Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 23(3): 753–775.
DOI: 10.3934/mbe.2026029
Received: 31 July 2025
Revised: 13 January 2026
Accepted: 20 January 2026
Published: 06 February 2026

Research article

Real-time responses to epidemics: A Reinforcement-Learning approach

Gabriele Gemignani[1,2], Alberto d'Onofrio[3], Alberto Landi[1] Giulio Pisaneschi[1,*] and Piero Manfredi [4]

[1] Department of Information Engineering, University of Pisa, via G. Caruso 16, Pisa, 56122, Italy

[2] Department of Electrical and Information Engineering, Polytechnic of Bari, via Re David 200, Bari, 70125, Italy

[3] Department of Mathematics, Informatics and Geosciences, University of Trieste, via A. Valerio 12, Trieste, 34127, Italy

[4] Department of Economics and Management, University of Pisa, via C. Ridolfi 10, Pisa, 56124, Italy

* **Correspondence:** Email: giulio.pisaneschi@phd.unipi.it.

Abstract: Open-loop optimal control applied to epidemic outbreaks is a valuable tool to develop control principles and inform future preparedness guidelines. A drawback of this approach is its assumption of complete knowledge of both transmission dynamics and the effects of policy measures. As a result, such methods lack responsiveness to real-time conditions, since they do not integrate feedback from the evolving epidemic state. Overcoming this requires a closed-loop approach. We propose a novel closed-loop method for real-time social distancing responses using a general Reinforcement Learning (RL)-based decision-support framework. It enables adaptive management of social distancing policies during an epidemic, thereby balancing direct health costs (e.g., hospitalizations, deaths) with indirect (economic, social, psychological) costs from prolonged interventions. The framework builds on and compares with a COVID-19 model that was previously used for open-loop assessments, thereby capturing key disease characteristics like asymptomatic transmission, healthcare saturation, and quarantine. We test the framework by evaluating optimal real-time responses for a severe outbreak under varying priorities of indirect costs by public authorities. The full spectrum of policy strategies—elimination, suppression, and mitigation—emerges depending on the cost prioritization as a result of closed-loop adaptability. The framework supports timely, informed decisions by governments and health authorities during current or future pandemics.

Keywords: Real-time epidemic response; direct and indirect epidemic costs; epidemic modelling; closed-loop optimal control; Reinforcement Learning

1. Introduction

The COVID-19 pandemic brought a major challenge for global public health policy and preparedness science to the forefront: how to effectively balance the enforcement and duration of control interventions against the wide-ranging societal costs they incur. This tension between the direct and indirect costs of an outbreak became dramatic during the first pandemic wave in early 2020, when governments worldwide implemented strict lockdowns in an attempt to curb the rapid spread of the virus [1, 2]. While these interventions were crucial to mitigate immediate health risks, such as surging hospital admissions, increasing mortality rates, and the potential collapse of overburdened healthcare infrastructures, they also caused substantial side consequences.

The societal and economic toll of such large-scale restrictions was dramatic, giving rise to disruptions across multiple domains. Beyond the direct impact on public health, prolonged lockdowns contributed to rising levels of economic instability, increased mental health disorders, and widened pre-existing social and economic disparities. Furthermore, extended periods of isolation strained interpersonal relationships and placed psychological and relational stress on individuals and communities [2].

These realities have underscored the urgent need to better understand and model the intricate trade-offs involved in epidemic response strategies. Policymakers were faced with a delicate balancing act: on one side, the imperative to contain viral transmission and preserve healthcare system functionality; and on the other, the necessity to minimize the adverse societal consequences of restrictive measures. Designing effective response strategies not only requires real-time epidemiological data availability, but also requires analytical tools capable of evaluating the interactions between health outcomes and social resilience.

In the last decades, mathematical transmission models have become critical tools in the design of public health policies [3], thanks to their ability to describing, understanding, and forecasting the spread of infectious diseases. This role has been dramatically emphasised since the emergence of the COVID-19 pandemic. A diverse range of modeling frameworks has been proposed, varying in scope, mathematical structure, and the level of abstraction of biological and epidemiological details. These models have been tailored to address different research questions—primarily estimating transmission rates to evaluating the strength and duration of policy interventions—and exhibit considerable variation in terms of complexity, from rather simple compartmental structures to highly detailed stochastic individual-based models.

One of the most critical lessons drawn from the pandemic is the inherent difficulty in constructing models that robustly and accurately capture the essential dynamics of a novel infectious disease. Basic compartmental formulations, such as the SIR (Susceptible-Infected-Recovered) and SEIR (Susceptible-Exposed-Infected-Recovered) models, while foundational, have shown substantial limitations when confronted with the complex, multifactorial, and rapidly evolving nature of the COVID-19 crisis. Indeed, the latter showed an endless list of possible intervention actions, including vaccinations, a rapidly changing epidemiology, and the continued appearance of new virus strains, there was substantial disagreement on which overall policy option (i.e., elimination, suppression or mitigation, should be used to pursue and contrast the virus [4, 5]).

In light of these complexities, the scientific community has responded by proposing a broad spectrum of enhanced modeling approaches, thereby reflecting the growing awareness of the need to ac-

count for the dynamic interplay of biological, social, and policy-related factors during pandemics [6,7].

Prior to the COVID-19 pandemic, the application of the optimal control theory in the management of communicable diseases was relatively narrow in scope, typically focusing on predefined interventions such as vaccinations, without explicitly considering their broader systemic consequences. This ad-hoc approach, while useful in theoretical scenarios, often overlooked the indirect societal and infrastructural impacts of intervention strategies. For example, the strain placed on public healthcare systems, disparities in access to care, and behavioral feedback loops were frequently excluded from the modeling process [8–11].

The far-reaching consequences of the COVID-19 pandemic have renewed interest in the application of the optimal control theory to epidemiological decision-making. This renewed attention stems from the awareness that effective pandemic response strategies must grapple with the complex task of dynamically allocating and prioritizing a variety of non-pharmaceutical and pharmaceutical interventions, such as non-pharmaceutical measures, testing protocols, and vaccination rollouts, while considering trade-offs with the indirect (i.e., social, economical, psychological) costs of such measures [11–14]. The strategic coordination of response measures is critical to minimize both health-related and socio-economic costs over the course of an outbreak.

From a systems engineering perspective, the control of infectious disease models—particularly through the modulation of social behavior and mobility restrictions—has traditionally been approached using model-based optimization techniques. These methods, rooted in the classical control theory [15], typically generate a sequence of optimal control inputs based on a predefined model of the epidemic's dynamics. However, a significant limitation of this open-loop paradigm is its lack of a feedback mechanism: the computed control trajectory is only optimal under the assumption that the model accurately represents the evolving system dynamics over its entire horizon [11, 14, 16–19].

The aforementioned limitation does not remove the importance of open-loop approaches, which-as argued in detail in [11]- maintain a central role in the identification of broad 'apriori' control principles that are key for the design of epidemic preparedness guidelines. However, the absence of real-time adaptability severely restricts the practical utility of open-loop control to design and implement real-time responses to an ongoing epidemic challenge. Real-time responses face the intrinsic uncertainty of the underlying biological (e.g., virus structure), epidemiological, and behavioral (e.g., first of all, the agents' policy compliance) processes that are both highly nonlinear and subject to rapid change.

To address this limitation, one widely explored alternative is the implementation of receding-horizon strategies, most notably Model Predictive Control (MPC). In MPC, the optimization problem is repeatedly solved over a moving time window, thereby leveraging the most recent data to accurately update the predictions and revise the control actions [20–22]. While this approach introduces a degree of feedback and adaptability, it remains heavily reliant on the accuracy of the underlying model and often requires substantial computational resources, particularly in high-dimensional or nonlinear settings. Consequently, while MPC offers a more flexible framework than open-loop optimization, it still faces notable barriers to real-time deployment, particularly in the early stages of a pandemic, when data is sparse and the system's behavior is poorly understood.

Reinforcement Learning (RL), a key branch of artificial intelligence, has seen increasing interest in recent years, largely owing to its ability to uncover feature representations from high-dimensional, real-time data. By engaging in direct trial-and-error interactions with its environment, a neural network learns a closed-form control policy that naturally accommodates the model's uncertainties. In practice,

RL methods have demonstrated outstanding performance in the sequential control of dynamic systems, including video games [23], mobile robotics [24], autonomous driving [25], and even epidemiological modeling [26, 27].

Recently, RL - as a model-free approach suitable to tune multi-objective reward functions - has also attracted the interest of scholars in the field of epidemic control by non-pharmaceutical interventions [28]. For instance, [29] suggested that an RL-derived strategy that implements shorter but sustained lockdowns can be very effective in reducing mortality compared to a number of alternative control policies.

Other works frame epidemic control policies as sequential decision problems: [30] couples RL with Long-Short-Term-Memory (LSTM) forecasts to trade off health and economic objectives, and [31] treated staged interventions as learned policies across disease phases. Related contributions ([32], [33]) expanded the state/action space via metapopulation dynamics and resource-allocation models, thus posing practical challenges (larger state spaces, partial observability, and complex cost tradeoffs) for RL-based controllers.

In this article, we consider a previous framework [11, 34], which applied open-loop optimal control to a realistically parametrized model of the COVID-19 epidemic [7], to define the optimal preparedness principles, and reformulate it to investigate the shape of real-time epidemic responses that emerge from a RL, closed-loop approach.

2. Materials and methods

2.1. Epidemic model

The adopted transmission model follows the one in [11], which integrates the initially developed framework to analyze the first wave of COVID-19 in Italy [7], thus enhancing it by integrating the impact of social distancing interventions [12, 13], the constraints imposed by the finite hospital capacity, and the differential role of asymptomatic and symptomatic recovered individuals.

It tracks ten state variables: S (susceptible), E (exposed), P (presymptomatic), I (infected), A (asymptomatic), H (hospitalization demand), Q (quarantined), R_1 (recovered symptomatic), R_2 (recovered asymptomatic), and D (deceased), each of them $\in \mathbb{R}^+$ in the range $[0, 1]$. A standard population that has a total size equal to one at time $t = 0$ is adopted. The social distancing control L takes four discrete values, $L \in \{0, 0.25, 0.50, 0.75\}$, thus indicating the proportion of individuals isolated by policy: a higher L reduces the transmission and eases healthcare demand but increases the economic costs. Finally, when the hospitalized population H surpasses the threshold $H_{\max}$, we define a compartment to capture those who cannot be admitted to the hospital and consequently suffer a higher fatality rate; this new class is denoted as U (untreated).

The dynamics are governed by the following system of Ordinary Differential Equations (ODEs), augmented by the algebraic relation (2.11), which accounts for untreated people. (see flowchart in Figure 1):

$$\dot{S} = -\lambda S\,(1-\theta L)^2, \tag{2.1}$$

$$\dot{E} = \lambda S\,(1-\theta L)^2 - \delta_E E, \tag{2.2}$$

$$\dot{P} = \delta_E E - \delta_P P, \tag{2.3}$$

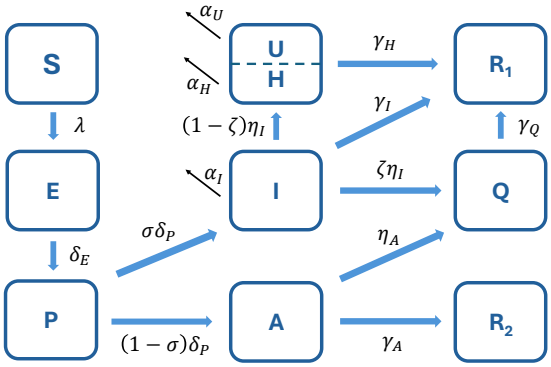


Figure 1. Compartment model flowchart.

$$\dot{I} = \sigma\,\delta_P\,P - (\eta_I + \gamma_I + \alpha_I)\,I, \tag{2.4}$$

$$\dot{A} = (1-\sigma)\,\delta_P\,P - (\gamma_A + \eta_A)\,A, \tag{2.5}$$

$$\dot{H} = (1-\zeta)\,\eta_I\,I - (\gamma_H + \alpha_H)\,\min(H, H_{\max}) - \alpha_U\,U, \tag{2.6}$$

$$\dot{Q} = \zeta\,\eta_I\,I + \eta_A\,A - \gamma_Q\,Q, \tag{2.7}$$

$$\dot{R}_1 = \gamma_I\,I + \gamma_H\,H + \gamma_Q\,Q, \tag{2.8}$$

$$\dot{R}_2 = \gamma_A\,A, \tag{2.9}$$

$$\dot{D} = \alpha_I\,I + \alpha_H\,\min(H, H_{\max}) + \alpha_U\,U, \tag{2.10}$$

$$U = \max(0,\, H - H_{\max}). \tag{2.11}$$

An explanation of the meaning of each model parameter and their literature sources is reported in Table 1. Absent of intervention measures, the force of infection λ — the per-susceptible rate of acquiring infection from presymptomatic (P), symptomatic (I), and asymptomatic (A) individuals — is expressed as follows:

$$\lambda = \frac{\beta_P\,P + \beta_I\,I + \beta_A\,A}{S + E + P + I + A + R_1 + R_2}, \tag{2.12}$$

where the denominator represents the pool of socially active individuals, excluding H and Q compartments. Deaths are accounted for among the H, U, and I groups.

Initial conditions mirror an outbreak initialized by a few exposed individuals (in a proportion of $E_0 = 10/N_{pop}$) in an otherwise fully susceptible population.

2.2. *Optimal control problem*

The optimal control framework [11, 45] seeks the social distancing trajectory that minimizes the aggregate cost C, subject to the dynamical constraints given by (2.1-2.10) with (2.11), (2.12), and the initial conditions. This total cost is a convex combination of indirect societal losses C_{Ind} and direct health-related expenses C_{Dir}:

Table 1. Model and cost parameters.

Parameter	Value	Units	Description	Source
		Model Parameters		
β_P	2.7272	week^{-1}	Pre-symptomatic transmission rate	[35–37]
β_I	4.2987	week^{-1}	Symptomatic transmission rate	[7,38]
β_A	1.9362	week^{-1}	Asymptomatic transmission rate	[7,36,37]
δ_E	2.1084	week^{-1}	Latency rate	[7,39]
δ_P	3.7234	week^{-1}	Post-latency rate	[7]
σ	0.25	–	Probability to manifest symptoms	[7,40]
η_I	1.7284	week^{-1}	Detection rate of symptomatic	[7]
η_A	0.8642	week^{-1}	Detection rate of asymptomatic	[7]
ζ	0.40	–	Probability of being hospitalized	[7,40]
γ_A	0.9779	week^{-1}	Recovery rate of asymptomatic	[7,41]
γ_H	0.4889	week^{-1}	Recovery rate of hospitalized	[7]
γ_I	0.4889	week^{-1}	Recovery rate of symptomatic	[7]
γ_Q	0.4889	week^{-1}	Recovery rate of quarantine	[7]
α_I	0.2888	week^{-1}	Death rate of symptomatic	[7]
α_H	0.2888	week^{-1}	Death rate of hospitalized	[7]
α_U	1.1552	week^{-1}	Death rate of untreated	[11]
θ	0.70	–	Adherence to social distancing	
N_{pop}	$60 \cdot 10^6$	–	Italian population	
$H_{\max}$	$1.95 \cdot 10^5/N_{pop}$	–	Maximum hospital capacity over population	[42]
		Cost Parameters		
ω	625.00	\$	Average weekly wage	[11]
Λ	[0,1]	–	Preference to indirect costs	
$1/l$	20	years	Life years lost per death	[12,43]
α_1	2275.20	\$	Hospitalization cost per patient	[44]
p_y	0.151	–	Fraction of under-65 in deceased	[42]

$$C = \Lambda\, C_{Ind} \;+\; (1-\Lambda)\, C_{Dir}, \tag{2.13}$$

where the weighting factor $\Lambda \in [0,1]$ reflects the policymakers' relative emphasis on economic impacts versus health outcomes, and is consistent with previous studies [11, 14].

The indirect cost C_{Ind} is defined as the per-capita Gross domestic product (GDP) shortfall attributable to distancing measures, and follows the formulation in [12]:

$$C_{Ind} = \omega \int_0^T \Big[L((1-D) - (W_Q + R_1)) \;+\; W_Q \Big]\, dt, \tag{2.14}$$

where $W_Q = Q + H$ is the non-working fraction in quarantine or hospitalized, ω denotes the average daily wage, $1 - D$ is the living population, and the control horizon T spans approximately one year.

The direct health cost C_{Dir} accounts for hospitalization and mortality burdens as in [11]:

$$C_{Dir} = \int_0^T \left[\alpha_1 \, \min(H, H_{\max}) \; + \; p_y \, \frac{365\,\omega}{l} \, \dot{D}\right] dt, \tag{2.15}$$

where α_1 is the average daily cost per hospitalized patient, p_y is the proportion of the working-age population, and $1/l$ represents the mean number of life-year lost per COVID-19 fatality.

A comprehensive catalogue of all model parameters and cost-function coefficients—most of which are sourced from Italian COVID-19 data studies [7, 11]—can be found in Table 1.

3. Reinforcement Learning-based optimization

An open-loop, continuous control optimization of the problem described in Section 2.2 is provided and thoroughly assessed in [11]. This initial approach enables validation of the epidemic model by analyzing how different social restrictions impact the population. If the model accurately reflects real-world dynamics, the resulting optimal controller can support effective preparedness to such crises. Because there is no feedback, an optimization process should be run every time the real-world phenomenon significantly deviates from the modeled behavior, which is likely to occur even with highly accurate models.

This section describes the theoretical and implementation-related aspects of RL, which nowadays stands as one consolidated closed-loop, learning-based control technique for problems of various kinds [46]. This choice notably enforces dynamic adaptability to various epidemic scenarios, as well as intrinsic stability to changes in the model's parameters.

3.1. Markov Decision Process design

RL is based on the framework of Markov Decision Processes (MDPs), which provide a mathematical model for stochastic sequential decision-making [47]. An MDP is described by the 5-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the set of states with $s \in \mathcal{S}$, $\mathcal{A}$ is the set of actions with $a \in \mathcal{A}$, $\mathcal{P}(s_{k+1} \mid s_k, a_k)$ are the transition probabilities, which reduce to a deterministic dynamics in the fully deterministic case, $\mathcal{R}(s_k, a_k, s_{k+1})$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor.

Our MDP design follows the scheme depicted in Figure 2. The environment is represented by the compartment model described in Section 2, and is numerically integrated over weekly time-steps, which, considering an optimization horizon of one-year in the integrals (2.14) and (2.15), gives an episode length of $N = 52$. The state space consists of the state vector of the ODEs, passed as input to the policy-maker, thus constituting the feedback path of the loop. The agent represents the policy maker—specifically, the public health authority that determines the weekly epidemic response. A feedforward neural network is used to approximate the input–output mapping and generate weekly updates of the policy. The neural agent outputs the social distancing level L, which is then applied in the subsequent integration of the compartmental model. This procedure is cyclically repeated.

The reward function represents the objective to maximize during the training of the neural policy-maker; more specifically, the neural network learns to maximize the cumulative discounted reward, during each episode and at each time step, i.e.,

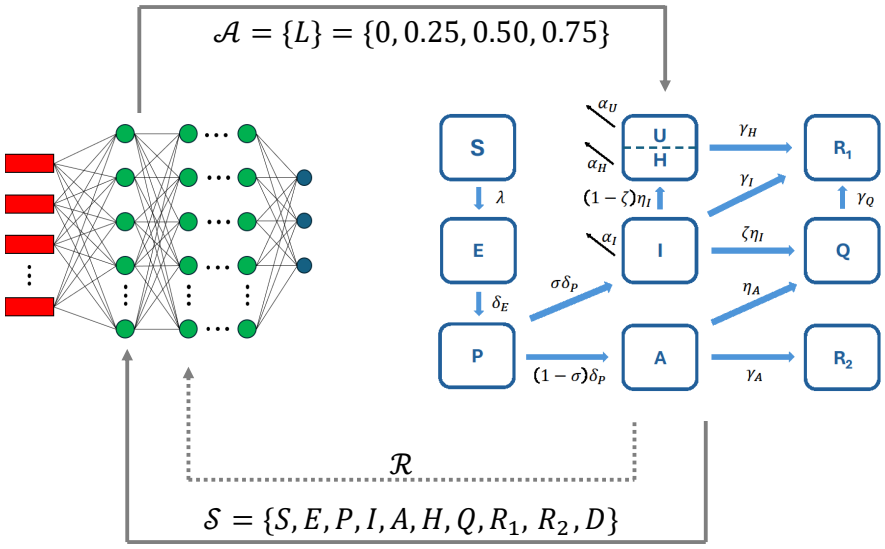


Figure 2. Markov Decision Process scheme. The reward function of our MDP is linked to the cost function definition in (2.14) and (2.15)

$$G_k = \sum_{n=0}^{N-k-1} \gamma^n \, r(s_{k+n}, a_{k+n}, s_{k+n+1}), \quad \forall\, k \in \{0, \dots, N-1\}. \tag{3.1}$$

When choosing values of the discount factor close to its upper bound 1, the long-term goal is prioritized over immediate rewards, and the policy-learning takes the whole optimization timespan into account, rather than the step-by-step achievements.

The cost functions in (2.14) and (2.15) are discretized (with sufficient accuracy, assuming moderate dynamics speed between consecutive time steps) by applying the forward-Euler approximation, with the time step equal to one week, obtaining the following:

$$\begin{aligned} C_{Ind} &= \sum_{n=0}^{N-1} c_{Ind}(s_n, a_n, s_{n+1}) = \sum_{n=0}^{N-1} \left(\omega L_n \left[(1 - D_n) - (W_{Q_n} + R_{1_n})\right] + \omega W_{Q_n}\right), \\ C_{Dir} &= \sum_{n=0}^{N-1} c_{Dir}(s_n, a_n, s_{n+1}) = \sum_{n=0}^{N-1} \left(\alpha_1 \min(H_n, H_{\max}) + p_y \frac{365\,\omega}{l} (D_{n+1} - D_n)\right). \end{aligned} \tag{3.2}$$

Now, it is evident that the similarity between the sums in (3.1) and (3.2) when $k = 0$ (that is, at the beginning of the control horizon) stands in the assumption that the reward discount factor is approximately one. At that point, following the cost equation (2.13), we take the reward as follows:

$$r = -\Lambda c_{Ind} - (1 - \Lambda) c_{Dir}. \tag{3.3}$$

By substituting (3.3) in (3.1) at the beginning of the outbreak ($k = 0$), the original, continuous optimal control problem becomes sequential and thus suitable to be solved with an RL algorithm.

3.2. Deep Q Network Learning

The neural network that formalizes the agent in the MDP goes through a training phase, which is guided by the reward function. Nowadays, there are many training algorithms that achieve this goal, and are based on the nature of the control (continuous vs discrete), characteristic architectures, and definitions of loss functions [47]. The choice is still heuristic and dependent of the reward function. In our case, the discrete control and the sparsity of the reward, namely the high frequency at which it is provided, make the problem suitable for the Deep-Q-Network (DQN) learning [48], which we briefly explain below.

The agent's policy is the generally probabilistic, state-to-action map; following a standard RL notation, it can be represented as follows:

$$\pi(a \mid s) = \mathbb{P}[A_k = a \mid S_k = s], \tag{3.4}$$

or in a particular case of deterministic environment, simply $\pi : \mathcal{S} \to \mathcal{A}$, in the same fashion of a closed-loop controller. In regards to the DQN, the policy is deterministic and "greedy", namely the highest-quality action is picked:

$$\pi_{\text{DQN}}(s) = \arg\max_{a \in \mathcal{A}} q(s, a), \tag{3.5}$$

where q is a *quality* measure that estimates how good it is in terms of the reward to be in a certain state s after applying the action a. In practice, to encourage exploration during training, an ϵ-greedy policy is applied: the neural network outputs $q(s, a)$ $\forall a \in \mathcal{A}$; then, the agent selects a random action with probability ϵ; or, otherwise, it follows the best policy defined in (3.5).

When using a neural network as the q-function approximator, the notation may include the dependency from the network weights ω, namely $q(s, a, \omega)$. The loss function is defined to minimize the Root Mean Square Error (RMSE) between the estimated and effective episodic reward, so that the policy efficiently interprets a state-action pair as advantageous or not and can choose the best action accordingly. In particular, we opted for the Double DQN (DDQN) algorithm that overcomes overestimation issues found in classical DQN [49]:

$$L(\omega) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{E}} \left[y - q(s, a; \omega) \right]^2 \tag{3.6}$$

with

$$y = r + \gamma q(s', \arg\max_{a'} q(s, a', \omega), \tilde{\omega}). \tag{3.7}$$

Note that the loss is averaged on a mini-batch of transitions of the kind (state, action, reward, next state) extracted from a memory buffer $\mathcal{E}$. This feature is present in offline DQN-like algorithms, in which past agent-environment interactions can influence the loss computation multiple times, before being discarded when the buffer maximum capacity is reached, and the older transitions are replaced by newer ones. Another key feature [49] lies in how the target output (3.7) is computed: it involves a forward pass through a separate, non-learning network with weights $\tilde{\omega}$. At a fixed frequency throughout the training, the weights of the main (learning) network are copied to this target network, thus providing a stable reference for the RMSE calculation. This approach helps to mitigate destabilizing oscillations that arise from constantly changing target values. Summing up, the training algorithm consists of the

following steps, repeated at each epoch. Initially, both the primary and target Q-networks are initialized with random weights, and a replay buffer of a fixed size is instantiated. At each time step, the agent observes the current state, selects an action according to an ϵ-greedy policy, and receives a reward along with the next state. This transition is stored in the replay buffer. Once a certain number of experiences have been collected, a mini-batch of data is randomly sampled from the buffer. Then, the loss function (3.6) is computed, and the network parameters are updated via backpropagation using a predefined *learning rate*. To improve the stability of the training, the weights of the target network are periodically updated to match those of the primary network.

3.3. Training validation

Our implementation follows a modular structure in which the environment, agent, and training algorithm are merged from separate code modules: the epidemic compartment model is coded following the Gymnasium Open AI library [50], the neural network is Pytorch-based, and the DDQN algorithm comes from the Tianshou framework [51]. After developing the whole architecture, a fine-tuning is necessary to achieve the desired training performance. The convergence of a machine learning application is in fact highly affected by tunable, non-trainable *Hyperparameters*, and this aspect is especially remarked in RL [52]. Moreover, our reward definition depends on the arbitrary weight Λ, which makes fine-tuning of the hyperparameters even more complex in trying to find a "sweet spot" (i.e., a set of hyperparameters that yields an acceptable performance for every Λ of interest). The results that we present below are obtained with the set reported in Table 2, which yielded the best results for various Λ values.

Table 2. DDQN Hyperparameters and Neural Network Structure

Hyperparameter	**Value**
Batch size	128
Replay buffer size	$5 \cdot 10^4$
Buffer warm-up (initial fill)	$5 \cdot 10^2$
Env. transitions per epoch	10
Discount factor γ	0.99
Target network update frequency	$5 \cdot 10^3$ epochs
Learning rate	$1 \cdot 10^{-5}$
Exploration factor ϵ	
Initial value	0.5
Exponential decay rate	$4 \cdot 10^5$ epochs
Final value	0.05
Neural Network Architecture	
Hidden layers	[128, 128]
Activation function	tanh
Backpropagation Optimizer	AdamW [53]

Once the training setup is finalized, a multi-seeds training is carried out in order to test our framework. This is a standard protocol in testing RL algorithms, which are inherently characterized by

several factors of randomness, such as the initialization of neural network parameters, the sampling of transitions from the replay buffer during updates, and the exploration mechanism driven by the ϵ-greedy policy.

4. Results

In this section, we report simulations carried out to validate RL as a social distancing policy optimizer. First, we give details on the training campaign and show how to assess the performance of a trained neural network.

4.1. Effectiveness of the trained network

We analyze the effectiveness of the trained networks to provide a closed-loop response to an ongoing outbreak, thereby addressing how the method coherently reacts to the prioritization of either the economic or epidemic costs.

Figure 3 shows metrics recorded from ten training runs for each value of Λ, where each run used a different random seed. All experiments were conducted on a system running Ubuntu 20.04.6, equipped with an AMD EPYC 7413 24-Core CPU, and four NVIDIA A100-80GB GPUs. Each training session required approximately 1.5 hours when utilizing a GPU.

The plotted data were gathered to test the network at a fixed frequency during training and used a greedy-only ($\epsilon = 0$) policy, until the termination of the episode (which, as already mentioned, occurs always at the 52^{nd} week). In particular, the left-hand plot displays an increasing trend in the episodic rewards for every Λ as proof of the agent's learning. A min-max normalization was done to equally scale the monitored metric for the various Λ, thus improving the readability of the graph. The true weighted and non-weighted costs, which are related to the non-normalized reward by (3.3), will be reported later. The graph on the right-hand side shows how the DDQN loss (3.6) decreases over time, once again validating the neural network training.

We set the training duration to 1 million epochs in this earlier stage to achieve a satisfactory convergence for every Λ. However, the RL-trained neural networks that were later evaluated on the epidemic model correspond to the best metric achieved during training; this is because, as it is evident for the normalized reward plot for $\Lambda = 0.6$, a fixed number of epochs may present large variability in the latest epochs and this is reflected in significant changes in the social distancing policy between networks trained on different seeds. By saving the weights of the best network tested during training, this problem is circumvented and reliable, well-posed social distancing policies can be later evaluated on the epidemic model.

4.2. Implications for real-time outbreak control: the role of costs prioritization

Based on the findings of Figure 3, we report our main results on the emerging shapes of the closed-loop, RL-based, social distancing policy for the model in Section 2.1. We focus on the control policies that result from different levels of parameter (Λ) that tune the prioritization attributed by policy-makers to indirect costs. The critical role of this parameter has already been discussed in previous studies [11, 14]. In the context of open-loop optimal social distancing problems, which are appropriate for preparedness analyses, [11] showed that, by setting Λ to vary over its $[0, 1]$ range, it is possible to

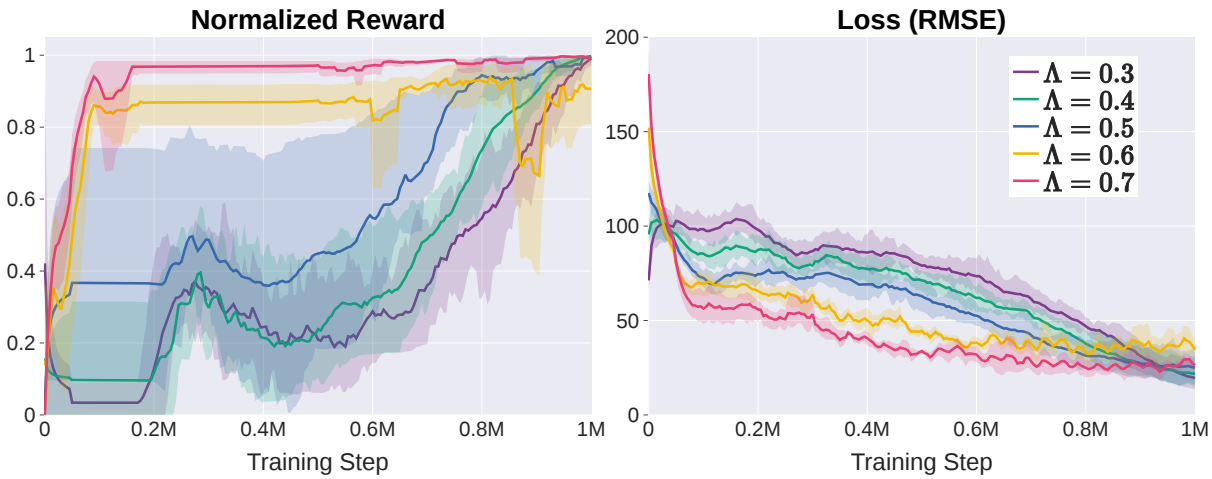


Figure 3. Training Metrics: min-max normalized episodic reward and DDQN loss for different economic/epidemic cost prioritization weight values. The solid line represents the sample mean computed over the random seeds, and the shaded band spans the 0.1–0.9 quantile interval (i.e. the 10th to 90th percentiles) across runs. The trend in these metrics suggests a proper learning of the social distancing policy.

robustly identify the entire spectrum of possible response actions to a severe epidemic, namely elimination (at low Λ values), suppression, mitigation, and doing nothing (at very high Λ values). We expect that such a range of response actions should also emerge, though in different forms, under a closed loop analysis to control an ongoing epidemic, similar to the one proposed here by the aid of RL tools.

Figure 4 reports the following outputs of epidemiological interest for different values of Λ that correspond to different prioritization of indirect costs: the temporal trends over the control horizon of (i) the selected social distancing action (first column), (ii) the demand for hospitalizations (second column), (iii) the weekly incidence of deaths (third column), and cumulative deaths (fourth column). Figure 5 provides additional details by comparing key epidemiological trends, namely the demand for hospitalizations and weekly deaths, for the different scenarios outlined by Λ.

The results reported in Figure 4 were obtained by executing the trained neural-network lockdown policy on the nominal pandemic model described in Section 2.1, under the same assumptions adopted during training. Specifically, the system was numerically integrated over the one-year optimization horizon from fixed initial conditions $E_0 = 10/N_{\text{pop}}$, with the remaining population being initially susceptible. No measurement noise, exogenous disturbances, or parameter/model uncertainties were considered in these simulations; thus, Figure 4 shows the performance under nominal conditions. Alternatively, Figure 5 assesses the robustness of the learned policy with respect to perturbations in the initial conditions.

When policymakers almost entirely target direct costs ($\Lambda = 0.05$, Figure 4, first row), the social distancing policy starts more than ten weeks after the outbreak ignition. This policy is capable of almost entirely preserving the susceptible population, with a negligible hospitalization demand and mortality, and therefore identifies an *elimination response*. This response is characterized by prolonged phases of harsh closures (at maximal intensity) intermitted with short-lasting reopening epochs, whose main purpose is that of weakening the pressure on indirect costs. Notably, this response dramatically

differs from the one resulting - other things being equal - from the open-loop solution [11], where elimination is achieved by acting early and aggressively for a prolonged phase, thus subsequently allowing completely release restrictions.

Increasing the level of priority to indirect costs to $\Lambda = 0.2$, Figure 4, second row, delays the response, which starts 15 weeks after outbreak ignition and, overall, makes it less intense. This allows for an initial epidemic wave, which is then effectively *suppressed*, thus keeping hospitalization demand well below the hospitals' capacity and achieving a declining incidence of fatalities over the entire horizon (see Figure 5). This outcome is achieved through a strong lockdown that lasts about one month and a half, followed by a rather regular sequence of intermittent epochs in which restrictions are switched between the two intermediate levels of closure intensities ($L = 0.25, L = 0.5$). These oscillations are necessary to balance the two components of costs. In other words, the underlying level of prioritization to indirect costs achieves suppression (and avoids further epidemic waves) by requiring that a certain degree level of social distancing is maintained for the majority of the control horizon.

Further increasing the prioritization to indirect costs ($\Lambda = 0.3$, Figure 4) maintains suppression; however, the regular pattern in the response found in the previous case becomes more fragmented. In particular, whilst the time at the control onset is almost the same, the maximum level of the initial harsh intervention is maintained for a shorter duration, and the subsequent epoch shows phases of complete relaxation of restrictions. This again complies with the need to put direct and indirect costs in a suitable balance. The overall epidemiological outcome clearly worsens compared to the previous case. Indeed, the intensity of suppression is weaker, and eventually results in epidemic activity at the reproduction level, with an almost constant incidence of hospitalizations and deaths over the second half of the control horizon (Figure 5). This, in turn, leads to the persistent linear growth of cumulative fatalities.

A further weakening of the prioritization on direct costs ($\Lambda = 0.4$, Figure 4) shifts the overall response from suppression to what we termed *effective mitigation* (see [11]). Effective mitigation describes a control action where epidemic reproduction is locked at threshold level when the hospitals capacity has been reached, therefore preventing hospitals from being overwhelmed. In fact, the adopted control policy suffers a short period, during the first phase of the response, during which the hospitals' capacity is surpassed and a certain number of individuals cannot be treated (see also Figure 5a). This "initial" failure essentially follows from the adaptive nature of the proposed control approach and from the weekly upgrade of the policy action, which are unable to fully mitigate the inertial growth of the epidemic once measures are set up. The intensity of the overall mitigation response is set at its lower level ($L = 0.25$) for most of the response epoch. Balancing costs forces a relaxation of measures in the second part of the horizon, which requires further restrictions to avoid the hospitals being overwhelmed. The measures are fully relieved when the decline in the susceptible population surpasses the herd immunity threshold. Notably, the (short-lasting) initial jump to the maximum control level is necessary to respond to the inertial growth of the epidemic that was bringing the hospitals out of control. Unlike previous cases, a significant mortality burden occurs (see also Figure 5b)).

Expectedly, further strengthening the importance attributed to indirect costs ($\Lambda = 0.5, 0.6, 0.7$, Figure 4) weakens the intensity of the response (notably, the time of onset of the response remains the same regardless of the value of Λ). This makes the mitigation regime more and more ineffective. Comparing the case $\Lambda = 0.5$ with the previous one ($\Lambda = 0.4$), the increased pressure to protect the indirect costs forces a relaxation of measures when the hospitals are already saturated. This brings

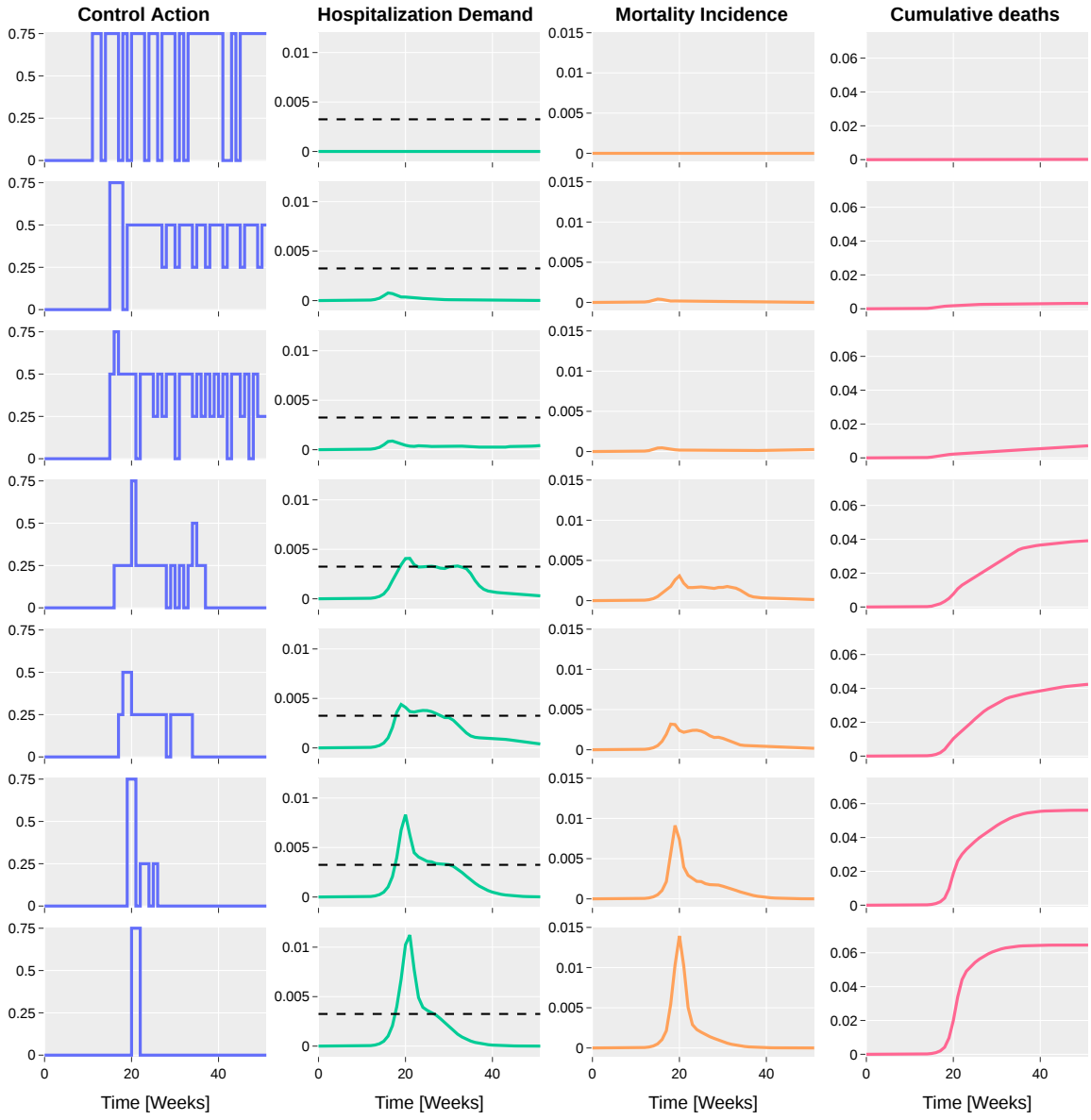
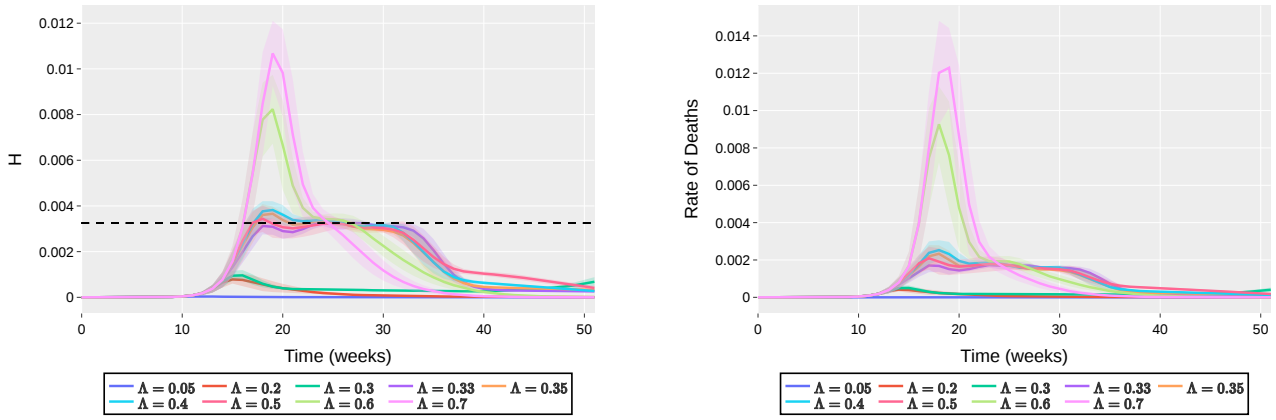


Figure 4. Closed-loop, RL based, social distancing responses to an ongoing epidemic (section 2.1), emerging for different levels of the prioritization to indirect costs Λ. Each row reports for a different Λ value (from top to bottom $\Lambda = \{0.05, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The temporal trend of the following outputs are shown: the social distancing action L (first column); the hospitalization demand H compared to the hospital saturation level (the dashed black line); when H exceeds its saturation level, the vertical difference $H - H_{max}$ identifies the size of the untreated U population; and the weekly incidence of deaths $\dot{D}$ (third column). Other epidemiological and cost parameters are described in Table 1.

the epidemic activity above the threshold, thus resulting in a second wave and therefore overwhelming the public health system, with a large mortality burden. For $\Lambda = 0.6$ (Figure 4), mitigation becomes largely ineffective. To cope with the increase in direct costs, a harsh but late and short lasting response



(a) **(b)**

Figure 5. a) Comparative temporal trends of the demand for hospitalization emerging under the closed-loop social distancing responses to an ongoing epidemic reported in Figure 4. Other epidemiological and cost parameters are described in Table 1. b) Temporal trends of deaths incidence under different values of Λ. The solid line represents the mean computed over a set of 25 different initial conditions for the exposed individuals, randomly chosen from a uniform distribution in the range $[1/N_{pop}, 100/N_{pop}]$. The shaded band spans the 0.1–0.9 quantile interval (i.e., the 10th to 90th percentiles).

is implemented only when the hospitals have been dramatically overwhelmed, where requests exceed the capacity threshold and a significant fraction of untreated individuals arise. For $\Lambda = 0.7$ (Figure 4), the response is short lasting and mostly palliative, with mild differences compared to the case of a free epidemic.

4.3. Comparing different policies

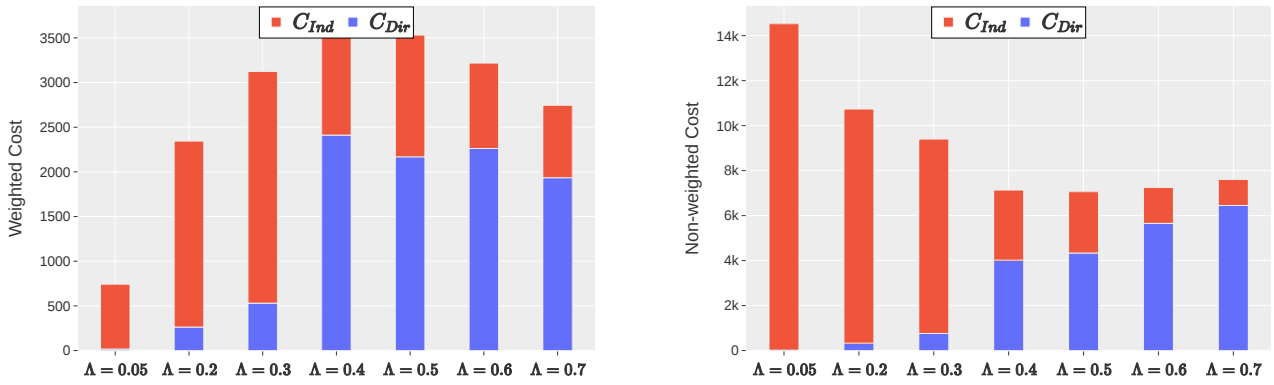
Figure 5 provides a more detailed comparative overview of the temporal trend in the demand for hospitalization and how this mirrors into the weekly pattern of mortality (for the same values of Λ). This summarizes our main findings and highlights the different responses and related epidemic outcomes for a set of 25 different initial conditions for the exposed individuals, which were randomly chosen from a uniform distribution in the range $[1/N_{pop}, 100/N_{pop}]$. When the preference for indirect costs is low (conversely: the prioritization to direct ones is high), the epidemic is either eliminated ($\Lambda = 0.05$) or suppressed ($\Lambda = 0.2, \Lambda = 0.3$) long before the hospital's admissions approach the saturation level. As Λ increases, the policy response switches in a quite abrupt manner (in [11], we termed this the 'razor blade' effect of epidemic responses; a finer Λ grid has been considered in the proximity of the switching point between suppression and mitigation) to an effective mitigation regime, where the full hospital's capacity is reached but is only slightly overwhelmed $\Lambda = 0.4$. Further increases in Λ increasingly weaken the effectiveness of the mitigation, thus resulting in major epidemics.

4.4. *Insights from the costs pattern*

Figure 6a reports direct and indirect costs, each of them weighted by the respective relative priority ($1-\Lambda$ and Λ, respectively), and the sum of both contribution as the total costs. Consistent with previous studies on open-loop control [11], weights heavily affect the policymaker's costs perceptions; this phenomenon is clearly evident from the case of very low preference to indirect costs when compared to the corresponding unweighted value (see Fig 6b). Seen through the lenses of policymaker preferences, the control action enacted in this case leads to the lowest weighted costs, and at the same time, to the highest unweighted costs. For Λ greater than 0.4, the direct component prevails, and is almost steady, due to the effect of a decreasing weight on direct costs. The pattern of unweighted costs reflects the real costs faced by policymakers: as shown in [11], suppressive solutions are characterized by the highest total costs, which are mainly driven by the indirect component. Instead, effective mitigation exhibits a balance between both components: the total costs are almost steady for higher Λ values, thus underscoring that solutions obtained in the ineffective mitigation regime are at least as expensive as those that effectively mitigate the disease spread.

5. Discussion

This work explored the applicability of RL techniques to the problem of real-time management of a threatening epidemic outbreak. With this aim, we translated a framework originally designed to inform preparedness guidelines through open-loop optimal control [11] into a closed-loop counterpart. A real-time response to an outbreak needs to systematically feedback from emerging measurements of the epidemic state to the upgrade of the currently adopted policy [20, 21, 54, 55]. Feedback is often able to intrinsically handle noise and uncertainties that affect measurements, and by consequence, epidemic state projections, thus practically achieving some degree of robustness. In particular, when the margin for effective mitigation is narrow, appropriately modulating the transmission may require adaptive policies, based on information on epidemic progression; this becomes especially relevant when a threshold on a scarce resource (e.g., ICU availability [54]) should not be overwhelmed. From these perspectives, RL offers a natural strategy for real-time responses [29–31, 34, 56].



(a) **(b)**

Figure 6. Weighted (a) and unweighted (b) costs for different values of Λ

Our RL-based closed loop analysis investigated the dependence of the epidemic response on the degree of prioritization on indirect costs (Λ). Our results show that as this parameter is varied, where the social distancing policy switches from near-elimination to suppression, to mitigation, and finally to near-inaction. Specifically, very low Λ values promote a strong response, which allows infection elimination and a very low epidemic burden; as Λ increases, the epidemic is first suppressed, thus allowing small scale peaks and avoiding the healthcare system to be overwhelmed, then effectively mitigated (hospitals saturated but never overwhelmed), and finally mitigated with progressively diminishing effectiveness until near-inaction.

Overall, our RL-based, closed-loop, control policies resulted in real time actions and epidemic outcomes that mirror, with all due differences, those obtained in the open-loop, a priori analysis [11]. Therefore, the philosophy at the backbone of the optimal control problem design proved to be robust when we moved from a complete knowledge of the epidemic trajectories (the open-loop problem) to a local optimization solely based on the knowledge of the current state of the system (the closed-loop one).

As a preliminary effort, this work is amenable to several refinements. For example, the rapid switching patterns exhibited by some RL-derived policies (see Figure 4) emerge from idealized behavioral assumptions which, for instance, neglect compliance decay due to population fatigue. The idea that fast closing-reopening policies can be effective in epidemic control and in avoiding the explosion of population distress, has been emphasized in several papers (e.g., [57]). However, behavioral feedback can substantially constrain the feasibility of frequent regime shifts. In the present baseline implementation, we did not include an explicit switching cost, to isolate the structural properties of the RL-based closed-loop control and maintain comparability with the open-loop framework [11]. Future developments should combine switching penalties and behavioral response mechanisms to assess the robustness of the resulting policies under more realistic assumptions.

In general, an important and subtle gap remains between RL and the classical optimal control theory (e.g., open-loop). Due to its local learning dynamics, RL does not perform a global optimization of the cost functional; rather, it optimizes the local costs at each time step. In other words, the RL agent optimizes single-step costs, rather than the cumulative costs over the entire time horizon. Despite this limitation, the response policies obtained through RL remain informative and practically valuable, as shown in Section 4. Therefore, RL offers a flexible and potentially data-driven alternative to classical approaches. The present study should be viewed as a preliminary, proof of concept for applying RL to epidemic control. The proposed framework is based on a model that was previously examined using open-loop optimal control techniques [11], and was deliberately chosen to ensure methodological clarity alongside a transparent comparison between the two approaches. Future work should expand the analysis by including comparisons with simpler control strategies, such as threshold-based or bang-bang controllers, as well as applying the RL framework to more complex, network-based epidemic models. These extensions will be essential to assess the robustness and generalization capacity of the proposed methodology beyond the assumptions adopted in the current formulation.

Overall, the adoption of RL for epidemic control represents a promising step towards adaptive, data-driven, real time strategies to manage epidemics under uncertainty.

Although the current RL implementation was trained on a low-dimensional epidemic model for clarity and benchmarking purposes, the *model-agnostic* nature of RL makes it particularly well-suited to high-dimensional or partially unknown epidemic systems, where uncertainty and structural variability

are intrinsic to outbreak dynamics.

Notably, RL applications are still uncommon in epidemiology, yet they have considerable potential in scenarios where policy decisions must be directly derived from data rather than from a predefined model [29, 30, 32, 33, 56].

In such contexts, where model-based approaches such as MPC [21, 54, 55] may become unreliable or computationally burdensome, RL can directly infer adaptive feedback policies from data or simulations. This enhances the robustness and responsiveness of real-time epidemic management, thus suggesting that RL-based control strategies could complement, or in some cases even surpass, classical techniques in guiding public health interventions against emerging infectious threats. This work is a preliminary exploratory demonstration of the potential of artificial intelligence tools (AI), of which RL is a major instance, to address optimal response problems during an ongoing epidemic. Despite its preliminary nature, we believe that the present results motivate future studies aimed at fully exploiting RL's potential in terms of robustness, handling uncertainty, and noise-tolerant control.

6. Conclusions

Within the numerous studies on optimal epidemic control triggered by the COVID-19 pandemic, our past work focused on the use of open-loop approaches to derive preparedness principles [11, 45]. Building on that foundation, we translated an open-loop framework into a closed-loop RL approach. We demonstrated that RL delivers adaptive, operationally interpretable control policies while retaining the qualitative control regimes identified by a previous open-loop analysis.

Crucially, RL's *model-agnostic* learning paradigm allows response policies to be directly inferred from simulated or empirical data, thus enabling rapid adaptation to partial observability, measurement noise, structural uncertainty, and high-dimensional epidemic dynamics. Our application to a realistically parametrized COVID-19 model with an explicit trade-off between healthcare and socio-economic costs showed that the RL agent produces responsive real-time interventions that balance competing objectives and preserve, under appropriate prioritization of direct costs, the healthcare capacity and resources.

These results make RL a practical, data-driven engine for real-time epidemic management: a complementary tool to deploy adaptive closed-loop strategies that integrate seamlessly into pandemic preparedness and responses.

Use of AI tools declaration

The authors declare that no Artificial Intelligence (AI) tools were used in the writing or analysis of this article.

Acknowledgments

This work was partially supported by the Italian National Recovery and Resilience Plan (NRRP) under grant S4-02.P0001 COC-1-2023-ISS-02 INF-Act ("BEHAVE-MOD"). It was also partially cofunded by the European Union – Next Generation EU – within the NRRP, Mission 4, Component 1, Investment 4.1, pursuant to Decree No. 118 (2 March 2023) and Concession Decree No. 2333 (22

December 2023) of the Italian Ministry of University and Research, Project Code D93C23000450005, under the Italian National PhD Program in Autonomous Systems (DAuSy).

Additional support was provided by MUR through the FoReLab project (Departments of Excellence).

The authors gratefully acknowledge the editor and the two reviewers for their thoughtful comments and constructive recommendations, which have substantially improved the quality and clarity of the manuscript.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. A. Brodeur, D. Gray, A. Islam, S. Bhuiyan, A literature review of the economics of covid-19, *J. Econ. Surv.*, **35** (2021), 1007–1044. https://doi.org/10.1111/joes.12423
2. R. Horton, *The COVID-19 catastrophe: What's gone wrong and how to stop it happening again*, John Wiley & Sons, 2021.
3. R. M. Anderson, R. M. May, *Infectious diseases of humans: dynamics and control*, Oxford university press, 1991. https://doi.org/10.1093/oso/9780198545996.001.0001
4. N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, et al., Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. imperial college covid-19 response team, *Imperial College COVID-19 Response Team*, **20** (2020), 77482. https://doi.org/10.25561/77482
5. M. G. Baker, N. Wilson, T. Blakely, Elimination could be the optimal response strategy for covid-19 and other emerging pandemic diseases, *BMJ*, **371** (2020), m4907. https://doi.org/10.1136/bmj.m4907
6. A. K. Sabherwal, A. Sood, M. A. Shah, Evaluating mathematical models for predicting the transmission of covid-19 and its variants towards sustainable health and well-being, *Discov. Sustain.*, **5** (2024), 38. https://doi.org/10.1007/s43621-024-00213-6
7. M. Gatto, E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi, et al. Spread and dynamics of the covid-19 epidemic in italy: Effects of emergency containment measures, *Proceed. Nat. Aca. Sci.*, **117** (2020), 10484–10491. https://doi.org/10.1073/pnas.2004978117
8. K. Wickwire, Mathematical models for the control of pests and infectious diseases: A survey, *Theor. Popul. Biol.*, **11** (1977), 182–238. https://doi.org/10.1016/0040-5809(77)90025-9
9. M. Betta, M. Laurino, A. Pugliese, G. Guzzetta, A. Landi, P. Manfredi, Perspectives on optimal control of varicella and herpes zoster by mass routine varicella vaccination, *Proceed. Royal Soc. B Biol. Sci.*, **283** (2016), 20160054. https://doi.org/10.1098/rspb.2016.0054
10. O. Sharomi, T. Malik, Optimal control in epidemiology, *Ann. Oper. Res.*, **251** (2017), 55–71. https://doi.org/10.1007/s10479-015-1834-4

11. G. Pisaneschi, M. Tarani, G. Di Donato, A. Landi, M. Laurino, P. Manfredi, Optimal social distancing in epidemic control: Cost prioritization, adherence and insights into preparedness principles, *Sci. Rep.*, **14** (2024), 4365. https://doi.org/10.1038/s41598-024-54955-4

12. F. Alvarez, D. Argente, F. Lippi, A simple planning problem for covid-19 lock-down, testing, and tracing, *Am. Econ. Rev. Insights*, **3** (2021), 367–382. https://doi.org/10.1257/aeri.20200201

13. D. Acemoglu, V. Chernozhukov, I. Werning, M. D. Whinston, Optimal targeted lockdowns in a multigroup sir model, *Am. Econ. Rev. Insights*, **3** (2021), 487–502. https://doi.org/10.1257/aeri.20200590

14. S. A. Nowak, P. Nascimento de Lima, R. Vardavas, Optimal non-pharmaceutical pandemic response strategies depend critically on time horizons and costs, *Sci. Rep.*, **13** (2023), 2416. https://doi.org/10.1038/s41598-023-28936-y

15. R. Bellman, Dynamic programming, *Science*, **153** (1966), 34–37. https://doi.org/10.1126/science.153.3731.34

16. L. Ó. Náraigh, Á. Byrne, Piecewise-constant optimal control strategies for controlling the outbreak of covid-19 in the irish population, *Math. Biosci.*, **330** (2020), 108496. https://doi.org/10.1016/j.mbs.2020.108496

17. T. A. Perkins, G. España, Optimal control of the covid-19 pandemic with non-pharmaceutical interventions, *Bull. Math. Biol.*, **82** (2020), 118. https://doi.org/10.1007/s11538-020-00795-y

18. D. H. Morris, F. W. Rossine, J. B. Plotkin, S. A. Levin, Optimal, near-optimal, and robust epidemic control, *Commun. Phys.*, **4** (2021), 78. https://doi.org/10.1038/s42005-021-00570-y

19. A. Kasis, S. Timotheou, N. Monshizadeh, M. Polycarpou, Optimal intervention strategies to mitigate the covid-19 pandemic effects, *Sci. Rep.*, **12** (2022), 6124. https://doi.org/10.1038/s41598-022-09857-8

20. R. Carli, G. Cavone, N. Epicoco, P. Scarabaggio, M. Dotoli, Model predictive control to mitigate the COVID-19 outbreak in a multi-region scenario, *Ann. Rev. Control*, **50** (2020), 373–393. https://doi.org/10.1016/j.arcontrol.2020.09.005

21. J. Köhler, L. Schwenkel, A. Koch, J. Berberich, P. Pauli, F. Allgöwer, Robust and optimal predictive control of the covid-19 outbreak, *Ann. Rev. Control*, **51** (2021), 525–539. https://doi.org/10.1016/j.arcontrol.2020.11.002

22. R. An, J. Hu, L. Wen, A nonlinear model predictive control model aimed at the epidemic spread with quarantine strategy, *J. Theor. Biol.*, **531** (2021), 110915. https://doi.org/10.1016/j.jtbi.2021.110915

23. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, et al., Human-level control through deep reinforcement learning, *Nature*, **518** (2015), 529–533. https://doi.org/10.1038/nature14236

24. H. Le, S. Saeedvand, C.-C. Hsu, A comprehensive review of mobile robot navigation using deep reinforcement learning algorithms in crowded environments, *J. Intell. Robot. Syst.*, **110** (2024), 1–22. https://doi.org/10.1007/s10846-024-02198-w

25. B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, et al., Deep reinforcement learning for autonomous driving: A survey, *IEEE Transact. Intell. Transport. Syst.*, **23** (2021), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

26. J. Weltz, A. Volfovsky, E. B. Laber, Reinforcement learning methods in public health, *Clin. Ther.*, **44** (2022), 139–154. https://doi.org/10.1016/j.clinthera.2021.11.002

27. A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, M. Guizani, Reinforcement learning for intelligent healthcare systems: A review of challenges, applications, and open research issues, *IEEE Int. Things J.*, **10** (2023), 21982–22007. https://doi.org/10.1109/JIOT.2023.3288050

28. E. Jordan, D. E. Shin, S. Leekha, S. Azarm, Optimization in the context of covid-19 prediction and control: A literature review, *Ieee Access*, **9** (2021), 130072–130093. https://doi.org/10.1109/ACCESS.2021.3113812

29. H. Khadilkar, T. Ganu, D. P. Seetharam, Optimising lockdown policies for epidemic control using reinforcement learning: An ai-driven control approach compatible with existing disease and network models, *Transact. Indian Nat. Acad. Eng.*, **5** (2020), 129–132. https://doi.org/10.1007/s41403-020-00129-3

30. A. Q. Ohi, M. Mridha, M. M. Monowar, M. A. Hamid, Exploring optimal control of epidemic spread using reinforcement learning, *Sci. Rep.*, **10** (2020), 22106. https://doi.org/10.1038/s41598-020-79147-8

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

32. S. N. Khatami, C. Gopalappa, Deep reinforcement learning framework for controlling infectious disease outbreaks in the context of multi-jurisdictions, medRxiv, 2022–10. https://doi.org/10.1101/2022.10.18.22281063

33. M. Arango, L. Pelov, Covid-19 pandemic cyclic lockdown optimization using reinforcement learning, arXiv preprint arXiv:2009.04647. https://doi.org/10.48550/arXiv.2009.04647

34. G. Gemignani, A. Landi, P. Manfredi, G. Pisaneschi, A reinforcement learning-based decision framework for assessing health/economics dilemma in pandemic control, in *2025 25th International Conference on Control Systems and Computer Science (CSCS)*, IEEE, 2025, 46–53. https://ieeexplore.ieee.org/document/11181599

35. D. Buitrago-Garcia, D. Egli-Gany, M. J. Counotte, S. Hossmann, H. Imeri, A. M. Ipekci, et al., Occurrence and transmission potential of asymptomatic and presymptomatic sars-cov-2 infections: A living systematic review and meta-analysis, *PLoS Med.*, **17** (2020), e1003346. https://doi.org/10.1371/journal.pmed.1003346

36. B. Nogrady, What the data say about asymptomatic covid infections, *Nature*, **587** (2020), 534–535. https://doi.org/10.1038/d41586-020-03141-3

37. M. K. Slifka, L. Gao, Is presymptomatic spread a major contributor to covid-19 transmission?, *Nat. Med.*, **26** (2020), 1531–1533. https://doi.org/10.1038/s41591-020-1046-6

38. W. C. Koh, L. Naing, L. Chaw, M. A. Rosledzana, M. F. Alikhan, et al., What do we know about sars-cov-2 transmission? A systematic review and meta-analysis of

the secondary attack rate and associated risk factors, *PloS One*, **15** (2020), e0240205. https://doi.org/10.1371/journal.pone.0240205

39. B. Rai, A. Shukla, L. K. Dwivedi, Incubation period for covid-19: A systematic review and meta-analysis, *J. Public Health*, **30** (2022), 2649–2656. https://doi.org/10.1007/s10389-021-01478-1

40. J. Zhang, M. Litvinova, W. Wang, Y. Wang, X. Deng, X. Chen, et al., Evolving epidemiology of novel coronavirus diseases 2019 and possible interruption of local transmission outside hubei province in china: A descriptive and modeling study, *MedRxiv*. Available from: https://pubmed.ncbi.nlm.nih.gov/32511424/

41. M. Cevik, M. Tate, O. Lloyd, A. E. Maraolo, J. Schafers, A. Ho, Sars-cov-2, sars-cov, and mers-cov viral load dynamics, duration of viral shedding, and infectiousness: A systematic review and meta-analysis, *Lancet Microbe*, **2** (2021), e13–e22. https://doi.org/10.1016/S2666-5247(20)30172-5

42. Istituto Superiore di Sanità (ISS), *COVID-19: Sorveglianza, impatto delle infezioni ed efficacia vaccinale. Report esteso — Aggiornamento nazionale 25 maggio 2022*, Istituto Superiore di Sanità (ISS), Roma, Italy, 2022, In Italian. Available from: https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_25-maggio-2022.pdf

43. R. E. Hall, C. I. Jones, P. J. Klenow, *Trading off consumption and COVID-19 deaths*, Technical report, National Bureau of Economic Research, 2020. https://doi.org/10.3386/w27340

44. R. L. Ohsfeldt, C. K.-C. Choong, P. L. Mc Collam, H. Abedtash, K. A. Kelton, R. Burge, Inpatient hospital costs for covid-19 patients in the united states, *Adv. Ther.*, **38** (2021), 5557–5595. https://doi.org/10.1007/s12325-021-01887-4

45. A. Landi, G. Pisaneschi, M. Laurino, P. Manfredi, Optimal social distancing in pandemic preparedness and lessons from covid-19: Intervention intensity and infective travelers, *J. Theor. Biol.*, **604** (2025), 112072. https://doi.org/10.1016/j.jtbi.2025.112072

46. K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, *IEEE Signal Process. Magaz.*, **34** (2017), 26–38. https://doi.org/10.1109/MSP.2017.2743240

47. C. Szepesvári, *Algorithms for reinforcement learning*, Springer nature, 2022. https://doi.org/10.1007/978-3-031-01551-9

48. V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, et al., Playing atari with deep reinforcement learning, arXiv preprint arXiv:1312.5602. https://doi.org/10.48550/arXiv.1312.5602

49. H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016. https://doi.org/10.1609/aaai.v30i1.10295

50. M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, et al., Gymnasium: A standard interface for reinforcement learning environments, arXiv preprint arXiv:2407.17032. https://doi.org/10.48550/arXiv.2407.17032

51. J. Weng, H. Chen, D. Yan, K. You, A. Duburcq, M. Zhang, et al., Tianshou: A highly modularized deep reinforcement learning library, *J. Mach. Learn. Res.*, **23** (2022), 1–6. Available from: `http://jmlr.org/papers/v23/21-1127.html`

52. T. Eimer, M. Lindauer, R. Raileanu, Hyperparameters in reinforcement learning and how to tune them, in *International conference on machine learning*, PMLR, 2023, 9104–9149. https://doi.org/10.48550/arXiv.2306.01324

53. I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101

54. T. Alleman, E. Torfs, I. Nopens, Covid-19: from model prediction to model predictive control, *Unpubl Preprint*. Available from: `https://www.researchgate.net/publication/360342734_Covid-19_from_model_prediction_to_model_predictive_control`

55. S. Beregi, K. V. Parag, Optimal algorithms for controlling infectious diseases in real time using noisy infection data, medRxiv, 2024–05. https://doi.org/10.1101/2024.05.24.24307878

56. K. Mitsopoulos, L. Baker, C. Lebiere, P. Pirolli, M. Orr, R. Vardavas, Cognitively-plausible reinforcement learning in epidemiological agent-based simulations, *Front. Epidemiol.*, **5** (2025), 1563731. https://doi.org/10.3389/fepid.2025.1563731

57. M. Bin, P. Y. Cheung, E. Crisostomi, P. Ferraro, H. Lhachemi, R. Murray-Smith, et al., Post-lockdown abatement of covid-19 by fast periodic switching, *PLoS Comput. Biol.*, **17** (2021), e1008604. https://doi.org/10.1371/journal.pcbi.1008604

AIMS Press