UNIVERSITY OF OKLAHOMA GRADUATE COLLEGE

SMART POLICY DESIGN FOR COVID-19:
A DEEP REINFORCEMENT LEARNING FRAMEWORK

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements
for the Degree of

MASTER OF SCIENCE

By Azmayn Inkishaf

Norman, Oklahoma

2025

SMART POLICY DESIGN FOR COVID-19:
A DEEP REINFORCEMENT LEARNING FRAMEWORK

A THESIS APPROVED FOR THE

GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Talayeh Razzaghi, Chair

Dr. Charles Nicholson

Dr. Andres Gonzalez

Contents

1 Introduction **1**

2 Methodology **2**
2.1 Dataset 2
2.2 Reinforcement Learning 6
2.3 Simulation 7

3 Results **9**
3.1 Interventions 9
3.2 Comparison 14
3.3 Evaluation 16

4 Discussion **18**

5 Conclusion **19**

Abstract

More than 700 million people became infected and about 7 million have died from COVID-19, making it one of the deadliest pandemics in history. It triggered severe economic recession around the world that disrupted multiple industries such as agriculture, manufacturing and tourism. Governments around the world responded by implementing interventions to control disease transmission, but their impact varied from one country to another. There have been numerous studies to understand the effectiveness of the interventions but there are considerable variations in the interpretation. However, it is evident that tailoring policies for individual regions makes them more impactful. In the research, we decided to focus on regions of the United States and treat each state individually because of their diversity. We employ deep reinforcement learning to handle the dynamic nature of the disease, specifically multi-agent reinforcement learning where individual agent handles distinct states in a shared space to mimic realistic environment. The agent intervention differed from the actual intervention most of the time. Overall suggestion is to implement intervention earlier with higher intensity then gradually reduce aggressiveness and maintain a moderate level of intensity. However, this will depend on how much we want to prioritize public health over national economy.

1 Introduction

The Coronavirus pandemic has been one of the biggest calamity in modern history, profoundly affecting countries worldwide. What began as a localized outbreak quickly escalated into a global emergency and the World Health Organization categorized it as a pandemic in March 2020. More than 700 million people became infected and about 7 million have died, marking it as one of the deadliest diseases in history. However, the crisis extended beyond health, triggering a severe economic recession that disrupted multiple industries. Sectors such as agriculture [1], manufacturing [2] and tourism [3] suffered immense losses, leading to widespread financial instability and job insecurity. The financial markets witnessed the biggest crash since 1987 [4] and the International Labor Organization estimated that around 400 million full-time positions were dissolved [5]. Many small businesses struggled to recover, with countless facing permanent closures due to financial strain from reduced consumer demand. The pandemic revealed the limitations in healthcare systems, business operations and governance structures, creating long-term challenges for policymakers. Addressing these crises required difficult trade-offs as no single solution could resolve the complex issues. As countries struggled to contain the disease, an immediate response became essential to mitigate both the short-term devastation and the long-lasting consequences.

Governments around the world have implemented various measures to combat disease transmission and mitigate the impact of the virus. Many countries enforced strict lockdowns, restricted movement and ordered closures to dampen the spread. Social distancing rules, mask-wearing mandates, and public health campaigns were introduced to encourage protective behaviors and reduce the spread of disease. However, these strategies varied significantly in timing, intensity and duration across the world, leading to differing outcomes. Although countries like New Zealand effectively contained the virus through strict border controls and a swift response, others such as Brazil struggled because of inconsistent governmental responses and lack of coordination. As restrictions became widespread, researchers analyzed their effectiveness, but conclusions were mixed based on regional differences [6–15]. Some studies found that lockdown significantly reduced cases in Europe but had minimal impact in less developed countries in Asia. Others suggested that school closures lowered transmission in Central Asia, whereas in North America, they paradoxically increased infections due to changing childcare responsibilities. Consequently, governments had to weigh the trade-offs between controlling the virus and managing social, political and economic disruptions. Evidence on policy effectiveness remains inconsistent, as outcomes varied depending on factors such as local climate, population density and healthcare capabilities.

Despite numerous studies on the disease in the United States, few have focused on a granular level of analysis [16–19]. Research indicates that addressing the pandemic at the regional level improves predictions, suggesting that it can allow for more effective policy implementation [20]. Each state in the U.S. is diverse, makes the policies independently and centralized decision-making will be suboptimal. Furthermore, most existing studies are based on epidemiological models which generally lack adaptability in changing environments [21–25]. Unlike traditional models, reinforcement learning optimizes decisions dynamically, considers multiple factors and evaluates different scenarios [26–30]. Q-learning (QL), Deep Q-Network (DQN), Deep Deterministic Policy Gradient (DDPG), Soft Actor Critic (SAC) and State Action Reward State Action (SARSA) are some examples of algorithms and their implementation depends on the application. This paper aims to leverage reinforcement learning to optimize government interventions for the COVID-19 pandemic across states of US. Unlike countries, states share strong interdependence, requiring a coordinated yet flexible approach to policymaking. We propose using decentralized multi-agent reinforcement learning, allowing states to make independent decisions while remaining aware of the policies of others in the environment. This method ensures adaptive responses, reduces policy conflicts and improves the overall effectiveness of intervention strategies. Table 1 outlines papers that employ reinforcement learning to optimize government interventions.

Table 1: Existing papers that uses RL to optimize policies for COVID-19

Paper	**Model**	**Dataset**	**Variable**
Kwak et al. [26]	DQN	216 countries	Daily cases, daily recovery, daily deaths
Bushaj et al. [27]	DQN	Simulation	Daily cases, daily deaths, economic impact
Padmanabhan et al. [28]	DQN	Simulation	Daily cases, hospital capacity, economic impact
Kompella et al. [29]	SAC	Simulation	Daily cases, hospital capacity, economic impact
Uddin et al. [30]	QL, DQN, DDPG, SARSA	Simulation	Daily cases, hospital capacity, economic impact

2 Methodology

2.1 Dataset

The first dataset used n this research consists of infection numbers that were collected by John Hopkins University [31]. The statistics are recorded at county level which can provide granular insight. It is a time series spanning from January 2020 to December 2020. Each of the entries represents a daily record of cumulative infections. Some of the variables related to this research are county, state and cases. The dataset was first aggregated to state level to make it compatible with other datasets. A 7-day moving average was performed on daily cases to account for daily fluctuations from reporting inconsistency [32]. The change in the modified number for daily cases was used to calculate the infection rate. Figure 1 below shows the trend curves of daily cases in the United States. The blue line represents the actual record of daily cases, which has a lot of sudden increases.

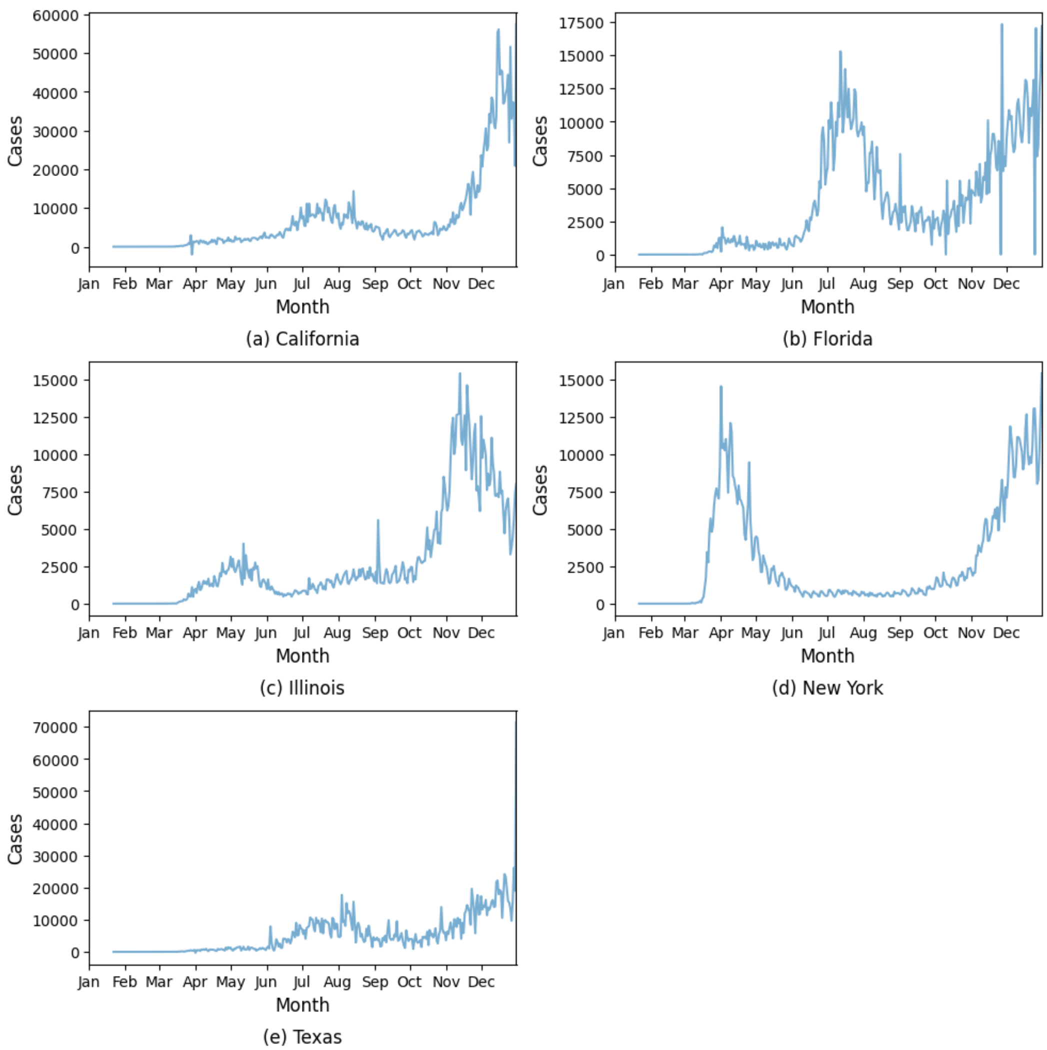
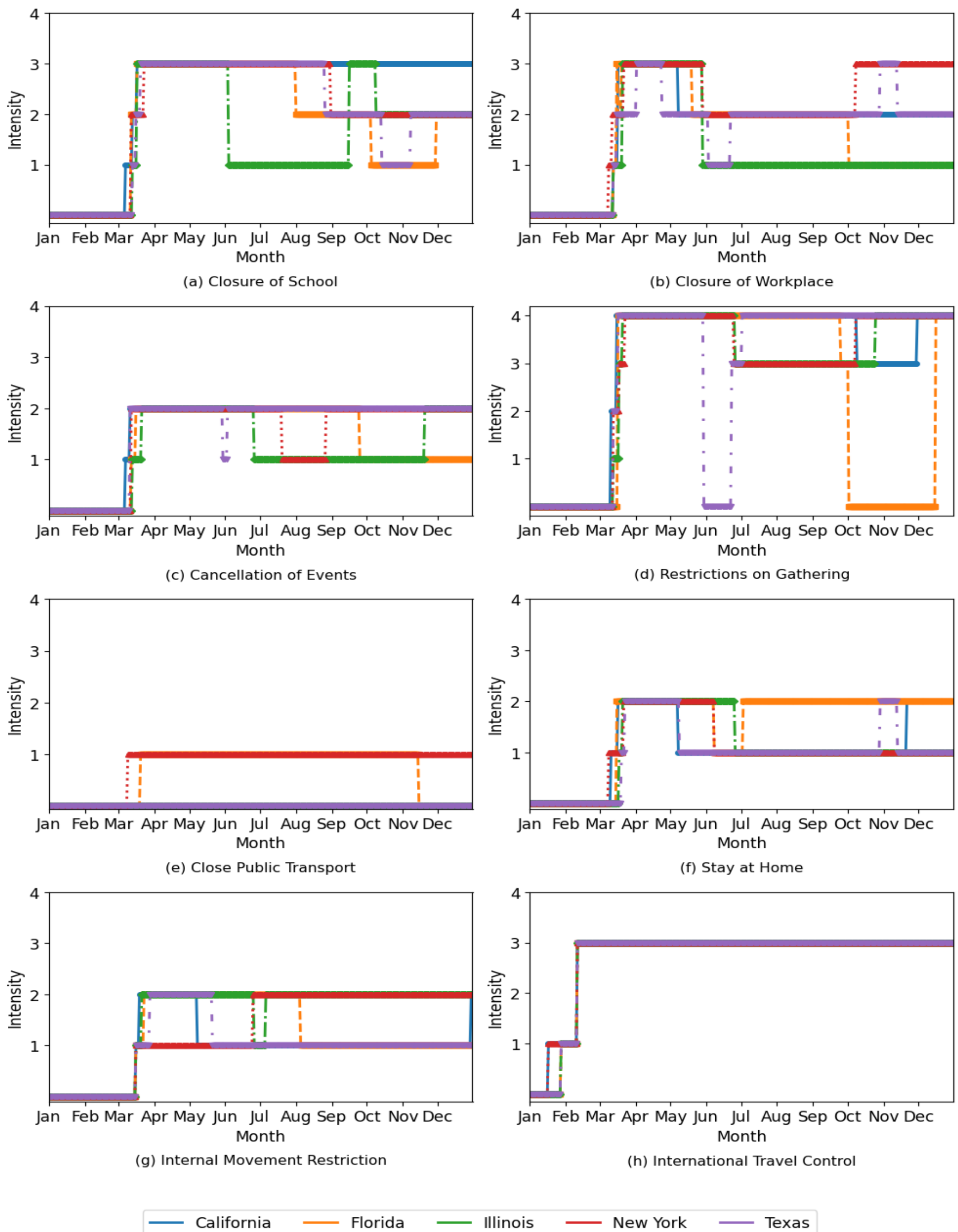


Figure 1: 2020 daily cases by state

The second dataset used in this research consists of government interventions that were collected by International Business Machines [33]. The statistics are recorded at state level since the policies are decided by the governor. It is a time series spanning from January 2020 to December 2020. Each of the entries represents a daily record of the intensity of the implemented interventions. Some of the variables related to this research are state, jurisdiction and interventions. There are eight different interventions, which are school closure, workplace closure, canceled events, gathering restrictions, stopping transportation, required quarantine, restricted movement and travel bans [34]. Each of these has five intensities represented by an ordinal number, with zero representing none and 4 representing high. The dataset was partitioned by state for compatibility. Figure 2 below shows the timeline of interventions at each state. The color of the lines differentiates the state of US.



(a) Closure of School
(b) Closure of Workplace
(c) Cancellation of Events
(d) Restrictions on Gathering
(e) Close Public Transport
(f) Stay at Home
(g) Internal Movement Restriction
(h) International Travel Control

Figure 2: 2020 government interventions by state

The third dataset used in this research consists of public expenditures that were collected by Affinity Solutions Incorporation [35]. The statistics are recorded at state level and contains information of different types of spending. It is a time series spanning from January 2020 to December 2020. Each of the entries represents a daily record of the change in spending in relation to 14th January, 2020. Some of the variables related to this research are FIPS, frequency and spending. The dataset includes total spending, spending by product and spending by income. We will focus on the total spending from each state for our purpose. The information on spending was missing for the first few days and these days were omitted from analysis. The FIPS codes were mapped to their corresponding state names and the dataset was organized by state. Figure 3 below shows the change in spending by state.

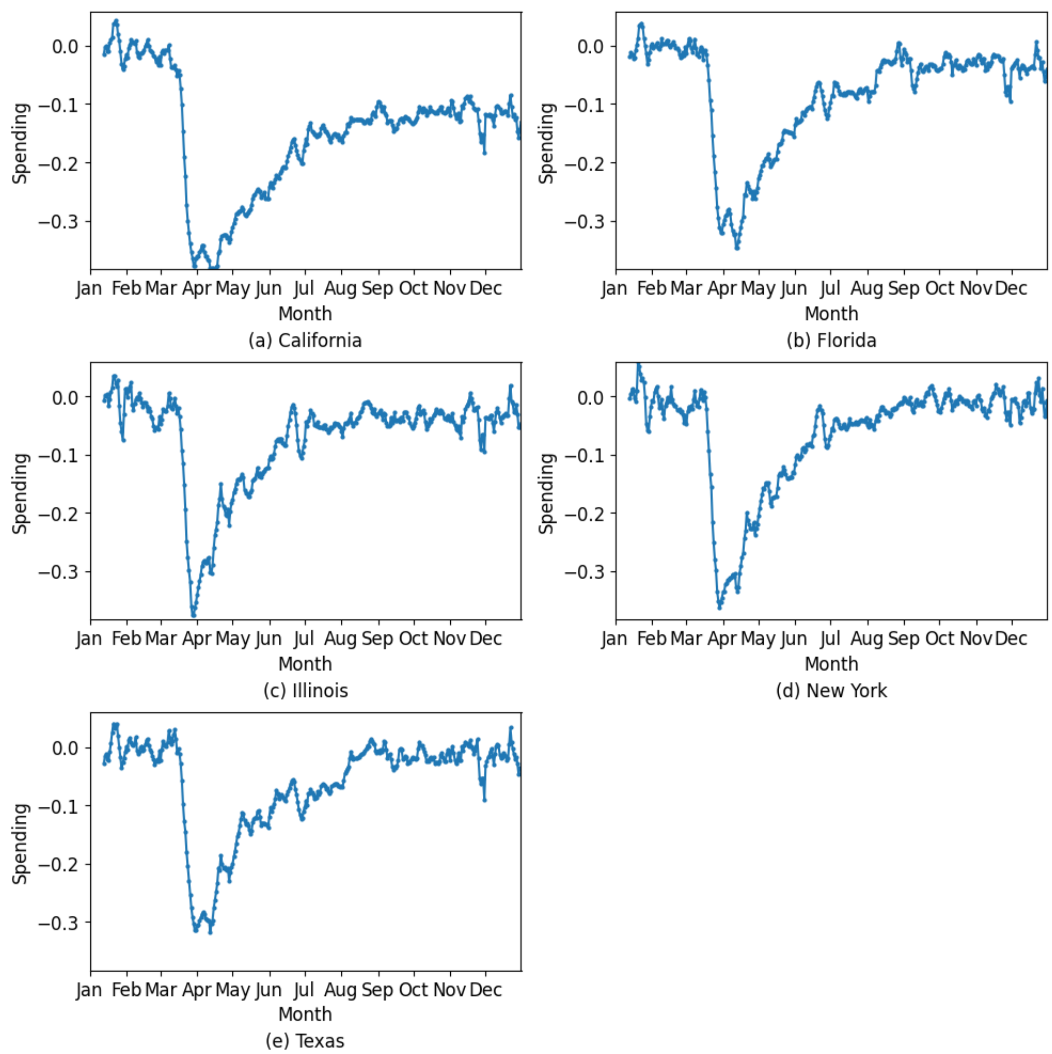


Figure 3: 2020 population spending by state

2.2 Reinforcement Learning

Edward Thorndike introduced the concept of the 'Law of Effect,' inspired by the foundational principle of trial-and-error [36]. This concept suggests that a response producing a satisfying outcome has higher probability of repeating than a response producing a unpleasant outcome. Claude Shannon presented one of the earliest practical demonstrations of trial and error learning with his maze-running mechanical mouse [37]. This marked a significant advancement in understanding automated learning mechanisms. Ivan Pavlov first introduced the term "reinforcement" in his studies on conditioned reflexes and behavioral conditioning [38]. The application of trial-and-error concepts in computer systems became crucial in the evolution of artificial intelligence. Michie and Chambers developed a reinforcement learning agent called GLEE and a controller named BOXES designed for tic-tac-toe [39]. Their work became a significant reference in reinforcement learning. Sutton and Barto expanded these ideas by developing classical conditioning models based on temporal difference [40, 41]. This integrated concepts from Klopf's research and animal psychology theories.

Reinforcement Learning is machine learning where an agent learns to optimize actions through interaction with the environment and receiving feedback as rewards. This learning is inspired by behavioral psychology, mimicking how animals learn from trial and error. The agent's objective is to choose actions that maximizes reward over time. The agent explores various actions and gradually improves its strategy based on overall experiences. This is useful in complex environments where explicit models are difficult to design [42]. Reinforcement learning draws from multiple disciplines which includes modern human psychology, optimal control theory and temporal difference learning. A key mathematical foundation of the algorithm is the Markov Decision Process which models decision-making in environments with stochastic transitions. It consists of states, actions and rewards for every possibility that is visited. The core of solving the problem lies in evaluating the value of states and action. This leads directly to the formulation of Bellman's equation [43].

$$Q(s,a) = R(s,a) + \gamma \max_{a'} Q(s',a') \tag{1}$$

where:

- $Q(s,a)$ = Value from action-state pair
- $R(s,a)$ = Reward resulting from the action
- γ = Choosen importance for future rewards
- s' = Next state after certain action
- a' = Possible action in next state

Multi-Agent Reinforcement Learning (MARL) is an extension of reinforcement learning where agents learn simultaneously by interacting with a shared environment. The collaboration between agents can be centralized, decentralized or mixed. Agents learn policies based on local observations, without access to the global state. The objective of each agent is to maximize individual cumulative reward. Learning dynamics in MARL are driven by joint actions, where the state transitions depend on the combined behavior of agents [44, 45]. In this study, each state is modeled as an independent agent in a decentralized framework. Our problem has a large number of states which cannot be handled by basic models. Hence, each agent employs Deep Q-Network, which uses neural network to approximate Q-function, mapping local state-action pairs to expected future rewards. The reward function encodes trade-offs between public health, economic impact and intervention cost to penalize aggressive intervention. The variables are normalized, which makes them unitless and comparable to each other.

$$R_t = -\alpha \cdot C_t + \beta \cdot E_t - \gamma \cdot I_t \tag{2}$$

where:

- R_t = Immediate reward at time t
- C_t = Infection rate at time t
- E_t = Spending change at time t
- I_t = Average intensity at time t
- α = Weight assigned to infection rates
- β = Weight assigned to spending change
- γ = Weight assigned to average intensity

2.3 Simulation

The environment model is a data-driven approximation of real-world disease dynamics. It simulates how government interventions affect infection rates and economic conditions. The agent learns by interacting with the environment model to observe the effects of different intervention strategies. This enables repeated simulations to evaluate different action combinations under controlled conditions. Agents are trained using historical datasets from the five states to create a realistic environment for individual locations. The dataset was modified by introducing a column for infection rate seven days later to account for the delayed effect of the chosen actions. A fully-connected feedforward neural network is used to understand any non-linear relationships between actions and outcomes [46]. This can allow generalization across different policy patterns. It contains two nodes for input layer, 32 nodes for hidden layer and

3 nodes for output layer. Performance of the model is evaluated using root mean squared error (RMSE) for Rt, spending and reward.

Reinforcement learning model is trained through repeated rollouts in the simulated environment by conducting a series of forward interactions [47]. In each rollout, state is observed, action is selected and reward is given. An episode is defined as a 365-day simulation where the agent makes a new decision daily based on the current state. We train the model for 3,000 episodes to provide exposure to varied conditions. Epsilon controls exploration, starting at 1.00 and decaying to 0.1, shifting behavior from random to optimal. The discount factor gamma is set to 0.95 to balance importance of current reward and future reward [48]. State is current infection rate and spending change, action are the eight available interventions and reward is delayed infection rate and spending change. At every step, the model records state, action and reward in a replay buffer. Learning involves minimizing the difference between predicted and target Q-values based on reward. Over time, the model converges as loss decreases and policy decisions stabilize [49].

Simulation is the process of applying a trained agent to a controlled environment to evaluate its policy. It mimics deployment by allowing the agent to make decisions using real inputs and predicted outcomes [50]. The purpose of this simulation is to assess the effectiveness of the learned policy. It also provides a basis for comparing suggested and government interventions. The first seven days are initialized using observed Rt values to provide the agent with realistic starting inputs. On day one, the trained agents suggest action based on current state, the environment model predicts delayed Rt and the reward is calculated based on that information. This is repeated for consecutive 364 days. Cumulative reward is tracked and used to evaluate overall performance. A higher cumulative reward indicates the agent effectively reduced Rt with minimal economic impact. Results showing sustained Rt control will confirm the agent's ability to optimize its strategy over time.

3 Results

3.1 Interventions

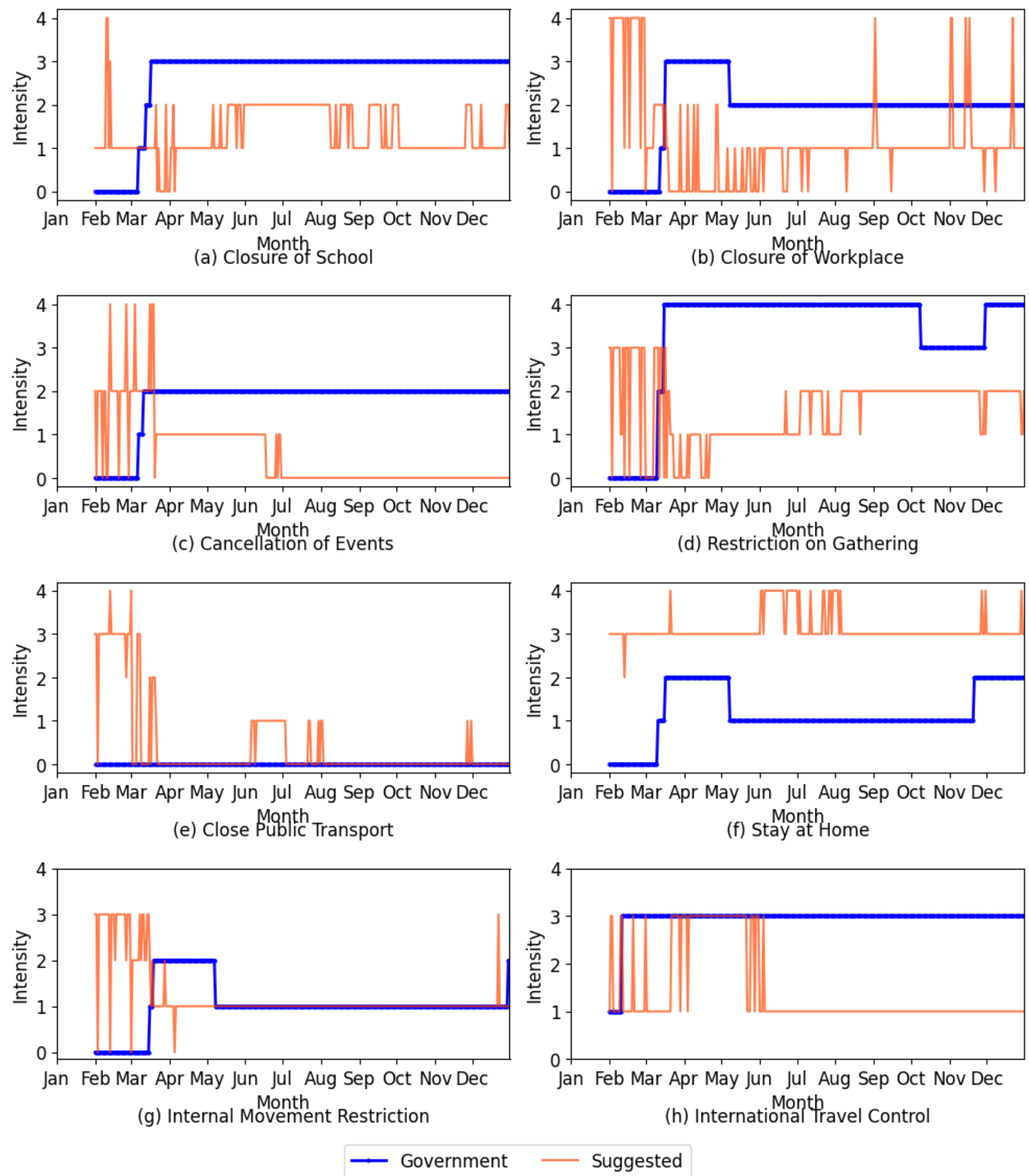


Figure 4: 2020 intervention comparison for California

Figure 4 presents a comparison between the agent's suggested interventions and the actual government interventions for California during 2020. The government adopted a higher intensity approach early in the pandemic and maintained relatively stable intervention levels throughout the year. This likely reflects California's consistently high case numbers, driven by its dense urban population which promotes extensive social mobility. In contrast, the agent proposed a more dynamic intervention strategy by frequently adjusting policy intensity over time. These adjustments appear to represent an effort to strike a more flexible balance between effective disease control and preserving economic activity across different pandemic phases.

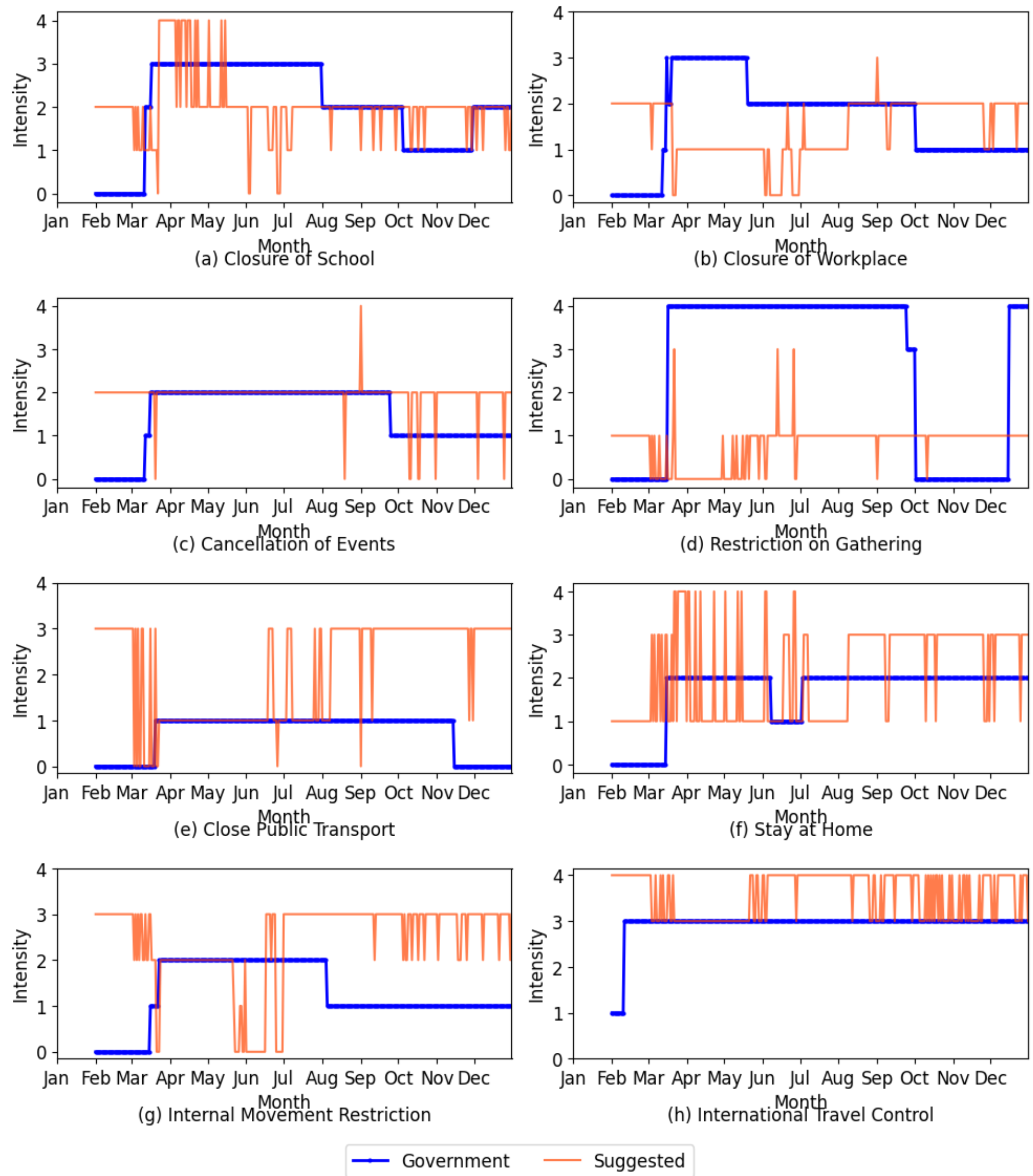


Figure 5: 2020 intervention comparison for Florida

Figure 5 illustrates the comparison between the agent's suggested interventions and the actual government interventions in Florida during 2020. The state began with relatively high-intensity measures which were gradually reduced as the year progressed. This decline was largely influenced by the governor's stance that strict interventions were ineffective which resulted in loosening the restrictions to enable growth in the economy. In contrast, the model suggested increasing intervention intensity in the second half of the year. This recommendation likely stems from the seasonal increase in tourism during that period which contributed to a surge in daily cases and heightened transmission risk.

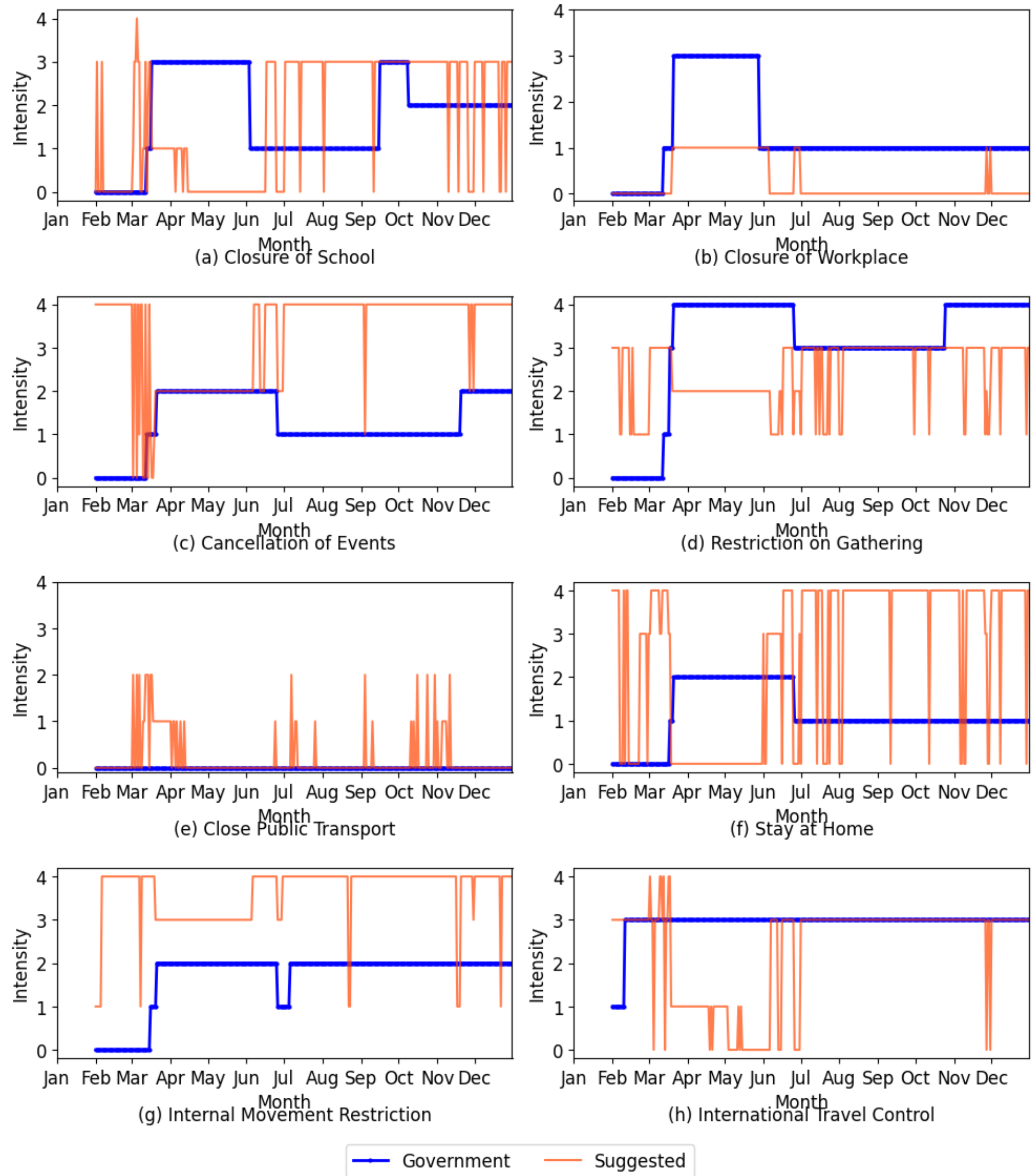


Figure 6: 2020 intervention comparison for Illinois

Figure 6 displays a comparison between the agent's suggested interventions and the actual government interventions in Illinois throughout 2020. The government adopted a stepped approach by starting at lower intensity and gradually increasing the intensity of interventions over time. This likely reflects the absence of sudden spikes in case numbers which allowed for a more controlled response. Interestingly, the agent's suggested interventions closely mirrored this pattern in terms of timing. This alignment suggests that Illinois effectively balanced public health concerns with economic impact considerations in a manner consistent with the model's optimization strategy, indicating sound decision-making under evolving pandemic conditions.

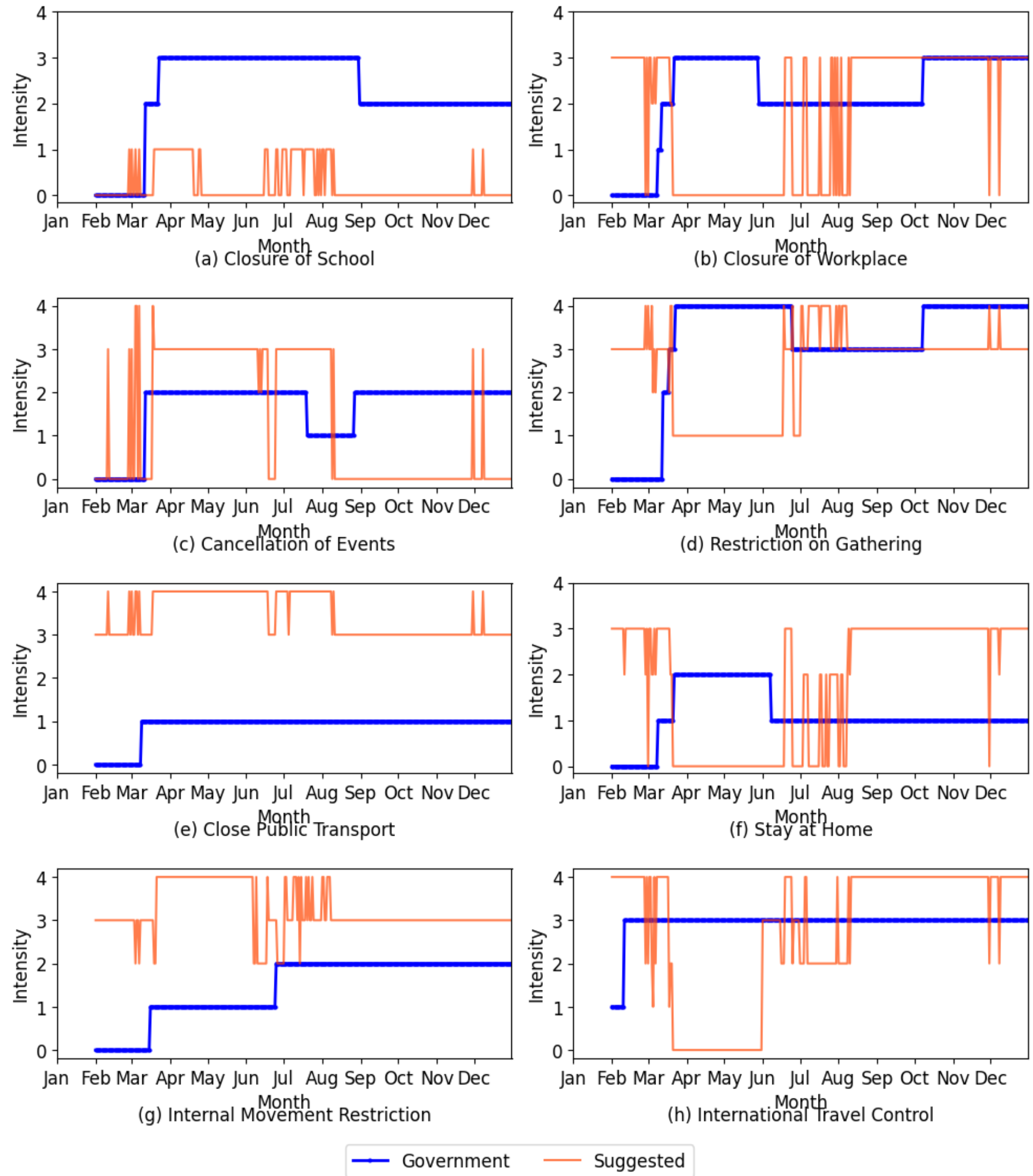


Figure 7: 2020 intervention comparison for New York

Figure 7 compares the agent's suggested interventions with the actual government intervention in New York during 2020. The government maintained a consistently high level of intervention throughout the year in response to persistently high case numbers. This approach was likely driven by the urgent need to control widespread transmission in densely populated areas. However, the agent recommended even higher intervention intensity than what was enacted. New York City was the epicenter for the disease because of the increased mobility from various activities that resulted in higher infections and this extreme burden warranted even stricter control measures according to the model.

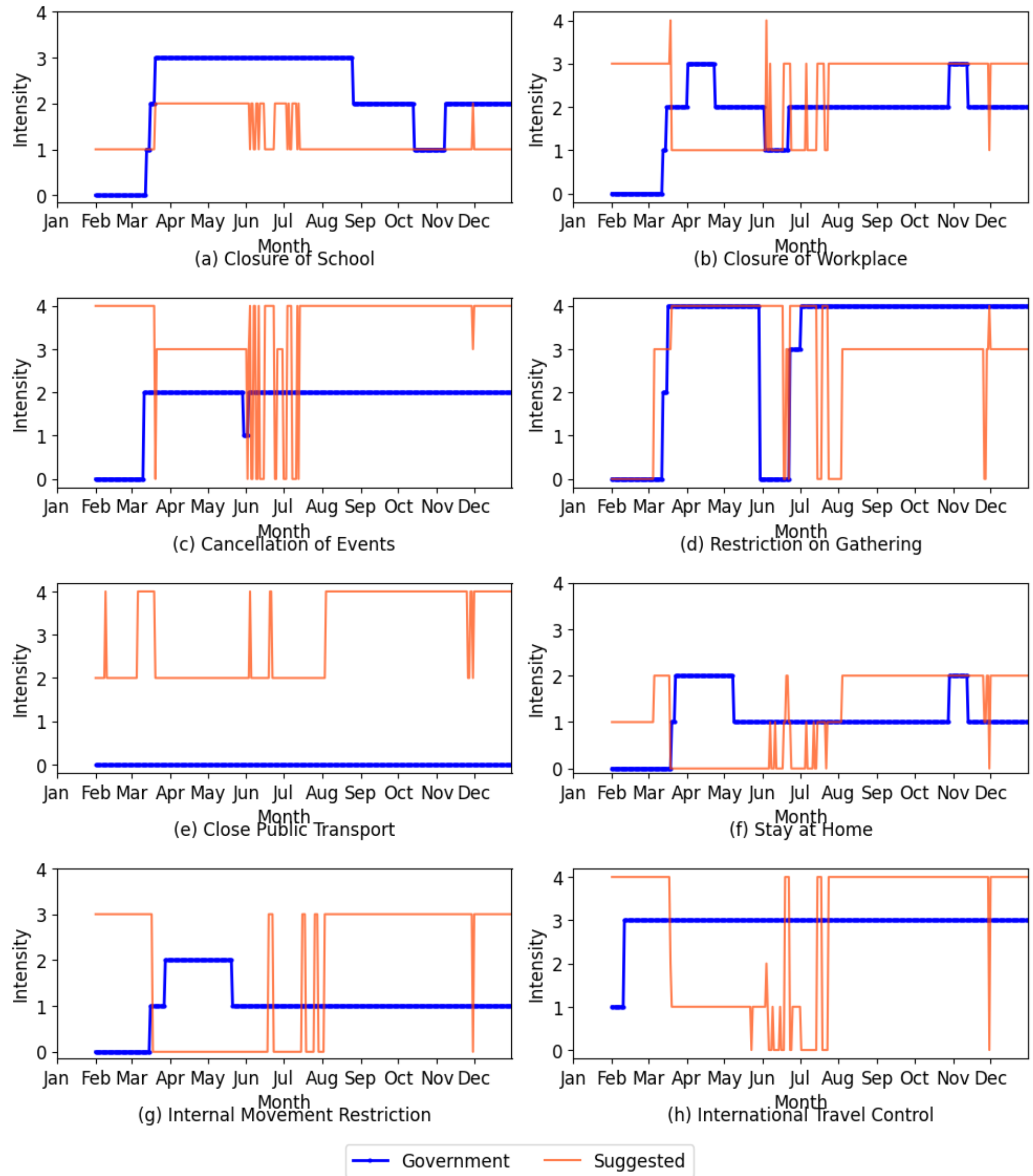


Figure 8: 2020 intervention comparison for Texas

Figure 8 shows a comparison between the agent's recommended interventions and the actual government intervention in Texas during 2020. The government maintained a relatively consistent level of intervention with minimal adjustments over time. This approach likely reflects Texas's decentralized strategy where local authorities were given the responsibility to manage the transmission of disease. In contrast, the model suggested more intense interventions that varied dynamically in response to the evolving situation. These fluctuations in policy intensity likely correspond to the model's reaction to changing infection numbers which indicated a need for greater responsiveness to local disease trends than what was implemented.

3.2 Comparison

Table 2: RMSE values for the target variables in the reward function

Variables	California	Florida	Illinois	New York	Texas
Infection Rate	0.269	0.087	0.152	0.141	0.185
Spending Change	0.023	0.021	0.019	0.029	0.022
Reward	0.067	0.066	0.047	0.061	0.066

Table 2 displays RMSE values for predicted infection rate, spending change, and expected reward. RMSE quantifies prediction accuracy by measuring average error magnitude. Lower RMSE suggests better model performance. This metric is especially useful for comparing prediction reliability across various setups. The table helps evaluate the model's generalization across different contexts. Florida has the lowest and California has the highest RMSE for infection rate. Illinois has the lowest and California has the highest RMSE for spending change. It is the same for expected reward. Overall, the values are very low. This indicates that the environment model can create realistic situations.

Table 3: Finding the optimal number of episodes for the model training

Total Episodes	Final Reward	Average Loss	Converged Status
1000	-0.57	0.11	No
2000	-0.39	0.05	Yes
3000	-0.22	0.03	Yes
4000	-0.19	0.03	Yes
5000	-0.17	0.02	Yes

Table 3 shows the effect of increasing episode number. In one episode, agents observed 364 days of intervention. Average loss is calculated as the average of the last five iterations from five agents. Objective is to identify the optimum number of episodes. This is done by balancing ending rewards, resource usage and loss convergence. Usually, 1000 episodes are enough but loss value does flatten. The loss value first converged at 2000 episodes. At 3000 episodes, the agents have that improved final rewards. This is optimum because higher value does not improve reward significantly. At this point the model is trained properly.

Table 4: Change in reward form altering the weights in reward function

Alpha (α)	**Beta (β)**	**Gamma (γ)**	**Cumulative Reward**
1.0	0.0	0.25	-76.66
1.0	0.0	0.75	-94.49
1.0	1.0	0.00	-133.64
1.0	0.0	0.00	-142.31
1.0	0.0	0.50	-210.99
1.0	0.0	1.00	-225.03
1.0	0.25	0.00	-344.99
1.0	0.50	0.00	-396.81
1.0	0.75	0.00	-575.25
1.0	1.00	0.00	-623.14

Table 4 shows the reward change from altering weights. This is total reward from training. The values from agents were averaged. It indicates the capacity of agents to optimize the problem. Lower value means more difficulty. Weight for infection rate is kept constant and others are modified in relation to alpha. Increasing the weight for spending change decreases the total reward. Intervention cost has lower impact. This does not imply agents are trained worse but proves that variables are conflicting and the weights will depend on prioritization. We set alpha as 1, beta as 0.5 and gamma as 0.25.

3.3 Evaluation

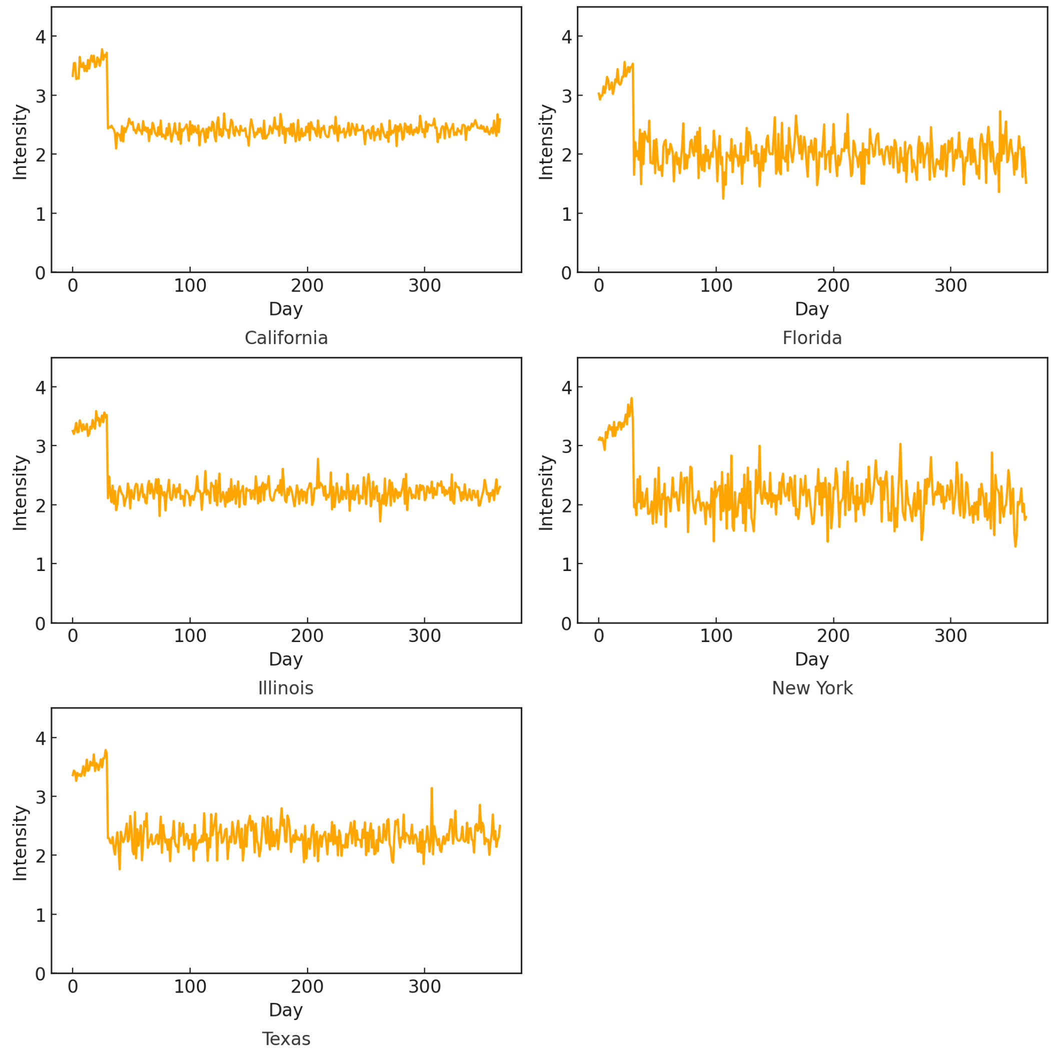


Figure 9: Average intensity of simulated intervention

Figure 9 illustrates the average intervention intensity over the 364-day simulation period for five states. The curves represent how the agent adjusts actions in response to the state. Higher values indicate stricter interventions and vice versa All the states begin with elevated intensities, reflecting strong early responses to high initial higher infection rate. Eventually, the agents maintain moderate intensity. California does not see significant changes. Florida experiences waves of elevated intensity. Illinois also does not experience major spikes. New York is the most volatile with the changes in the intensity of intervention. Texas observes bursts of increase throughout the year.

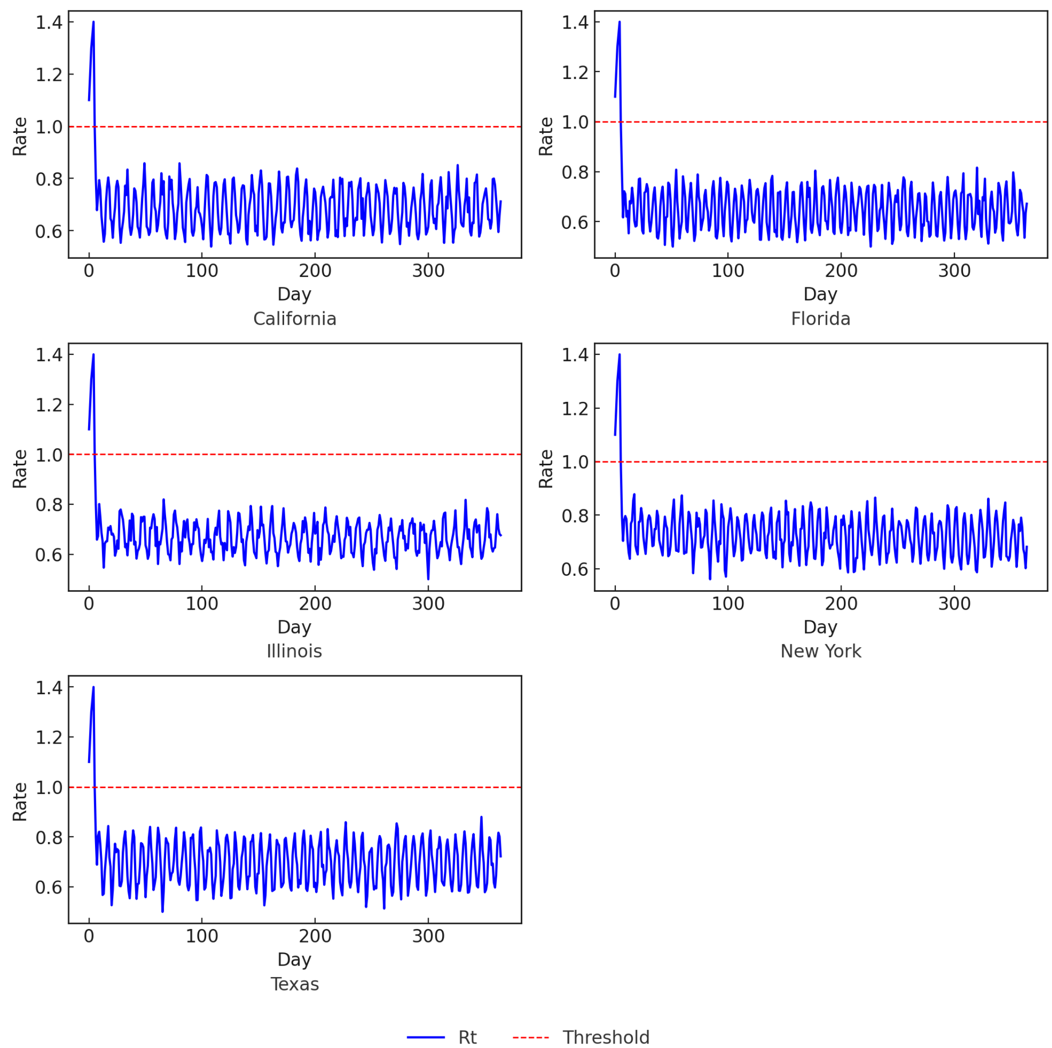


Figure 10: Infection rate from model simulation

Figure 10 shows the results from simulations for the five states. Each graph demonstrates how the agent adapted to control the transmission. The initial outbreak was introduced to provide realistic scenario. The agents handled this effectively by reducing and maintaining the infection rate below one. This would shrink the daily cases over time. California shows moderate fluctuations with Rt between 0.58 and 0.82. Florida exhibits higher variability bouncing between 0.55 and 0.80. Illinois maintains stable infection rate levels between 0.60 and 0.75. New York displays wider swings between 0.60 and 0.85. Texas achieves control between infection rate 0.55 and 0.85.

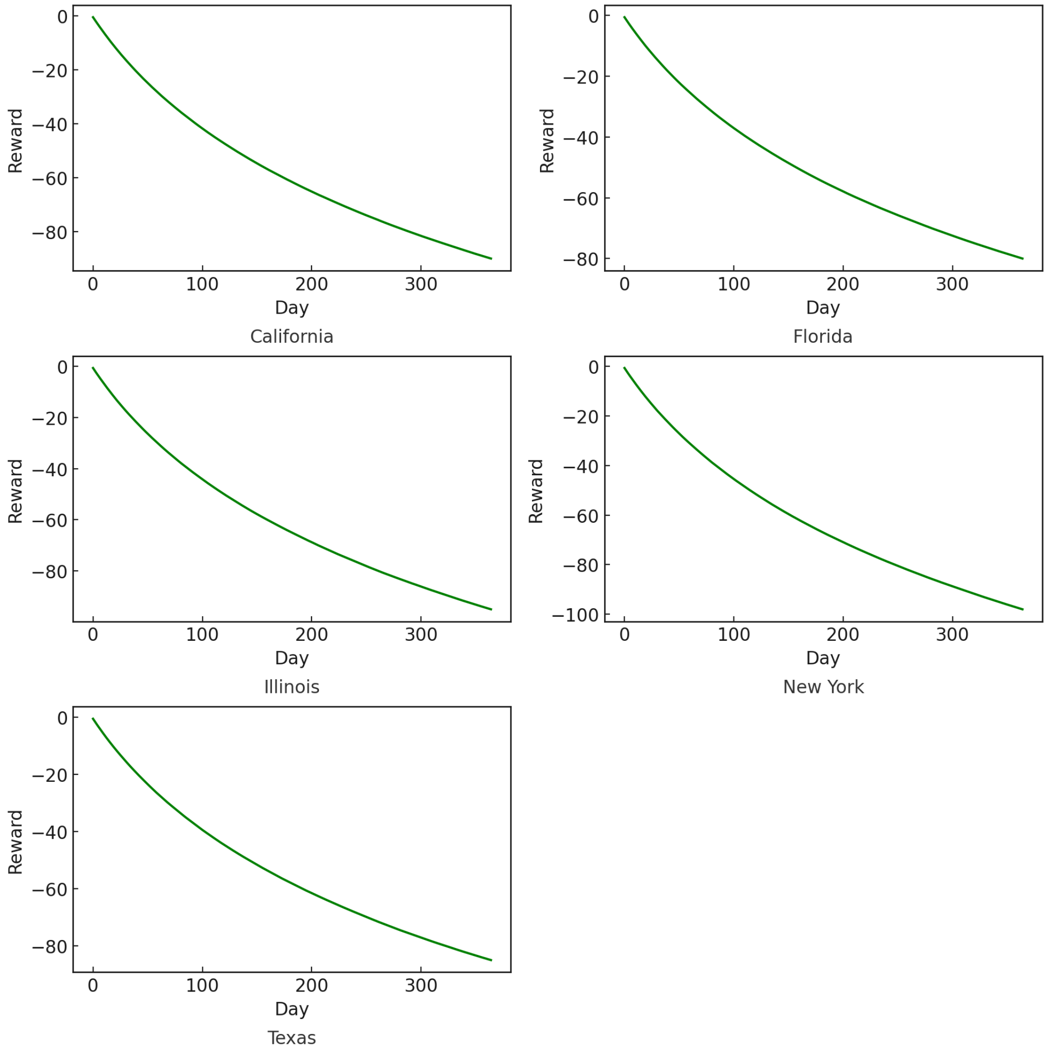


Figure 11: Cumulative reward from model simulation

Figure 11 shows the cumulative reward from agent simulations. Each graph demonstrates the change in total reward for the state. Initially, the curve is steeper indicating lower reward. This is because the infection rate is higher. Eventually, the agents maintain the infection below one. The curve becomes linear indicating this stability. The cumulative reward is negative because of the way the reward function is designed. Infection rate cannot be below zero which means there are daily cases. It is also unrealistic to have positive spending change during a pandemic. But the average negative reward is a low 0.22 each day.

4 Discussion

The environment models were successful in creating realistic situation based on the datasets of the states in the US. This allowed the reinforcement learning agents to be trained properly to control the transmission of the disease by suggested effective combination of interventions. Although the government was effective in handling the pandemic, the agents differed in the actions during much of the time. Overall, the suggested intervention started earlier at a higher intensity, slowly reduced in aggressiveness and maintained a moderate level for the rest of the year which aligns with other findings [26]. However, this depends on the weight assigned to the variables in the reward function. For school closure, the suggested was generally lower than the government which might be because shift in childcare increased cases [51]. Suggested control for workplace closure was lower to none compared to government and that can be because it promotes other behavior that increases transmission. The agent suggested higher intensity for public transport closure even if the government suggested none and this might be because it observed this intervention be effective in other states. On a granular level, the timing, intensity and duration of the interventions differed in states [19]. This might be influenced by regional factors.

5 Conclusion

The proposed model demonstrates strong potential in effectively controlling the transmission of disease similar to COVID-19. It can be adapted for use in different regions by adjusting the weighting parameters to reflect local priorities. While the current model captures key dynamics, we acknowledge that real-world implementation of government interventions involves numerous additional variables. These factors can be integrated into the reward function but incorporating too many variables may reduce model efficiency due to trade-offs between conflicting objectives. Since many interventions are complementary— where implementing one may enhance or enable the effects of another— the absence of certain options in a region would require the model to be retrained to account for those limitations. Moreover, frequent changes in policy are often impractical in real-world governance, but this can be addressed by introducing a transition penalty to discourage overly dynamic decision-making. This adjustment encourages more stable policy suggestions while maintaining responsiveness to public health needs. In further research, it can be valuable to analyze the impact of interventions on regional mobility, as mobility plays a crucial role in transmission dynamics. Understanding this relationship will offer insights into public compliance with government directives which will help guide effective distribution of resources like vaccines.

References

[1] P. B. Poudel, M. R. Poudel, A. Gautam, S. Phuyal, C. K. Tiwari, N. Bashyal, and S. Bashyal. Covid-19 and its global impact on food and agriculture. *J. Biol. Today's World*, 9(5):221–225, 2020.

[2] M. S. Tareq, T. Rahman, M. Hossain, and P. Dorrington. Additive manufacturing and the covid-19 challenges: An in-depth study. *J. Manuf. Syst.*, 2021.

[3] M. Sigala. Tourism and covid-19: Impacts and implications for advancing and resetting industry and research. *J. Bus. Res.*, 117:312–321, 2020.

[4] L. Jones, D. Palumbo, and D. Brown. Coronavirus: How the pandemic has changed the world economy. *BBC News*, 2021. [Online]. Available: https://www.bbc.com/news/business-51706225.

[5] V. McKeever. The coronavirus is expected to have cost 400 million jobs in the second quarter, un labor agency estimates. *CNBC*, 2020. [Online]. Available: https://www.cnbc.com/2020/06/30/coronavirus-expected-to-cost-400-million-jobs-in-the-second-quarter.html.

[6] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, and H. Coupland. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[7] Z. Qian, A. M. Alaa, and M. van der Schaar. When and how to lift the lockdown? global covid-19 scenario analysis and policy assessment using compartmental gaussian processes. *Adv. Neural Inf. Process. Syst.*, 33, 2020.

[8] N. Haug, L. Geyrhofer, A. Londei, E. Dervic, A. Desvars-Larrive, and V. Loreto. Ranking the effectiveness of worldwide covid-19 government interventions. *Nat. Hum. Behav.*, 4(12):1303–1312, 2020.

[9] N. Banholzer, E. Van Weenen, A. Lison, A. Cenedese, A. Seeliger, and B. Kratzwald. Estimating the effects of non-pharmaceutical interventions on the number of new infections with covid-19 during the first epidemic wave. *PLoS One*, 16(6):e0252827, 2021.

[10] E. Bendavid, C. Oh, J. Bhattacharya, and J. P. Ioannidis. Assessing mandatory stay-at-home and business closure effects on the spread of covid-19. *Eur. J. Clin. Invest.*, 51(4):e13484, 2021.

[11] J. M. Brauner, S. Mindermann, M. Sharma, D. Johnston, J. Salvatier, and T. Gavenčiak. Inferring the effectiveness of government interventions against covid-19. *Science*, 371(6531), 2021.

[12] Y. Liu, C. Morgenstern, J. Kelly, R. Lowe, M. Jit, and CCW Group. The impact of non-pharmaceutical interventions on sars-cov-2 transmission across 130 countries and territories. *medRxiv*, 2020.

[13] Y. Li, H. Campbell, D. Kulkarni, A. Harpur, M. Nundy, and X. Wang. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (r) of sars-cov-2: A modelling study across 131 countries. *Lancet Infect. Dis.*, 21(2): 193–202, 2021.

[14] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, and T. Chong. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584(7820):262–267, 2020.

[15] V. Barros, I. Manes, V. Akinwande, C. Cintas, O. Bar-Shira, M. Ozery-Flato, Y. Shimoni, and M. Rosen-Zvi. A causal inference approach for estimating effects of non-pharmaceutical interventions during covid-19 pandemic. *PLOS One*, 2022.

[16] V. Chernozhukov, H. Kasahara, and P. Schrimpf. Causal impact of masks, policies, behavior on early covid-19 pandemic in the us. *Journal of Econometrics*, 220(1):23–62, 2021.

[17] S. Singh, M. Shaikh, K. Hauck, and M. Miraldo. Impacts of introducing and lifting nonpharmaceutical interventions on covid-19 daily growth rate and compliance in the united states. *Proceedings of the National Academy of Sciences*, 118(12), 2021.

[18] S. Pei, S. Kandula, and J. Shaman. Differential effects of intervention timing on covid-19 spread in the united states. *Science Advances*, 6(49): eabd6370, 2020.

[19] W. Pan, D. Fernández, S. Tyrovolas, G. V. Iago, R. R. Dasgupta, and B. F. Zaitchik. Heterogeneity in the effectiveness of non-pharmaceutical interventions during the first sars-cov2 wave in the united states. *Frontiers in Public Health*, page 1857, 2021.

[20] S. Mohanty, A. Shimamura, C. D. Nicholson, A. D. Gonzalez, and T. Razzaghi. Hierarchical time series forecasting of covid-19 cases using county-level clustering data. *Operations Research Forum*, 6:28, 2025.

[21] S. Moein, M. S. Nickaeen, A. Roointan, M. Borhani, F. Heidary, S. Javanmard Shahraki, and M. Safavi-Naeini. Inefficiency of sir models in forecasting covid-19 epidemic: a case study of isfahan. *Scientific Reports*, 11(1):4725, 2021.

[22] R. Liu, J. Li, Y. Wen, H. Li, P. Zhang, B. Sheng, and D. D. Feng. Deep dynamic epidemiological modelling for covid-19 forecasting in multi-level districts. *arXiv preprint arXiv:2306.12457*, 2023.

[23] A. Spannaus, T. Papamarkou, S. H. Erwin, and J. B. Christian. Inferring the spread of covid-19: The role of time-varying reporting rate in epidemiological modelling. *Scientific Reports*, 12(1):10907, 2022.

[24] S. Hong and Y. Li. Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the covid-19 pandemic. *PLOS ONE*, 15(7):e0236464, 2020.

[25] D. Papageorgiou and G. Tsaklidis. A dynamic seihcrdv model for covid-19 with unscented kalman filter parameter estimation. *arXiv preprint arXiv:2207.05011*, 2022.

[26] G. H. Kwak, L. Ling, and P. Hui. Deep reinforcement learning approaches for global public health strategies for covid-19 pandemic. *PLOS ONE*, 16 (5):e0251550, 2021.

[27] S. Bushaj, X. Yin, A. Beqiri, D. Andrews, and İ. E. Büyüktahtakın. A simulation-deep reinforcement learning (sirl) approach for epidemic control optimization. *Annals of Operations Research*, 328(1):245–277, 2022.

[28] R. Padmanabhan, N. Meskin, T. Khattab, M. Shraim, and M. Al-Hitmi. Reinforcement learning-based decision support system for covid-19. *Applied Artificial Intelligence*, 36(1):1–25, 2022.

[29] V. Kompella, R. Capobianco, S. Jong, J. Browne, S. Fox, and L. Meyers. Reinforcement learning for optimization of covid-19 mitigation policies. *arXiv preprint arXiv:2010.10560*, 2020.

[30] M. I. Uddin, S. A. A. Shah, E. A. Alsolami, and A. K. Sangaiah. Optimal policy learning for covid-19 prevention using reinforcement learning. *Journal of Information Science*, 47(5):709–726, 2021.

[31] CSSEGISandData. Cssegisanddata github repository. `https://github.com/CSSEGISandData`. GitHub.

[32] S. J. Watson, O. J. Watson, P. Winskill, and A. C. Ghani. A causal inference approach for estimating effects of non-pharmaceutical interventions during covid-19 pandemic. *medRxiv*, 2020.

[33] IBM. Ibm github repository. `https://github.com/IBM`. GitHub.

[34] Federal Emergency Management Agency (FEMA). Key planning factors for response to and recovery from a biological incident. `https://www.fema.gov/cbrn-tools/key-planning-factors-bio/kpf-3/1`. FEMA.

[35] Opportunity Insights. Economic tracker github repository. `https://github.com/OpportunityInsights/EconomicTracker`. GitHub.

[36] E. L. Thorndike. *Animal Intelligence*. Macmillan Company, New York, 1911.

[37] Anonymous. 1952 – 'theseus' maze-solving mouse – claude shannon (american). `http://cyberneticzoo.com/mazesolvers/1952-{-theseus-maze-solving-mouse-{-claude-shannon-american/`. cyberneticzoo.com.

[38] I. P. Pavlov and G. V. Anrep. *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press, London, 1927.

[39] D. Michie and R. A. Chambers. Boxes: An experiment in adaptive control. *Machine Intelligence*, 2:137–152, 1968.

[40] R. S. Sutton and A. G. Barto. An adaptive network that constructs and uses an internal model of its world. *Cognition and Brain Theory*, 4(3): 217–246, 1981.

[41] R. S. Sutton and A. G. Barto. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2):135–170, 1981.

[42] A. K. Shakya, G. Pillai, and S. Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 213, 2023.

[43] R. E. Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, 16(3/4):221–229, 1956.

[44] S. Gronauer and K. Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55:895–943, 2021.

[45] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*. Springer, 2021.

[46] X. Li, W. Wang, Y. Fan, Y. Liu, and D. Li. Deep neural networks excel in covid-19 disease severity prediction: A systematic review and meta-analysis. *NPJ Digital Medicine*, 7(1):1–13, 2024.

[47] B. Frauenknecht, T. Lüdecke, R. S. Wahby, and R. Hutter. On rollouts in model-based reinforcement learning. *arXiv preprint arXiv:2501.16918*, 2025.

[48] L. Miralles-Pechuán, F. Jiménez, H. Ponce, and L. Martínez-Villaseñor. A methodology based on deep q-learning/genetic algorithms for optimizing covid-19 pandemic government actions. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, page 1135–1144, 2020.

[49] J. Fan, Z. Wang, Y. Xie, and Z. Yang. A theoretical analysis of deep q-learning. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control*, volume 120, page 486–489, 2020.

[50] C. C. Kerr, R. M. Stuart, D. Mistry, R. G. Abeysuriya, K. Rosenfeld, and G. Hart. Covasim: An agent-based model of covid-19 dynamics and interventions. *PLoS Comput Biol*, 17(7):e1009149, 2021.

[51] R. M. Viner, S. J. Russell, and H. Croker. School closure and management practices during coronavirus outbreaks including covid-19: a rapid systematic review. *The Lancet Child and Adolescent Health*, 4(5):397–404, 2020.