Simulation of the spread of diseases in the population

Wojciech Suski

April, 2025

Abstract

In this project, we develop an agent-based simulation framework to study the spread of infectious diseases in a population and evaluate the effectiveness of dynamic policy interventions. By combining agent-based modeling (ABM) with reinforcement learning (specifically Q-learning), we simulate individual behavior and policy-driven restrictions in a realistic and adaptive environment. Each agent follows state-based health transitions, including infection, recovery, and reinfection, while interacting with others in shared spaces such as homes and public venues.

Our system incorporates a multi-objective reward function that balances economic output, public health, and psychological well-being. A Q-learning agent acts as a decision-making body, adjusting lockdown severity over time to optimize long-term outcomes. Simulation results show that early, stricter measures such as increased mask-wearing and staying at home significantly reduce the overall infection rate and allow for safer relaxation of restrictions in later phases. In contrast, random or static policies lead to higher long-term costs and more volatile infection dynamics.

The proposed approach demonstrates that reinforcement learning can be a valuable tool for guiding public health decisions, allowing adaptive, data-driven responses to epidemic scenarios.

Keywords: disease, virus, spread, population, reinforcement learning, simulation, agent-based modeling, Q-learning

1 Introduction and review of related research

1.1 Introduction

Infectious diseases can spread rapidly across the globe, with COVID-19 being one of the most recent examples. Given its highly contagious nature, controlling the infection and fatality rates of COVID-19 has largely depended on the effective implementation of public health measures such as social distancing, isolation, quarantine, and mask usage.

We selected SARS-CoV-2 as our primary example of viral transmission due to our first-hand experience with its impact. Moreover, as one of the most extensively documented global pandemics, it provides valuable insights into the effectiveness of various containment strategies.

Different countries have taken various measures in response to the COVID-19 pandemic. While the government enforced restrictions on gatherings and closed certain venues, some individuals disregarded stay-at-home directives. Consequently, the government moderately altered the implemented measures: certain regions were allowed to reopen, leading to suboptimal pandemic control. We would like to avoid such situations by providing Agent-Based-Model simulation for virus.

Traditionally, statistical and mathematical models like the susceptible–infectious–recovered (SIR) model have been used to simulate and predict epidemic trends. However, these models often oversimplify the complexities of disease spread by overlooking factors such as human behavior, social interactions, population diversity, and individual variations. Moreover, policy interventions typically experience delays, and their immediate effectiveness is not always guaranteed.

With the rapid advancement of AI, Agent-Based Models (ABMs) have emerged as a promising alternative to address these limitations. ABMs simulate complex systems by modeling individual entities and their interactions within an environment. To improve epidemic modeling, we developed a pandemic simulator using ABM techniques, incorporating individual behaviors, interactions, and environmental influences to assess the spread of disease and the impact of control measures.

Although ABMs have been widely used in epidemic analysis, overly simplistic simulations often fail to provide meaningful insights for real-world decision-making. To enhance their effectiveness, we adopted a data-driven approach that integrates real-world data into the simulations. Specifically, we utilized reinforcement learning technique, to optimize decision-making in complex scenarios. By leveraging feedback mechanisms, Q-learning can identify optimal control strategies—such as vaccine distribution and social distancing—to minimize infection rates and mortality in epidemic management.

1.2 Related work

Over the past decades, researchers have developed a range of models to simulate the spread of infectious diseases. Traditional approaches are based on compartmental models—such as the SIR, SEIR, and SIRS models—which use systems of differential equations to divide the population into susceptible, infected, and recovered (or removed) compartments [2] . These models provide important insights into the overall dynamics of epidemics and have been extensively used to study disease transmission and intervention strategies.

In recent years, agent-based models (ABMs) have emerged as a powerful alternative. ABMs simulate individual interactions and can capture the heterogeneity in behavior and contact patterns that influence disease spread. For example, Perez and Dragicevic (2009) demonstrated how ABMs can realistically reproduce urban epidemic dynamics by incorporating detailed movement and interaction rules [1].

Furthermore, data-driven and spatially explicit simulation frameworks such as STEM and the SimInf R package have been developed to handle large-scale simulations that integrate real-world demographic and mobility data [3] , [4] . These tools allow researchers to evaluate the impact of public health interventions under various scenarios, making them essential for both theoretical studies and practical decision-making in epidemiology.

2 Problem formulation and proposed solution

As mentioned before the outbreak of COVID-19 led to a global public health crisis, prompting governments to implement various lockdown measures to curb the spread of the virus. However, determining the most effective strategies for balancing public health and economic stability remains a significant challenge. Traditional mathematical and statistical models provide macro-level epidemic forecasts but fail to account for individual variations, human behavior, and the impact of government policies. As a result, there is a need for a more sophisticated approach that can simulate real-world interactions and optimize decision-making in epidemic control.

To address these limitations, we propose a simulation framework that integrates Agent-Based Modeling (ABM) with Reinforcement Learning (RL) techniques, specifically the Deep Q-Network (DQN) method. Our approach leverages NetLogo as the ABM platform to create a detailed simulation of individual behaviors, interactions, and virus transmission dynamics. By incorporating real-world data collected during the COVID-19 pandemic, we enhance the RL model's learning process, allowing it to make data-driven decisions - it incorporates some more unpredictability in contrary to mathematical and statistical models. With all this in mind we could be able to conduct experiments with other types of viruses which could have much different parameters e.g. higher mortality or contagiousness rates. Such experiments could potentially show whether the actions taken during COVID-19 pandemic would be successful in quite different situations or lead to a disaster.

2.1 Preliminaries

2.1.1 Agent-Based Model

Agent-based modeling (ABM) is a computational approach that represents complex systems as assemblies of independent agents that interact with each other and their surroundings. Each agent follows its own set of rules, exhibits distinct behaviors, and makes decisions based on individualized processes. The interactions among these agents and their environment give rise to unexpected, system-wide patterns and phenomena.

More recently, ABMs have been extensively utilized to simulate the spread of COVID-19 and evaluate the impact of government policies. These models are capable of mimicking complex interactions among individuals—such as social contacts, compliance with preventive measures, and movement patterns—by incorporating real-world data and assumptions about individual behaviors. This allows them to shed light on how the virus spreads, the effects of various interventions, and the overall effectiveness of different policy measures. Additionally, some researchers have employed ABMs to assess resource requirements during the pandemic's peak and to forecast the demand for medical resources and vaccine supplies over time.

In summary, ABMs have become an essential tool for understanding the dynamics of COVID-19 and for analyzing policy effectiveness, thanks to their ability to capture the diversity and complexity of individual behaviors and interactions.

2.1.2 Q-Learning

Reinforcement Learning (RL) is a machine learning approach in which an agent interacts with an environment to learn decision-making strategies that maximize the total reward over time. In this paradigm, the agent learns through trial and error by taking various actions and receiving rewards or penalties as feedback. The ultimate goal is to derive an optimal policy— a strategy that maps states to actions to maximize expected cumulative rewards.

One popular RL algorithm is Q-learning. This method learns a value function that estimates the expected cumulative reward for taking a specific action in a given state and then following the optimal policy thereafter. The value function is defined as:

$$Q(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} Q(s',a') \mid s,a\right],$$

where s is the current state, a is the action taken, R_{t+1} is the reward received at the next step, s' is the next state, and a' is the subsequent action. The Q-learning algorithm updates this function using the rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(R_{t+1} + \gamma \max_{a'} Q(s',a') - Q(s,a) \right),$$

with α being the learning rate that determines the influence of new experiences relative to past experiences.

2.2 Algorithms

In our work, we designed a pandemic simulator (PS) that models virus transmission through an agent-based simulation framework. The simulator mimics the spread of a contagious disease under different behavioral policies and government interventions. It provides a platform to analyze how restrictions such as lockdowns or hospital capacity limits influence infection dynamics.

The simulator is structured around three core components:

- E (Environment) – representing the spatial and logical structure of the world,
- A (Agents) – representing individuals that move, interact, and transition through various health states,
- I (Infection mechanism) – governing the rules of disease transmission, progression, recovery, and immunity.

The environment is a grid of size 500 × 500, chosen to be easily visualizable while large enough to support population dynamics. Each grid cell represents a spatial patch in which agents can reside or travel through. Houses, points of interest (POIs such as workplaces, schools, and shops), and hospitals are placed randomly across the grid before the simulation begins. These elements are spread evenly across four predefined regions, simulating urban partitioning.

Agents simulate individual humans and are described by a tuple <Action, State>. Actions include traveling (e.g., staying home, visiting a POI, or going to the hospital) and optional behaviors such as wearing masks. States capture their health condition: healthy, latent, asymptomatic, symptomatic, immune, or dead. The transition between these states follows predefined probabilities derived from empirical data.

Each agent is randomly assigned a residence (house) and a primary point of interest within the same region. Every simulated day is divided into discrete time units (ticks). Agents follow a schedule: they stay home from midnight to 8:00 a.m., travel to their destination (if allowed or needed) from 8:00 a.m. to 4:00 p.m., and return home or stay at the hospital afterward. This division results in varied infection risk across time periods—commuting periods are associated with higher exposure.

The simulation time is discretized into 82 ticks per day: 30 ticks for each commute window and 1 tick per hour otherwise. Infection probability is calculated at each tick, based on proximity to infected agents. The infection model includes:

- A latent period during which newly infected agents are not yet contagious,
- Transition to either asymptomatic or symptomatic phases, with differing infectiousness,
- A recovery or death outcome after the infectious period, based on real-world infection-fatality rates,

- A possibility of reinfection for recovered agents.

To ensure realistic modeling, infection-related parameters are drawn from statistical research on COVID-19 (e.g., incubation periods, infection radius, and reinfection probability), as detailed in Table 2 of this report.

The goal of this simulator is to serve as an environment for reinforcement learning. It is integrated with the Q-learning algorithm, enabling the exploration of policies that optimize a balance between minimizing infections and preserving socioeconomic activity. Through this integration, the simulator supports the development of data-driven, intelligent pandemic response strategies.

2.3 Environment

To effectively train a Q-learning-based agent in our pandemic simulation, we designed the environment to reflect real-world conditions that policymakers face during public health crises. The environment is composed of three key components: state space, action space, and reward design. Each component is carefully constructed to capture the trade-offs between public health, economic stability, and societal well-being.

The state space is formulated to reflect the distribution of population health statuses, which are directly tied to both disease progression and economic productivity. The action space models a range of government interventions, from minimal to strict lockdowns, based on real-world policy implementations. Finally, the reward function adopts a multi-objective approach, balancing economic, health, and psychological impacts. The design allows policymakers to prioritize different aspects of the pandemic response by adjusting weight parameters, making it possible to simulate diverse policy strategies and their consequences.

This structured environment enables reinforcement learning agents to explore and evaluate a wide range of epidemic control strategies under realistic constraints. As a result, it provides a valuable framework for studying optimal decision-making during dynamic and uncertain pandemic scenarios.

Implementation

This chapter presents the details of the implementation of a disease spread simulator based on agent-based modeling and the Q-learning algorithm. It describes the system architecture, the structure of the environment, and the integration of the learning component with the simulation process.

3 Experimental research results

3.1 Infection Mechanism and Reward Function

The simulator uses an agent-based infection process in which each agent can reside in one of several mutually exclusive health states: healthy, latent (infected but not yet contagious), asymptomatic or symptomatic (contagious), recovered (with partial immunity), or dead. The sequence of state transitions for any newly infected agent is as follows:

1. **Latent Period:** Immediately upon infection, an agent enters a latent state of duration drawn from a distribution (e.g. log-normal with mean 5.8days for incubation). During this phase, the agent cannot infect others.

2. **Infectious Phase:** After latency, the agent becomes contagious. With a fixed probability (e.g. 50% asymptomatic rate), it transitions into either:
    - *Asymptomatic*: remains contagious without showing symptoms, with reduced viral shedding.
    - *Symptomatic*: develops clear symptoms after the incubation period and becomes fully infectious.
3. **Outcome:** Once the infectious period ends (drawn from empirical data, e.g. normal with mean 8.5days), the agent either:
    - *Recovers:* moves to the immune state, with a given probability (complement of the fatality rate, e.g. 0.9% IFR).
    - *Dies:* transitions to the dead state according to the infection-fatality rate.
4. **Reinfection:** Agents in the immune state can, with a low reinfection probability (e.g. 11%), return to the susceptible class after some time.

At each time tick, any susceptible agent checks for nearby contagious agents within a specified infection radius (e.g. 2m). If at least one contagious agent is within that radius, the susceptible agent becomes infected with a probability that depends on the current government policy (mask usage, staying at home, etc.; see Sec. 3.2).

Reward Function. The Q-learning agent receives a scalar reward at every decision step that balances three competing factors:

- **Economic contribution (R_c)**: Each healthy (or working) individual contributes 1 unit per day; if an agent is at home or ill, their economic contribution is reduced according to a parameter $p_i \in \{0.8, 1, 0\}$ (home, point of interest, hospital) and a health weight $s_i \in \{0.8, 0.2, 0\}$ (asymptomatic, symptomatic, dead). Formally:

$$R_c^{(t)} = \frac{1}{N}\sum_{i=1}^{N} p_i \cdot s_i \,,$$

  where N is the total number of agents.

- **Health contribution (R_h)**: Each agent's health state $h_i \in [-2, 1]$ is mapped to a daily health score (+1 = healthy, 0.9 = latent, 0.5 = asymptomatic, 0 = symptomatic, -2 = dead). Then:

$$R_h^{(t)} = \frac{1}{N}\sum_{i=1}^{N} h_i \,.$$

- **Psychological factor (R_p)**: Based on the lockdown level $l \in \{0,\ldots,5\}$, where $l = 0$ means minimal intervention and $l = 5$ maximum intervention, the public's psychological well-being is approximated by:

$$R_p^{(t)} = 1 - \frac{l}{l_{\max} - 1}, \quad l_{\max} = 5.$$

  Strict lockdowns (higher l) reduce R_p.

These partial contributions are aggregated into a single multi-objective reward:

$$R^{(t)} = \alpha R_c^{(t)} + \beta R_h^{(t)} + \gamma R_p^{(t)}, \qquad \alpha + \beta + \gamma = 1.$$

By adjusting α, β, γ, one can trade off between economic, health, and psychological priorities.

3.2 Government Policy

At each decision point (e.g. once per day or every fixed number of timesteps), the "government agent" selects a lockdown level $l \in \{0,\dots,5\}$. Each level l is a bundle of five policy levers:

1. **Mask Wearing (W)**: The fraction $W \in [0,1]$ of the population required to wear masks when outside. Higher W lowers the per-contact infection probability.

2. **Stay at Home (S)**: The probability $S \in [0,1]$ that healthy individuals remain at home rather than visit points of interest (POIs). A higher S reduces overall contact rates.

3. **Gathering Limit (G)**: The maximum number of people allowed to gather in public. $G = \infty$ means no limit; $G = 0$ prohibits all gatherings.

4. **Area Lockdown (A)**: A binary flag $A \in \{0,1\}$. If $A = 1$, certain POIs (e.g. shops, schools) are closed and no one may leave home for non-essential reasons.

5. **Isolation if Infected (I)**: A binary flag $I \in \{0,1\}$. If $I = 1$, any detected infected individual is forced to isolate in hospital and does not travel.

Table 1 summarizes the five policy levers for each discrete lockdown level:

Level l	Mask W	StayHome S	GatherLimit G	AreaLock A	Isolate I
0	0.1	0.0	∞	0	0
1	0.3	0.2	∞	0	0
2	0.3	0.5	∞	0	0
3	0.3	0.5	8	0	0
4	0.3	0.5	4	0	1
5	0.5	0.8	0	1	1

Table 1: Lockdown levels $l = 0,\dots,5$ and their corresponding policy parameters (W,S,G,A,I).

Higher levels impose stricter measures (more mask usage, more home staying, smaller gatherings, closing POIs, and forced isolation). The Q-learning agent's action space is thus simply choosing l at each decision epoch.

In this section, we present the outcomes of our simulations, analyze how the learned policy affects infection dynamics, and illustrate the learning process of the Q-learning agent. We focus on two main aspects: (1) how infection counts evolve under the learned control strategy, and (2) how the agent's cumulative reward grows over time during training.

3.3 Figures and tables

Below we include all figures and tables referenced in the text. Figure 1 shows the agent's reward as training progresses. Figure **??** illustrates a representative infection curve under the learned policy versus a baseline scenario with no interventions.

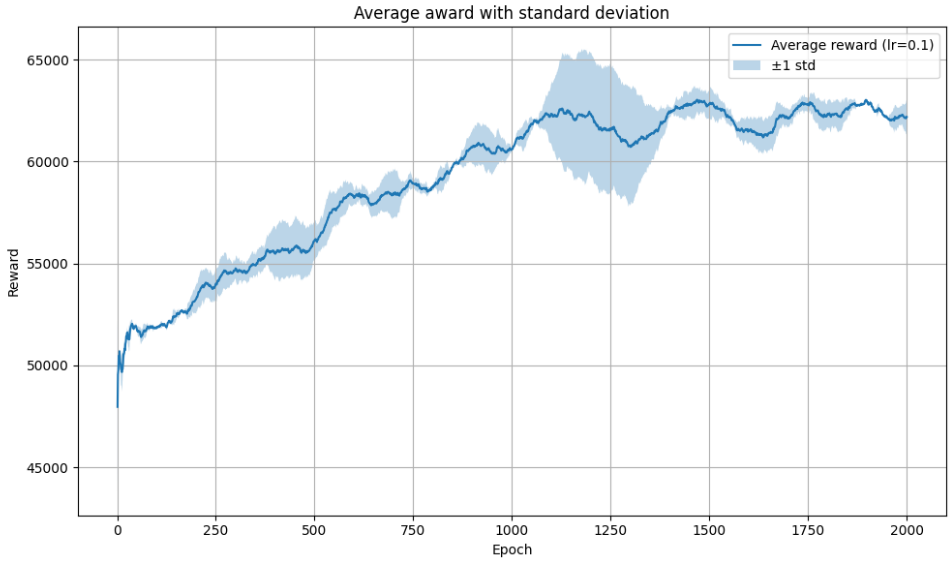

Figure 1: Learning curve of the Q-learning agent: average reward per epoch as training progresses. Early epochs show large fluctuations, but as training continues the mean reward converges to a higher, stable value.

3.3.1 Figures

This plot depicts how the agent's cumulative reward evolves over the course of training. The horizontal axis represents epoch number (each epoch corresponds to a simulated day/week of interaction), while the vertical axis shows the average reward obtained in that epoch. Initially, rewards are low and highly variable, reflecting random exploration. Around epoch 50, the curve begins to rise steadily, indicating that the agent has discovered that enforcing stricter measures (mask-wearing and staying at home) yields higher long-term returns. By epoch 1000, the reward plateaus, suggesting convergence to an optimal policy.

This plot shows the average reward per epoch under a strictly random policy, where at each decision step the "government agent" selects a lockdown level $\ell \in \{0,1,2,3,4,5\}$ with equal probability. The horizontal axis is the epoch number, and the vertical axis is the cumulative reward obtained in that epoch. Unlike the learned policy (Fig. 1), the random policy's reward does not improve over time and exhibits large fluctuations throughout all 2000 epochs. This behavior confirms that arbitrary action selection yields poor outcomes, as the system fails to learn to protect public health or preserve economic activity.

This figure shows the progression of infections and deaths under a random policy where actions are selected without learning. Unlike the learned strategy, there is no sustained decline in the number of infected individuals. The curve fluctuates or remains high, indicating persistent virus spread. The number of deceased agents continues to rise throughout the simulation. The absence of early containment and policy consistency leads to prolonged health system burden and uncontrolled epidemic dynamics.

This figure illustrates the time evolution of the infected and deceased population when applying the policy learned by the Q-learning agent. During the early epochs, the number of infected agents is high, but it steadily decreases as the agent enforces restrictive measures such as stay-at-home behavior and mask-wearing. As the infection subsides, the agent relaxes restrictions, maintaining low infection levels without triggering a second wave. The number of deceased individuals increases initially but quickly stabilizes, showing that the system success-

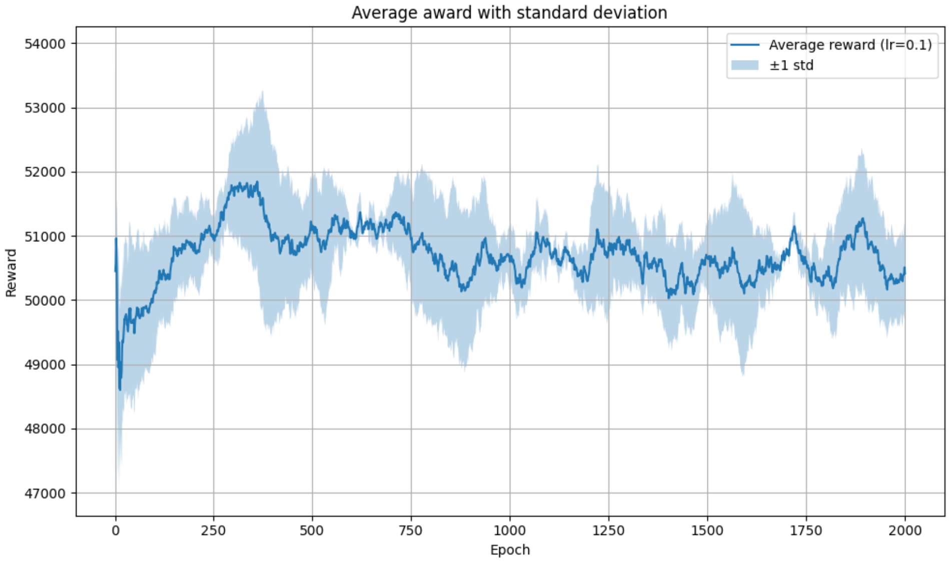

Figure 2: Reward trajectory when actions are chosen uniformly at random. The curve remains low and highly variable, indicating that without learning, the policy fails to balance infection reduction and economic/social costs.

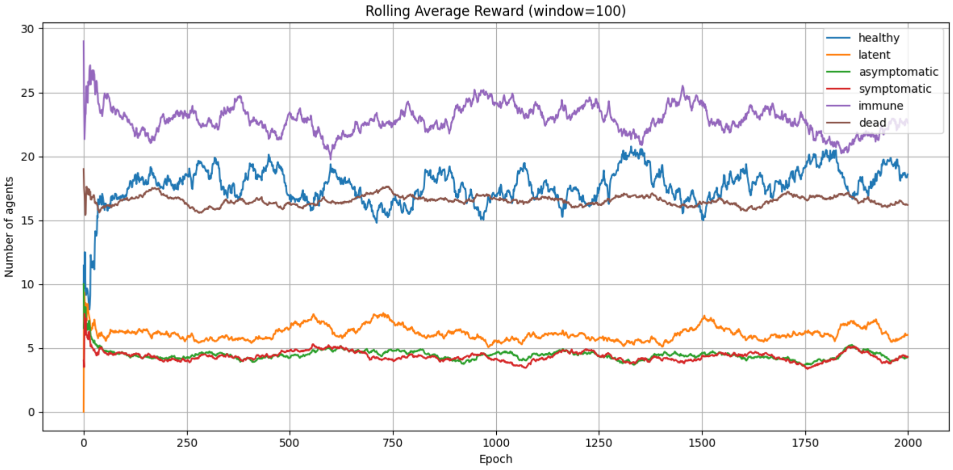

Figure 3: Number of infected and deceased individuals over time under a random policy. The lack of consistent intervention prevents infection rates from declining.

fully flattens the curve and prevents further escalation.

3.3.2 Tables

Parameter	Value	Description
Learning rate (α)	0.10	Step size for Q-learning updates
Discount factor (γ)	0.95	Future reward discounting
Exploration (ε)	1.0 → 0.01	Annealed over epochs
Number of epochs	2000	Total training epochs
epochs per simulation	300	Time steps per epoch

Table 2: Key Q-learning parameters used in the experiments.

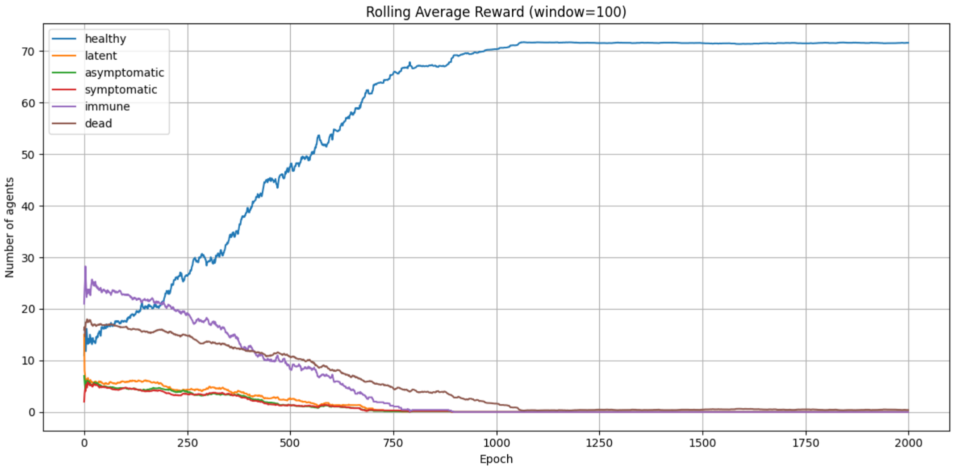

Figure 4: Number of infected and deceased individuals over time under the learned Q-learning policy. The policy gradually reduces infections and deaths through early strict intervention followed by relaxation.

This table lists the main hyperparameters for the Q-learning agent. The learning rate and discount factor control how quickly the agent updates its Q-values and how much it values future rewards. The exploration parameter ε starts at 1 (full exploration) and is linearly annealed to 0.01 by the final epoch. Each epoch consists of 300 time steps (representing daily cycles), and a total of 2000 epochs were run.

3.4 Results and Analysis

The Q-learning agent was trained over 2000 epochs. In each epoch, the agent interacted with the environment for 300 time steps. At each decision point (once per time step), the agent selected one of three actions (no restrictions, moderate restrictions, or full lockdown) for each sector based on its current Q-table.

Reward Growth. As shown in Figure 1, early epochs exhibit large variance in reward, as the agent largely explores random actions. Around epoch 50–60, a clear upward trend begins, indicating that the agent has discovered that enforcing stricter protective measures initially (encouraging agents to stay home and wear masks) leads to higher long-term rewards by reducing infection counts. By epoch 1000, the curve stabilizes, meaning the agent's policy has converged to a near-optimal strategy.

Key Findings.

- **Early strict measures are optimal.** The Q-learning agent learns that applying stricter controls (more home-stays and mask usage) at the beginning minimizes total infections and thus maximizes cumulative reward.
- **Gradual relaxation.** Once the infection prevalence drops below a threshold, the agent shifts to fewer restrictions, balancing economic/social costs.
- **Avoiding oscillations.** The learned policy does not oscillate rapidly between lockdown and no-lockdown. Instead, it stays in the strict regime until the infection is well under control, then transitions to minimal restrictions.

These results confirm that an adaptive, learning-based approach yields better outcomes than a static policy (e.g., constant moderate restrictions). Proactive containment—early home-stays and mask usage—proves critical to suppressing the outbreak and enabling safe reopening later.

4 Summary and conclusions

The main objective of the project was to simulate the spread of infectious diseases within a population using an agent-based model combined with Q-learning, in order to study the effectiveness of different policy strategies in limiting infections while maintaining individual freedom and social activity.

The simulation model included agents representing individuals moving through a shared environment, each capable of transitioning between health states (susceptible, infected, recovered). The agents' behaviors and the environment were influenced by simple pandemic policies such as mask-wearing and staying at home. A Q-learning agent was trained to select optimal policies at each simulation step based on the current state of the system.

The results of the experiments indicate that the optimal long-term strategy involves enforcing stricter protective measures at the beginning of the epidemic. In particular, encouraging agents to stay at home more frequently and to wear masks in the early stages significantly reduces the infection rate. This, in turn, makes it possible to relax these measures in the later phases without leading to a resurgence of the disease. On the other hand, prematurely loosening restrictions typically results in prolonged periods of high infection rates, requiring more drastic and sustained interventions later.

This suggests that proactive and early containment measures are not only more effective from a public health perspective but also minimize the need for long-term restrictions. The learning-based approach confirms that intelligent, adaptive policy-making can lead to better outcomes than static, rule-based strategies.

Plans for further work

Several directions for future research have been identified:

- Incorporating additional agent heterogeneity (e.g., age, profession, health conditions) to better reflect real populations.
- Extending the state and action spaces to include vaccination policies, testing, and quarantine.
- Integrating mobility networks and geographic zones for more realistic spatial dynamics.
- Exploring multi-objective Q-learning to explicitly balance competing goals such as minimizing infections, economic loss, and psychological distress.
- Implement DQN algorithm

Overall, the project demonstrates the usefulness of reinforcement learning techniques in guiding public health policy decisions during pandemics.

References

[1] R. Blanco, G. Patow, and N. Pelechano. “Simulating Real-life Scenarios to Better Understand the Spread of Diseases Under Different Contexts”. In: *Scientific Reports* 14.2694 (2024). Accessed: 2025-04-01. DOI: 10.1038/s41598-024-52903-w. URL: https://www.nature.com/articles/s41598-024-52903-w.

[2] C. for Disease Control and Prevention. *Technical Explainer: Infectious Disease Transmission Models*. Accessed: 2025-04-01. 2023. URL: https://www.cdc.gov/cfa-modeling-and-forecasting/about/explainer-transmission-models.html.

[3] E. Foundation. *Spatiotemporal Epidemiological Modeler (STEM)*. Accessed: 2025-04-01. 2006. URL: http://www.eclipse.org/stem/.

[4] S. Widgren, P. Bauer, R. Eriksson, and S. Engblom. “SimInf: An R package for Data-driven Stochastic Disease Spread Simulations”. In: *arXiv preprint arXiv:1605.01421* (2016). Accessed: 2025-04-01. URL: https://arxiv.org/abs/1605.01421.