reduce contact between the individuals in a population, thus reducing infections and keeping the COVID-19 curve under control.
3. *Mandatory Mask*: A mandatory mask can be enforced on a population.
4. *Testing, Contact Tracing, Quarantine, Close Schools, and Non-Essential Workplaces*: Because actions one and two are not exclusive, governments can choose to enforce them at the same time to have a higher impact on slowing the disease spread.
5. *Testing, Contact Tracing, Quarantine, and Mandatory Mask*: Action one can also be enforced in combination with action three. This action does not close schools or businesses, but it enforces mandatory mask usage to control the spread.
6. *Vaccination*: When vaccines become available, it is a form of action that can be combined with any non-pharmaceutical measure. This action considers only vaccination in case governments decide to only use vaccination and reopen without any other enforced intervention.
7. *Vaccination and Mandatory Mask*: This action consists of a combination of actions three and six. It is seen as a probable reopening strategy as with vaccines, the population will become more protected, and masks will reduce the transmission of disease.
8. *Total Lockdown*: In an extreme situation, where the healthcare system has failed and the government did not intervene timely, a full lockdown might be applied, which enforces all non-pharmaceutical interventions together with vaccination. This measure can result in economic hardships and failures due to the closure of workplaces and businesses.

5 Integrated simulation-RL

Using the ABM simulation and DRL environment presented in Sects. 3 and 4, respectively, we create an Integrated Simulation - RL (SiRL) framework. Figure 2 shows how Covasim agent-based simulation interacts with the DRL procedure. We start by creating an RL and an agent-based model environment where compartmental statistics for the population are stored. At the first step, we have initial information about the compartments. So, the DRL agent takes an action based on the initial proportion of the population in each health compartment. Once the simulation starts, it picks up the decision from the DRL agent, applies the respective intervention, and runs for s days, where s is the step size of the simulation. After s days, the compartmental statistics from the simulation are used to formulate the DRL state and feed it to the DRL agent. Based on the DRL state, we calculate a reward using Eq. (2). This reward is the evaluation of the last intervention applied by the DRL agent. At this point, we check if the end date of the simulation is reached, and based on that, the execution of the SiRL framework ends, or the DRL agent will take another decision to be applied in the next s days on the simulation environment and follow the same cycle until the end date. When the end date is reached, the episode terminates. Note that the take action in green and dashed yellow boxed in Fig. 2 represent the same compartment. The difference is that the action compartment in yellow is executed once in the beginning and then called inside the cycle until the SiRL episode is terminated.

Figure 3 describes how the agent-based simulation and the DRL agent interact and exchange information. At time $t = 1$ after we have started the environments (RL and agent-based simulation) and taken the first action, we declare the initial data and start the agent-based simulation incorporating the first action. The simulation will run for $s = 15$ days (our defined

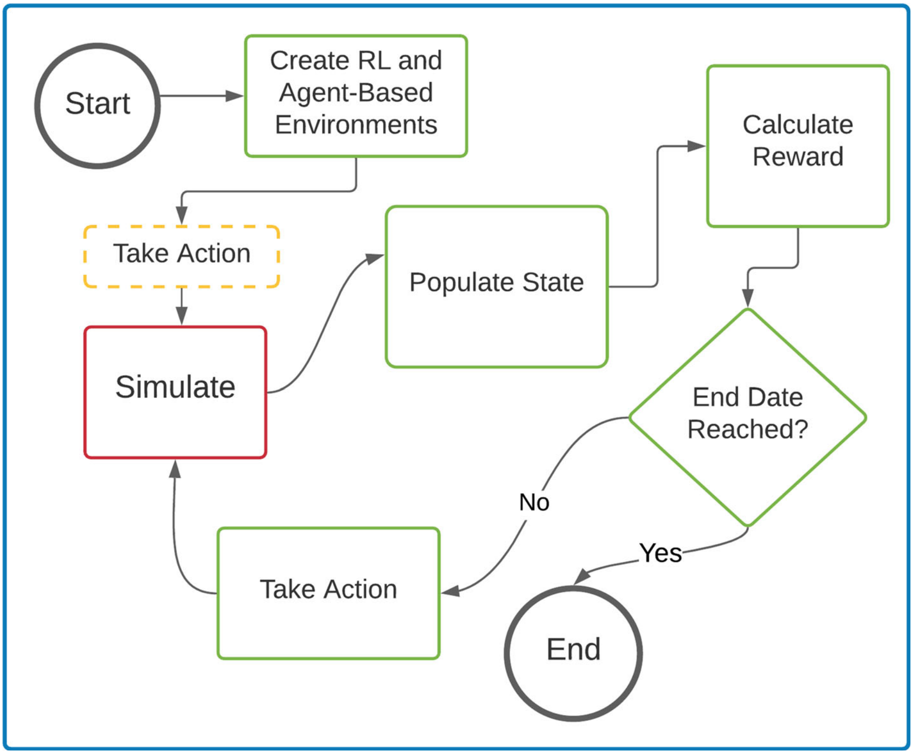


Fig. 2 The SiRL Framework, which is an agent-based simulation integrated at the heart of a DRL framework

step size) and, at the end of the simulation step, will feed compartmental statistics and the economic index to the DRL agent. Our assumption for a step size of 15 days is based upon the incubation period, which is at most up to two weeks (Lauer et al. 2020). The initial interventions of the government were initially put in place for two weeks, in most of the countries (Ngonghala et al. 2020). Based on the state, the DRL agent takes decision x_1, which enforces interventions on the simulation for the next simulation period. In turn, after this simulation period ends, it will again give the new compartmental statistics after the intervention where the agent's action is evaluated according to the reward function. This process continues until the entire simulation ends.

5.1 Training algorithm

Algorithm 5.1 describes the general steps and data used to train an agent. First, we create the respective simulation and RL environments. At the start of a simulation, we decide on the total population, the total length of the simulation, and the step size at which we enforce interventions. At the start of the RL environment, we initialize the agent and define the weights of the reward function. Then for each simulation period, we extract the compartmental statistics from the simulation and feed them to the DRL agent. Compartmental statistics include the percentage of the population in each compartment and the economic index at the end of the simulation period. Having this information, the DRL agent will decide on an action $x_j \in \mathcal{X}$ at simulation run j. Based on this action, the simulation is run for a one-step size, and then a reward is generated to quantify how good the action of the agent was.

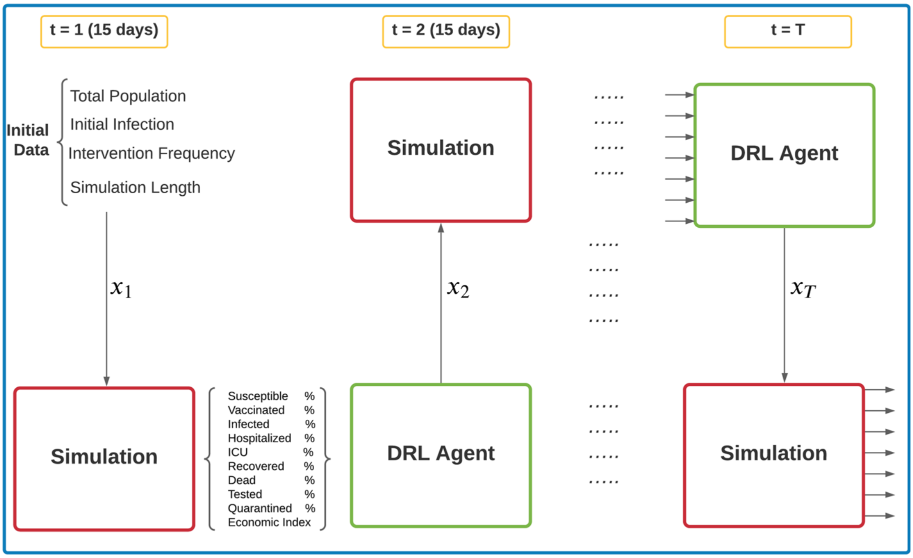


Fig. 3 Agent-based simulation and DRL agent information exchange between simulation periods

Algorithm 1 Simulation DRL Training Algorithm

```
Procedure: SiRL
Input: $\Delta$, $s$, $\sigma$, $\Theta^0$, $\theta^0$   {We start with the total population, step size, simulation periods, initial compartmental statistics, and initial state.}
Output: $\Omega$   {Trained Deep Q-Network Model.}
Initiate SiRL with $\Delta$, $s$, $\sigma$   {Create DRL and simulation environments.}
Take an initial action $x_0$   {First interventions.}
for $j \in \mathcal{J}$ do {for each simulation period}
    $\theta^j \longleftarrow \Theta^j$   {Update state with compartmental statistics}
    Take action $x_j \in \mathcal{X}$
    $R(\theta^j, x_j) \longleftarrow$   {Calculate reward.}
end for
```

6 Experiments

In our experiments, we want to be flexible and generalize over different possible epidemics. Therefore, we test different models: one without a vaccine available (no-vaccination model as **NVM**), another after the vaccine is discovered and an age-based vaccination is applied (age-based vaccination **AVM**), and another after the vaccine is discovered, but everyone is eligible to be vaccinated above the age of 12 (random vaccination **RVM**). We gather data for the COVID-19 epidemic in New Jersey and address the management of the disease. In this section, we assume each objective of the multi-objective reward function in Eq. (2) has equal importance. Hence, we use the same weight for all parameters ($\lambda = \mu = \rho = \pi = 1$), except Sect. 6.3.3, where we tune the objective weights in the reward function (2) for multi-objective analysis. We train our agents using a population of 500,000 benefiting from the scaling properties of Covasim, and test our trained agents on the population of New Jersey (8,882,190). For time efficiency, we use a lower population for training. After training, testing with the SiRL framework on a population of nearly 9 million takes 15 min. We report training times and the number of episodes for each model in “Section A1”.

6.1 Data gathering

We collect bi-weekly compartmental data from the start of the Covid-19 epidemic until the beginning of our project (March 1, 2020, to April 15, 2021). The compartments we consider are Susceptible (S), Infected (I), Hospitalized (H), ICU (C), Dead (D), Recovered (R), Tested (T), Vaccinated with 1st shot (V_1), and Vaccinated with the 2nd shot (V_2) obtained from CDC database (CDC 2022) and crosschecked with the NJ COVID-19 dashboard (NJ 2021). In addition, to compare and understand decision-making at any point of the pandemic, we also collect government decisions to identify what interventions are active and a specific date. We collect these data for the whole US and the state of NJ in particular.

Figure 4 presents a decision timeline for the US during the beginning period of the COVID-19 (March 1, 2020, to June 30, 2020). This timeline also corresponds in close dates with the responses that each state has taken to control the spread.

To provide a robust framework, SiRL can be used to extract control measures for different epidemics. In our case, we want to draw conclusions at any point during the COVID-19 pandemic. That is why we calibrate our model and train our DRL agent in different stages of the pandemic. We consider the start of the COVID-19 pandemic where a vaccine is not available, and we refer to this model as the no-vaccination model, **NVM**. We also investigate the COVID-19 dynamics after the vaccines are introduced. With the vaccination models, to be consistent with the reality, we calibrate our model using age and comorbidity-based

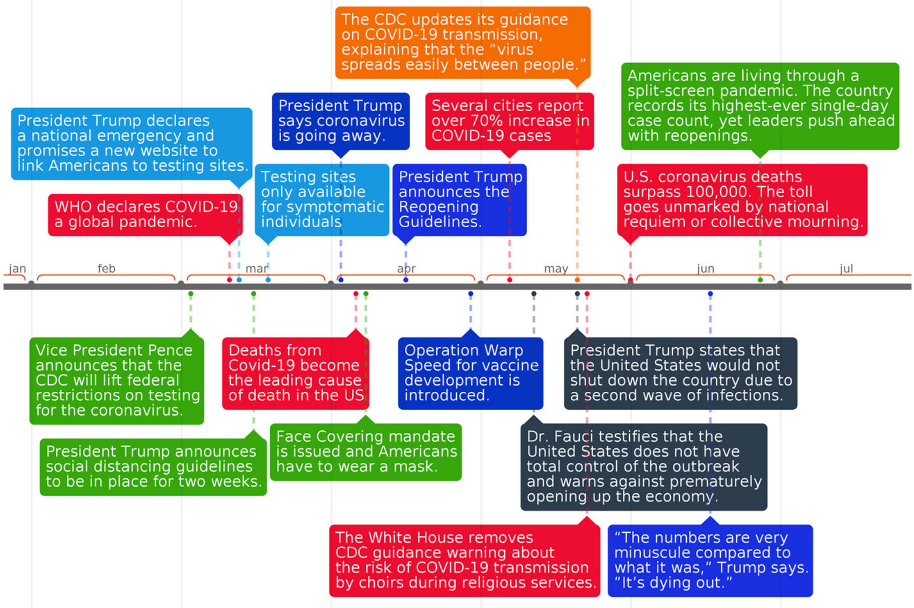

The CDC updates its guidance on COVID-19 transmission, explaining that the “virus spreads easily between people.”
Americans are living through a split-screen pandemic. The country records its highest-ever single-day case count, yet leaders push ahead with reopenings.
President Trump declares a national emergency and promises a new website to link Americans to testing sites.
President Trump says coronavirus is going away.
Several cities report over 70% increase in COVID-19 cases
Testing sites only available for symptomatic individuals
President Trump announces the Reopening Guidelines.
U.S. coronavirus deaths surpass 100,000. The toll goes unmarked by national requiem or collective mourning.
WHO declares COVID-19 a global pandemic.
jan
feb
mar
apr
may
jun
jul
Vice President Pence announces that the CDC will lift federal restrictions on testing for the coronavirus.
Deaths from Covid-19 become the leading cause of death in the US
Operation Warp Speed for vaccine development is introduced.
President Trump states that the United States would not shut down the country due to a second wave of infections.
President Trump announces social distancing guidelines to be in place for two weeks.
Face Covering mandate is issued and Americans have to wear a mask.
Dr. Fauci testifies that the United States does not have total control of the outbreak and warns against prematurely opening up the economy.
The White House removes CDC guidance warning about the risk of COVID-19 transmission by choirs during religious services.
“The numbers are very minuscule compared to what it was,” Trump says. “It’s dying out.”

Fig. 4 COVID-19 timeline from April 1, 2020, to June 30, 2020, created based on the information provided in Thebault et al. (2021)

vaccination strategy, **AVM**. We then use these calibrated actions to implement also a random vaccination strategy, **RVM**.

6.2 Validation of intervention effect for NVM

We use our simulation to measure the effect of non-pharmaceutic risk measures on the COVID-19 progress. Thus, we can use these quantified effects to train our agents. To calibrate between Covasim and New Jersey environments, we calibrate each no-vaccination model action by reproducing the COVID-19 spread during its first four months when vaccines were not available. In these four months, we mimic governmental actions in the same period that they were enforced.

During the first four months (March 1 to June 30), the government suffered from resources, and not many effective interventions were implemented. On March 3, Vice President Pence announced that CDC would lift federal restrictions on testing for COVID-19. Despite that, until April 12, 2020, it was not easy to get tested. On March 11, 2020, WHO declared COVID-19 a global pandemic. Two days later, on March 13, 2020, President Trump declared a national emergency and promised to increase efforts to make testing available and accessible for Americans. On March 16, 2020, the Trump government also announced social distancing guidelines to be in place for two weeks initially (Thebault et al. 2021).

Around the same time, on March 18, 2020, Governor Murphy of New Jersey, in an attempt to slow down the spread of the disease, ordered the closure of all pre-K, K-12, higher education institutions, casinos, theaters, gyms, and non-essential retail, recreational and entertainment businesses also banning gatherings of people more than 50.

In addition, we apply the paired t-test to investigate the difference between the mean bi-weekly compartmental values obtained from the simulation and the mean actual corresponding data values provided by the CDC. According to the statistical analysis shown in Table 1, our validation is statistically similar to the actual data reported by the CDC since all p-values are greater than 0.05. We also demonstrate the validation of the **NVM** model where we exclude vaccination as an intervention in Fig. 14 in "Appendix A2". Similarly, we validate the intervention effect of the age-based vaccination **AVM** in "Appendix A3". Figure 16a–f in "Appendix A4" show comparisons between real values and simulated results from the SiRL framework for cumulative Infections, hospitalizations, and recoveries (with fixed interventions to represent the reality).

6.3 Results

We show results for different periods of the pandemic in the state of New Jersey. We emphasize the usability and flexibility of the framework by a comparative study of different strategies for

Table 1 Paired t-test analysis comparing the compartmental data from the NVM simulation (Predicted) with the actual data from the CDC (Actual) with the 95% confidence level

Compartment	Mean		Two-tailed paired t-test		
	Actual	Predicted	t-stat	t-critical	p-value
Infected	107, 650	107, 569	0.72	2.2	0.49
Hospitalized	22, 523	20, 890	0.76	2.1	0.46
ICU	966	964	0.73	2.2	0.48
Dead	9251	9242	0.67	2.3	0.51

decision-making during the pandemic. Details of training results are presented in “Appendix Section A1”.

6.3.1 Comparison to government actions

In Sect. 6.2, we calibrate the government decisions during the first four months of the pandemic. After we train, we allocate our DRL agent the same resources and actions to observe what the agent deems optimal. Based on the agent’s response, the government closes schools and non-essential workplaces and uses tests to identify infected individuals and then trace their contacts and quarantine them for the first 45 days from March 1 to April 15, 2020. After that is done in the first months, our agent suggests a reopening but enforcing mandatory masks. It is even more interesting that this strategy is very similar to what the government did, but there is a shift in time. Our model suggests contact tracing and stay-at-home orders must have been enforced exactly at the beginning of March and then start reopening around mid-April. This result implies that the government was late in any of the actions except for the reopening date. Since measures were not executed timely and sufficiently to control the outbreak, reopening backfired in more cases after the April reopening, keeping most educational institutions closed for the rest of the year. Figure 5 compares the decisions taken by the government with the decisions suggested by the trained DRL agent for the NVM model. Above the x-axis, we map the government decision. Notice that there is a delay in action. Testing is announced that it will be available in the first week of March, but it was made widely available for symptomatic people around mid-April. Below the x-axis, we describe the suggested actions from the DRL agent. Notice that during the first months, testing and contact tracing are important, while also schools and workplaces are temporarily closed to slow the spread down. After that, a reopening is suggested by only enforcing masks.

Figure 6 illustrates the comparison between the NJ and Federal government interventions enforced during the first four months after vaccines were introduced, specifically from December 15, 2020, to April 15, 2021, with the interventions suggested from the DRL agent trained using the age-based vaccination model. Above x-axis notice that the government implemented all available interventions in combination with vaccination. During this period, the government gave priority to older people and those with pre-existing conditions. Month after month, the age bar for vaccination was reduced, allowing more younger people to get vaccinated. On April 19, 2021, all individuals older than 16 became eligible for vaccination in New Jersey. In our age-based vaccination strategy, we follow a similar pattern. We give a

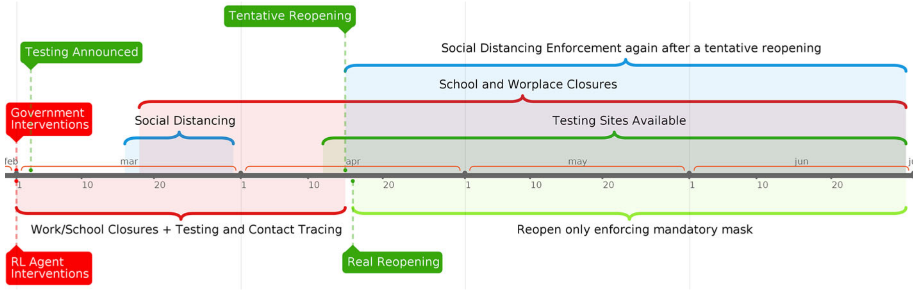

Tentative Reopening
Testing Announced
Social Distancing Enforcement again after a tentative reopening
School and Worplace Closures
Government Interventions
Social Distancing
Testing Sites Available
feb
mar
apr
may
jun
ju
1
10
20
1
10
20
1
10
20
1
10
20
Work/School Closures + Testing and Contact Tracing
Reopen only enforcing mandatory mask
RL Agent Interventions
Real Reopening

Fig. 5 Comparison of government actions and DRL agent actions during the first four months of the COVID-19 pandemic in the US. Above the x-axis, we describe the government actions, and below x-axis the DRL agent suggestions are shown

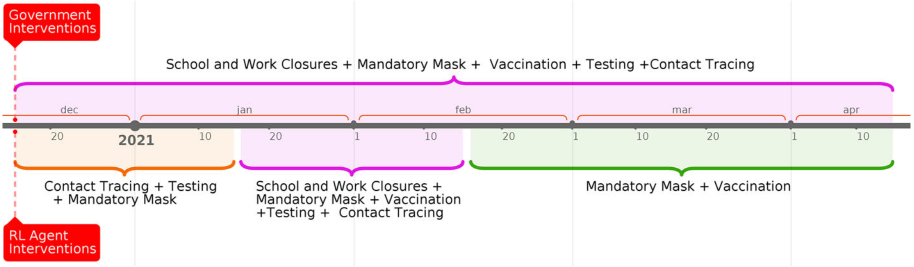

Government Interventions
School and Work Closures + Mandatory Mask + Vaccination + Testing +Contact Tracing
dec
jan
feb
mar
apr
20
2021
10
20
1
10
20
1
10
20
1
10
Contact Tracing + Testing + Mandatory Mask
School and Work Closures + Mandatory Mask + Vaccination +Testing + Contact Tracing
Mandatory Mask + Vaccination
RL Agent Interventions

Fig. 6 Comparison of government and DRL agent actions for the first four months after vaccines were introduced for the COVID-19 using an age group vaccination strategy. Above the x-axis we describe the government actions, and below x-axis the DRL agent suggestions are shown

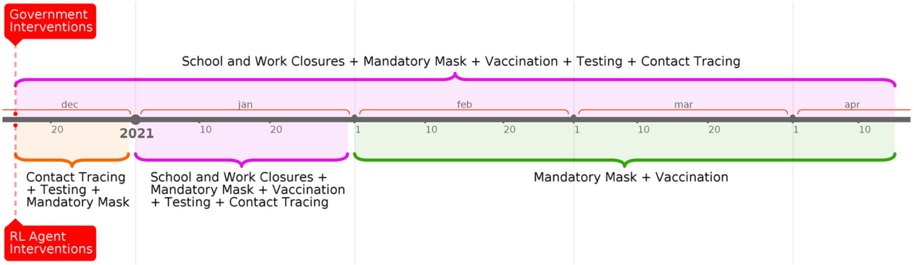

Government Interventions
School and Work Closures + Mandatory Mask + Vaccination + Testing + Contact Tracing
dec
jan
feb
mar
apr
20
2021
10
20
1
10
20
1
10
20
1
10
Contact Tracing + Testing + Mandatory Mask
School and Work Closures + Mandatory Mask + Vaccination + Testing + Contact Tracing
Mandatory Mask + Vaccination
RL Agent Interventions

Fig. 7 COVID-19 timeline allowing all age groups access to vaccination. Above the x-axis we describe the government actions and below x-axis the DRL agent suggestions are shown

higher probability to the older ages while reducing it after each month. Our model suggests that testing, contact tracing, and mandatory mask be enforced for the second half of December 2020 and the first half of January 2021. It is interesting that the model does not suggest an immediate vaccination. This is because, at the beginning, the vaccine supply was small. So, the DRL agent does not see it as highly beneficial since the number of vaccine doses available was very low when vaccines were first offered. When equal weights are assigned to each sub-objective in the reward function (2), the DRL agent cannot capture that even a very small number of vaccines should be used. From the second half of January 2021, the agent suggests the enforcement of all measures while vaccination should also be applied with those interventions. After only a month, in mid-February, the DRL agent suggests lifting the closures but recommends a continued vaccination while also enforcing the use of masks.

Figure 7 compares decisions suggested by the DRL agent trained with random vaccination strategy towards those enforced by the government. Differently also from the **AVM** strategy, **RVM** suggests only 15 days of the mandatory mask, testing, and contact tracing followed by full closures with vaccination for the next month up until January 31. A reopening is suggested at the beginning of February, combining the mandatory mask with vaccination, which is earlier than that suggested by the **AVM**.

6.3.2 Economic standing

To compare the economic situation in different simulations, we use the formula in Eq. 1. We calculate the economic index by assigning a weight to each of the compartments. Figure 8 compares the economic situation between the simulation with the no-vaccination model

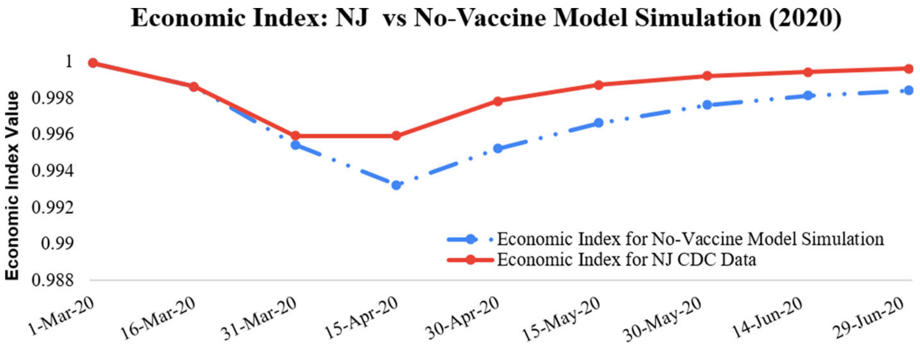


Fig. 8 Comparison of the economic index between NJ CDC Data and simulation data using NVM

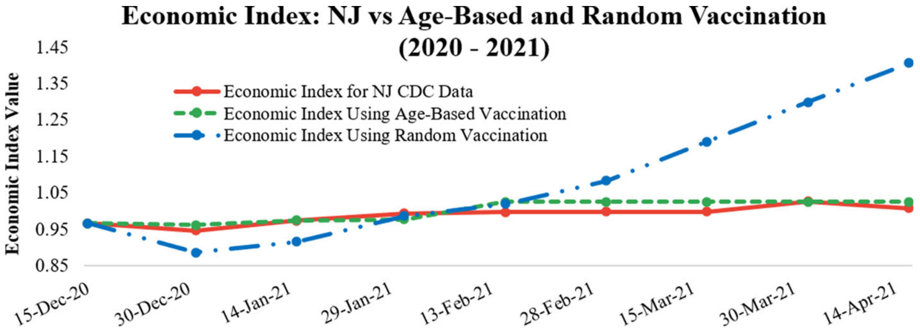


Fig. 9 Comparison of the economic index between NJ CDC Data and simulation data using AVM and RVM

(**NVM**) and the CDC data for the state of New Jersey. Notice that NVM simulation allows for a weaker economic situation since we use equal weights for each sub-objective in the reward function (2). Based on our comparison, our model did decide on reopening, but it is doing slightly worse from the economic point of view. This is due to the multi-objective function as we give equal weight to each of the sub-objectives. If a manager wants to focus more on a flowing economy, a larger weight can be given to the economic sub-objective of the reward function in Eq. (2).

We also compare the economic situation between the CDC-reported data and our vaccination strategies based on the SiRL framework. Figure 9 compares the economic standing at 15-day intervals. We notice that the age-based vaccination maintains slightly the same economy as the CDC data. That is because our results for using an age-based vaccination strategy provide a very good estimation of the real situation. Surprisingly, a better economic standing would be achieved using a model with available vaccination for all individuals older than 12 years old, corresponding to the random vaccination strategy.

6.3.3 Multi-objective analysis

In this section, we consider modifying the reward function shown in Eq. (2). Tuning this function will shift importance and suggest decisions to reduce the worst outcomes of different situations. For example, if a government wants to prevent the infection rate and severity of the epidemic, it can give more weight to the μ parameter. To analyze how each tuning parameter affects the compartmental statistics, we consider four formulations with respect to the reward

Table 2 Comparison of different tuning parameters in the reward function for age-based vaccination

Compartment $(\lambda, \mu, \rho, \pi)$	Economy $(5, 1, 1, 1)$	Total infections $(1, 5, 1, 1)$	Death toll $(1, 1, 5, 1)$	Healthy or vaccinated $(1, 1, 1, 5)$
Infected	6, 762, 388	2, 078, 877	2, 348, 258	1, 057, 088
Hospitalized	385, 494	102, 179	127, 259	66, 881
ICU	119, 828	28, 796	37, 156	21, 365
Dead	39, 014	10, 218	11, 147	8, 360

function: economy, death toll, total infections, and healthy individual, each with a weight of λ, μ, ρ, and π, respectively. We modify the reward function for each of these models by increasing the respective tuning parameter or weight five times and retraining the model to perform tests. For example, in our experiments, we give a value of one to each of the tuning parameters. When we want to emphasize the economy, we use $\lambda = 5$, while other tuning parameters are still one.

Table 2 compares how the number of individuals in each compartment changes with different focus on the reward multi-objective function. We observe that a weight distribution of 1, 1, 1, and 5 for the tuning parameters λ, μ, ρ, and π, respectively, results in the minimum number of infections and deaths. This means that strategies aiming to keep individuals healthy and vaccination have the highest impact on the infections and the death rate. Parameters below the headers in Table 2 show the respective values for each tuning parameter of the multi-objective reward function shown in Eq. (2). Each specific objective also effects decisions taken. For example, emphasizing the economy would shift decisions from a full closure to reopening and vaccinating while everyone can move freely. When the economy is emphasized, we notice a massive disease spread. If a government only aims to maintain a healthy economy, then the number of individuals, who are hospitalized, in critical condition, and dead sharply increases. Among each part of the reward function, although emphasizing the total infections or the death toll reduces the spread in each compartment, emphasizing the portion of healthy and vaccinated individuals seems to show the best situation with respect to the total infections and death toll. Between the total infections and the death toll, emphasizing the total infections seems to be more beneficial because of the transmission mechanism. Emphasizing the death rate does not directly affect infections; therefore, the higher spread still contributes to more deaths.

6.3.4 Vaccine decisions and distributions

Vaccination restrictions have been among important discussions during the pandemic. There is definitely good in giving priority to individuals with pre-existing conditions or older people who might be more endangered from the pandemic. But older ages are among individuals who have the least amount of contact during the day. Therefore, their contribution to the spread is generally low. Hence, we want to analyze the trade-off between age-based vaccination and random vaccination. In random vaccination, we do not prioritize super-spreaders to vaccinate, but we allow them and individuals with pre-existing conditions and of older age to get vaccinated with the same probability. Comparing our **AVM** and **RVM** models can give us insight into the benefits of each. We notice from comparing Figs. 6 and 7 that the **RVM** strategy offers a faster closure and earlier vaccination start, hence also improving the economy. In another situation where a government tends to be cautious about the total

Table 3 Comparison of vaccine distribution among age groups and the compartmental values for each model

Subgroup	NJ CDC data	RVM	AVM	pdiff^1 (%)	pdif^2 (%)
Vaccination under 50 (%)	45	67	55	– 18	– 33
Vaccination 50 to 75 (%)	79	66	94	42	20
Vaccination over 75 (%)	78	63	96	51	22
Infected	851, 485	675, 310	892, 672	32	26
Hospitalized	668, 201	652, 431	665, 185	2	2
ICU	153, 657	142, 122	153, 268	8	8
Recovered	828, 283	845, 298	898, 803	6	– 2
Dead	24, 702	25, 080	24, 151	– 4	– 2

pdiff^1 shows the percentage difference between the random and age-based vaccination models and **pdiff**^{2} calculates the percentage difference between the random vaccination model and the NJ COVID-19 reported data

number of deaths, they can give higher weight to the ρ parameter, representing the weight of the death toll. This would cause lower rewards when death rates increase; thus, the DRL agent will optimize the decision while focusing on minimizing the death rates. In addition, Fig. 9 shows a comparison between the age-based and random models. Allowing super-spreaders to get vaccinated as early as possible during an outbreak reduces infections in general, hence explaining the suggested faster reopening and better economic performance.

Table 3 compares the vaccination percentage and compartmental statistics for random vaccination and age-based vaccination strategies and real data. To clarify, the NJ CDC Data and AVM columns do not report data from the validation experiments. The NJ CDC Data is obtained from the CDC, and AVM represents the results obtained from the DRL agent suggestions. Column **pdiff**^{1} represents the percentage difference between the random vaccination and age-based vaccination models, while **pdiff**^{2} calculates the percentage difference between the random vaccination data and CDC reported data for NJ. Notice that the vaccine distribution for the age groups differs between the two methods. Random vaccination slightly suggests that some portion of the younger people should get vaccinated as soon as vaccines are available, while the age-based model vaccinates almost all older-age groups first. Compartmental data shows that the total number of infected individuals reduces by 32% when using a random vaccination strategy over the age-based vaccination strategy. Due to this, the number of hospitalized and critical cases and recovered individuals slightly reduce as well. The total number of dead individuals, though, is slightly increased by 4%. This shows that a random vaccination strategy can offer earlier reopening and slower spread, but fast reopening and not focusing on an age-based vaccination strategy comes with a cost. Figure 3 also compares the random vaccination strategy with the real data reported from the CDC for NJ. We notice a similar trend to that of the age-based vaccination. The experiments show that random vaccination could reduce the number of infections but still reports a 2% higher death rate than that of the CDC statistics for NJ.

6.3.5 Vaccine decisions with epidemic behavior change

To incorporate the impacts of behavioral change into our simulations, we raise the assumption that due to the fast spread of COVID-19, a portion of older-aged individuals take further measures to self-quarantine and reduce contacts to a minimum. We train our simulation-deep

Table 4 Comparison of vaccine distribution among age groups and the compartmental values for each model with older age-group's self-protection assumption

Subgroup	NJ CDC data	RVM	AVM	pdiff^1 (%)	pdif^2 (%)
Vaccination under 50 (%)	45	72	55	– 24	– 38
Vaccination 50–75 (%)	79	72	94	30	10
Vaccination over 75 (%)	78	72	96	34	7
Infected	851, 485	686, 766	755, 505	10	24
Hospitalized	668, 201	603, 785	613, 074	2	11
ICU	153, 657	123, 544	130, 046	5	24
Recovered	828, 283	787, 420	898, 803	1	5
Dead	24, 702	19, 507	22, 294	14	27

pdiff^1 shows the percentage difference between the random and age-based vaccination models, and **pdiff**^{2} calculates the percentage difference between the random vaccination model and the NJ COVID-19 reported data

reinforcement learning framework by enforcing the assumption that 10% of the individuals belonging to the risky groups (over the age of 50) reduce their contacts to zero.

Similar to Tables 3, 4 presents the vaccination percentage for three distinct age groups. We notice that the vaccination percentages among different age groups have not been affected using the age-based vaccination strategy, while the random vaccination strategy now suggests an equal distribution of vaccines among different age groups. This is due to the reduction in disease spread and death rate as a result of the behavioral change in risky groups.

The protection of individuals at risk is quite effective also in the case of the age-based vaccination strategy as results show smaller values for infections, hospitalization, critical condition, death compartments when compared to the results in Table 3 and with respect to the NJ reported data where an age-based vaccination strategy is used. Compartmental results in Table 4 suggest that if 10% of individuals belonging to higher risk groups (individuals over the age of 50) were fully quarantined and a random vaccination strategy was used, then we would have 10% fewer infected cases and death toll reduced by 14% compared to the age-based vaccination strategy.

6.3.6 Model flexibility

To address the flexibility of our approach to be applied to other locations, we validate and apply our model to the COVID-19 case in the state of Kansas, using CDC data (CDC 2022). When applying the model in another location, some adjustment needs to be made because populations in different locations differ in their cultures, proximity, and community involvement which highly affects disease spread. Similarly, when looking at other epidemics, the disease's biological properties change the course of an epidemic. In both cases, an initial validation must be performed to ensure that the interventions and simulations are synchronized and represent the dynamics of the disease in a real situation. Appendix Section A5 describes the validation process and results for the Kansas case study. We consider a situation where each part of our multi-objective function in Eq. (2) has the same weight of 1 ($\lambda = 1$, $\mu = 1$, $\rho = 1$, and $\pi = 1$). Figure 17 in "Appendix A5" shows the validation of interventions for Kansas in the time period from September 15, 2021, to January 15, 2022. With the validated actions, an agent can be trained by studying different states of the compartmental statistics.

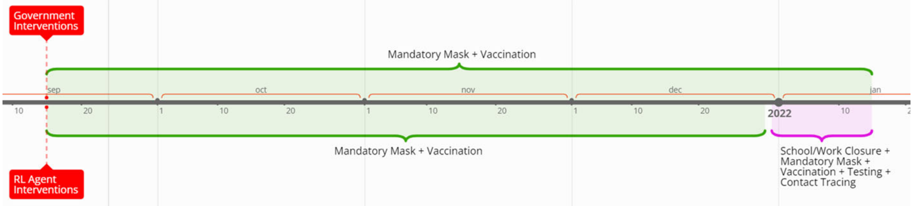

Government Interventions
Mandatory Mask + Vaccination
sep
oct
nov
dec
jan
10
20
1
10
20
1
10
20
1
10
20
2022
10
Mandatory Mask + Vaccination
School/Work Closure + Mandatory Mask + Vaccination + Testing + Contact Tracing
RL Agent Interventions

Fig. 10 Comparison of government actions and DRL agent actions for the Kansas case from September 15, 2021, to January 15, 2022. Above the x-axis, we describe the government actions, and below the x-axis, the DRL agent's suggestions are shown

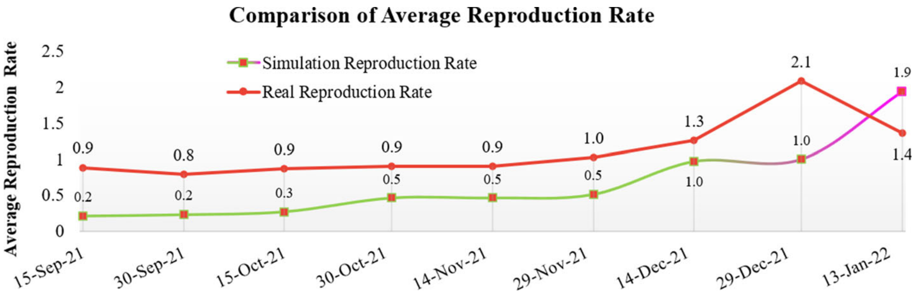


Fig. 11 Average reproduction rate for each 15-day period starting September 15, 2021, to January 15, 2022. The green line represents the average reproduction rate for the decision Mandatory Mask + Vaccination, and the pink line represents the average reproduction rate for Mandatory Mask + Vaccination + Work/School Closure + Contact Tracing. The Red line represents the average reproduction rate calculated as is defined as the average number of secondary cases per primary case on a certain date by Gu (2022)

Figure 10 plots the government actions (above x-axis) against actions suggested by the trained RL agent (below x-axis) over a four-month period. During the same period, the government removed the lockdown policy and only enforced masks in indoor environments, while several institutions pushed their employees to vaccinate. Our DRL agent suggests that in addition to mandatory masks and vaccination, a lockdown should start beginning of January 2022. This suggestion is reasonable considering the increase in infection at the end of 2021.

To observe how actions suggested by the SiRL agent related to the effective reproduction number, R_e, we use the "instantaneous reproductive number" presented in Gostic et al. (2020). R_e for the SiRL is computed by dividing the new number of infections on day t by the number of actively infectious individuals on day t, then multiplied by the average duration of infectiousness. Since we do not have access to the actual number of actively infectious individuals, we calculate the R_e for the real case using the formulation provided by Gu (2022), which is defined as the average number of secondary cases per primary case on a certain date. Figure 11 shows the average reproduction rate for every 15-days (we observe and intervene every 15 days) for both the real and SiRL data. The Red line color represents the average R_e resulting by the decisions suggested by the government for each period, whereas green gives the average R_e by the SiRL, which suggests a combination of *Mandatory Mask* and *Vaccination* and pink means in addition to those, the average R_e reflects that *Contact Tracing* and *School/Work Closures*. We observe an agreement in trend in the R_e for both the real data and the SiRL, despite using two different formula. We notice that decision is adapted to become more conservative as the R_e gets larger, according to both calculations shown. Specifically, as the R_e increases up to around 0.9, the DRL agent started suggesting closures.

7 Conclusion

We present a Simulation-Deep Reinforcement Learning (SiRL) framework for epidemic decision-making. In SiRL, an agent can be trained to take actions based on different available interventions and epidemic infection situations. Our results show that more could be done in handling the COVID-19 pandemic spread in the US. Our DRL agent identifies situations in which government agencies should have acted faster toward slowing the spread of the virus. In addition, we compare different vaccination strategies and provide insights on the trade-off between random and age-based vaccination strategies.

7.1 Managerial insights

- Our approach demonstrates that learning algorithms can be trained to understand an epidemic situation based on compartmental statistics and take decisions in effect to improve a certain objective, such as reducing infections, keeping people healthy, maintaining a healthy economy, or reducing the death toll.
- Our experiments show that strategies aiming to keep individuals healthy result in lower infections and a lower death rate.
- In a situation where the economy is highly prioritized, infections at any level sharply increase. This specifically indicates that closures and reopening should be done carefully as they can result in a higher disease spread.
- Our trade-off analysis between age-based and random vaccination suggests that vaccinating super-spreaders can help in faster reopening but can increase deaths.

7.2 Future directions

Future directions can include extensions from the DRL and agent-based simulation as well. Further experiments could tell us more if we consider racial or geographical data. The simulation model can be extended to account for additional costs, different virus strains, or an economic value of a current infestation, including the interventions active at a point in time. The DRL framework can also be extended to another epidemic disease. To achieve this, the biological properties of the epidemic, such as the disease spread, the incubation period, and transition probabilities, need to be adjusted, and the DRL agent needs to be trained with new characteristics of the disease and the population considered. Another DRL algorithm could be used to study how that changes agent performance. Furthermore, new approaches could be developed to calculate the Pareto frontier with a 4-dimensional objective. Finally, to study the flexibility of the framework, other epidemic data in different regions of the US and the world can be validated.

Acknowledgements We gratefully acknowledge the support of the National Science Foundation CAREER Award co-funded by the CBET/ENG Environmental Sustainability program and the Division of Mathematical Science in MPS/NSF under Grand No. CBET-1554018.

Appendices

A1: Training results

No-vaccination model training

Figure 12 shows the reward agent gets during training. We train by simulating around 30k episodes, where each is a four-month simulation (March 1 to June 30, 2020). The approximate training time was 30.2 h. As the agent goes through more episodes, we notice that it builds a behavior to improve rewards. The learning trend in Fig. 12 is calculated using a moving average of 15 periods.

Age-based vaccination model

Figure 13 shows the progress of the training agent for around 30k episodes. An episode is done once a four-month simulation is run (December 15, 2020, to April 15, 2021). The approximate consumed time to train for the **AVM** is 32.7 h. The trend calculated using a moving average shows an increase as the agent trains in more episodes. This signifies that the agent is learning to win higher rewards, therefore, building knowledge of what action is good in a certain state.

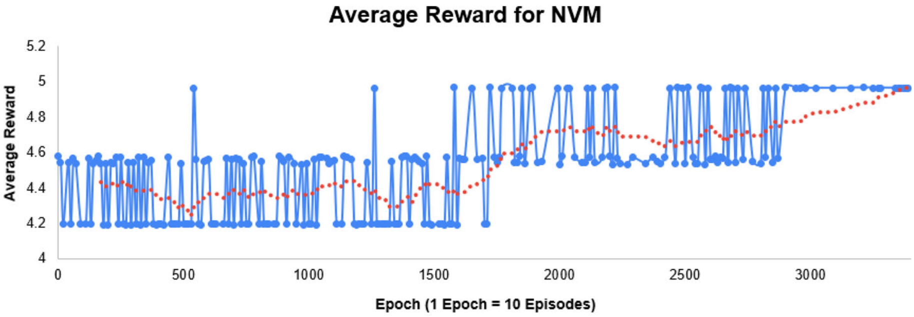


Fig. 12 Reward for each training epoch for the **NVM** model

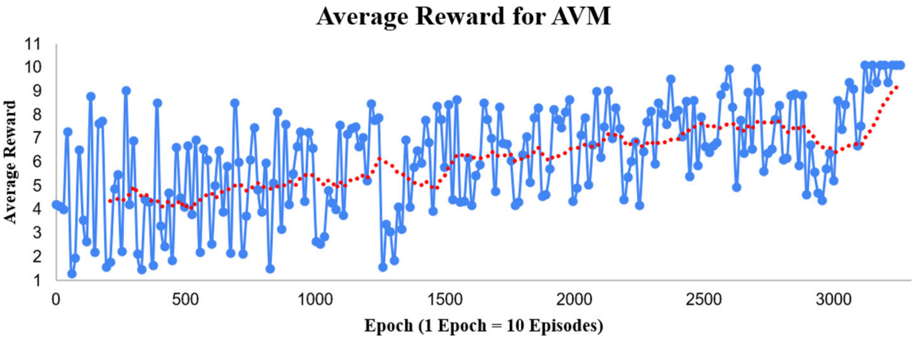


Fig. 13 Reward for each training epoch for the **AVM** model

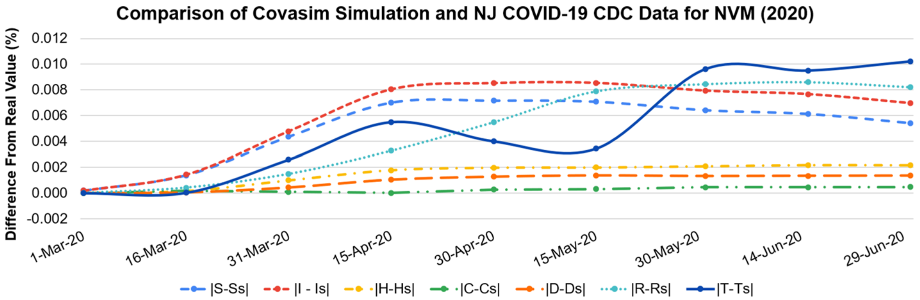


Fig. 14 Comparison of the Covasim agent-based simulation data and CDC data for the state of New Jersey for the period from March 1, 2020, to June 30, 2020. The x-axis shows the timeline where each dot on the trend lines corresponds to a decision. The y-axis shows the absolute value of the difference between the real number of individuals in each compartment (S, I, H, C, D, R, T) at a point in time and the number of individuals estimated from the simulation in that compartment ($S_S, I_S, H_S, C_S, D_S, R_S, T_S$). For example, trend line $|S - S_S|$ represents the absolute difference between the real susceptible proportion of the population (S) and simulated susceptible proportion (S_S) at bi-weekly dates starting March 1 to June 30. Similarly, the trend lines for each compartment are plotted

A2: Validation of intervention effect for NVM

Figure 14 shows the validation of the **NVM** model where we exclude vaccination as an intervention. We present the absolute value of the difference for each compartment between our simulation and the CDC data. The y-axis represents the absolute value of the difference in percentage between the value of each compartment of CDC data and the respective value of the compartment in the simulation. Notice that we are 0.01% away from the real data on a four-month simulation based on the $|T - T_s|$ metric, which refers to the absolute difference between the real treated proportion of the population (T) and simulated treated proportion of the population (T_s) in the worst case. In the best case, the percent difference between the final compartmental statistics of the SiRLat the end of each simulation period and the real data based on the $|C - C_s|$ metric is 0.001%.

A3: Validation of intervention effect for AVM

To model vaccination intervention in our model, we study the period when vaccines became available. On December 11, 2020, The Food and Drug Administration authorized the Pfizer-BioNTech vaccine for emergency use. A week later, on December 18, the Moderna vaccine was also authorized with the same status. Despite this, for the rest of the year 2020, the vaccination campaign is off to a chaotic, confused, and slower-than-expected start, ending up with less than the planned 20 million doses. We model our vaccination models based on the vaccine availability data provided by CDC for Pfizer-BioNTech, Moderna, and Johnson & Johnson vaccines and their respective protection levels by does as research shows. Figure 15 shows the validation for the age-based vaccination strategy (**AVM**) compared to the real compartmental data for NJ. Each line represents the absolute value of the difference between the agent-based simulation model and the real CDC Data reported for NJ, including vaccination. We simulate for four months starting from December 15, 2020, to April 15, 2021. During this period, vaccinations were done according to age and comorbidity-based priority. Our validation is off only 0.12 % during the whole four-month simulation period. The actions

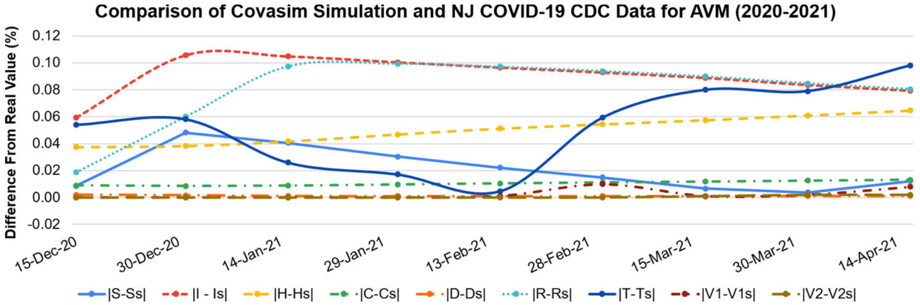


Fig. 15 Comparison of the Covasim agent-based simulation data and CDC date for the state of New Jersey for the period from December 15, 2020, to April 15, 2021. The x-axis shows the timeline where each dot on the trend lines corresponds to a decision. The y-axis shows the absolute value of the difference between the real number of individuals in a compartment at a point in time and the number of individuals estimated in that compartment from the simulation. For example, trend line $|H - H_S|$ represents the absolute value of the difference between the hospitalized individuals in New Jersey and the hospitalized individuals estimated by the simulation at bi-weekly dates starting December 15, 2020, to April 15, 2021. Similarly, the trend lines for each compartment are plotted

Table 5 Paired t-test analysis comparing the bi-weekly compartmental data from the AVM simulation with the actual data from the CDC with confidence interval 95%

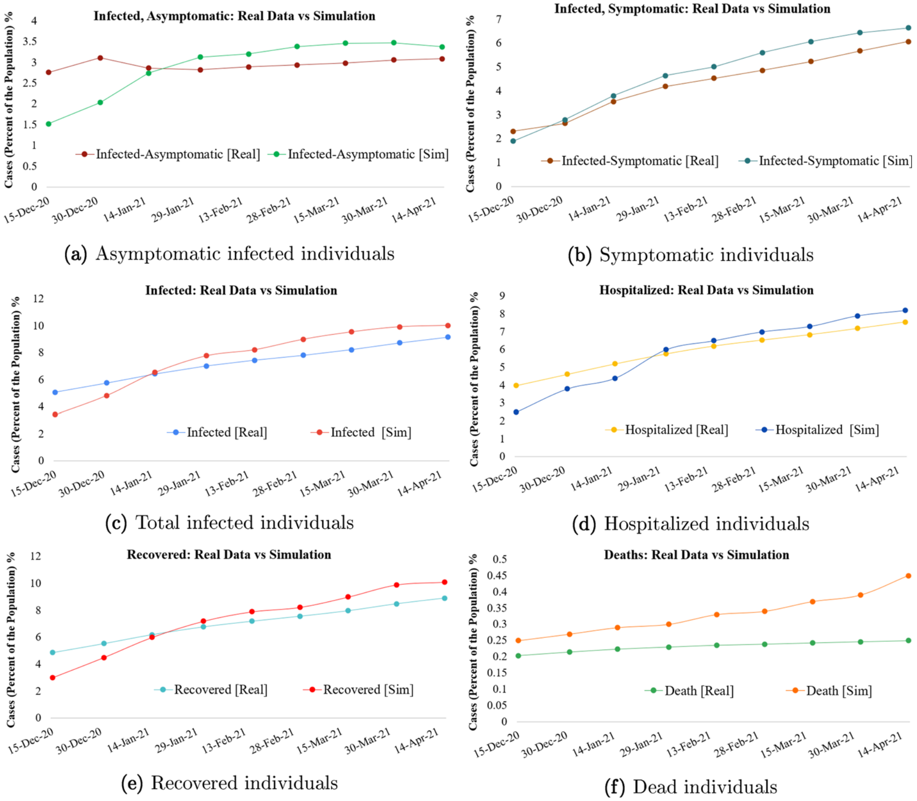
Compartment	Mean		Two-tailed paired t-test		
	Actual	Predicted	t-stat	t-critical	p-value
Infected	677, 437	667, 457	0.03	2.3	0.97
Hospitalized	556, 169	556, 452	0.36	2.3	0.72
ICU	124, 795	125, 920	1.01	2.3	0.40
Dead	21, 525	21, 510	0.27	2.3	0.79
Vaccinated 1	1, 373, 739	1, 368, 578	1.04	2.3	0.35
Vaccinated 2	631, 651	630, 619	0.31	2.3	0.75

calibrated using the agent-based simulation for the age-based vaccination strategy (AVM) will also be used to run a simulation regarding the random vaccination strategy **RVM**.

Further, we apply the paired t-test to investigate the difference between the mean of the bi-weekly compartmental values of the simulation, including one-shot and two-shot vaccinated individuals, and the mean actual values provided by the CDC. According to the statistical analysis shown in Table 5, the mean biweekly AVM simulation DRL results and the mean CDC data values are not statistically different.

A4: Visual comparison of cumulative infections, hospitalizations, and recoveries

Figure 16a–f show comparisons between real values and simulated results from the SiRL framework (with fixed interventions to represent the reality). In almost all cases (Fig. 16a–f) the simulation slightly overestimates the real data. This is possibly due to the under-reporting or biases involved in the reported data. For example, we observe a constant slight overestimate in the death rate comparison (Fig. 16f). On May 5, 2022, WHO published a report where



(a) Asymptomatic infected individuals

(b) Symptomatic individuals

(c) Total infected individuals

(d) Hospitalized individuals

(e) Recovered individuals

(f) Dead individuals

Fig. 16 Comparison between the real (reported) data and the results from the simulation on a population of 9,288,994 (estimated population of NJ on April 1, 2020, by US Census Bureau). **a** Number of infected people with no symptoms. **b** Number of infected people with symptoms. **c** Total number of infected individuals. **d** Number of hospitalized individuals. **e** Total number of recovered individuals. **f** Number of reported deaths

they estimated excess of 14.9 million deaths associated with COVID-19 in 2020 and 2021 (see who.int). This brings forward the point that more effort should be put into modeling the systems as realistically as possible to avoid transferable bias/error coming with data.

A5: Validation for Kansas case study

During the validation process, we calibrate the DRL interventions in our agent-based environment according to the real interventions and their effect over four months starting from September 15, 2021, to January 15, 2022. We collect the bi-weekly data from the CDC and map the same decisions that were active during this considered simulation period. Figure 17 shows the validation for the age-based vaccination strategy (**AVM**) compared to the real compartmental data for the state of Kansas. Each line represents the absolute value of the difference between the agent-based simulation model and the real CDC Data reported for Kansas, including vaccination. In Fig. 17, it is shown that the difference from the real value for all compartmental predictions is less than 0.09%.

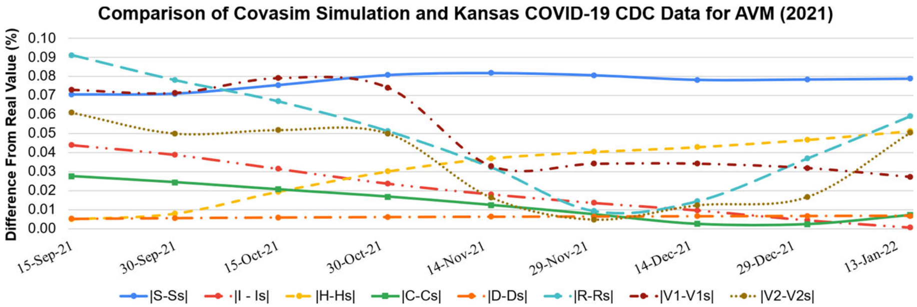


Fig. 17 Comparison of the Covasim agent-based simulation data and CDC data for the state of Kansas for the period from September 15, 2021, to January 15, 2022. The x-axis shows the timeline where each dot on the trend lines corresponds to a decision. The y-axis shows the absolute value of the difference between the real number of individuals in a compartment at a point in time and the number of individuals estimated in that compartment from the simulation. For example, trend line $|H - H_S|$ represents the absolute value of the difference between the hospitalized individuals in Kansas and the hospitalized individuals estimated by the simulation at bi-weekly dates starting September 15, 2021, to January 15, 2022. Similarly, the trend lines for each compartment are plotted

References

Alzu'bi, A. A., Alasal, S. I. A., & Watzlaf, V. J. (2021). A simulation study of coronavirus as an epidemic disease using agent-based modeling. *Perspectives in Health Information Management* **18**.

Ashraf, B. N. (2020). Economic impact of government interventions during the COVID-19 pandemic: International evidence from financial markets. *Journal of Behavioral and Experimental Finance, 27*, 100371.

Awasthi, R., Guliani, K. K., Khan, S. A., Vashishtha, A., Gill, M. S., Bhatt, A., Nagori, A., Gupta, A., Kumaraguru, P., & Sethi, T. (2020). Vacsim: Learning effective strategies for COVID-19 vaccine distribution using reinforcement learning. *arXiv preprint* arXiv:2009.06602.

Bednarski, B. P., Singh, A. D., & Jones, W. M. (2020). On collaborative reinforcement learning to optimize the redistribution of critical medical supplies throughout the COVID-19 pandemic. *Journal of the American Medical Informatics Association, 28*(4), 874–878.

Bell, D. N., & Blanchflower, D. G. (2020). US and UK labour markets before and during the COVID-19 crash. *National Institute Economic Review, 252*, R52–R69.

Bilinski, A., Salomon, J. A., Giardina, J., Ciaranello, A., & Fitzpatrick, M. C. (2021). Passing the test: a model-based analysis of safe school-reopening strategies. *Annals of Internal Medicine*.

Bushaj, S., Büyüktahtakın, İ. E. (2021). A deep reinforcement learning approach for solving multi-dimensional knapsack problem. *Under Review*.

Bushaj, S., Büyüktahtakın, İ. E., & Haight, R. G. (2022). Risk-averse multi-stage stochastic optimization for surveillance and operations planning of a forest insect infestation. *European Journal of Operational Research, 299*(3), 1094–1110.

Bushaj, S., Büyüktahtakın, İ. E., Yemshanov, D., & Haight, R. G. (2020). Optimizing surveillance and management of emerald ash borer in urban environments. *Natural Resource Modeling, 34*(1), e12267.

Büyüktahtakın, İ. E. (2022). Stage-t scenario dominance for risk-averse multi-stage stochastic mixed-integer programs. *Annals of Operations Research, 309*(1), 1–35.

Büyüktahtakın, İ. E., de Bordes, E., & Kıbış, E. Y. (2018). A new epidemics-logistics model: Insights into controlling the Ebola virus disease in West Africa. *European Journal of Operational Research, 265*(3), 1046–1063.

Büyüktahtakın, İ. E., & Haight, R. G. (2018). A review of operations research models in invasive species management: State of the art, challenges, and future directions. *Annals of Operations Research, 271*(2), 357–403.

CDC (2022). COVID data tracker. https://covid.cdc.gov/covid-data-tracker/#datatracker-home. Accessed 20 May 2022.

Chen, I.-M., & Chan, C.-Y. (2021). Deep reinforcement learning based path tracking controller for autonomous vehicle. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235*(2–3), 541–551.

Contreras, S., Dehning, J., Loidolt, M., Zierenberg, J., Spitzner, F. P., Urrea-Quintero, J. H., Mohr, S. B., Wilczek, M., Wibral, M., & Priesemann, V. (2021). The challenges of containing SARS-CoV-2 via test-trace-and-isolate. *Nature Communications, 12*(1), 1–13.

Coşgun, Ö., & Büyüktahtakın, İE. (2018). Stochastic dynamic resource allocation for HIV prevention and treatment: An approximate dynamic programming approach. *Computers & Industrial Engineering, 118*, 423–439.

Dan, J. M., Mateus, J., Kato, Y., Hastie, K. M., Yu, E. D., Faliti, C. E., Grifoni, A., Ramirez, S. I., Haupt, S., Frazier, A., et al. (2021). Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science* **371**(6529).

Dasaklis, T. K., Pappis, C. P., & Rachaniotis, N. P. (2012). Epidemics control and logistics operations: A review. *International Journal of Production Economics, 139*(2), 393–410.

De Mooij, J., Dell Anna, D., Bhattacharya, P., Dastani, M., Logan, B., & Swarup, S. (2021). Quantifying the effects of norms on COVID-19 cases using an agent-based simulation. In *Proceedings of the 22nd international workshop on multi-agent-based simulation (MABS)*.

Delarue, A., Anderson, R., & Tjandraatmadja, C. (2020). Reinforcement learning with combinatorial actions: An application to vehicle routing. *arXiv preprint* arXiv:2010.12001.

D'Orazio, M., Bernardini, G., Quagliarini, E. (2020). How to restart? an agent-based simulation model towards the definition of strategies for COVID-19" second phase" in public buildings. *arXiv preprint* arXiv:2004.12927.

Epstein, J. M. (2009). Modelling to contain pandemics. *Nature, 460*(7256), 687.

Galanakis, C. M., Rizou, M., Aldawoud, T. M., Ucak, I., & Rowan, N. J. (2021). Innovations and technology disruptions in the food sector within the COVID-19 pandemic and post-lockdown era. *Trends in Food Science & Technology*.

Ghaffarzadegan, N., & Rahmandad, H. (2020). Simulation-based estimation of the early spread of COVID-19 in Iran: Actual versus confirmed cases. *System Dynamics Review, 36*(1), 101–129.

Gharakhanlou, N. M., & Hooshangi, N. (2020). Spatio-temporal simulation of the novel coronavirus COVID-19 outbreak using the agent-based modeling approach (case study: Urmia, Iran). *Informatics in Medicine Unlocked, 20*, 100403.

Gillisa, M., Saifa, A., Kamala, N., & Murphy, M. (2021). A simulation-optimization framework for optimizing response strategies to epidemics.

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine, 26*(6), 855–860.

Giuntella, O., Hyde, K., Saccardo, S., & Sadoff, S. (2021). Lifestyle and mental health disruptions during COVID-19. *Proceedings of the National Academy of Sciences* **118**(9).

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., et al. (2020). Practical considerations for measuring the effective reproductive number, r t. *PLoS Computational Biology, 16*(12), e1008409.

Grix, J., Brannagan, P. M., Grimes, H., & Neville, R. (2021). The impact of COVID-19 on sport. *International Journal of Sport Policy and Politics, 13*(1), 1–12.

Gu, M. (2022). Effective reproduction number. https://covid19-study.pstat.ucsb.edu/#tab-9987-4. Accessed 27 May 2022.

Gupta, R., & Morain, S. R. (2021). Ethical allocation of future COVID-19 vaccines. *Journal of Medical Ethics, 47*(3), 137–141.

Hasselt, H. (2010). Double q-learning. *Advances in Neural Information Processing Systems, 23*, 2613–2621.

Higazy, M. (2020). Novel fractional order SIDARTHE mathematical model of COVID-19 pandemic. *Chaos, Solitons & Fractals, 138*, 110007.

Hinch, R., Probert, W. J. M., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Cruz, A. B., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Finkelstein, A., Bonsall, D. G., Abeler-Dörner, L., & Fraser, C. (2020). Openabm-covid19 - an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *medRxiv*.

Joe, W., & Lau, H. C. (2020). Deep reinforcement learning approach to solve dynamic vehicle routing problem with stochastic customers. In: *Proceedings of the international conference on automated planning and scheduling*, Vol. 30, pp. 394–402.

Jones, L., Palumbo, D., & Brown, D. (2021). Coronavirus: How the pandemic has changed the world economy. https://www.bbc.com/news/business-51706225. Accessed 06 July 2021.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, 115*(772), 700–721.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Rosenfeld, K., Hart, G. R., Núñez, R. C., Cohen, J. A., Selvaraj, P., Hagedorn, B., et al. (2021). Covasim: An agent-based model of covid-19 dynamics and interventions. *PLOS Computational Biology, 17*(7), e1009149.

Khalilpourazari, S., & Doulabi, H. H. (2021a). Designing a hybrid reinforcement learning based algorithm with application in prediction of the covid-19 pandemic in quebec. *Annals of Operations Research*, pp. 1–45.

Khalilpourazari, S., & Doulabi, H. H. (2021b). Using reinforcement learning to forecast the spread of covid-19 in france. In *2021 IEEE international conference on autonomous systems (ICAS)*, pp. 1–8. IEEE.

Kıbış, E. Y., & Büyüktahtakın, İE. (2019). Optimizing multi-modal cancer treatment under 3d spatio-temporal tumor growth. *Mathematical Biosciences, 307*, 53–69.

Kıbış, E. Y., Büyüktahtakın, İ. E., Haight, R. G., Akhundov, N., Knight, K., & Flower, C. (2020). A multi-stage stochastic programming approach to the optimal surveillance and control of emerald ash borer in cities. *INFORMS Journal on Computing*, pp. 1–36.

Kieu, L.-M., Malleson, N., & Heppenstall, A. (2020). Dealing with uncertainty in agent-based models for short-term predictions. *Royal Society Open Science, 7*(1), 191074.

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. *arXiv preprint* arXiv:2010.10560.

Kong, W., Liaw, C., Mehta, A., & Sivakumar, D. (2018). A new dog learns old tricks: Rl finds classic optimization algorithms. In *International conference on learning representations*.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine, 172*(9), 577–582.

Li, J., Giabbanelli, P., et al. (2021). Returning to a normal life via COVID-19 vaccines in the USA: A large-scale agent-based simulation study. *JMIR Medical Informatics, 9*(4), e27419.

Lin, Y., McPhee, J., & Azad, N. L. (2020). Comparison of deep reinforcement learning and model predictive control for adaptive cruise control. *IEEE Transactions on Intelligent Vehicles, 6*(2), 221–231.

Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems, 29*(6), 2063–2079.

McKeever, V. (2020). The coronavirus is expected to have cost 400 million jobs in the second quarter, un labor agency estimates. https://www.cnbc.com/2020/06/30/coronavirus-expected-to-cost-400-million-jobs-in-the-second-quarter.html. Accessed 06 July 2021.

Mehrotra, S., Rahimian, H., Barah, M., Luo, F., & Schantz, K. (2020). A model of supply-chain decisions for resource sharing with an application to ventilator allocation to combat COVID-19. *Naval Research Logistics (NRL), 67*(5), 303–320.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint* arXiv:1312.5602.

Moghadas, S. M., Vilches, T. N., Zhang, K., Nourbakhsh, S., Sah, P., Fitzpatrick, M. C., & Galvani, A. P. (2021). Evaluation of COVID-19 vaccination strategies with a delayed second dose. *PLoS Biology, 19*(4), e3001211.

Müller, S. A., Balmer, M., Charlton, W., Ewert, R., Neumann, A., Rakow, C., Schlenther, T., Nagel, K. (2021). Predicting the effects of COVID-19 related interventions in urban settings by combining activity-based modelling, agent-based simulation, and mobile phone data. *medRxiv*.

Ngonghala, C. N., Iboi, E. A., & Gumel, A. B. (2020). Could masks curtail the post-lockdown resurgence of covid-19 in the us? *Mathematical Biosciences, 329*, 108452.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research, 290*(1), 99–115.

NJ (2021). COVID-19 information hub. https://covid19.nj.gov/forms/datadashboard. Accessed 06 July 2021.

Ohi, A. Q., Mridha, M., Monowar, M. M., & Hamid, M. A. (2020). Exploring optimal control of epidemic spread using reinforcement learning. *Scientific Reports, 10*(1), 1–19.

Onal, S., Akhundov, N., Büyüktahtakın, İ. E., Smith, J., & Houseman, G. (2020). An integrated simulation-optimization framework to optimize search and treatment path for controlling a biological invader. *International Journal of Production Economics, 222*, 107507.

Onal, S., Bushaj, S., Büyüktahtakın, İ. E., & Houseman, G. (2021). A Gaussian dispersal approach to capture long-term and long-distance dispersal through simulation-optimization. *Working Paper*.

Poudel, P. B., Poudel, M. R., Gautam, A., Phuyal, S., Tiwari, C. K., Bashyal, N., & Bashyal, S. (2020). COVID-19 and its global impact on food and agriculture. *Journal of Biology and Today's World, 9*(5), 221–225.

Queiroz, M. M., Ivanov, D., Dolgui, v, & Wamba, S. F. (2020). Impacts of epidemic outbreaks on supply chains: mapping a research agenda amid the COVID-19 pandemic through a structured literature review. *Annals of Operations Research*, pp. 1–38.

Rahmandad, H., Lim, T. Y., & Sterman, J. (2021). Behavioral dynamics of covid-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations. *System Dynamics Review, 37*(1), 5–31.

Rocha, R. (2020). What countries did right and wrong in responding to the pandemic. https://www.cbc.ca/news/canada/covid-19-coronavirus-pandemic-countries-response-1.5617898. Accessed 06 July 2021.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint* arXiv:1511.05952.

Shamil, M. S., Farheen, F., Ibtehaz, N., Khan, I. M., & Rahman, M. S. (2021). An agent-based modeling of COVID-19: Validation, analysis, and recommendations. *Cognitive Computation*, pp. 1–12.

Sigala, M. (2020). Tourism and covid-19: Impacts and implications for advancing and resetting industry and research. *Journal of Business Research, 117*, 312–321.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science, 362*(6419), 1140–1144.

Tanner, M. W., Sattenspiel, L., & Ntaimo, L. (2008). Finding optimal vaccination strategies under parameter uncertainty using stochastic programming. *Mathematical Biosciences, 215*(2), 144–151.

Tareq, M. S., Rahman, T., Hossain, M., & Dorrington, P. (2021). Additive manufacturing and the COVID-19 challenges: An in-depth study. *Journal of Manufacturing Systems*.

Thebault, R., Meko, T., & Alcantara, J. (2021). Sorrow and stamina, defiance and despair. It's been a year. https://www.washingtonpost.com/nation/interactive/2021/coronavirus-timeline/. Accessed 06 July 2021.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

Wan, R., Zhang, X., & Song, R. (2020). Multi-objective reinforcement learning for infectious disease control with application to COVID-19 spread. *arXiv preprint* arXiv:2009.04607.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292.

Wu, Y., Mansimov, E., Grosse, R. B., Liao, S., & Ba, J. (2017). Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. *Advances in Neural Information Processing Systems, 30*, 5279–5288.

Yin, X., & Büyüktahtakın, İE. (2021). A multi-stage stochastic programming approach to epidemic resource allocation with equity considerations. *Health Care Management Science, 24*, 597–622.

Yin, X., & Büyüktahtakın, İE. (2022). Risk-averse multi-stage stochastic programming to optimizing vaccine allocation and treatment logistics for effective epidemic response. *IISE Transactions on Healthcare Systems Engineering, 12*(1), 52–74.

Yin, X., Büyüktahtakın, İ. E., & Patel, B. P. (2021). Covid-19: Data-driven optimal allocation of ventilator supply under uncertainty and risk. *European Journal of Operational Research, 304*(1), 255–275. https://doi.org/10.1016/j.ejor.2021.11.052.

Zhou, S. K., Le, H. N., Luu, K., Nguyen, H. V., & Ayache, N. (2021). Deep reinforcement learning in medical imaging: A literature review. *Medical Image Analysis, 73*, 102193.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.