SIRV-RL: Reinforcement Learning for Optimized Policy Control during Epidemiological Outbreaks in Emerging Market and Developing Economies

Maeghal Jain[1,*] **and Ziya Uddin**[1]

[1]BML Munjal University
[*]e-mail: maeghaljain@gmail.com

ABSTRACT

The outbreak of COVID-19 has highlighted the intricate interplay between public health and economic stability on a global scale. This study proposes a novel reinforcement learning framework designed to optimize health and economic outcomes during pandemics. The framework leverages the SIR model, integrating both lockdown measures (via a stringency index) and vaccination strategies to simulate disease dynamics. The stringency index, indicative of the severity of lockdown measures, influences both the spread of the disease and the economic health of a country. Developing nations, which bear a disproportionate economic burden under stringent lockdowns, are the primary focus of our study. By implementing reinforcement learning, we aim to optimize governmental responses and strike a balance between the competing costs associated with public health and economic stability. This approach also enhances transparency in governmental decision-making by establishing a well-defined reward function for the reinforcement learning agent. In essence, this study introduces an innovative and ethical strategy to navigate the intricate challenge of balancing public health and economic stability amidst infectious disease outbreaks.

1 Introduction

In the past, global spread of infectious diseases was largely due to colonization, slavery, and war, leading to widespread illness and death from diseases like tuberculosis, polio, smallpox, and diphtheria. Medical advancements, better access to health care, and improved sanitation have worked towards improving the situation of mortality and morbidity linked to infectious diseases in the past twenty years. However, in low and lower-middle income countries the burden of infectious diseases still persists. The rapid pace of urbanization in low and middle-income countries, along with the rise in populations living in crowded, poor-quality homes, has led to new conditions that favor the emergence of infectious diseases[1,2].

Recently, the COVID-19 pandemic caused a havoc worldwide. Till date there have been 772 million cases and more than 6 million deaths[3]. The pandemic triggered the sharpest economic recession in modern history with a 3% decline, much worse than during the 2008-09 financial crisis[4]. As nations grappled with the immediate health crisis, the economic fallout disproportionately affected vulnerable populations and exacerbated existing inequalities. Lockdowns and restrictions imposed to curb the spread of the virus led to widespread unemployment, business closures, and disruptions in global supply chains[5]. The challenges faced by low and lower-middle income countries were particularly acute, highlighting the intricate interplay between public health and economic stability on a global scale[6].

The need for a nuanced understanding of how interventions impact both health outcomes and economic indicators became increasingly evident, prompting a comprehensive examination by epidemiologists to assist policy makers[7]. The outbreak of COVID-19 has prompted epidemiologists to research on various aspects, including mobility control[8,9], vaccination strategies[10,11], stringency measures/non-pharmaceutical interventions (NPIs)[12–14], and financial considerations[15]. Despite the numerous studies conducted, very few explore how common interventions meet multiple policy objectives or how a precise articulation of the main policy goals directs the selection of the most effective interventions in terms of health and economic results[8,16–22]. The economic impact of the COVID-19 pandemic varied between rich and poor countries. Although COVID-19 deaths had a slightly larger negative effect on the Gross Domestic Product (GDP) in advanced economies, this difference was not statistically significant. However, lockdown restrictions were found to have a more damaging impact on economic activity in emerging and developing economies[6,23,24]. It's also suggested that an increase in COVID-19 cases was associated with the introduction of harsher NPIs and lockdown measures could be relaxed once vaccination rates increase[23,25].

Many economists have studied the effect of COVID-19 on the economy of nations[6,26–28]. In advanced economies like Korea, where the stringency index was below the median the recession was milder than other advanced economies like the United Kingdom where the stringency was much higher[26], they achieved it mostly with very aggressive testing, contact tracing, and enforced quarantines and isolations[29,30]. In India, social distancing and containment measures have been effective in reducing the number of COVID-19 cases but have come with economic costs. Social distancing had the most adverse effect on the economy in areas with high urbanization[27].

In this paper we optimize the government policies regarding stringency. Therefore, alongside epidemiological data, we use the measures of globally comparable government responses[31]. We use the simple SIR model without vital dynamics[32–34] as it is assumed that the timescale is small enough that it can be neglected[35]. By lesioning the model, as opposed to proposing a new mathematical model with more specialized compartments to more accurately represent the actual environment[36,37], we can effectively address the real-world conditions and propose a solution that is both effective and extendable. The current model (SIRV with lockdown) accounts for the recovery reached through vaccination[38–42] and the effects of lockdown[21,43,44]. Traditional SIR model neglects the time-varying property of parameters like β and γ, but in this paper since, lockdown is a parameter of the proposed model, and it changes with time therefore, we propose a time-dependent SIR model[45]. However, the study has limitations; First, the deterministic SIR model (predecessor to our proposed model) fails to account for chance in disease spread and lacks confidence intervals on results and while stochastic models incorporate chance, they are typically more challenging to analyze than their deterministic counterparts[33]. Second, the underreporting of cases during the period selected by our study. Lastly, the reinforcement learning agent should be resistant to how the vaccination rate changes – or to when it is first introduced because then it scopes the environment for succumbing to wishful thinking which can be potentially dangerous.

After modelling the disease with lockdown (via stringency index) and vaccination, we try to understand the effects of lockdown on the GDP. Therefore, decisions made by the government regarding the level of lockdown to be enforced plays a role on both the public health outcomes and economic stability during a pandemic. On one hand, stringent lockdown measures can effectively slow the spread of the disease, thereby improving public health outcomes. However, these measures often come at the cost of significant economic disruption, leading to job losses, business closures, and reduced economic growth. On the other hand, relaxing lockdown measures may help to mitigate the economic impact of the pandemic, but could result in increased disease transmission and worsened public health outcomes. In order to capture competing costs within the environment and achieve a balance between health and economic outcomes, we intend to employ reinforcement learning[8,19,20,46,47]. Not only does the formulation of the model deal better with competing costs, but it also offers more transparency behind the reasoning of the decisions being made in such circumstances. When we conceptualize our problem as a reinforcement learning task, an agent is tasked with making decisions in an environment with the aim of optimizing cumulative rewards (i.e., the total amount of reward it receives over the long run). Simply put, given discrete time steps $t = 0, 1, 2, 3, \ldots$, at each time step the agent receives a representation of the environment's state, $s_t \in \mathcal{S}$, and selects an action $a_t \in \mathcal{A}(s_t)$, where $\mathcal{A}(s_t)$ is the set of actions available in state s_t, and one step later receives a reward $r_{t+1} \in \mathcal{R}$ and the state is updated[48]. The way we define the way these rewards that are given to the agent, makes this decision process more transparent, however, it has its limitations. A universal optimal policy may not suit diverse socioeconomic contexts due to variations in healthcare resources and economic vulnerabilities across countries, regions, or cities and a comprehensive consideration of decision factors, extending beyond pure reinforcement learning results is needed[8,49,50].

Additionally, since most modern reinforcement learning achievements are due to a combination of deep learning[51], in the following framework we make the use of this. Deep reinforcement learning is an advancement to reinforcement learning which helps normalize the input and reduce its dimensionality[51–54]. We use a Long Short Term Memory recurrent neural network for time-series data[55,56] and a simple fully connected network for the non-time-series data.

In summary, by using reinforcement learning augmented with deep learning techniques for the SIRV with lockdown environment, we can better understand the effects of lockdown measures on both public health outcomes and economic stability during a pandemic. However, it is crucial to consider the limitations of this approach and take into account a comprehensive set of decision factors in order to make informed policy decisions that are tailored to specific socioeconomic contexts.

2 Mathematical Formulation and Numerical Computation

In SIRV-RL, we use a compartmental model to model infectious disease environment. We iteratively develop this model to fit the actual data better. In an SIR model people of the population are divided based on whether they are yet to come into contact with an infected person (Susceptible), are infectious themselves (Infectious), or have recovered from the infection (Recovered). These compartments create the SIR model which can be represented as follows:

2.1 Simple SIR Model

$$\boxed{S} \xrightarrow{\lambda} \boxed{I} \xrightarrow{\gamma} \boxed{R}$$

$$\frac{dS}{dt} = -\lambda S \tag{1}$$

$$\frac{dI}{dt} = \lambda S - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

Here, λ is the force of infection, it is the rate at which susceptible individuals acquire an infectious disease[57]. It depends on other factors:

$$\lambda = pc\frac{I}{N} \tag{4}$$

Here, c is the average number of contacts a susceptible person makes per day. p is the probability of the susceptible person becomes infectious after coming into contact with an infectious person. $\frac{I}{N}$ is the proportion of the contacts that are infectious.

And, β the effective transmission rate is defined as:

$$\beta = pc \tag{5}$$

During an epidemic, the fundamental drivers of an epidemic growth is the rate of infection β i.e. the average number of infections per infected case and the infectious period $1/\gamma$ i.e. the average period for which the infected case is infected for. Epidemics can only happen if the case is infectious enough for long enough and this defined by $R_0 = \beta/\gamma$. Here, R_0 is The average number of secondary infections caused by each infected case, in an otherwise fully susceptible population.

At the peak of an epidemic R_0 declines as there are no more susceptible people left in the pool, therefore, R_e (effective reproductive number) comes into play. R_e is defined as the average number of secondary cases arising from an infected case, at a given point in an epidemic, therefore, it takes into account the existing immunity of the system.

$$R_e = R_0 \frac{S}{N} \tag{6}$$

S is the number of susceptible people, N is the total population. At the start of an epidemic when everyone is susceptible, $R_e = R_0$ as, $S = N$ (i.e., the whole population is susceptible). β and γ are also used to define probability of and infectious individual infecting another individual $\beta/(\beta+\gamma)$ and the probability of recovery, $\gamma/(\beta+\gamma)$.

Most government policies look at the value of R_e to come us with an effective strategy to combat the disease as the fate of the evolution of the disease depends upon it. When R_e is less than one, the infected population I will steadily decline to zero. Conversely, if R_e is greater than one, the infected population will increase. In other words, when $\frac{dI(t)}{dt} < 0 \Rightarrow R_e < 1$ and $\frac{dI(t)}{dt} > 0 \Rightarrow R_e > 1$, therefore, the effective reproductive rate R_e serves as a critical threshold that determines whether an infectious disease will rapidly extinguish or escalate into an epidemic[35].

To estimate the parameters β and γ for India, based on the data from May, 2020, to October, 2022, we simply define two cost functions eqs. (8) and (9) to calibrate the model with the use of huber loss[58]. We generally calibrate the model using eq. (8). However, sometimes we must strike a balance between modeling all three population groups (susceptible, infected, and recovered) and focusing just on the infected group, which is most important because it drives the spread of the disease, therefore, in that case we consider the loss obtained from eqs. (8) and (9) (see section 2.4).

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \cdot (|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{7}$$

In the above equation, y is the actual data and $f(x)$ is the prediction.

$$\begin{aligned} \text{loss\_SIR} &= \text{cost\_function\_SIR}(S, \hat{S}, I, \hat{I}, R, \hat{R}) \\ &= L_{\delta=1}(S, \hat{S}) + L_{\delta=1}(I, \hat{I}) + L_{\delta=1}(R, \hat{R}) \end{aligned} \tag{8}$$

$$\text{loss\_I} = \text{cost\_function\_I}(I, \hat{I}) = L_{\delta=1}(I, \hat{I}) \tag{9}$$

Where, S is the number of susceptible people and $\hat{S}$ is the predicted number of susceptible people, similarly, I for infected and $\hat{I}$ for the predicted number of infected people, and R for recovered. Using the equations Equations (1)–(3) , we minimize the cost function eq. (13) using the Nelder-Mead method[59] to estimate the parameters like β, γ, and fit the model to actual data. The following parameters and loss is obtained:

$$\beta_{optimal} = 0.042 \tag{10}$$

$$\gamma_{optimal} = 0.024 \tag{11}$$

$$R_{0_{optimal}} = \frac{\beta_{optimal}}{\gamma_{optimal}} = 1.762 \tag{12}$$

$$\text{loss\_SIR} = 85051490.533 \tag{13}$$

$$\text{loss\_I} = 45187665.281 \tag{14}$$

See fig. 3a to see how the model compares with the actual data.

2.2 SIR Model with Lockdown

Now that, a simple SIR model has been established – we need to model the effects of the stringency index (measure for the strictness of lockdown) on β (the effective transmission rate). To do this, we say the flow of susceptibles not only depend on β but also $s(t)$ the stringency index at time[22,43,44,60,61]. The stringency index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest)[62]. This index simply records the strictness of government policies and does not measure or imply the appropriateness or effectiveness of a country's response i.e. a higher score does not necessarily mean that a country's response is "better" than others lower on the index. To define the new time-varying beta that is dependent on the current stringency index, the following equations have been formulated:

$$\frac{dS}{dt} = -\beta(1 - s(t)/100)\frac{SI}{N} \tag{15}$$

$$\frac{dI}{dt} = \beta(1 - s(t)/100)\frac{SI}{N} - \gamma I \tag{16}$$

$$\frac{dR}{dt} = \gamma I \tag{17}$$

Where, $s(t)$ is the stringency index at time t and is scaled down by a factor of 100 to normalize it and bring it in the range $s(t)/100 \in [0,1]$. Multiplying the rate of flow from S to I compartment with $1 - s(t)/100$ allows us to account for the effect that stringency has on the disease progression. A higher stringency index can theoretically, stop the flow from the susceptible population to the infected population entirely. Optimizing these equations with eq. (8) using the Nelder-Mead method, we get the following parameters and loss:

$$\beta_{optimal} = 0.401 \tag{18}$$

$$\gamma_{optimal} = 0.090 \tag{19}$$

$$\begin{aligned}
\overline{R_0} &= 1.693 \quad \text{(Mean)} \\
\widetilde{R_0} &= 1.624 \quad \text{(Median)} \\
\text{Mode}(R_0) &= 0.804 \quad \text{(Mode)} \\
\sigma_{R_0} &= 0.786 \quad \text{(Standard Deviation)} \\
R_0 &\in [0.16467, 3.0497] \quad \text{(Range)}
\end{aligned} \tag{20}$$

$$\text{loss\_SIR} = 98438821.456 \tag{21}$$

$$\text{loss\_I} = 11345389.686 \tag{22}$$

2.3 SIRV Model with Lockdown

Lastly, an additional flow from the susceptible population to the recovered population can be shown by adding a vaccination rate ν in the model.

$$\frac{dS}{dt} = -\beta(1 - s(t)/100)\frac{SI}{N} - \nu S \tag{23}$$

$$\frac{dI}{dt} = \beta(1 - s(t)/100)\frac{SI}{N} - \gamma I \tag{24}$$

$$\frac{dR}{dt} = \gamma I + \nu S \tag{25}$$

Optimizing these equations with eq. (8) using the Nelder-Mead method:

$$\beta_{optimal} = 0.409 \tag{26}$$

$$\gamma_{optimal} = 0.092 \tag{27}$$

$$\nu_{optimal} = 2.904 \times 10^{-5} \tag{28}$$

$$
\begin{aligned}
\overline{R_0} &= 1.691 \\
\widetilde{R_0} &= 1.623 \\
\text{Mode}(R_0) &= 0.803 \\
\sigma_{R_0} &= 0.785 \\
R_0 &\in [0.165, 3.047]
\end{aligned}
\tag{29}
$$

$$
\text{loss\_SIR} = 94636860.384 \tag{30}
$$

$$
\text{loss\_SIR} = 10840360.995 \tag{31}
$$

2.4 Optimizing Window Length for Time-varying Vaccination Rate

However, as observed by the value of $\nu_{optimal}$ from eq. (28) which is almost negligible and the overestimation of infected individuals in fig. 5b suggests that ν might be varying with time. This suggests to accurately estimate the infected population a time-varying vaccination rate should be used as the transition from susceptibility to direct recovery fluctuates with time[24,38]. Therefore, using the $\beta_{optimal}$ and $\gamma_{optimal}$ from eq. (26) and eq. (27), we first find the optimal window length[?] i.e., the window length that result in the least loss from eq. (8). Using different window lengths to estimate value of ν given a sub-interval of $[start, start + window\_length]$, we observe the following. It is crucial to note that the variable ν is constrained to be a positive integer, reflecting the inherent one-way nature of vaccination: individuals can only receive vaccinations, not return them.

$$
\frac{dS}{dt} = -\beta_{optimal}(1 - s(t)/100)\frac{SI}{N} - \nu S \tag{32}
$$

$$
\frac{dI}{dt} = \beta_{optimal}(1 - s(t)/100)\frac{SI}{N} - \gamma_{optimal} I \tag{33}
$$

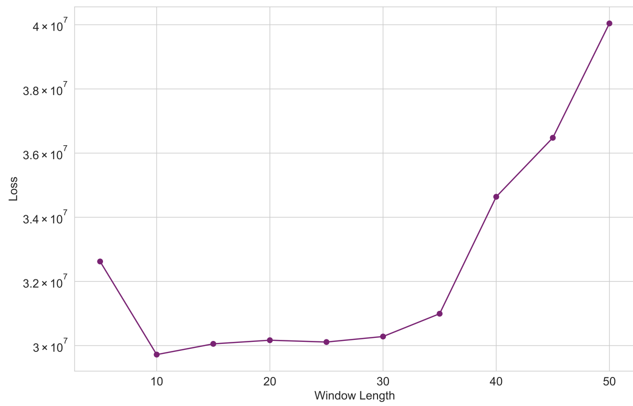
$$
\frac{dR}{dt} = \gamma_{optimal} I + \nu S \tag{34}
$$

The main difference here is instead of supplying these equations with the entire time series, only data lying within the sub-interval of $[start, start + window\_length]$ is given, then in the next step data of the next sub-inteval is provided to optimize for, i.e., $[start + window\_length, start + window\_length + window\_length]$. This model gives shows how ν varies with time. Since, the window length acts like a hyperparameter, we can try different window lengths in order to evaluate which would give us the best results. Therefore, the below graphs show us the loss using eqs. (8) and (9).

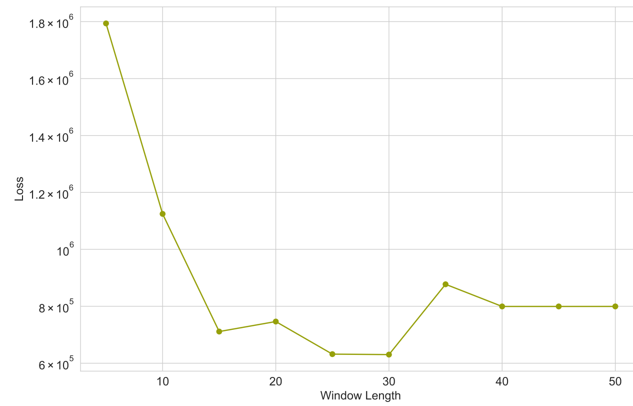
From the following graph, although the window length of 10 is best for all three groups of the population but this window length leads to a bad approximation for the infected population therefore, we choose a window length of 15.

This shows us that the ν coincides with the actual of data of when the vaccination drive was first launched in India[63–65]. Therefore, using these values we finally, recompute $\beta_{optimal}$ and $\gamma_{optimal}$ by supplying them into the equations for the SIRV Model with time varying ν.

Figure 1. Loss for Different Window Lengths

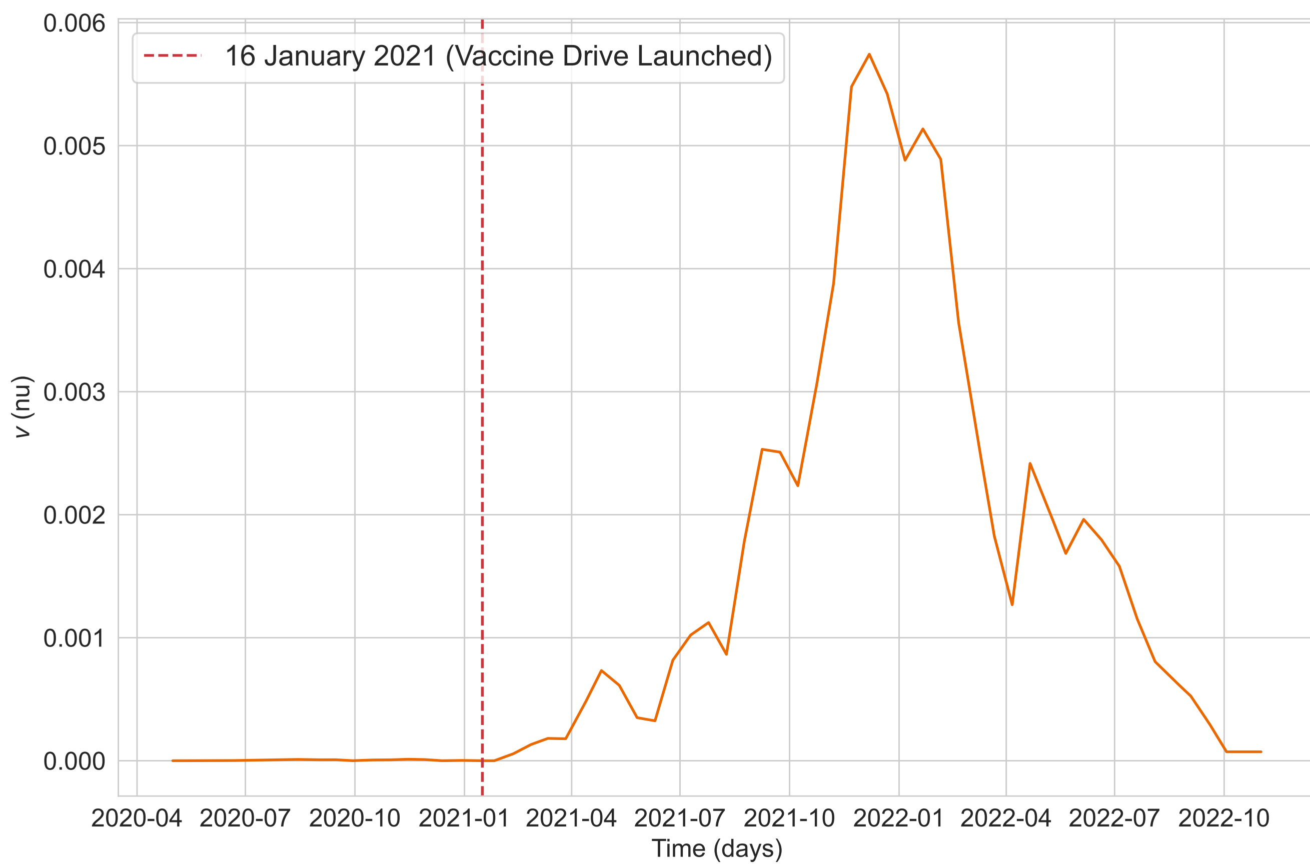


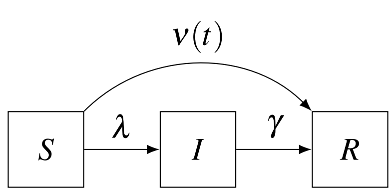
(a) Loss for Different Window Lengths for Susceptible, Infected and Recovered Population



(b) Loss for Different Window Lengths for Infected Population

Figure 2. ν Varying with Time





2.5 SIRV Model with Lockdown (Time-varying Vaccination Rate)

Lastly, we present our model using the following set of equations:

$$\frac{dS}{dt} = -\beta(1 - s(t)/100)\frac{SI}{N} - \nu(t)S \tag{35}$$

$$\frac{dI}{dt} = \beta(1 - s(t)/100)\frac{SI}{N} - \gamma I \tag{36}$$

$$\frac{dR}{dt} = \gamma I + \nu(t)S \tag{37}$$

$$\beta_{optimal} = 0.463 \tag{38}$$

$$\gamma_{optimal} = 0.114 \tag{39}$$

$$\begin{aligned} \overline{R_0} &= 1.546 \\ \widetilde{R_0} &= 1.483 \\ \text{Mode}(R_0) &= 0.734 \\ \sigma_{R_0} &= 0.718 \\ R_0 &\in [0.150, 2.785] \end{aligned} \tag{40}$$

$$\text{loss\_SIR} = 29116762.926 \tag{41}$$

$$\text{loss\_I} = 658537.443 \tag{42}$$

See Figures 3a–6b and **??** to see how the different models compare against each other.

Now, that a relation between β and $s(t)$ is set up, it must be investigated how stringency index affects the normalized Gross domestic product (GDP)[66,67]. To do this a polynomial equation of the third degree is fitted to the data points (x, y) where x is the $s(t)$ stringency at time t, and y the normalized GDP and minimize the squared error. For India after fitting a 3 degree polynomial, the following equation is obtained:

$$\begin{aligned} GDP = -5.96640236 \times 10^{-5} s^3 + 6.65064332 \times 10^{-3} s^2 - 2.23109924 \times 10^{-1} s^1 \\ + 1.01357226 \times 10^2 \end{aligned} \tag{43}$$

2.6 Reinforcement Learning

Given that the government is an agent that takes decisions in a deterministic environment defined above, we use reinforcement learning to model the competing costs of the environment. This environment is formally as a Markov decision process, and can be described as follows:

- Set of States $\mathcal{S}$: The state of the environment are described through the descriptors like the normalized GDP ($(GDP_{predicted} - GDP_{min})/(GDP_{max} - GDP_{min})$), R_e, a list of all the previous actions (in changing the stringency) and the proportion of the population that was susceptible, infected and recovered. The starting states are simply these values at the starting date and no previous actions.
- Actions $\mathcal{A}$: The stringency index variable was analyzed with a sample size of 915. The mean value was approximately 61.965049, with a standard deviation of 17.669831. The minimum value was 31.480000, while the maximum value reached 96.300000. And the differences between two consecutive stringencies had a mean of -0.070919, and standard deviation of 1.427145, with the minimum being -14.360000 and maximum 16.670000. Based on this we define the discrete action space. There are 7 actions for the agent, it can keep the stringency index same, reduce/increase by 2.5, reduce/increase by 5, and reduce/increase by 10 given that the stringency index doesn't exceed 100 or go below 0.
- Transition dynamics $\mathcal{T}\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)$ that map a state-action pair at time t onto a distribution of states at time $t+1$. This state transition is defined by the SIR model with lockdown and the model of how stringency index affects the GDP.
- Immediate reward $\mathcal{R}\left(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}\right)$. Here we define a reward strategy however it should be noted that this strategy can be easily changed to prioritize different needs.
- Discount Factor $\gamma \in [0,1]$, where lower values place more emphasis on immediate rewards. Here, we choose the default discount factor of 0.99.

In general, the policy π is a mapping from states to a probability distribution over actions: $\pi : \mathcal{S} \rightarrow p(\mathcal{A} = \mathbf{a} \mid \mathcal{S})$. If the MDP is episodic, i.e., the state is reset after each episode of length T, then the sequence of states, actions and rewards in an episode constitutes a trajectory or rollout of the policy. Every rollout of a policy accumulates rewards from the environment, resulting in the return $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$. The goal of RL is to find an optimal policy, π^*, which achieves the maximum expected return from all states. To achieve this, reinforcement learning start with an initial arbitrary policy, i.e., a Q-table with no entries. Q-table is a mapping from states $s_k \in \mathscr{S}$ to a predefined set of actions to increase or decrease the stringency at time t, which are the actions $a_k \in \mathscr{A}$. Each entry of the Q-table $\left(Q_k\left(s_k, a_k\right)\right)$ associates an action in the finite sequence $\left(\mathscr{A}_j\right)_{j \in \mathbb{J}^+}$ to a state of the finite sequence $\left(\mathscr{S}_i\right)_{i \in \mathbb{I}^+}$[48]. In this case of epidemic control by NPI based strategies this policy represents the series of stringencies to be imposed upon the population to shift the initial status of the environment to a targeted status which is equivalent to the desired set of system states. This is how the Q-table updates saying, if in state s_k the most ideal action is a_k. After having more and more experience with the environment and understanding which actions lead to a higher reward r an optimal policy is derived be maximizing the expected value of discounted reward $J\left(r_k\right) = \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^{(k-1)} r_k\right]$, where, discount factor $\gamma \in [0,1]$ and time steps $k = 1, 2, \ldots$.

The stringency index emerges as a critical factor influencing both the Gross Domestic Product (GDP) and the rate of infection spread. The decision to escalate or de-escalate the stringency index is a strategic one, with significant implications. Increasing the stringency decreases the spread of the infection.

Conversely, it must be noted that herd immunity can only be achieved when the epidemic reaches its peak i.e. when the effective reproductive number is equal to one ($R_e = 1$). This can only happen by lowering the stringency index which would allow the natural dynamics of the epidemic to transpire such that the population of susceptible individuals has depleted enough such that it is insufficient to propagate the disease further. Therefore, stringency is used to control the number of infected people and slow down the rate at which the epidemic reaches its peak, so that hospitals could house the number of infected people.

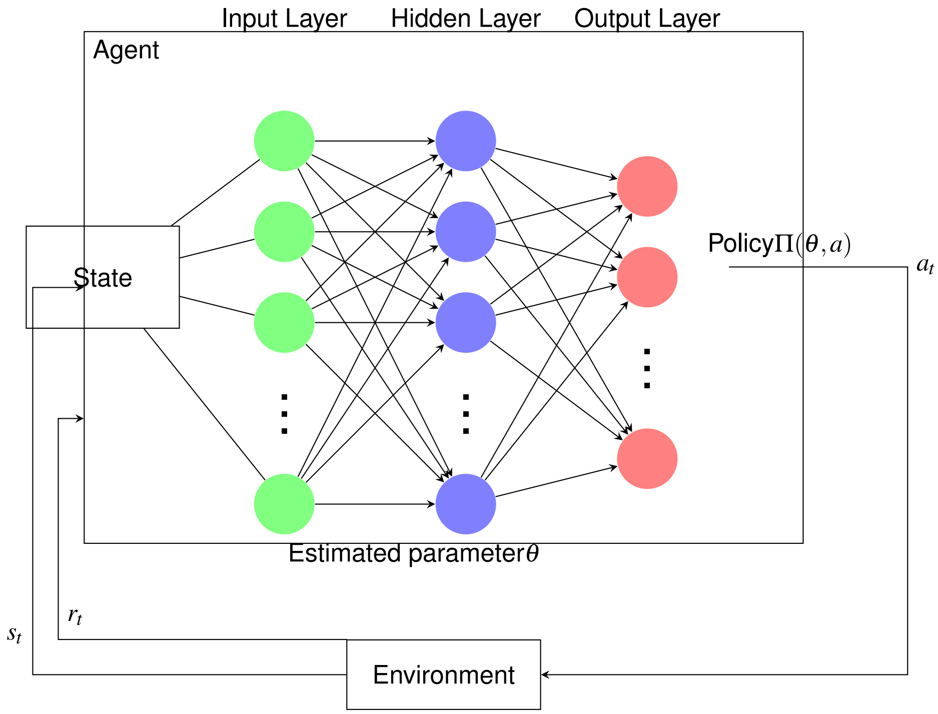
With this basis we define the reward function[19]. In Deep Reinforcement Learning (DRL), positive rewards promote and negative rewards demote actions. The agent tries to generate such a policy/knowledge to avoid the discouraging situation by following the policy. By designing a proper reward function, it is possible to generate such an agent that may follow the human desired situation.

The reward function is parameterized to account for key factors influencing decision-making. To incentivize reduction of $Reff$ (effective reproductive number) a negative reward is imposed of $-20 * R_e$, but as the R_e is between $[1.9, 1, 5]$ we start to positively reward it with a multiple of the GDP ($100 *$ min_max_normalized_GDP) and if the R_e is below 1.5 then $200 *$ min_max_normalized_GDP. Furthermore, there's a simple positive reward of 10 if R_e is below the threshold value (1.9) and negative reward of -10 otherwise. To reward not changing the stringencies frequently, we reward the absolute different between the previous stringency and the current stringency negatively ($|s(t) - s(t-1)| * -1 * 5$). If the proportion of the infected population were to rise above 0.003 the model is punished (-5000) and otherwise rewarded ($+20$). As a reward for herd immunity, another reward is defined using the proportion of the infected population. If it lies between $[0.0005, 0.003]$ a reward of 2000 is given - else 0.

It should be realized there can be an infinite number of ways to design the reward function to be more human and upgrade the way a decision is taken given the situation[68]. Therefore, this research act as a framework for promoting the development of more efficient reward strategies for the same.

The agent observes the percentage of the population that is susceptible, infected, recovered, and the GDP which extrapolated from the stringency index and all the past moves played, i.e., all the past stringency indexes decided by the agent. The following values are fed into a simple network. Stable Baselines3 supports multiple inputs by using Dict Gym space. This can be done using MultiInputPolicy, which by default uses the CombinedExtractor features extractor to turn multiple inputs into a single vector, handled by the network arch network. For data that varies with time (stringency, normalized GDP, R_e) we use a simple LSTM architecture[56]. For other data like, the current proportion of the population that is susceptible, infected, and recovered we use a simple fully connected layer. The output from both these networks are concatenated and used by the reinforcement learning agent to train on.

We train the model for 2742 time steps for 200 iterations and the best results are presented below.



3 Results

Using the simple SIR model from Equations (1)–(14), to model the disease dynamics we get: Here, it can be observed that the SIR model accurately fits the susceptible population and recovered population but overestimates the infected population by a significant margin. Although overestimating the infected population may not always be problematic, in our specific case, it can create complications. This is because our research involves rewarding the agent when the proportion of infected individuals falls below a predetermined threshold. Consequently, an overestimation of the infected population could lead to incorrect decision-making and undesirable outcomes.

Combining the lockdown dynamics in the SIR model using Equations (15)–(22), we get the following: Here, it can be observed there's an overestimation of infected individuals, but, the two stages of the epidemic are being accounted for. This is what suggests that there might be depletion of infected individuals through vaccination.

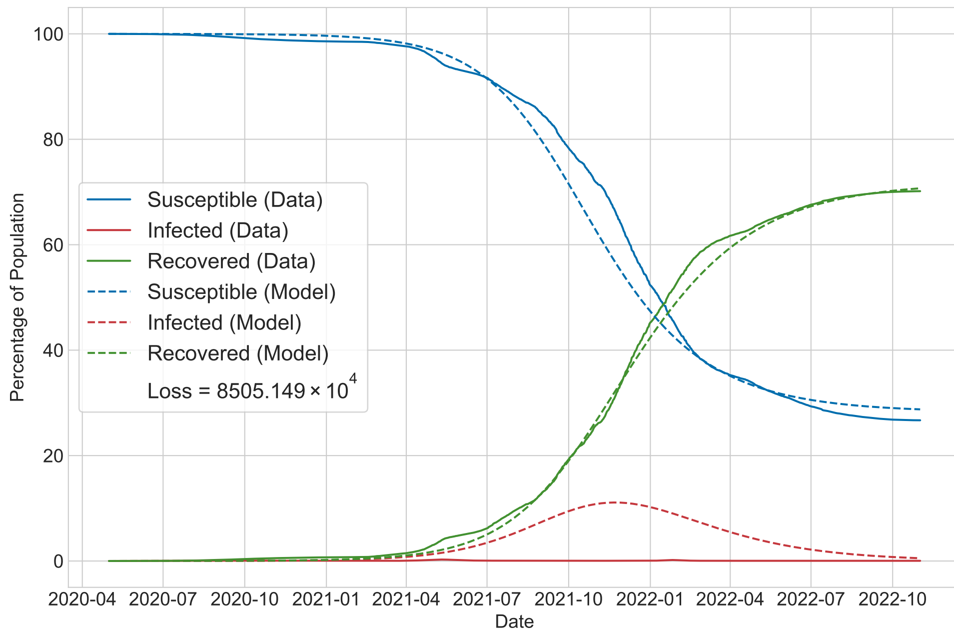
Combining vaccination dynamics with SIR with lockdown model using Equations (23)–(31), we get the following: Here, because the value of ν eq. (28) is negligible, it makes no change. However, time-varying ν shall be able to better account for the these dynamics.

For time-varying ν and SIRV model with lockdown from Equations (35)–(42), we get the following:

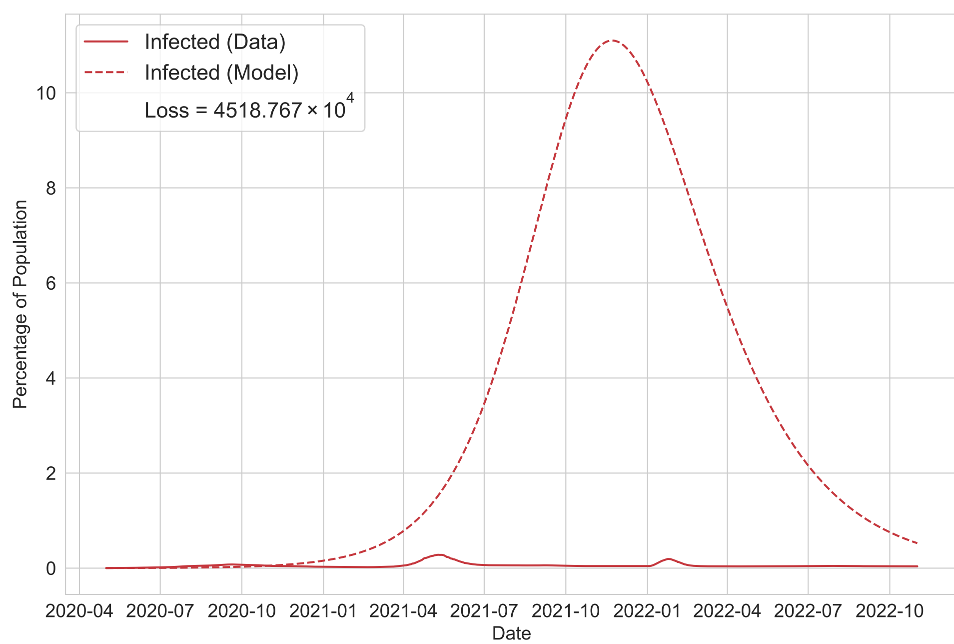
With a time-varying ν (vaccination rate) and the effect of lockdown, our model is able to account for the infected individuals and reduce the cost in comparison to all the previously formalized models for the data. This shows how interventions and changes in the way people behave in response of a epidemic[13] play a major role in the way the epidemic unfolds.

However, Non-pharmaceutical Interventions (NPIs) come with costs for developing nations. Below are scatter plots with pearson correlation, for three countries (India, Mexico, Brazil) which are Emerging Market and Developing Economies[4] from May, 2020 to October, 2022. It can be observed that stringency has a negative impact on the normalized Gross domestic product (GDP). Therefore, in some countries policies made during an epidemic have competing costs. This is not the case for Advanced Economies like (USA, Japan, Canada). From this we conclude, that maybe there are other factors that explain the dip

Figure 3. SIR Model with lockdown for India

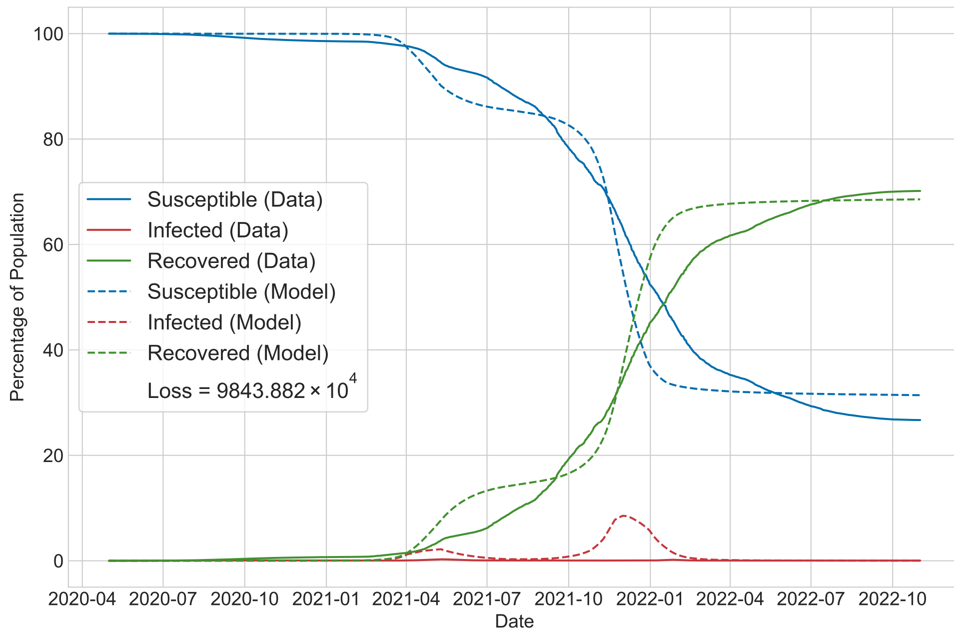


(a) SIR model with lockdown

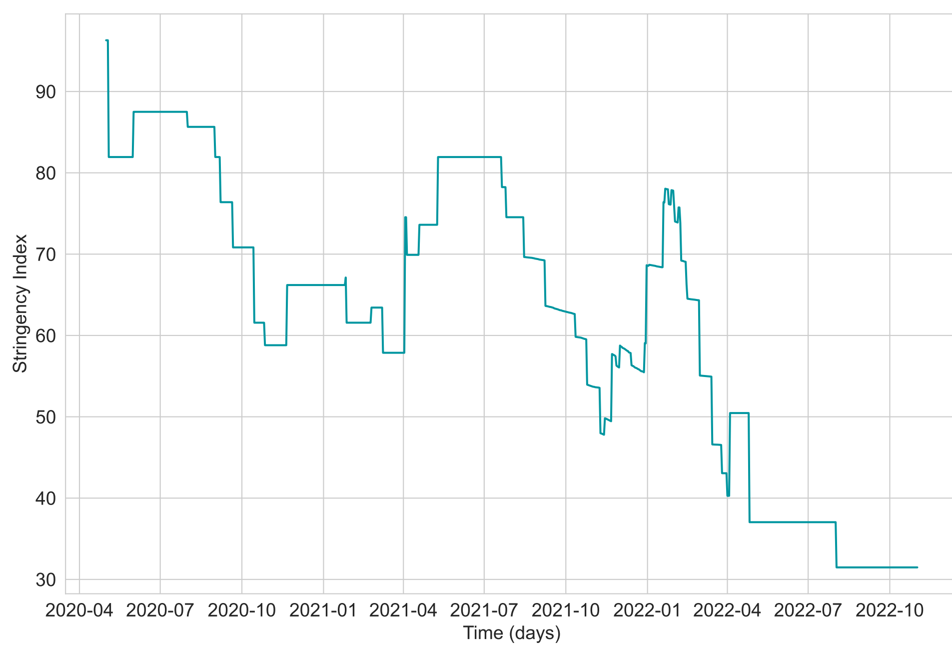


(b) Infections modelled with SIR model

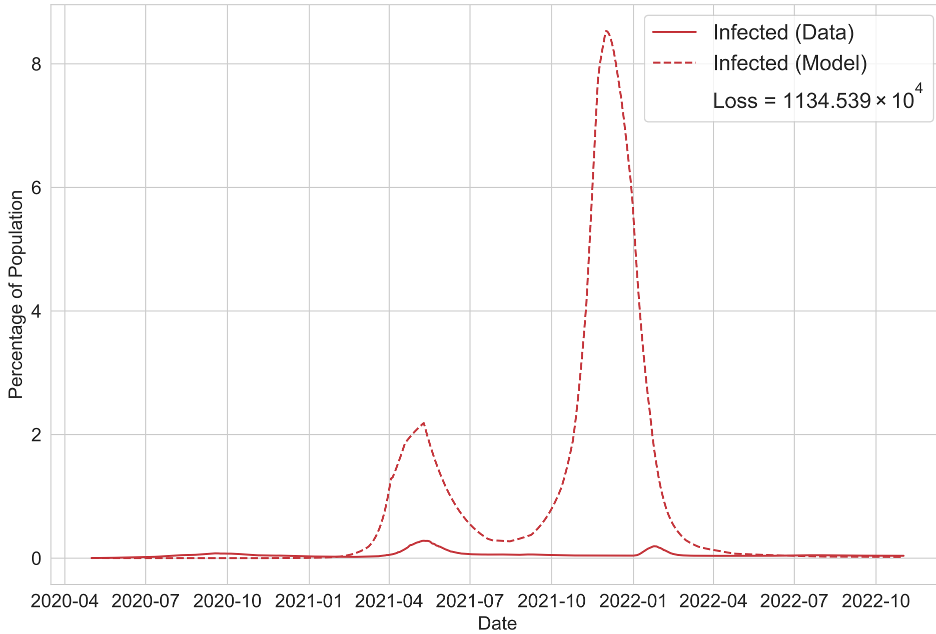
Figure 4. SIR Model with lockdown for India



(a) SIR model with lockdown



(b) Stringency varying with Time



(c) Infections modelled with SIR model with Lockdown

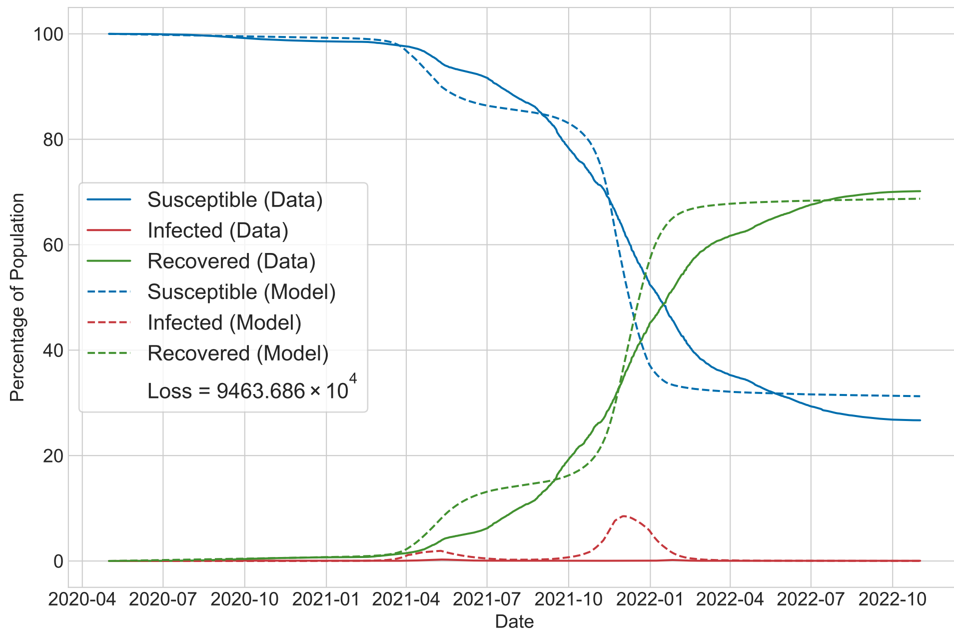
in the normalized GDP than just an increase in the stringency.

After median filtering to smooth the output (to reinforce the negative reward from changing the stringencies) from the trained reinforcement learning agent, here are some of the results obtained: In the presented result, we can see the reinforcement learning agent outperform the modelled outcome. A strategic decision is made by the agent to maintain the stringency index below 80 throughout the course of the study. This approach allows for the natural progression of disease dynamics, resulting in a rapid reduction of the effective reproduction number R_e to 1.2, significantly ahead of the projected timeline (refer to figure fig. 10). A subsequent decrease in stringency in the later stages leads to an increase in the normalized Gross Domestic Product (GDP), indicating an economic upturn. While this strategy poses a higher risk in terms of infection rates during the initial phase of the epidemic (prior to vaccine rollout), it proves to be more beneficial for the nation's economy in the long run. It is suggested that a more balanced policy, falling between the actual and the proposed strategy in this study, could have been a more effective approach for the government to adopt. This would have potentially led to a more optimal balance between controlling the spread of the disease and minimizing economic impact. An alternative outcome is presented by the agent, demonstrating a gradual escalation in stringency over time. This strategy results in the majority of infections occurring prior to the vaccine's release, with a subsequent cessation of new infections. This approach, however, has implications for the normalized Gross Domestic Product (GDP), as seen by the observed decline in figure fig. 11. Furthermore, the effective reproduction number (R_e) does not achieve the target value of 1.2 under this strategy. These findings underscore the complexity of managing public health crises and the need for careful strategic planning to balance health outcomes with economic considerations.

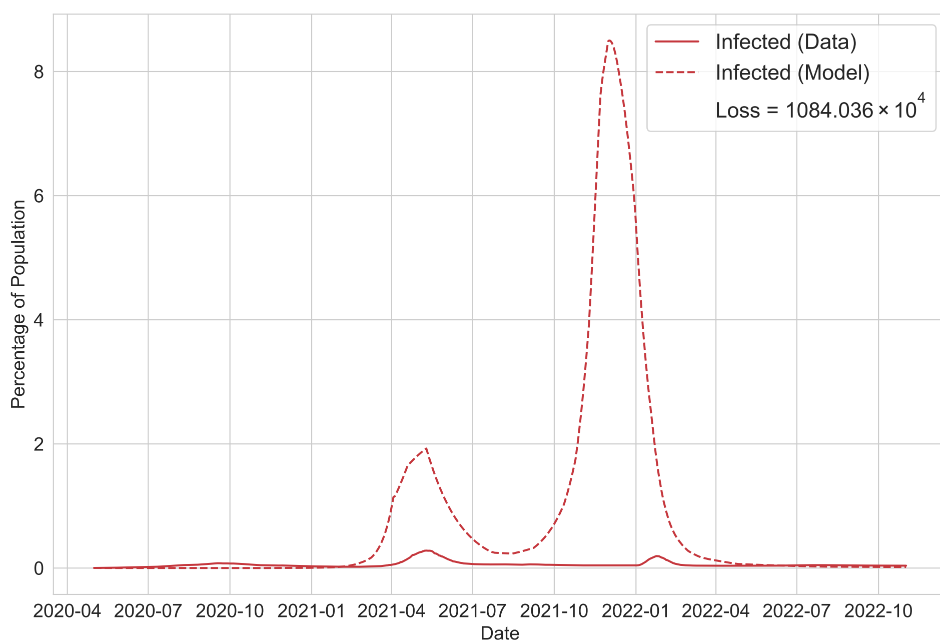
4 Discussion

The paper seeks to inspire epidemiologists by highlighting the advancements achieved through the application of reinforcement learning in policy making during the pandemic. We introduce a virtual environment that closely simulates a pandemic scenario and thoroughly explore innovative strategies

Figure 5. SIRV Model with lockdown for India

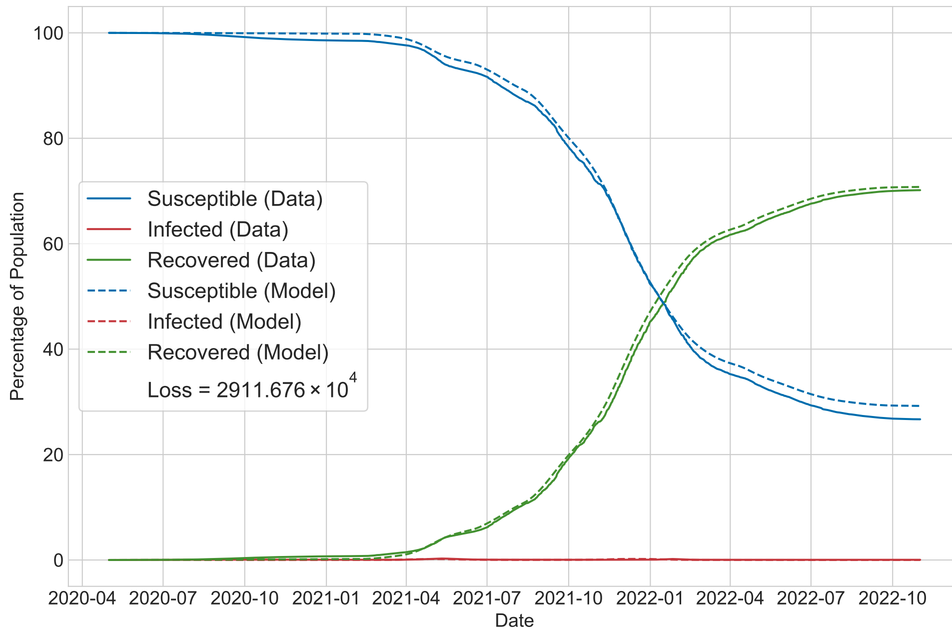


(a) SIRV model with lockdown

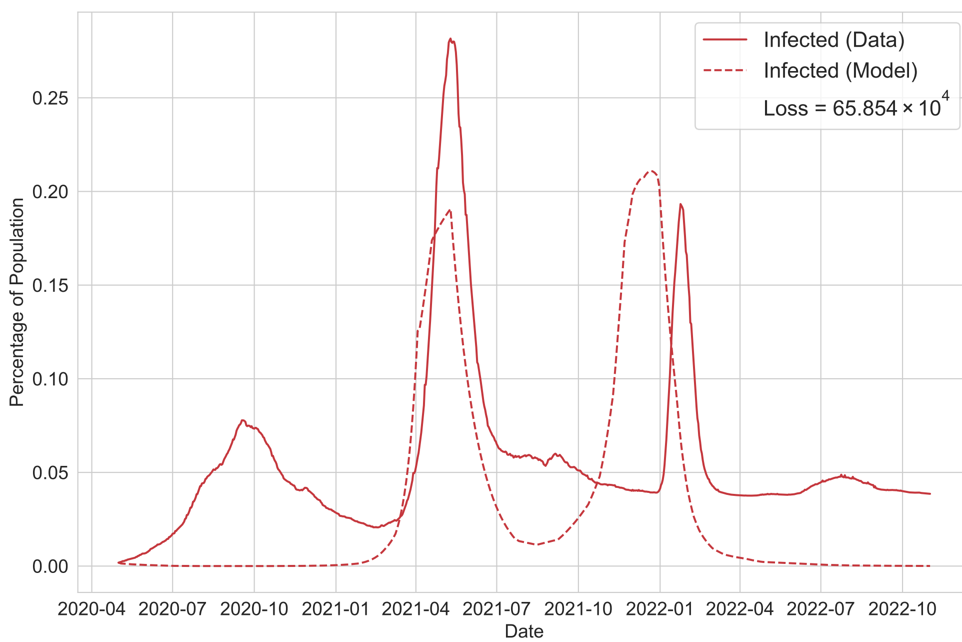


(b) Infections modelled with SIRV model with Lockdown

Figure 6. SIRV Model with lockdown and time-varying ν for India

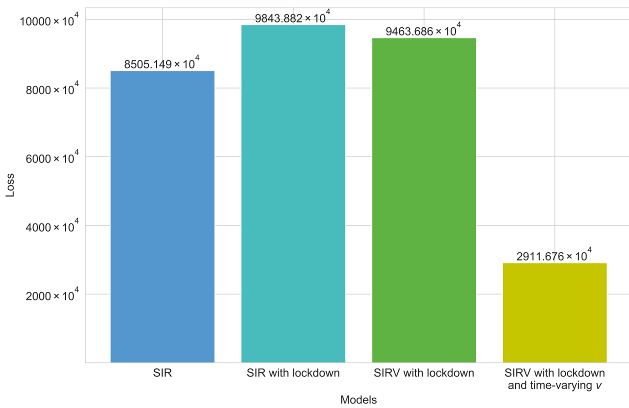


(a) SIRV model with lockdown and time-varying ν

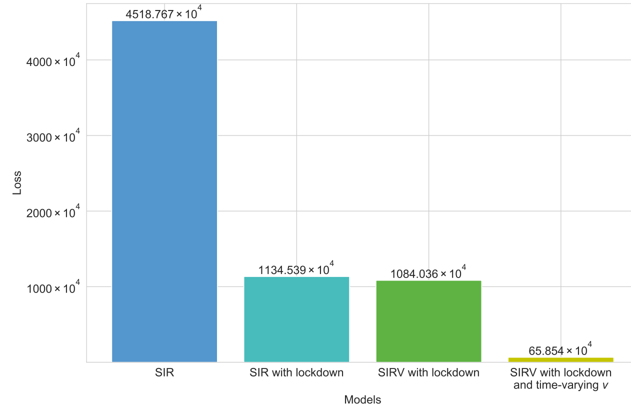


(b) Infections modelled with SIRV model with Lockdown and time-varying ν

Figure 7. Loss for Different Models



(a) Loss for Different Models for Susceptible, Infected and Recovered Population



(b) Loss for Different Models for Infected Population

for disease mitigation using reinforcement learning. Our proposed approach demonstrates compelling efficacy in achieving optimal decision-making, effectively balancing the formidable challenges posed by the pandemic and economic considerations. We are confident that this research contribution will forge a connection between epidemic studies and reinforcement learning, offering valuable insights to help humanity better defend against the ongoing pandemic crisis.

5 Experiment Settings

5.1 Dataset

The population-level epidemiological data can be obtained from the 'Our World In Data COVID-19' dataset: https://ourworldindata.org/coronavirus or more specifically: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv[62]. Data for the total cases, and recovered was acquired by scraping the Worldometers website[69] used Internet Archive[70]. Quaterly GDP data can be obtained from the 'Organisation for Economic Co-operation and Development': https://www.oecd-ilibrary.org/economics/data/main-economic-indicators/main-economic-indicators-complete-database_data-00052-en[71].

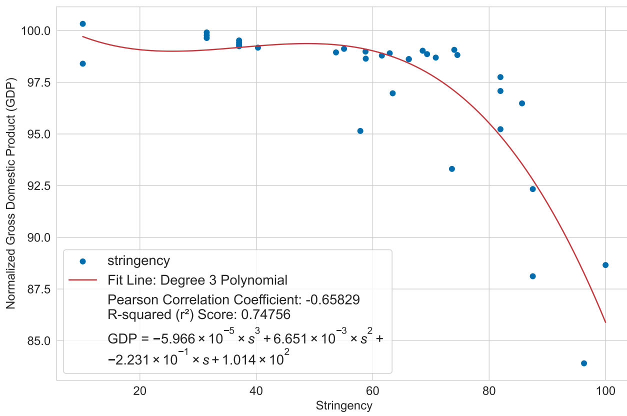
5.2 Code

We used stable-baseline3[72], Pytorch[73], Scipy[74], Pandas, Matplotlib, in Python[75]. Code: https://github.com/psymbio/sir_rl

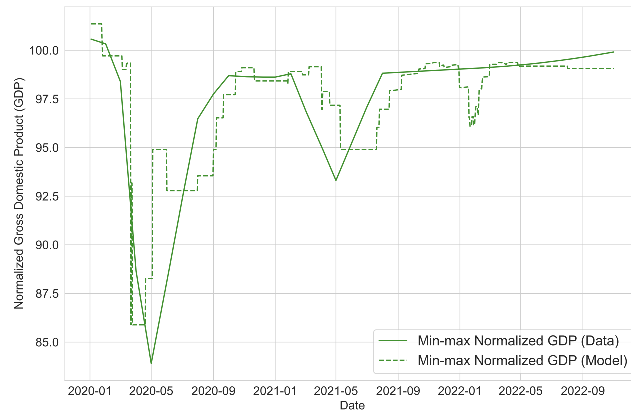
References

1. Baker, R. E. *et al.* Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205, DOI: 10.1038/s41579-021-00639-z (2022).
2. Tan, M. K. I. COVID-19 in an inequitable world: the last, the lost and the least. *Int. Heal.* **13**, 493–496, DOI: 10.1093/inthealth/ihab057 (2021). https://academic.oup.com/inthealth/article-pdf/13/6/493/41430650/ihab057.pdf.
3. Who coronavirus (covid-19) dashboard. https://covid19.who.int/. Accessed: 2024-01-12.
4. World economic outlook, april 2020: The great lockdown. Accessed: 2024-01-12.

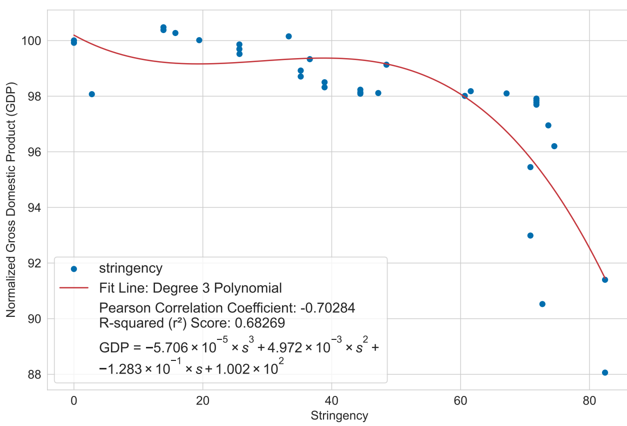
Figure 8. Stringency and GDP for Developing Economies



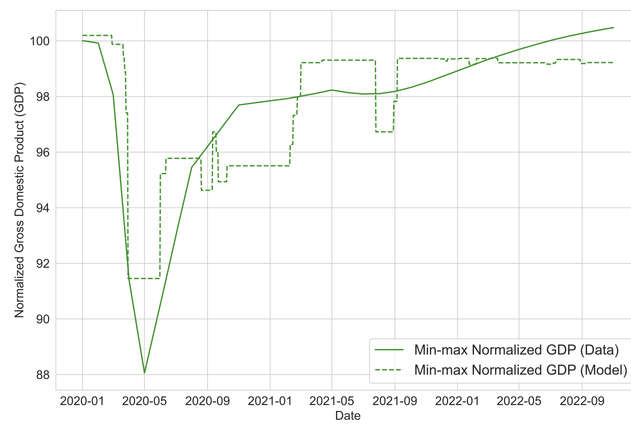
(a) Stringency and Normalized GDP for India



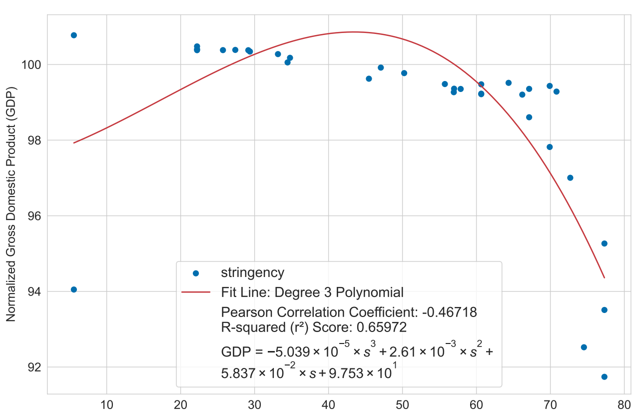
(b) Normalized GDP modelled with Stringency for India



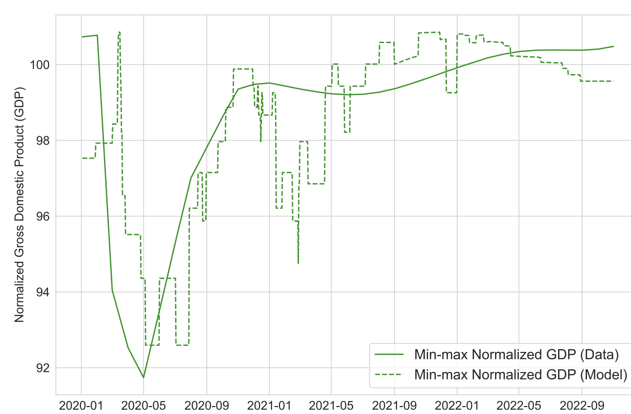
(c) Stringency and Normalized GDP for Mexico



(d) Normalized GDP modelled with Stringency for Mexico

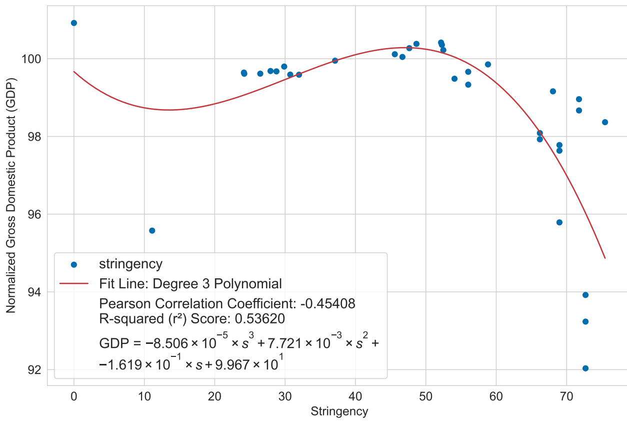


(e) Stringency and Normalized GDP for Brazil

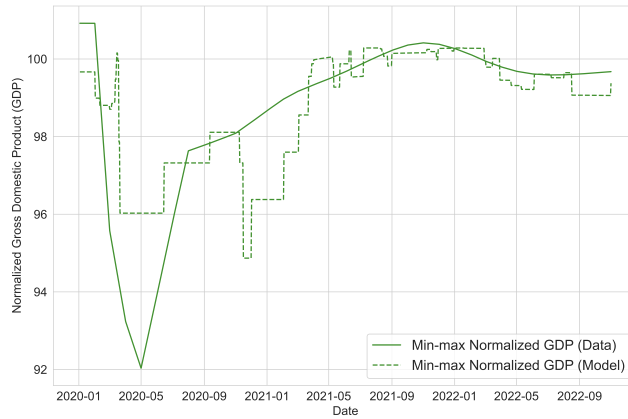


(f) Normalized GDP modelled with Stringency for Brazil

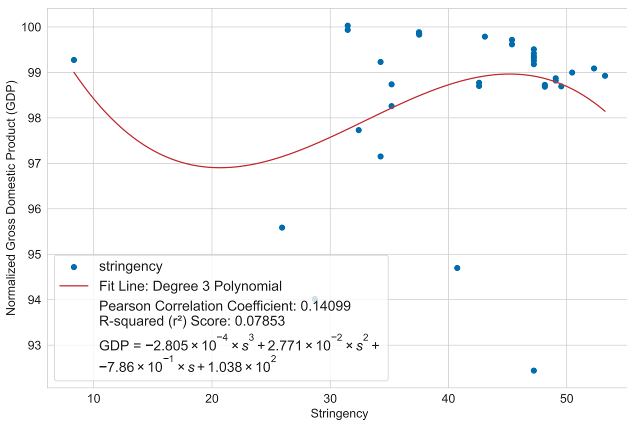
Figure 9. Stringency and GDP for Advanced Economies



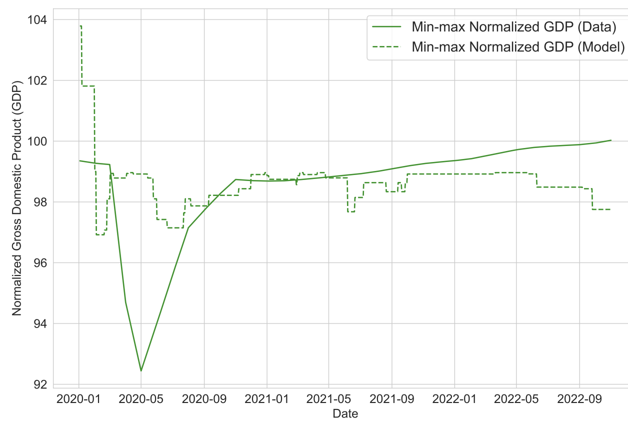
(a) Stringency and Normalized GDP for United States



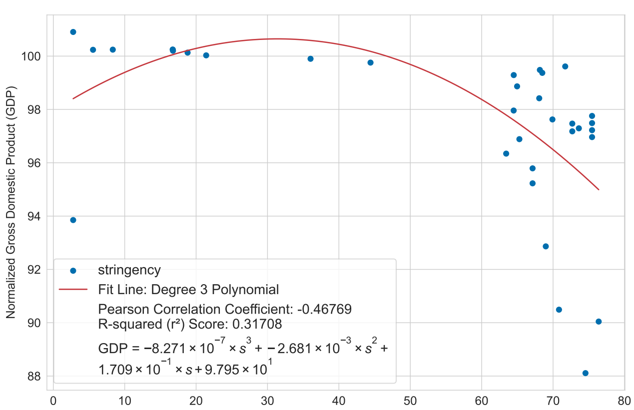
(b) Normalized GDP modelled with Stringency for United States



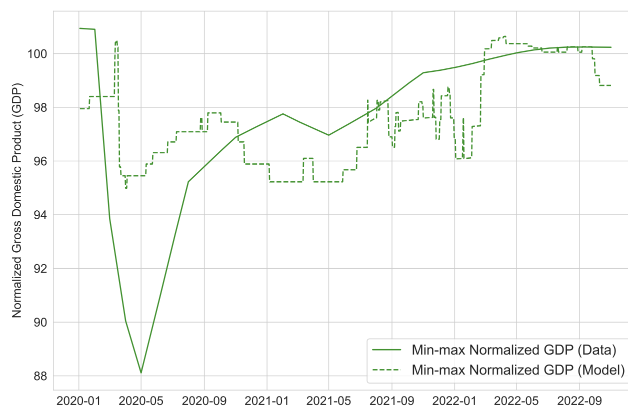
(c) Stringency and Normalized GDP for Japan



(d) Normalized GDP modelled with Stringency for Japan

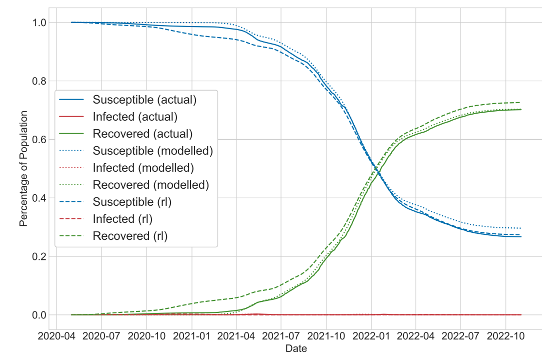


(e) Stringency and Normalized GDP for Canada

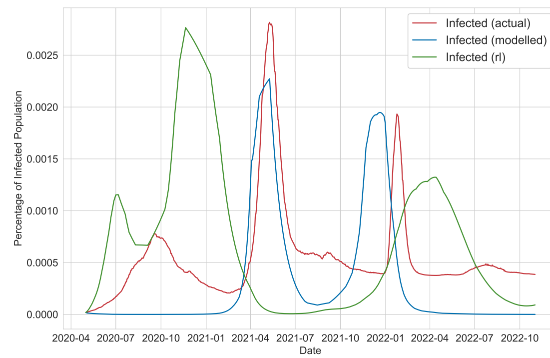


(f) Normalized GDP modelled with Stringency for Canada

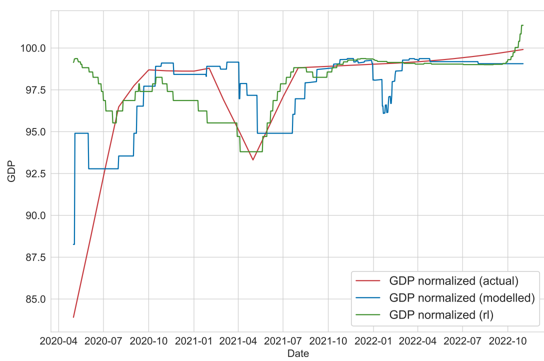
Figure 10. Strategy from Reinforcement Learning Agent 1



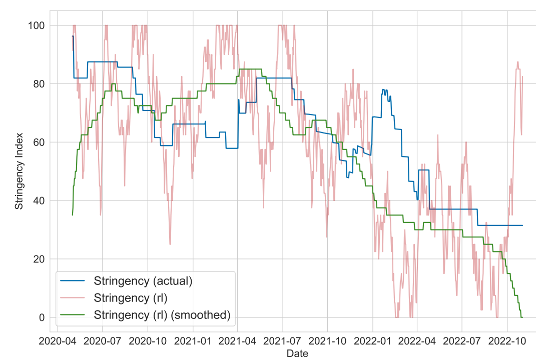
(a) SIR Dynamics



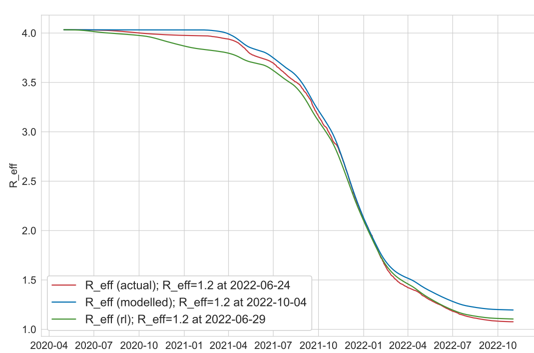
(b) Infected Population changing over Time



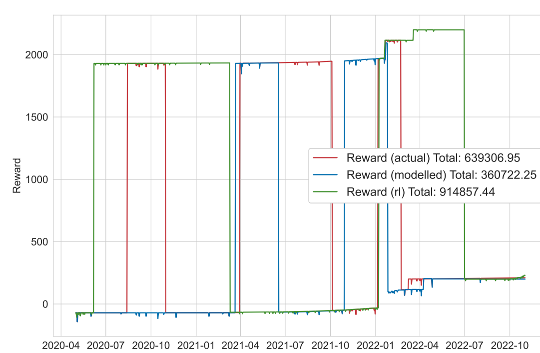
(c) Normalized GDP changing over Time



(d) Stringency changing over Time

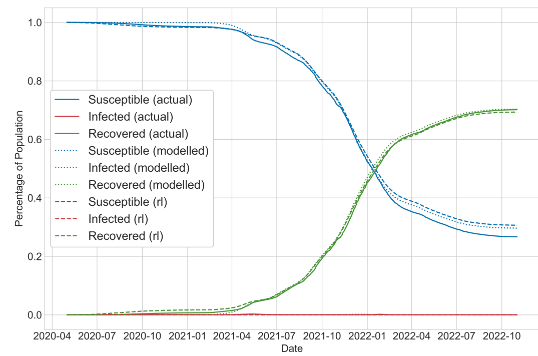


(e) R_e changing over Time

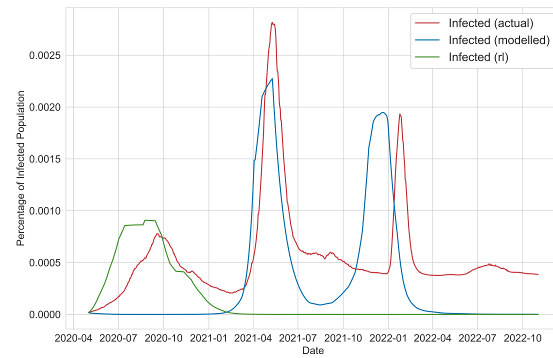


(f) Reward changing over Time

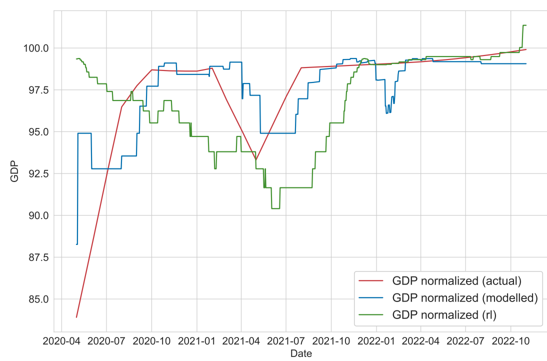
Figure 11. Strategy from Reinforcement Learning Agent 2



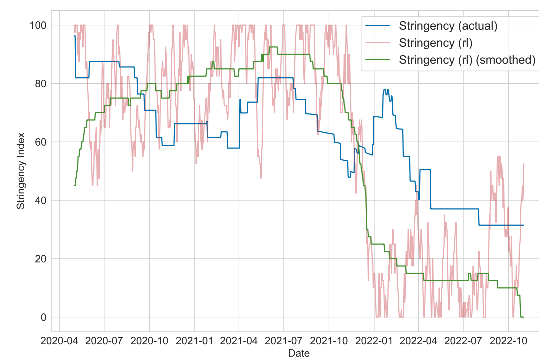
(a) SIR Dynamics



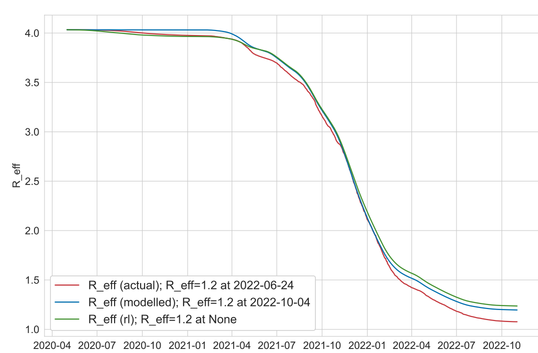
(b) Infected Population changing over Time



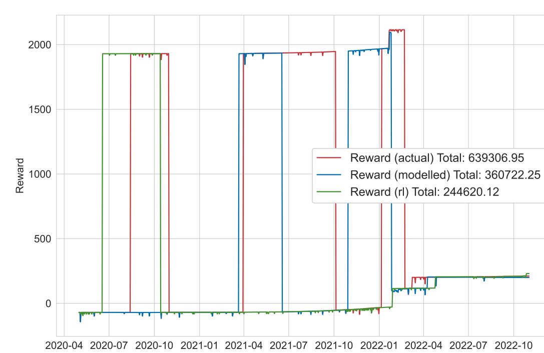
(c) Normalized GDP changing over Time



(d) Stringency changing over Time



(e) R_e changing over Time



(f) Reward changing over Time

5. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (covid-19): A review. *Int. J. Surg.* **78**, 185–193, DOI: 10.1016/j.ijsu.2020.04.018 (2020).

6. Gagnon, J. E., Kamin, S. B. & Kearns, J. The impact of the covid-19 pandemic on global gdp growth. *J. Jpn. Int. Econ.* **68**, 101258, DOI: 10.1016/j.jjie.2023.101258 (2023).

7. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet* **395**, 931–934, DOI: 10.1016/s0140-6736(20)30567-5 (2020).

8. Song, S., Liu, X., Li, Y. & Yu, Y. Pandemic policy assessment by artificial intelligence. *Sci. Reports* **12**, 13843, DOI: 10.1038/s41598-022-17892-8 (2022).

9. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science* **368**, 395–400, DOI: 10.1126/science.aba9757 (2020).

10. Nguyen, T. *et al.* Covid-19 vaccine strategies for aotearoa new zealand: a mathematical modelling study. *The Lancet Regional Heal. - West. Pac.* **15**, 100256, DOI: 10.1016/j.lanwpc.2021.100256 (2021).

11. Kim, D., Keskinocak, P., Pekgün, P. & Yildirim, The balancing role of distribution speed against varying efficacy levels of covid-19 vaccines under variants. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-11060-8 (2022).

12. Jalloh, M. F. *et al.* Drivers of covid-19 policy stringency in 175 countries and territories: Covid-19 cases and deaths, gross domestic products per capita, and health expenditures. *J. Global Heal.* **12**, DOI: 10.7189/jogh.12.05049 (2022).

13. Caldwell, J. M. *et al.* Understanding covid-19 dynamics and the effects of interventions in the philippines: A mathematical modelling study. *The Lancet Regional Heal. - West. Pac.* **14**, 100211, DOI: 10.1016/j.lanwpc.2021.100211 (2021).

14. Ferguson, N. *et al.* Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. DOI: 10.25561/77482 (2020).

15. De Foo, C. *et al.* Health financing policies during the covid-19 pandemic and implications for universal health care: a case study of 15 countries. *The Lancet Global Heal.* **11**, e1964–e1977, DOI: 10.1016/s2214-109x(23)00448-5 (2023).

16. Hollingsworth, T. D., Klinkenberg, D., Heesterbeek, H. & Anderson, R. M. Mitigation strategies for pandemic influenza a: Balancing conflicting policy objectives. *PLoS Comput. Biology* **7**, e1001076, DOI: 10.1371/journal.pcbi.1001076 (2011).

17. Pangallo, M. *et al.* The unequal effects of the health–economy trade-off during the covid-19 pandemic. *Nat. Hum. Behav.* DOI: 10.1038/s41562-023-01747-x (2023).

18. Ash, T., Bento, A. M., Kaffine, D., Rao, A. & Bento, A. I. Disease-economy trade-offs under alternative epidemic control strategies. *Nat. Commun.* **13**, DOI: 10.1038/s41467-022-30642-8 (2022).

19. Ohi, A. Q., Mridha, M. F., Monowar, M. M. & Hamid, M. A. Exploring optimal control of epidemic spread using reinforcement learning. *Sci. Reports* **10**, DOI: 10.1038/s41598-020-79147-8 (2020).

20. Padmanabhan, R., Meskin, N., Khattab, T., Shraim, M. & Al-Hitmi, M. Reinforcement learning-based decision support system for covid-19. *Biomed. Signal Process. Control.* **68**, 102676, DOI: https://doi.org/10.1016/j.bspc.2021.102676 (2021).

21. Alvarez, F. E., Argente, D. & Lippi, F. A simple planning problem for covid-19 lockdown. DOI: 10.3386/w26981 (2020).

22. Rachel, L. An analytical model of covid-19 lockdowns. (2020). Https://www.lse.ac.uk/CFM/assets/pdf/CFM-Discussion-Papers-2020/CFMDP2020-29-Paper.pdf.

23. Redlin, M. Differences in npi strategies against covid-19. *J. Regul. Econ.* **62**, 1–23, DOI: 10.1007/s11149-022-09452-9 (2022).

24. Liang, L.-L., Kao, C.-T., Ho, H. J. & Wu, C.-Y. Covid-19 case doubling time associated with non-pharmaceutical interventions and vaccination: A global experience. *J. Global Heal.* **11**, DOI: 10.7189/jogh.11.05021 (2021).

25. Patel, M. D. *et al.* The joint impact of covid-19 vaccination and non-pharmaceutical interventions on infections, hospitalizations, and mortality: An agent-based simulation. DOI: 10.1101/2020.12.30.20248888 (2021).

26. Gagnon, J. E. & Rose, A. 23-8 how did korea's fiscal accounts fare during the covid-19 pandemic? (2023).

27. Deb, P., Furceri, D., Ostry, J. & Tawk, N. The economic effects of covid-19 containment measures. *IMF Work. Pap.* **20**, DOI: 10.5089/9781513550251.001 (2020).

28. Eichenbaum, M. S., Rebelo, S. & Trabandt, M. The macroeconomics of epidemics. *The Rev. Financial Stud.* **34**, 5149–5187, DOI: 10.1093/rfs/hhab040 (2021).

29. Lim, S. & Sohn, M. How to cope with emerging viral diseases: lessons from south korea's strategy for covid-19, and collateral damage to cardiometabolic health. *The Lancet Regional Heal. - West. Pac.* **30**, 100581, DOI: 10.1016/j.lanwpc.2022.100581 (2023).

30. Coronavirus: South korea seeing a 'stabilising trend'. https://www.bbc.com/news/av/world-asia-51897979. Accessed: 2024-01-12.

31. Hale, T. *et al.* A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nat. Hum. Behav.* **5**, 529–538, DOI: 10.1038/s41562-021-01079-8 (2021).

32. Hethcote, H. W. *Three Basic Epidemiological Models*, 119–144 (Springer Berlin Heidelberg, 1989).

33. Hethcote, H. W. *The Basic Epidemiology Models: Models, Expressions for R0, Parameter Estimation, and Applications*, 1–61 (WORLD SCIENTIFIC, 2008).

34. Allen, L. J. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Model.* **2**, 128–142, DOI: https://doi.org/10.1016/j.idm.2017.03.001 (2017).

35. Cooper, I., Mondal, A. & Antonopoulos, C. G. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons amp; Fractals* **139**, 110057, DOI: 10.1016/j.chaos.2020.110057 (2020).

36. Bjørnstad, O. N., Shea, K., Krzywinski, M. & Altman, N. The seirs model for infectious disease dynamics. *Nat. Methods* **17**, 557–558, DOI: 10.1038/s41592-020-0856-2 (2020).

37. Mwalili, S., Kimathi, M., Ojiambo, V., Gathungu, D. & Mbogo, R. Seir model for covid-19 dynamics incorporating the environment and social distancing. *BMC Res. Notes* **13**, DOI: 10.1186/s13104-020-05192-1 (2020).

38. Marinov, T. T. & Marinova, R. S. Adaptive sir model with vaccination: simultaneous identification of rates and functions illustrated with covid-19. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-20276-7 (2022).

39. Maurício de Carvalho, J. P. S. & Rodrigues, A. A. Sir model with vaccination: Bifurcation analysis. *Qual. Theory Dyn. Syst.* **22**, DOI: 10.1007/s12346-023-00802-2 (2023).

40. Thäter, M., Chudej, K. & Pesch, H. J. Optimal vaccination strategies for an seir model of infectious diseases with logistic growth. *Math. Biosci. Eng.* **15**, 485–505, DOI: 10.3934/mbe.2018022 (2018).

41. Turkyilmazoglu, M. An extended epidemic model with vaccination: Weak-immune sirvi. *Phys. A: Statistical Mech. its Appl.* **598**, 127429, DOI: 10.1016/j.physa.2022.127429 (2022).

42. Yaladanda, N., Mopuri, R., Vavilala, H. P. & Mutheneni, S. R. Modelling the impact of perfect and imperfect vaccination strategy against sars cov-2 by assuming varied vaccine efficacy over india. *Clin. Epidemiology Global Heal.* **15**, 101052, DOI: https://doi.org/10.1016/j.cegh.2022.101052 (2022).

43. Lockdowns in sir models. https://benjaminmoll.com/wp-content/uploads/2020/05/SIR_notes.pdf. Accessed: 2023-12-26.

44. Atkeson, A. What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios. DOI: 10.3386/w26867 (2020).

45. Chen, Y.-C., Lu, P.-E., Chang, C.-S. & Liu, T.-H. A time-dependent sir model for covid-19 with undetectable infected persons. *IEEE Transactions on Netw. Sci. Eng.* **7**, 3279–3294, DOI: 10.1109/TNSE.2020.3024723 (2020).

46. Nguyen, Q. D. & Prokopenko, M. A general framework for optimising cost-effectiveness of pandemic response under partial intervention measures. *Sci. Reports* **12**, DOI: 10.1038/s41598-022-23668-x (2022).

47. Bastani, H. *et al.* Efficient and targeted covid-19 border testing via reinforcement learning. *Nature* **599**, 108–113, DOI: 10.1038/s41586-021-04014-z (2021).

48. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

49. Dunn, W. N. *Public policy analysis* (Routledge, London, England, 2017), 6 edn.

50. Demir, T. & Miller, H. Policy communities. In *Handbook of Public Policy Analysis*, 137–147 (CRC Press, 2006).

51. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533, DOI: 10.1038/nature14236 (2015).

52. Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G. & Pineau, J. An introduction to deep reinforcement learning. DOI: 10.48550/ARXIV.1811.12560 (2018).

53. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **34**, 26–38, DOI: 10.1109/msp.2017.2743240 (2017).

54. Henderson, P. *et al.* Deep reinforcement learning that matters. *Proc. AAAI Conf. on Artif. Intell.* **32**, DOI: 10.1609/aaai.v32i1.11694 (2018).

55. Bakker, B. Reinforcement learning with long short-term memory. In Dietterich, T., Becker, S. & Ghahramani, Z. (eds.) *Advances in Neural Information Processing Systems*, vol. 14 (MIT Press, 2001).

56. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735 (1997).

57. HENS, N. *et al.* Seventy-five years of estimating the force of infection from current status data. *Epidemiology Infect.* **138**, 802–812, DOI: 10.1017/S0950268809990781 (2010).

58. Huber, P. J. Robust Estimation of a Location Parameter. *The Annals Math. Statistics* **35**, 73 – 101, DOI: 10.1214/aoms/1177703732 (1964).

59. Gao, F. & Han, L. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* **51**, 259–277, DOI: 10.1007/s10589-010-9329-3 (2010).

60. Alvarez, F., Argente, D. & Lippi, F. A simple planning problem for covid-19 lock-down, testing, and tracing. *Am. Econ. Rev. Insights* **3**, 367–82, DOI: 10.1257/aeri.20200201 (2021).

61. Lockdowns in sir models (code). https://benjaminmoll.com/wp-content/uploads/2020/05/SIR_lockdown.m. Accessed: 2023-12-26.

62. Mathieu, E. *et al.* Coronavirus pandemic (covid-19). *Our World Data* (2020). Https://ourworldindata.org/coronavirus.

63. Covid-19 vaccine launch in india. https://www.unicef.org/india/stories/covid-19-vaccine-launch-india. Accessed: 2024-01-12.

64. Press information bureau, government of india, prime minister's office, pm launches pan india rollout of covid-19 vaccination drive. https://pib.gov.in/Pressreleaseshare.aspx?PRID=1689021. Accessed: 2024-01-12.

65. Covid-19 vaccination in india. https://en.wikipedia.org/wiki/COVID-19_vaccination_in_India. Accessed: 2024-01-12.

66. Oecd system of composite leading indicators. https://www.oecd.org/sdd/41629509.pdf. Accessed: 2024-01-12.

67. Oecd system of composite leading indicators. https://www.oecd.org/sdd/leading-indicators/oecd-composite-leading-indicators-clis.htm. Accessed: 2024-01-12.

68. Aws deepracer. https://aws.amazon.com/deepracer/league/. Accessed: 2024-01-12.

69. Covid-19 coronavirus pandemic. https://www.worldometers.info/coronavirus/. Accessed: 2024-01-12.

70. Internet archive. https://archive.org. Accessed: 2024-01-12.

71. OECD. Main economic indicators - complete database. DOI: https://doi.org/https://doi.org/10.1787/data-00052-en (2015).

72. Raffin, A. *et al.* Stable-baselines3: Reliable reinforcement learning implementations. *J. Mach. Learn. Res.* **22**, 1–8 (2021).

73. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library (2019). 1912.01703.

74. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).

75. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9**, 10–20, DOI: 10.1109/MCSE.2007.58 (2007).