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

Maeghal Jain[1,*], **Ziya Uddin**[1], **and Wubshet Ibrahim**[2]

[1]SoET, BML Munjal University, Gurugram, Haryana, 122413, India
[2]Department of Mathematics, Ambo University, Ambo, Ethiopia
[*]e-mail: wubshet.ibrahim@ambou.edu.et

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 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[?,?].

Recently, the COVID-19 pandemic caused a havoc worldwide. Till date there have been 772 million cases and more than 6 million deaths[?]. The pandemic triggered the sharpest economic recession in modern history with a 3% decline, much worse than during the 2008-09 financial crisis[?]. 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[?]. 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[?].

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 policymakers[?]. The outbreak of COVID-19 has prompted epidemiologists to research on various aspects, including mobility control[?,?], vaccination strategies[?,?], non-pharmaceutical interventions (NPIs) like restricting population movements and gatherings, closing schools and businesses, and requiring masks indoors[?,?,?], and financial considerations[?]. 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[?,?,?,?,?,?,?,?,?]. 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[?,?,?]. 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[?,?].

Many economists have studied the effect of COVID-19 on the economy of nations[?,?,?,?,?]. 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[?], they achieved it mostly with very aggressive testing, contact tracing, and enforced quarantines[?,?]. 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[?].

In this paper we optimize the government policies regarding stringency as it controls both the spread of the disease and the economy. To model the epidemiological data[?] we use the simple SIR model without vital dynamics[?,?,?] as it is assumed that the timescale is small enough that it can be neglected[?]. By lesioning the model, as opposed to proposing a new mathematical model with more specialized compartments to more accurately represent the actual environment[?,?], we effectively model the disease progression. Our model (SIR with lockdown and time-varying vaccination rate) builds on the foundational SIR model, by accounting for the recovery reached through vaccination[?,?,?,?,?] and the effects of lockdown[?,?,?,?,?]. Although the traditional SIR model is a valuable tool for understanding the spread of infectious diseases, it assumes that parameters like the transmission rate (β) and recovery rate (γ) are constant over time, which may not always be the case. In this paper, we propose a more sophisticated approach by introducing a time-dependent SIR model[?], enabling us to account for the changing dynamics of the pandemic due to factors such as lockdowns and vaccination rates. This proposed model effectively addresses the real-world conditions acts as a solution that is both effective and extendable. 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[?]. Secondly, 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, and different values for β and γ – keeping them the same scopes the environment for succumbing to wishful thinking which can be potentially dangerous. Therefore, before an actual deployment of the model, it would be a good measure to introduce stochasticity to these parameters (β and γ) and the vaccination rate (ν).

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[?,?,?,?,?]. 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[?]. 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[?,?,?].

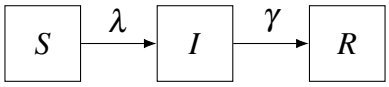
Additionally, since most modern reinforcement learning achievements are due to a combination of deep learning[?], in the following framework we make the use of this. Deep reinforcement learning is an advancement to reinforcement learning which helps normalize the input and reduce its dimensionality[?,?,?,?]. We use a Long Short Term Memory recurrent neural network for time-series data[?,?] and a simple fully connected network for data points that don't vary with time.

In summary, by using reinforcement learning augmented with deep learning techniques for the SIR with lockdown and time-varying vaccination rate 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 this paper, we use a compartmental model to model infectious disease environment. We iteratively develop this model, starting with the foundational SIR 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



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

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

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

Here, λ is the force of infection, it is the rate at which susceptible individuals acquire an infectious disease[?]. 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 there is a decline 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[?].

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

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[?].

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 **????** to calibrate the model with the use of huber loss[?]. Our model typically uses **??**, considering all three compartments: susceptible, infected, and recovered. However, there are instances where we need to balance modeling all groups and focusing on the infected group (population that drives disease spread). In such cases, we consider losses from both **????**. A weighted sum of the loss functions allows for trade-offs between comprehensive modeling and focusing on infected group dynamics (see **??**, where a hyperparameter (window length) is selected keeping both these losses in mind).

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

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 **??–??**, we minimize the cost function **??** using the Nelder-Mead method[?] to estimate the parameters like β, γ, and fit the model to actual data. The following parameters and loss is obtained:

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

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

$$R_0 = \frac{\beta_{optimal}}{\gamma_{optimal}} = 1.762 \tag{12}$$

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

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

See **??** 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[?,?,?,?,?]. 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)[?]. 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 **??** using the Nelder-Mead method, we get the following parameters and loss:

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

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

$$\begin{aligned} R_0 &= \frac{\beta_{optimal}}{\gamma_{optimal}}(1 - s(t)) \\ \overline{R_0} &= 1.693 \quad \text{(Mean)} \\ \widetilde{R_0} &= 1.624 \quad \text{(Median)} \\ \text{Mode}(R_0) &= 0.804 \quad \text{(Mode)} \\ \sigma_{R_0} &= 0.786 \quad \text{(Standard Deviation)} \\ R_0 &\in [0.16467, 3.0497] \quad \text{(Range)} \end{aligned} \tag{20}$$

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

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

See **??** to see how the model compares with the actual data.

2.3 SIR Model with Lockdown and Vaccination

Lastly, an additional flow from the susceptible to 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 **??** using the Nelder-Mead method:

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

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

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

$$\begin{aligned} R_0 &= \frac{\beta_{optimal}}{\gamma_{optimal}}(1 - s(t)) \\ \overline{R_0} &= 1.691 \\ \widetilde{R_0} &= 1.623 \\ \text{Mode}(R_0) &= 0.803 \\ \sigma_{R_0} &= 0.785 \\ R_0 &\in [0.165, 3.047] \end{aligned} \tag{29}$$

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

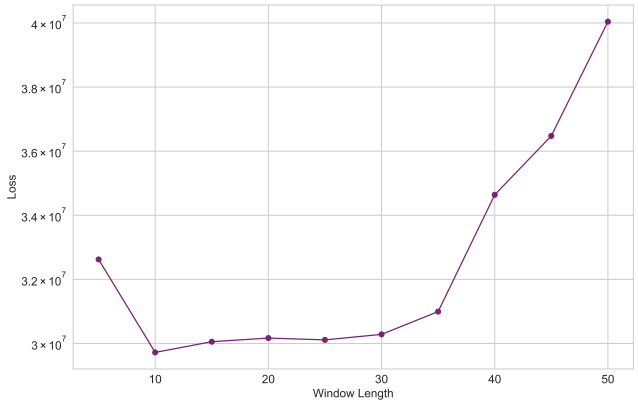
$$\text{loss\_I} = 10840360.995 \tag{31}$$

2.4 Optimizing Window Length for Time-varying Vaccination Rate

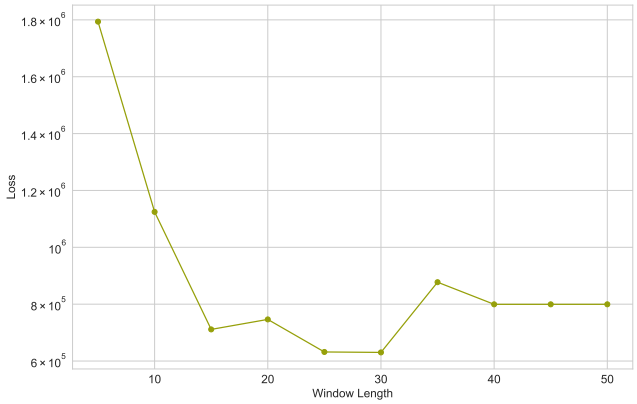
However, as observed by the value of $\nu_{optimal}$ from **??** which is almost negligible and the overestimation of infected individuals in **??** suggests that ν might be varying with time. This suggests to accurately estimate the infected population a time-varying vaccination rate should be used as the transition from susceptibility to direct recovery fluctuates with time[?,?]. Therefore, using the $\beta_{optimal}$ and $\gamma_{optimal}$ from **??** and **??**, we first find the optimal window length[?] for which the value of ν is constant and results in the least loss from **????**. Using different window lengths (window_length $= 5, 10, 15 \ldots 40, 45, 50$days), we estimate value of ν given a sub-interval of $[start, start + window\_length]$ going through the entire data, i.e., $\nu_{optimal}$ is constant only for a specific time window. For example, if the entire data period is 100 days, and the window length is 10 days, then ν is estimated for days $1-10$, then for days $11-20$, and so on. 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.

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

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



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



(b) Loss for Different Window Lengths for Infected Population

Figure 1. Loss for Different Window Lengths. We try different window lengths to find the optimal loss for both cases, either when predicting all three populations(susceptible, infected, recovered) or just the infected population.

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

The results in **??**, indicate that a window length of 10 days yields the least overall loss for all three population groups. However, this window length results in a poor approximation for just the infected group, which is crucial for accurately modelling the spread of the disease. Consequently, we have decided to use a window length of 15 days, which provides a more accurate approximation for the infected population while still maintaining reasonable loss for the other groups.

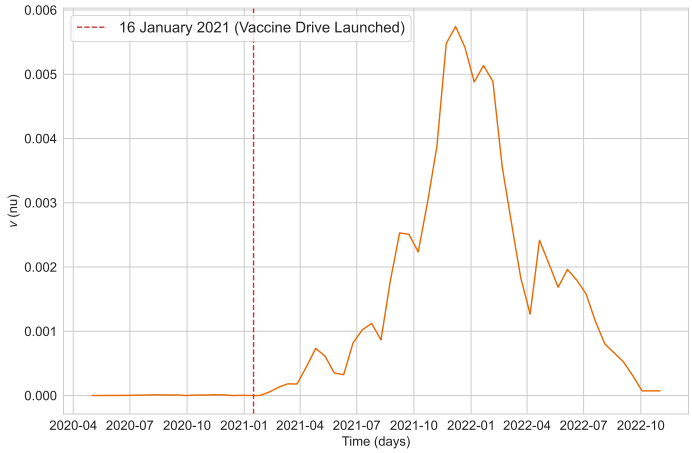


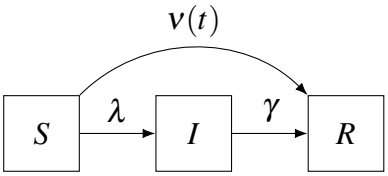
Figure 2. ν **Varying with Time.** This depicts how the vaccination rate (ν) changes over time and highlights the introduction of the vaccination campaign in India.

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

See
fig:SIR$_{m}odel_{w}ith_{l}ockdown_{w}ith_{v}accination_{I}ND_{p}arenttoseehowthemodelcompareswiththeactualdata$.

2.5 SIR Model with Lockdown and Time-varying Vaccination Rate

Finally, we integrate the time-varying vaccination rate (ν) into the SIR model that includes lockdown measures, resulting in the following set of equations, which represents our final model:



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

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

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

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

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

$$\begin{aligned} \overline{\nu_{optimal}} &= 0.001 \\ \widetilde{\nu_{optimal}} &= 0.001 \\ \text{Mode}(\nu_{optimal}) &= 0.000 \\ \sigma_{\nu_{optimal}} &= 0.002 \\ \nu_{optimal} &\in [0.000, 0.006] \end{aligned} \tag{40}$$

$$\begin{aligned} R_0 &= \frac{\beta_{optimal}}{\gamma_{optimal}}(1 - s(t)) \\ \overline{R_0} &= 1.546 \\ \widetilde{R_0} &= 1.483 \\ \text{Mode}(R_0) &= 0.734 \\ \sigma_{R_0} &= 0.718 \\ R_0 &\in [0.150, 2.785] \end{aligned} \tag{41}$$

$$\text{loss\_SIR} = 29116762.926 \tag{42}$$

$$\text{loss\_I} = 658537.443 \tag{43}$$

See **??** to see how the model compares with the actual data and **??** to see how the different models compare against each other.

2.6 Modelling Normalized GDP with Stringency

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)[?,?]. To do this a polynomial equation of the third degree is fitted to the data points $f(x) = ax^3 + bx^2 + cx + d = y$, here, x is the stringency (s), and y the normalized GDP, and we minimize the squared error to find the values of coefficients a, b, c, d. For India after fitting a 3 degree polynomial, the following equation is obtained:

$$\begin{aligned}\text{normalized\_GDP} = -5.96640236 \times 10^{-5} s^3 + 6.65064332 \times 10^{-3} s^2 - 2.23109924 \times 10^{-1} s \\ +1.01357226 \times 10^{2}\end{aligned} \tag{44}$$

2.7 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 known as a Markov decision process (MDP) and is characterized by the Markov property. To possess the Markov property is to create a compact state signal that retains all relevant information from past sensations without requiring the complete history. The Markov property ensures that the probability of transitioning to the next state and receiving a reward depends only on the current state and action, without requiring the entire history[?]. Our MDP is defined as follows:

- Set of States $\mathcal{S}$: The state of the environment are described through the descriptors like the normalized GDP, R_e, a list of all the previous actions (in changing the stringency) and the proportion of the population that was susceptible (S), infected (I) and recovered (R). 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 61.96505, with a standard deviation of 17.66983. The minimum value was 31.48, while the maximum value reached 96.3. And the differences between two consecutive stringencies had a mean of -0.070919, and standard deviation of 1.42715, with the minimum being -14.36 and maximum 16.67. 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 $\mathcal{T}(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ 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 $\mathcal{R}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$. The agent observes the state of the environment $\mathbf{s}_t$ at time t and takes an action $\mathbf{a}_t$, after which the state transitions to $\mathbf{s}_{t+1}$ and the agent receives a reward $\mathbf{r}_{t+1}$ as feedback. In **??** we define a reward strategy however it should be noted that this work serves as a framework where the strategy can be easily swapped for another 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.

Given that at each timestep t the agent has to choose an action a_t to maximize the reward r_{t+1} a policy is formulized by the agent. The policy π is a mapping from states to a probability distribution over actions: $\pi : \mathcal{S} \to p(\mathcal{A} = \mathbf{a} \mid \mathcal{S})$. Reinforcement learning methods specify how the agent changes its policy as a result of its experience. 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 $R = \sum_{t=0}^{t=T-1} \gamma^t r_{t+1}$. 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 $s_t \in \mathscr{S}$ to a predefined set of actions to increase or decrease the stringency at time t, which are the actions $a_t \in \mathscr{A}$. Each entry of the Q-table $(Q_t(s_t, a_t))$ associates an action in the finite sequence $\left(\mathscr{A}_j\right)_{j \in \mathbb{J}^+}$ to a state of the finite sequence $(\mathscr{S}_i)_{i \in \mathbb{I}^+}$**?**.

In this case of epidemic control by non-pharmaceutical interventions (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_t. 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 $J(r_t) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{(t-1)} r_t\right]$, where, discount factor $\gamma \in [0,1]$ (in our case $\gamma = 0.99$) and time steps $k = 1, 2, \ldots$.

2.7.1 Defining the Reward Function

The stringency index emerges as a critical factor influencing both the normalized 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.

In reinforcement learning, 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. While designing a reward function it is important to note that the rewards we set up truly indicate what we want accomplished. In particular, the reward signal is not the place to impart to the agent prior knowledge about how to achieve what we want it to do**?**. Taking inspiration from similar work**?**, we define the reward function.

The reward function is parameterized to account for key factors influencing decision-making. To incentivize reduction of R_e (effective reproductive number) and the increase of the normalized GDP after R_e is below 1.5. The reward is defined as follows:

$$\text{Reward} = \begin{cases} -20 \times R_e & \text{if } R_e > 1.5 \\ 100 \times \text{min\_max\_normalized\_GDP} & \text{if } 1.25 \leq R_e \leq 1.5 \\ 200 \times \text{min\_max\_normalized\_GDP} & \text{if } R_e < 1.25 \end{cases}$$

This reward function is parameterized to account for key factors influencing decision-making. When the effective reproductive number (R_e) exceeds 1.5, indicating a high transmission rate of the disease, the reward is negatively impacted to incentivize a reduction in R_e. As R_e decreases within the range $1.25 \leq R_e \leq 1.5$, indicating a moderate transmission rate, the reward is directly proportional to the normalized GDP, reflecting the importance of both controlling the spread of the disease and maintaining economic stability. Notably, when R_e drops below 1, signaling a declining transmission rate and potential containment of the disease, the reward function shifts focus towards economic recovery. In this scenario, the reward incentivizes an increase in the normalized GDP, emphasizing the need to stimulate economic activity and promote recovery efforts following successful control measures.

Additionally, if the proportion of the infected population were to rise above 0.003 (peak in the actual data) the model is punished (-2000) and otherwise rewarded (50). To reward not changing the stringencies frequently, we reward the absolute different between the previous stringency and the current stringency negatively ($|s(t) - s(t-1)| \times -12$).

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**?**. Therefore, this research act as a framework for promoting the development of more efficient reward strategies for the same.

2.7.2 Deep Reinforcement Learning and Training

The agent observes the percentage of the population that is susceptible, infected, recovered, and time-varying data like the GDP, and previous actions taken to change the stringency, and R_e. Since Stable Baselines3 can support multiple inputs (time-series data, single data points and images) by using Dict Gym space. For data that varies with time (stringency, normalized GDP, R_e) we use a simple Long Short Term Memory (LSTM) architecture**?**. 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 and some of the best results are presented.

3 Results

Using the simple SIR model from **??**–**??**, 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 which can create complications. This is because disease dynamics are controlled by this population and our work involves rewarding the agent when the proportion of infected individuals falls below a predetermined threshold. Therefore, an overestimation of the infected population could lead to incorrect decision-making and undesirable outcomes.

Combining the lockdown dynamics in the SIR model using **??**–**??**, 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.

Incorporating vaccination dynamics into the SIR model with lockdown measures, as described by **??**–**??**, we get the following **??**. Here, because the value of ν **??** is negligible, it doesn't change the results

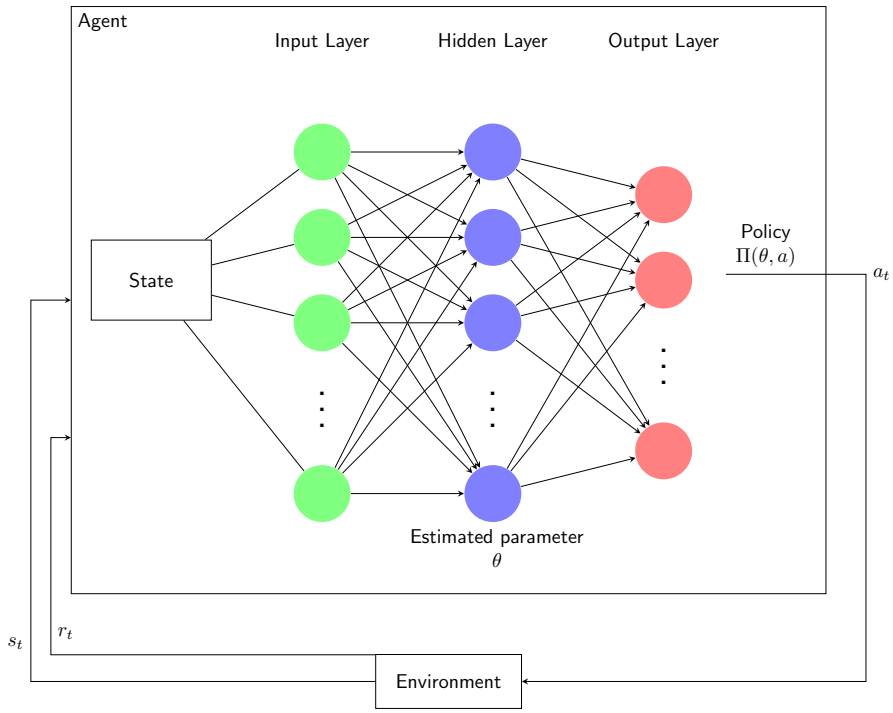


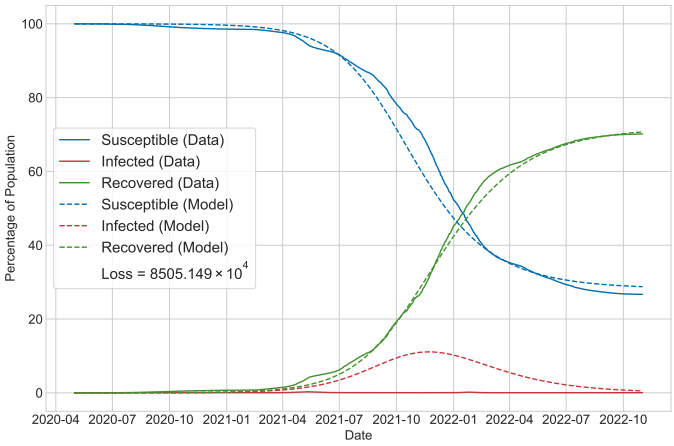
Figure 3. Deep Reinforcement Learning. Deep learning algorithms used in reinforcement learning enables more complex decision-making.

significantly compared to the previous model (**??**–**??** and **??**). Therefore, a time-varying ν shall be able to better account for the these dynamics.

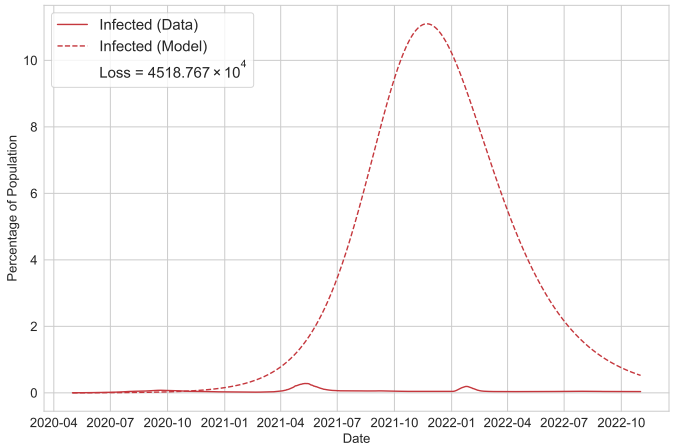
For SIR model with lockdown and time-varying vaccination rate from **??**–**??**, we get the following **??**. With a time-varying ν (vaccination rate) and the effect of lockdown, our model is able to account for the infected individuals and reduce the cost in comparison to all the previously formalized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

While non-pharmaceutical interventions (NPIs) can effectively manage the epidemic, they impose economic burdens on developing nations. In **??**, we plot the normalized GDP against the stringency and calculate various metrics like the Pearson correlation coefficient, coefficient of determination (r^2), and p-value for three countries (India, Mexico, Brazil) which are Emerging Market and Developing Economies[?] from May 2020 to October 2022. It can be observed from **??** that strict policies have a negative effect on the normalized GDP in these economies. However, this trend is not uniformly seen in advanced economies like the USA, Japan, and Canada as shown in **??**. In these countries, other factors could be contributing to the decrease in normalized GDP besides the implementation of stricter policies.

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 after the April, 2020 **??**. This approach allows for the natural progression of disease dynamics, resulting in a rapid reduction of the effective reproduction number R_e to below 1.2 after October, 2020 (refer to figure **??**). After October, 2021 there's a decrease in the stringency which leads to an increase in the normalized GDP, indicating an

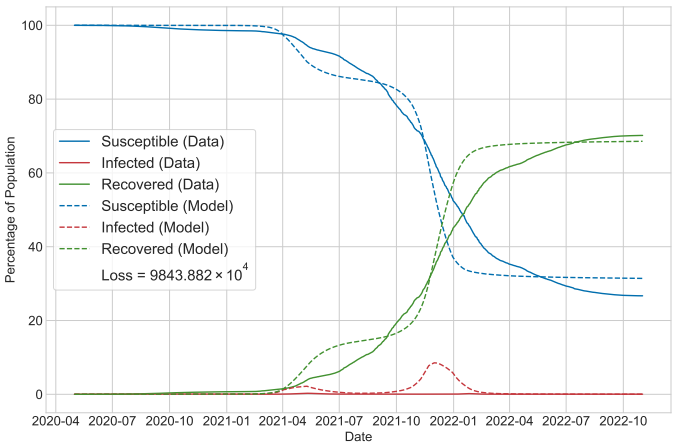


(a) SIR Model

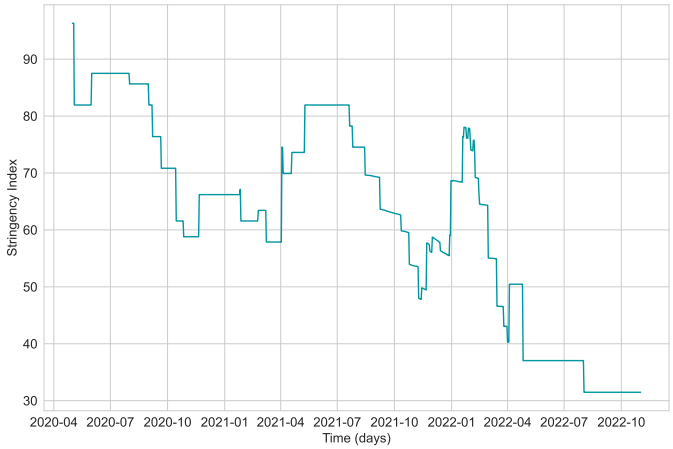


(b) Infections Modelled with SIR Model

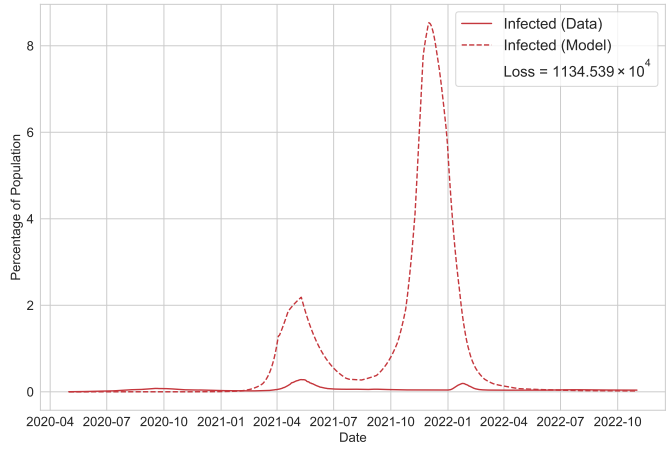
Figure 4. SIR Model Comparison for India. The figure presents a comparison between the fitted simple SIR model (**??–??**) and real data. Here an evident overestimation of the infected population is observed.



(a) SIR Model with Lockdown

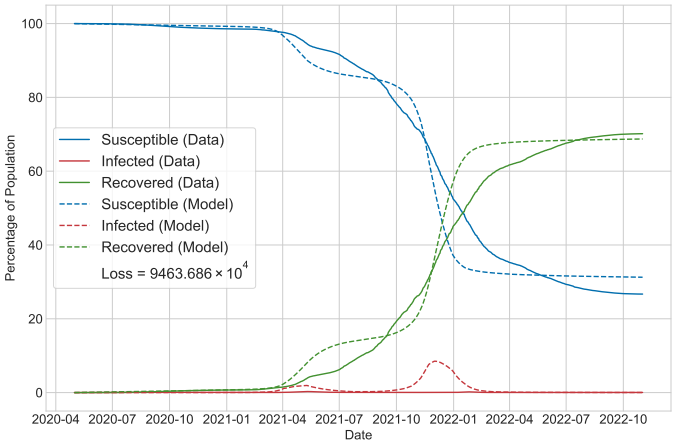


(b) Stringency Varying with Time

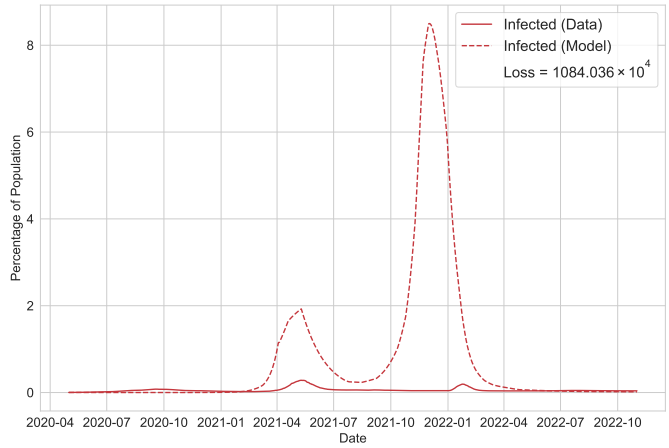


(c) Infections Modelled with SIR Model with Lockdown

Figure 5. SIR Model with Lockdown Analysis for India. This figure illustrates the fitting of the SIR model with lockdown (**??**–**??**) in comparison to real data. The introduction of lockdown measures showcases discernible effects on the dynamics of disease progression. While an overestimation persists, the model's peaks now closely align with the observed data and are able to capture key trends.

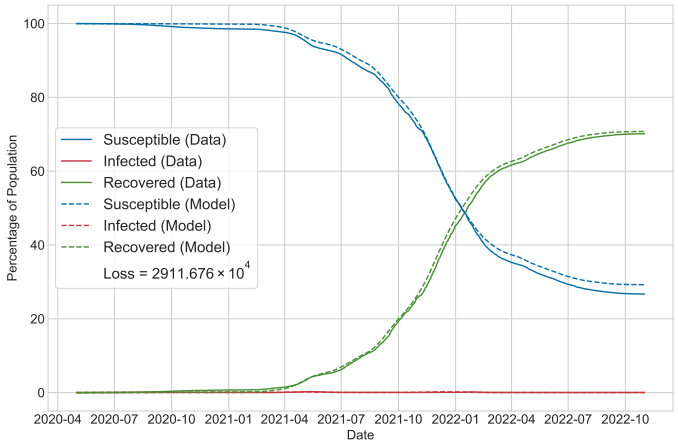


(a) SIR Model with Lockdown and Vaccination

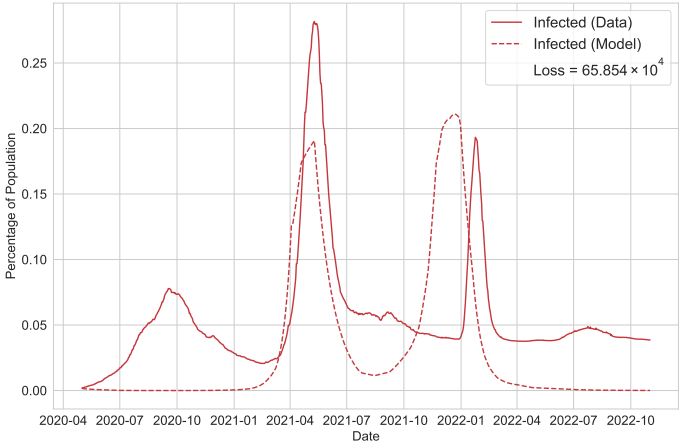


(b) Infections Modelled with SIR Model with Lockdown and Vaccination

Figure 6. SIR Model with Lockdown and Vaccination for India. This figure displays the fitting of the SIR model with lockdown (**??**–**??**) compared to the real data. The infection trends closely resemble those depicted by the SIR model with lockdown, as illustrated in **??** and this is because the rate of vaccination is negligible **??**. This is suggestive of a rate of vaccination that varies with time.

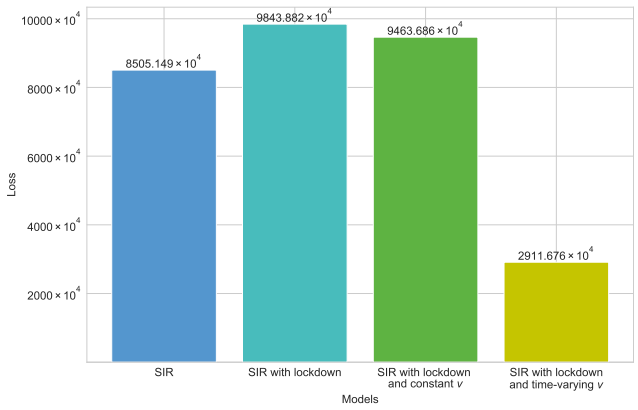


(a) SIR Model with Lockdown and Time-varying Vaccination Rate

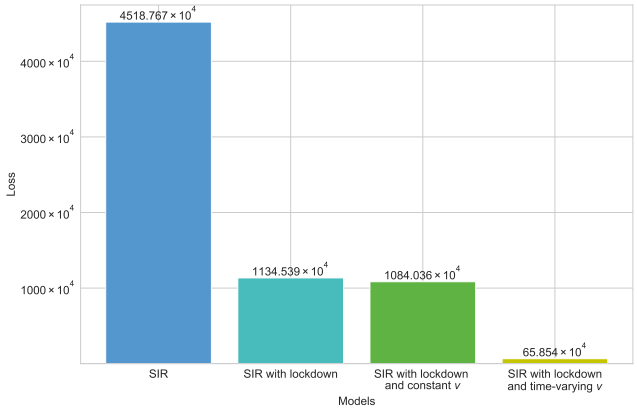


(b) Infections Modelled with SIR Model with Lockdown and Time-varying Vaccination Rate

Figure 7. SIR Model with Lockdown and Time-varying Vaccination Rate. This figure displays the fitting of the SIR model with lockdown and time-varying vaccination rate (**??**–**??**) compared to the real data. Incorporating a time-varying vaccination rate enhances the model's ability to capture variations in the infected population over time.



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



(b) Loss for Different Models for Infected Population

Figure 8. Loss for Different Models. Here, we can observe that the loss is the least for the SIR model with lockdown and time-varying vaccination rate.

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) as well as the later phase (second peak of infected individuals **??**), but it proves to be more beneficial for the nation's economy in the long run. Despite the economic benefits in the long run, this strategy is not the most effective for the government to adopt due to the high number of infected individuals. Therefore, we propose an alternative strategy that involves some loss in terms of the economic impact.

In contrast, an alternative output is presented in **??**, which demonstrates a gradual increase in stringency from October, 2020 till April, 2021. This strategy results in a decline of infections occurring prior to the vaccine's release, and a subsequent cessation of new infections. This approach, however, has implications for the normalized GDP, as seen by the observed decline in **??**. While both these approaches (**????**) outperform the actual strategy, they underscore the complexity of managing public health crises and the need for careful strategic planning to balance health outcomes with economic considerations.

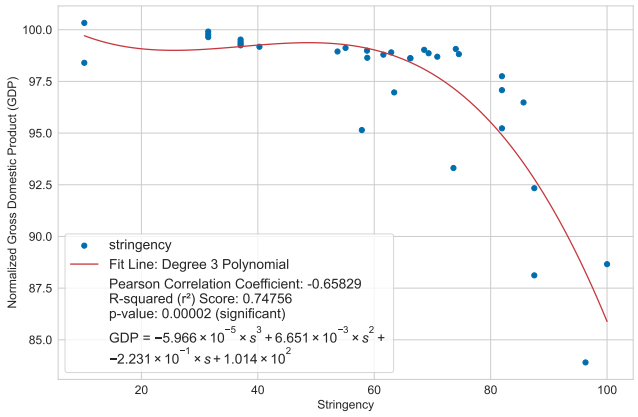
4 Discussion

The paper seeks to inspire epidemiologists by highlighting the advancements achieved through the application of reinforcement learning in policymaking during the pandemic. We introduce a virtual environment that closely simulates a pandemic scenario and thoroughly explore innovative strategies 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 that will help humanity better defend against potential pandemic crises in the future.

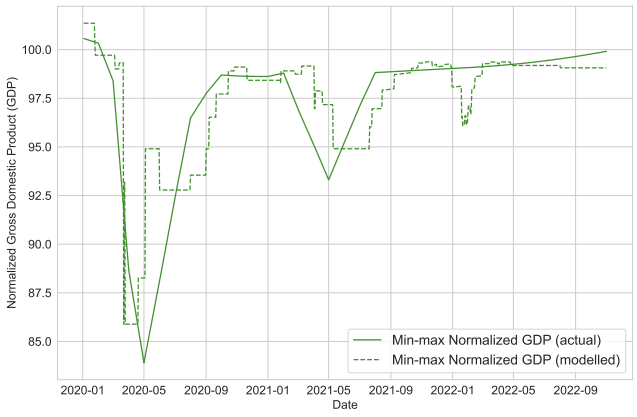
5 Experiment Settings

5.1 Dataset

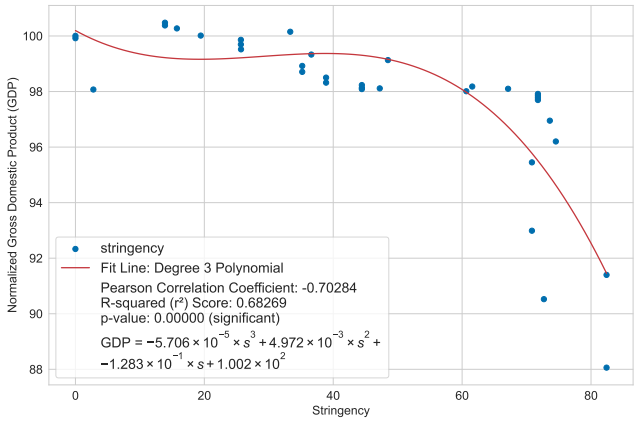
The population-level epidemiological data can be obtained from the "Our World In Data COVID-19" dataset: https://ourworldindata.org/coronavirus or more specifically: https://github.com/owid/covid-19-d



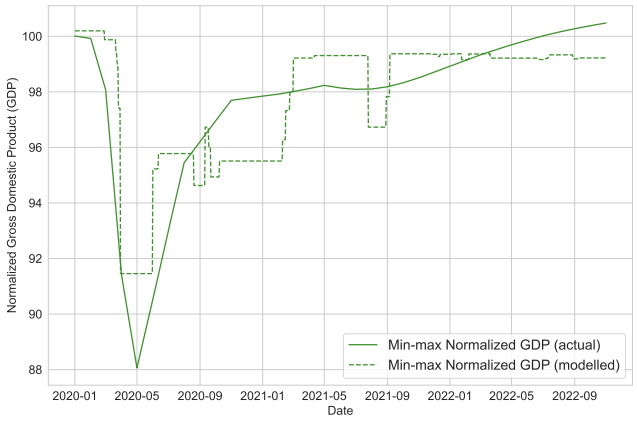
(a) Stringency and Normalized GDP for India



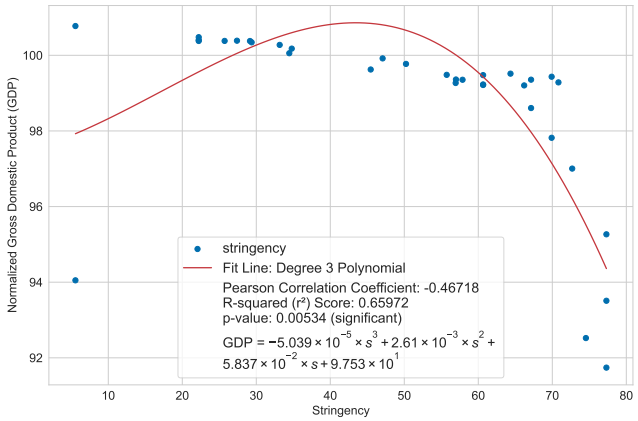
(b) Normalized GDP modelled with Stringency for India



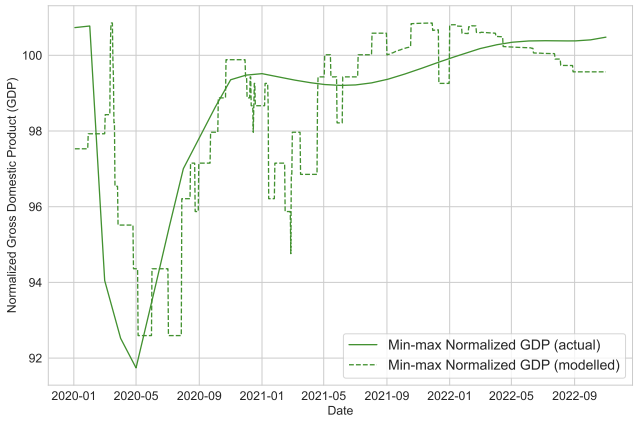
(c) Stringency and Normalized GDP for Mexico



(d) Normalized GDP modelled with Stringency for Mexico

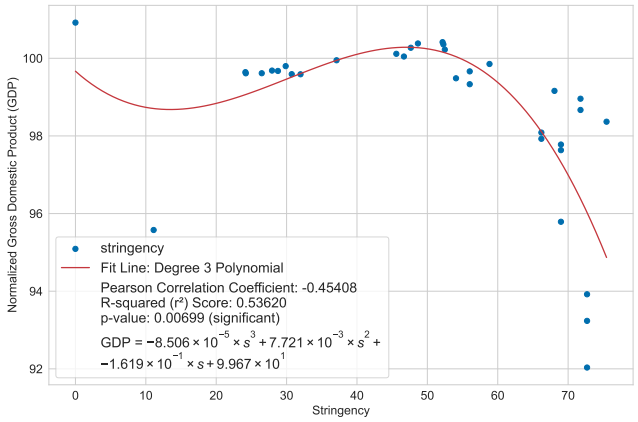


(e) Stringency and Normalized GDP for Brazil

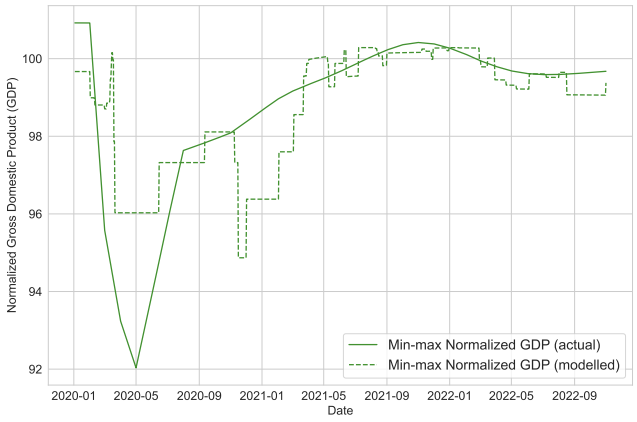


(f) Normalized GDP modelled with Stringency for Brazil

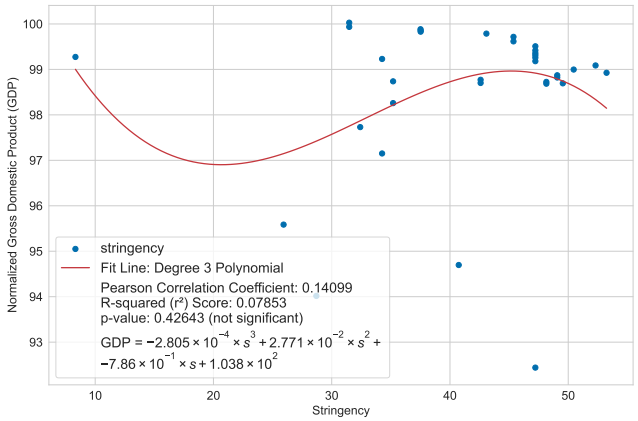
Figure 9. Stringency and GDP for Developing Economies. Here, "(actual)" is the real data, "(modelled)" is the model for normalized GDP given the stringency. For countries with developing economies, when we model the normalized GDP with stringency we see significant p-values and high r2 scores.



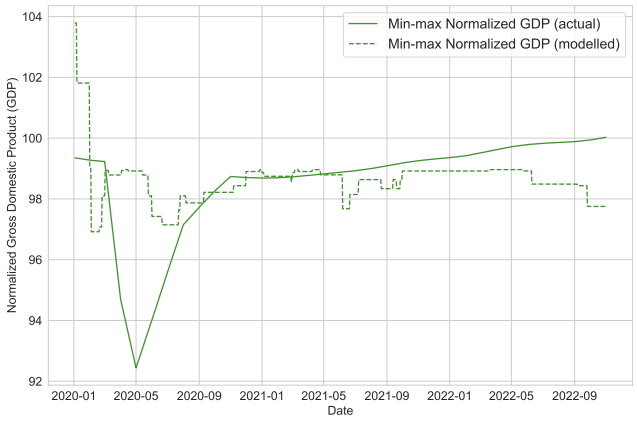
(a) Stringency and Normalized GDP for United States



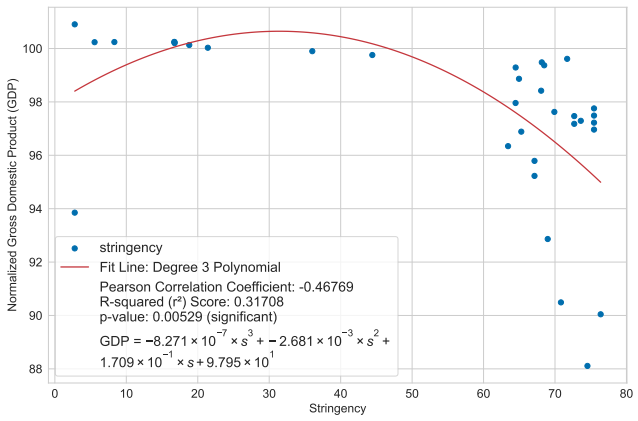
(b) Normalized GDP modelled with Stringency for United States



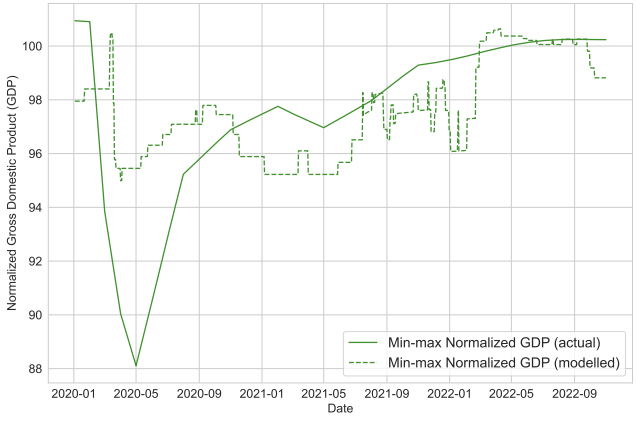
(c) Stringency and Normalized GDP for Japan



(d) Normalized GDP modelled with Stringency for Japan

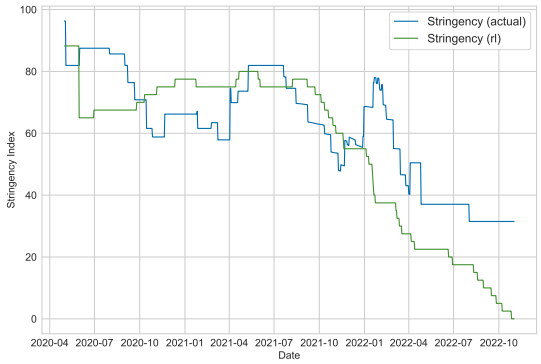


(e) Stringency and Normalized GDP for Canada

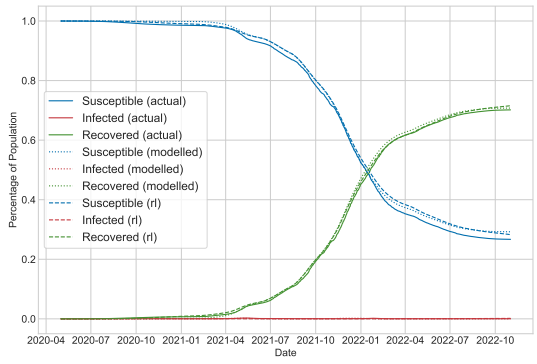


(f) Normalized GDP modelled with Stringency for Canada

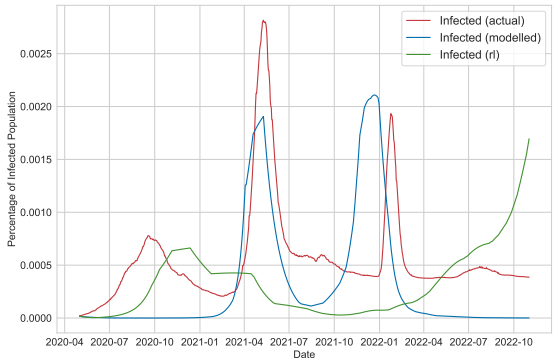
Figure 10. Stringency and GDP for Advanced Economies. Here, "(actual)" is the real data, "(modelled)" is the model for normalized GDP given the stringency. In economically advanced nations, when modeling the normalized GDP against stringency measures, we observe substantial p-values, providing evidence against the null hypothesis. However, the significance levels are not as high as those found for developing economies (**??**–**??**). This is reflected in the lower R-squared scores, which indicates that the relationship between these variables may be less pronounced when compared to developing economies.



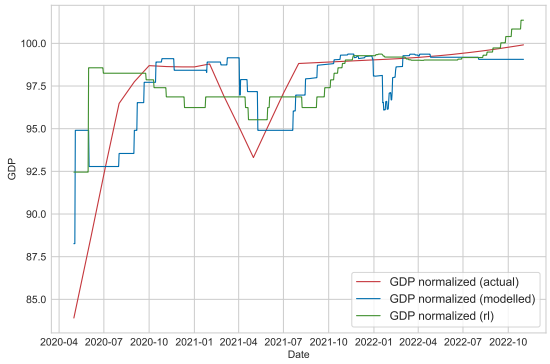
(a) Stringency changing over Time



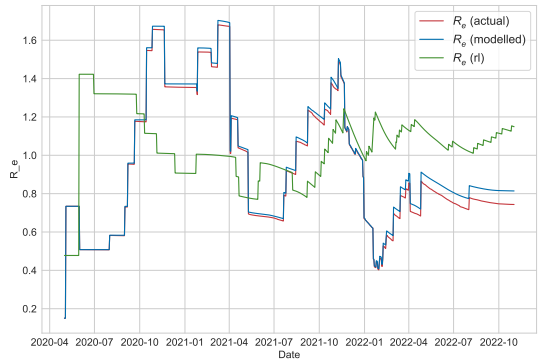
(b) SIR Dynamics



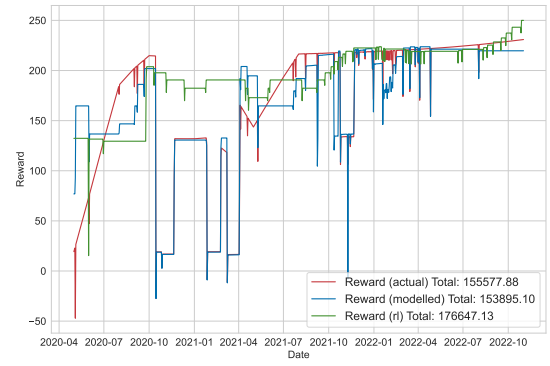
(c) Infected Population changing over Time



(d) Normalized GDP changing over Time

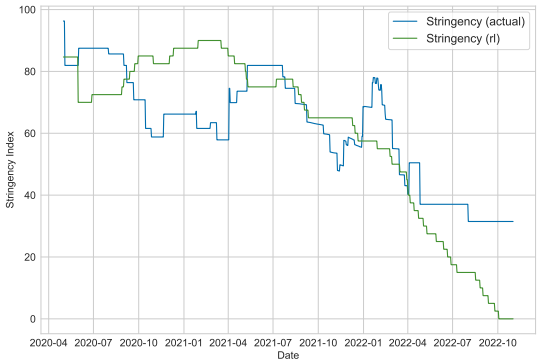


(e) R_e changing over Time

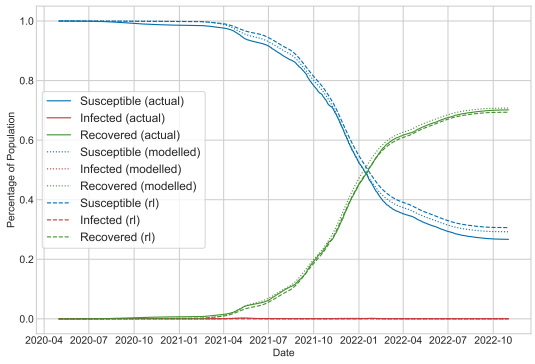


(f) Reward changing over Time

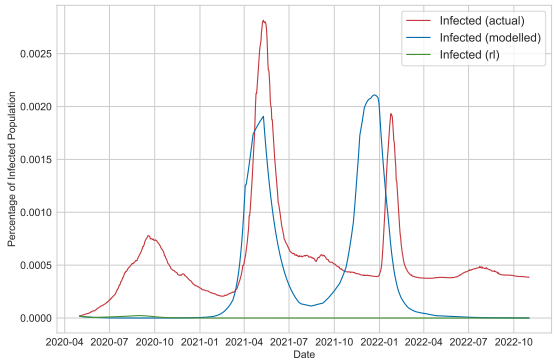
Figure 11. Strategy from Reinforcement Learning Agent. Here, "(actual)" is the real data, "(modelled)" is the result from real world stringency imposed with the use of the SIR model with lockdown and time-varying vaccination rate, and "(rl)" is the new stringency strategy we propose. **(a)** The new strategy proposed highlights a decrease in stringency from July, 2020 till October, 2020 compared to the actual data. After October, 2020 there's an increase and then a steady decline towards the end. **(c)** There's a peak in the number of infected people around October, 2020 and then a second peak after October, 2022 **(d)** The normalized GDP is also maintained and doesn't show a dip during April, 2022. **(e)** The R_e is maintained below 1.5 throughout, and below 1.2 after October, 2020. **(f)** A higher reward is achieved by the reinforcement learning agent.



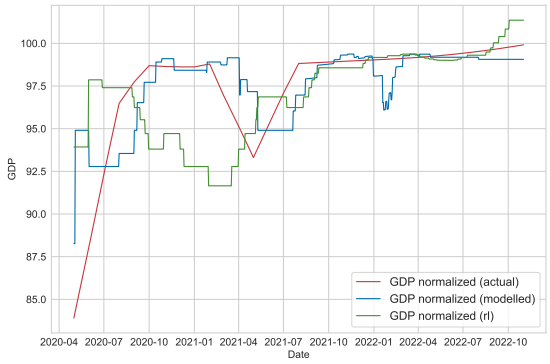
(a) Stringency changing over Time



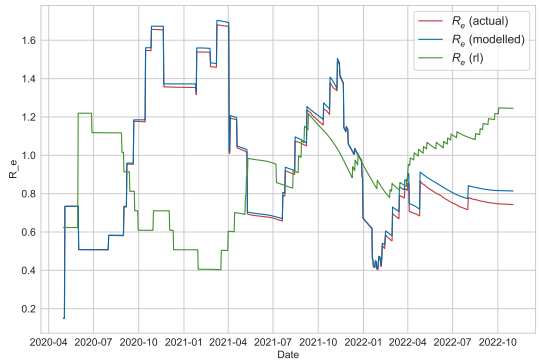
(b) SIR Dynamics



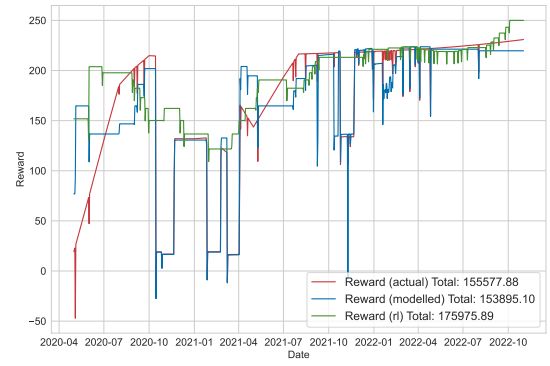
(c) Infected Population changing over Time



(d) Normalized GDP changing over Time



(e) R_e changing over Time



(f) Reward changing over Time

Figure 12. Strategy from Reinforcement Learning Agent. Here, "(actual)" is the real data, "(modelled)" is the result from real world stringency imposed with the use of the SIR model with lockdown and time-varying vaccination rate, and "(rl)" is the new stringency strategy we propose. **(a)** The new strategy proposed highlights an increase in stringency from October, 2020 till April, 2021 compared to the actual data and then a steady decline towards the end. **(c)** No sharp peaks in the infected population is observed. **(d)** The normalized GDP is affected by the increase in the stringency from October, 2020 till April, 2021. **(e)** The R_e is maintained below 1.2 throughout. **(f)** A higher reward is achieved by the reinforcement learning agent.

ata/blob/master/public/data/owid-covid-data.csv[?]. Data for the total cases, and recovered was acquired by scraping the Worldometers website[?] used Internet Archive[?]. 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[?].

5.2 Code

All code and data will be made open source upon acceptance of the paper.

References

1. Baker, R. E. *et al.* Infectious disease in an era of global change. *Nat. Rev. Microbiol.* **20**, 193–205 (2022).
2. Tan, M. K. Covid-19 in an inequitable world: the last, the lost and the least (2021).
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.
5. Nicola, M. *et al.* The socio-economic implications of the coronavirus pandemic (covid-19): A review. *Int. journal surgery* **78**, 185–193 (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 (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 (2020).
8. Song, S., Liu, X., Li, Y. & Yu, Y. Pandemic policy assessment by artificial intelligence. *Sci. Reports* **12**, 13843 (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 (2020).
10. Nguyen, T. *et al.* Covid-19 vaccine strategies for aotearoa new zealand: a mathematical modelling study. *The Lancet Reg. Heal. Pac.* **15** (2021).
11. Kim, D., Keskinocak, P., Pekgün, P. & Yildirim, I. The balancing role of distribution speed against varying efficacy levels of covid-19 vaccines under variants. *Sci. reports* **12**, 7493 (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. Glob. Heal.* **12** (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 Reg. Heal. Pac.* **14** (2021).
14. Ferguson, N. M. *et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*, vol. 16 (Imperial College London London, 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 Glob. Heal.* **11**, e1964–e1977 (2023).
16. Hollingsworth, T. D., Klinkenberg, D., Heesterbeek, H. & Anderson, R. M. Mitigation strategies for pandemic influenza a: balancing conflicting policy objectives. *PLoS computational biology* **7**, e1001076 (2011).

17. Pangallo, M. *et al.* The unequal effects of the health–economy trade-off during the covid-19 pandemic. *Nat. Hum. Behav.* 1–12 (2023).

18. Ash, T., Bento, A. M., Kaffine, D., Rao, A. & Bento, A. I. Disease-economy trade-offs under alternative epidemic control strategies. *Nat. communications* **13**, 3319 (2022).

19. Ohi, A. Q., Mridha, M., Monowar, M. M. & Hamid, M. A. Exploring optimal control of epidemic spread using reinforcement learning. *Sci. reports* **10**, 22106 (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 (2021).

21. Alvarez, F., Argente, D. & Lippi, F. A simple planning problem for covid-19 lock-down, testing, and tracing. *Am. Econ. Rev. Insights* **3**, 367–382 (2021).

22. Lukasz, R. An analytical model of covid-19 lockdowns (2020).

23. Redlin, M. Differences in npi strategies against covid-19. *J. Regul. Econ.* **62**, 1–23 (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 health* **11** (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. *MedRxiv* (2021).

26. Gagnon, J. & Rose, A. How did korea's fiscal accounts fare during the covid-19 pandemic? *Peterson Inst. for Int. Econ. Policy Brief* 23–8 (2023).

27. Deb, P., Furceri, D., Ostry, J. D. & Tawk, N. The economic effects of covid-19 containment measures. (2020).

28. Eichenbaum, M. S., Rebelo, S. & Trabandt, M. The macroeconomics of epidemics. *The Rev. Financial Stud.* **34**, 5149–5187 (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 Reg. Heal. Pac.* **30** (2023).

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

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

32. Hethcote, H. W. Three basic epidemiological models. In *Applied mathematical ecology*, 119–144 (Springer, 1989).

33. Hethcote, H. W. The basic epidemiology models: models, expressions for r0, parameter estimation, and applications. In *Mathematical understanding of infectious disease dynamics*, 1–61 (World Scientific, 2009).

34. Allen, L. J. A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infect. Dis. Model.* **2**, 128–142 (2017).

35. Cooper, I., Mondal, A. & Antonopoulos, C. G. A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals* **139**, 110057 (2020).

36. Bjørnstad, O. N., Shea, K., Krzywinski, M. & Altman, N. The seirs model for infectious disease dynamics. *Nat. methods* **17**, 557–559 (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**, 352 (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**, 15688 (2022).

39. Maurício de Carvalho, J. P. & Rodrigues, A. A. Sir model with vaccination: bifurcation analysis. *Qual. theory dynamical systems* **22**, 105 (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 (2017).

41. Turkyilmazoglu, M. An extended epidemic model with vaccination: Weak-immune sirvi. *Phys. A: Stat. Mech. its Appl.* **598**, 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. Epidemiol. Glob. Heal.* **15**, 101052 (2022).

43. Hale, T. *et al.* A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nat. human behaviour* **5**, 529–538 (2021).

44. Lockdowns in sir models. (2020).

45. Atkeson, A. What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios. Tech. Rep., National Bureau of Economic Research (2020).

46. 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 network science engineering* **7**, 3279–3294 (2020).

47. Bajra, U. Q., Aliu, F., Aver, B. & Čadež, S. Covid-19 pandemic–related policy stringency and economic decline: was it really inevitable? *Econ. research-Ekonomska istraživanja* **36**, 499–515 (2023).

48. Cilloni, L. *et al.* The potential impact of the covid-19 pandemic on the tuberculosis epidemic a modelling analysis. *EClinicalMedicine* **28** (2020).

49. Arinaminpathy, N. & Dye, C. Health in financial crises: economic recession and tuberculosis in central and eastern europe. *J. Royal Soc. Interface* **7**, 1559–1569 (2010).

50. Nguyen, Q. D. & Prokopenko, M. A general framework for optimising cost-effectiveness of pandemic response under partial intervention measures. *Sci. Reports* **12**, 19482 (2022).

51. Bastani, H. *et al.* Efficient and targeted covid-19 border testing via reinforcement learning. *Nature* **599**, 108–113 (2021).

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

53. Dunn, W. N. *Public policy analysis* (routledge, 2015).

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

55. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *nature* **518**, 529–533 (2015).

56. Francois-Lavet, V. *et al.* An introduction to deep reinforcement learning. *Foundations Trends Mach. Learn.* **11**, 219–354 (2018).

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

58. Henderson, P. *et al.* Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32 (2018).

59. Bakker, B. Reinforcement learning with long short-term memory. *Adv. neural information processing systems* **14** (2001).

60. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

61. Hens, N. *et al.* Seventy-five years of estimating the force of infection from current status data. *Epidemiol. & Infect.* **138**, 802–812 (2010).

62. Massad, E. Ethical and transborder issues. In *Global Health Informatics*, 232–263 (Elsevier, 2017).

63. Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, 492–518 (Springer, 1992).

64. Gao, F. & Han, L. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* **51**, 259–277 (2012).

65. Alvarez, F., Argente, D. & Lippi, F. A simple planning problem for covid-19 lock-down, testing, and tracing. *Am. Econ. Rev. Insights* **3**, 367–382 (2021).

66. Lockdowns in sir models (code) (2020).

67. Mathieu, E. *et al.* Coronavirus pandemic (covid-19). *Our world data* (2020).

68. Liao, Z., Lan, P., Liao, Z., Zhang, Y. & Liu, S. Tw-sir: time-window based sir for covid-19 forecasts. *Sci. reports* **10**, 22454 (2020).

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

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

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

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

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

74. OECD. Main economic indicators - complete database. (2015).