Received: 15 July 2025 | Revised: 4 November 2025 | Accepted: 27 November 2025 | Published online: 30 December 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1–11
DOI: 10.47852/bonviewAIA52026822

BON VIEW PUBLISHING

Optimal Policy Strategy for Pandemic Outbreak Control: A Deep Reinforcement Approach

Raphael Ibraimoh[1,*], **Mohammed Saraee**[1,2], **Kaveh Kiani**[1,2] **and Danial Saraee**[3]

[1] *School of Science, Engineering and Environment, University of Salford, UK*

[2] *Data Science and AI Hub, University of Salford, UK*

[3] *Hall University Teaching Hospitals, NHS Trust, UK*

Abstract: Since the global spread of COVID-19 pandemic in January 2020, residents in the United Kingdom (UK) have altered their daily routines due to the transmissibility of the virus. Sanitisation, quarantine, contact tracing, mass testing, and vaccination are implemented, affecting virus control, quality of life, resources, and economic development. From January 2020 to January 2021, data from repositories from the Office for National Statistics, NHS England, and the WHO provided statistics on confirmed cases, recoveries, and mortality. Wikipedia and Our World In Data provided the UK lockdown and travel restriction timelines. Deep reinforcement learning, a Dueling Q-learning algorithm, and a well-defined reward function determined the optimal lockdown and travel restriction timings. Initially, our agent (model) suggested strict lockdown and travel restrictions. By mid-March, advisories decreased significantly. In late March, key public health initiatives were introduced. Over the initial three months, the recommendations of our agent had gained support, which proposed slightly smaller lockdown measures than the public health policy but stricter travel restrictions. Our agent advised lockdown and travel limitations, generally suggesting measures before public health authorities or the government approved them. Our agent recommended implementing policies in late January, while authorities delayed until late March. Furthermore, our agent (model) advised against postponing UK policy implementation.

Keywords: COVID-19, deep reinforcement learning, discrete action space, dueling Q-network, lockdown, travel restrictions

1. Introduction

Since the World Health Organization (WHO) reported SARS-CoV-2 in late December 2019, the United Kingdom (UK) has taken at least 143 steps to halt its spread [1]. Some research analysed these interventions using [2] the classification framework and Oxford Stringency index. The policy classification system displays a spectrum of COVID-19 pandemic intervention options that intensify and then diminish as governments scale back on response efforts [2]. These classifications are essential for harm reduction and healthcare improvement through containment and mitigation. This area includes economic and health technology initiatives. In 2020, English pubs and restaurants reopened on July 4, but the Scottish waited until July 15. Northern Ireland allowed indoor restaurants and pubs to reopen on July 3, but Wales waited until August 3. The UK government initially focused on viral containment [3]. During this lockdown, people had to stay at home and work remotely, with limited exceptions for exercise, food shopping, and prescription retrieval, under strict social distancing guidelines. The number of daily COVID-19 case confirmations in the UK surged, plateaued, and reduced. The number of laboratory-confirmed cases peaked on May 1, while the number of symptomatic patients peaked on April 1. As confirmed cases decreased after this peak, the rules were relaxed [4].

Also in 2020, social distancing procedures and penalties escalated the scale and intensity of measures, resulting in severe restrictions on non-essential services on March 16 and a statewide lockdown on March 23. The UK government prioritised emergency planning for influenza over coronaviruses, resulting in inadequate reserves, testing procedures, and lockdown protocols when COVID-19 struck. This discrepancy rendered the country unprepared for the magnitude and characteristics of the actual pandemic [5]. The culture of groupthink and a convoluted distribution of responsibility between departments impeded decision-making, resulting in delays in essential activities. The absence of decisive leadership and the failure to question established assumptions undermined the urgency and efficacy of the response [6]. However, some policies recorded success: The UK Health Security Agency was created to consolidate responses to health hazards and enhance data-driven decision-making, replacing the less-responsive Public Health England. By 2024, it improved the nation's ability to identify outbreaks and respond to them with greater speed and efficacy [7]. A new Cabinet-level committee has been established to oversee civil emergency planning, facilitating swifter and more coordinated government responses. This structural modification was intended to avert the disjointed leadership observed during COVID-19 [8].

2. Literature Review

COVID-19 has numerous challenges. Vaccinations in the UK do not fully prevent the virus. Enhanced testing and immunisation, despite their laborious nature, typically stop transmission. Disease control and economic recovery are challenges for governments [9, 10]. Increasing sanitisation in densely populated areas to sterilise public spaces and reduce transmission is another common method, although it is resource-intensive and difficult to implement globally [10]. Lockdown and quarantine demand fewer medical and physical resources. Pandemic

[*] **Corresponding author:** Raphael Ibraimoh, School of Science, Engineering and Environment, University of Salford, UK. Email: R.Ibraimoh2@salford.ac.uk

strains worldwide healthcare systems. Travel restrictions and lockdowns have reduced tension. Study [11] investigated how these methods boost healthcare capacity and treat viruses. Although crucial for public health, these methods raise ethical issues [12, 13]. Lockdowns and travel restrictions need ethical decision-making that balances community health and individual rights with openness, equity, and compassion. Study [10] found that these constraints harm mental health and access to non-COVID medical care and general well-being, even if they prevent viral spread. Trade-offs are assessed and minimised using this paradigm. In the current economy, it is unreasonable to stop all economic activities and rely on government handouts. Serious infections require ventilators, while mild infections require critical supplies.

Transmission risks prevent unrestricted social gatherings without masks. Some techniques may stay unexplored [14]. Testing, sanitisation, and social distancing are necessary, but their optimal levels must be found to safeguard the public while balancing health with daily activities and reducing negative effects on quality of life and the economy [15]. To solve this challenge, our research uses quantitative, model-driven methodologies. Our research aims to develop a Deep Reinforcement Learning (DRL) agent capable of identifying the optimal combination of lockdown and travel-restriction policies for the UK. Conventional approaches, such as cost-benefit and risk analysis and epidemiological model, and the DRL approach are utilised to design optimal strategies for COVID-19 policy [11, 16]. The rapidly developing study of DRL can transform human history [17].

Academics and industry are interested in its autonomous optimisation. Reinforcement learning in intelligent systems may illuminate human intelligence. This definition of intelligence is learning from experience. Knowledge of the best algorithm is needed to make optimal decisions in diverse issue situations. Optimal decisions may mean sacrificing short-term profits for long-term success. Research [16, 18] examined regional lockdowns and travel restrictions to stop the spread of infectious diseases. Many studies replicate epidemic transmission to accommodate various disease traits and localised control. Their impact on domestic viral spread is small. Preventive lockdowns may limit local transmission. Lockdowns reduce mortality when infection rates are low. The COVID-19 pandemic has made demographic and socioeconomic factors vital; therefore, these interventions are effective but require a country-specific strategy. According to global studies, public health regulations [16], such as travel limits, reduced fatalities by 68% but were inefficient in the control of domestic transmission. Proactive lockdown reduces localised transmission. Study [19] promoted worldwide COVID-19 lockdowns and travel restrictions using reinforcement learning and the Deep Deterministic Policy Gradient (DDPG) algorithm [20]. The broad range of the proposed measures does not provide global recommendations for pandemic prevention and control. A more flexible and focused policy solution can be developed for different areas of action. Our research proposes a novel DRL discrete action space construction method. Using this strategy, the agent can choose the optimal lockdown and travel restrictions within a certain range. To overcome discrete spatial restrictions, it is crucial to establish a strong reward system that balances economic efficiency and quality of life. Study [17, 20] employed a DRL approach, Double Deep Q-Network (D3QN), which uses a discrete state and action representation space. The continuous action space of DDPG may lead to erroneous forecasts due to sudden shifts in epidemic situations. Under constrained training conditions, Deep Q-Network (DQN) outperforms DDPG [20].

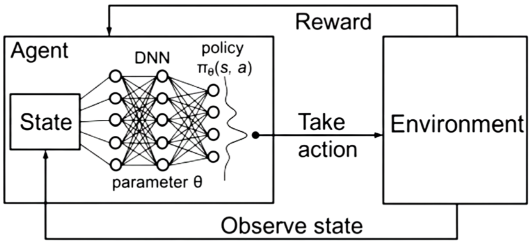
3. Methods

3.1. Deep reinforcement learning method

DRL is an algorithm that falls under the category of machine learning. Its primary goal is to address issues related to artificial intelligence (AI). This approach is achieved by developing computer programmes called agents, which are designed to solve complex issues that typically require intelligence. DRL stands out from other types of machine learning algorithms due to its unique learning framework. The learning process of the system involves iterative experimentation and adjustment, during which it interacts with its environment and receives feedback in the form of state and reward signals [21]. This suggests the lack of labelled data or a clearly defined appropriate response to use, which distinguishes it from supervised learning. Environmental feedback can occur concurrently or sequentially and can be evaluated or sampled from an initial probability distribution. By applying deep neural networks, agents are trained to approximate the underlying reward distribution through the use of nonlinear function approximation methods. The overarching objective is to optimise cumulative returns over an extended time period.

Figure 1 illustrates how DRL works: an agent engages with the environment by perceiving the current state, selecting the action based on a policy (often facilitated by a neural network), and receiving a reward and the subsequent state as feedback. The neural network functions as a function approximator to generalise over extensive state-action spaces. Through this feedback loop, the agent perpetually refines its strategy to optimise long-term rewards by learning from its experiences. The environment defines the task dynamics, including the state and action spaces, and provides a reward function. At each time step, the agent observes the current state and selects an action according to its policy, typically represented by a deep neural network. This network functions either as a policy approximator or a value estimator, depending on the learning paradigm. Upon executing an action, the environment transitions to a new state and returns a scalar reward, indicating the immediate outcome of the action. The experience tuple—including state, action, reward, and next state—is stored in a replay buffer. This buffer enables the agent to sample past experiences and update its neural network using gradient-based optimisation techniques. The use of deep networks enables the agent to operate effectively in complex, high-dimensional environments, while the replay buffer mitigates temporal correlations and enhances learning stability. Through repeated interaction, memory, and training, the agent progressively improves its decision-making to maximise long-term rewards. DRL has demonstrated success across domains such as robotics, game playing, and autonomous control.

Figure 1
Structure of deep reinforcement learning



3.2. D3QN architecture

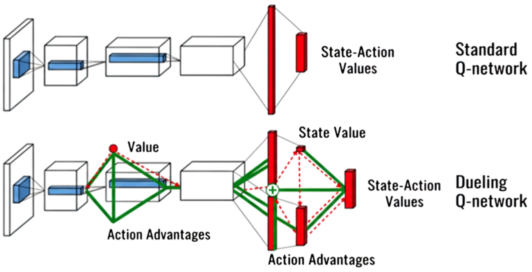
To this point, the deep learning models under consideration have predominantly employed a sequential architecture. Here, the term "models" refers to supervised learning frameworks, distinguishing them from models defined in the formalism of the Markov Decision Process. It is worth noting that while the terms "sequential architectures" and "sequential models" may have different meanings in other disciplinary contexts, in deep learning, they refer to a specific structural property of the network. In these architectures, each neuron in a given layer is restricted to forming connections only with neurons in the immediately preceding and subsequent layers. This constraint applies uniformly across all neurons within the layer. In other words, there were no branches or loops in these model structures. Although DQN and Double DQN each contain two Q networks, there is only one deep learning model, and the values of the other (target) network are periodic duplicates of the active (online) network [17, 21]. Dueling DQN is a non-sequential deep learning architecture in which the model layers are divided into two independent streams (sub-networks), each with its own fully connected layer and output layers. This enables the model to train more efficiently than conventional sequential designs.

D3QN architecture is an advanced reinforcement learning model designed to improve the stability and accuracy of value-based decision-making. It builds upon the standard DQN by incorporating two key innovations: the dueling network architecture and double Q-learning. In D3QN, the agent first processes input from the environment, such as images or numerical states, through a feature extraction layer, often a convolutional neural network if the input is visual. This layer captures essential patterns and compresses the input into a compact representation. Next, the network splits into two separate streams. One stream estimates the value of being in a particular state, regardless of the action taken. The other stream estimates the advantage of each possible action in that state. These two estimates are subsequently integrated to generate the final action-value predictions. This decomposition enables the agent to discern which states are intrinsically valuable even when the choice of action has minimal influence, thereby enhancing learning efficiency. To reduce overestimation bias, a common issue in Q-learning, D3QN uses double Q-learning, where one network selects the best action and another evaluates it. This decoupling leads to more accurate value estimates. Experience replay and periodic target network updates further stabilise training [21].

The first branch, often referred to as the value network, estimates the value of a given state and produces a single scalar output representing this value. The second branch, referred to as the advantage network, estimates the relative benefit of selecting a particular action compared to the baseline value associated with remaining in the current state. Figure 2 illustrates the overall architecture. It is important to emphasise that the Q-function in a Dueling DQN corresponds to the same fundamental Q-function used in conventional Q-learning algorithms. Consequently, the Dueling DQN framework is expected to operate conceptually in the same manner as standard Q-learning, producing absolute action-value estimates (Q-values) for each state–action pair.

The design in Figure 2 uses two streams to segregate state values and state-dependent action advantages. For feature learning, it estimates value and advantage functions with convolutional layers. The last module integrates these streams to estimate the state-action value function Q. The proposed function operates analogously to a conventional Q-learning algorithm by estimating absolute action values, commonly referred to as Q-values. Consequently, it becomes necessary to compute these action-value estimates for each state-action pair. Conceptually, an action value represents the expected utility of executing a specific action within a given state. To accomplish this, the model integrates two components: (i) the baseline value of the state, computed by the first network branch, and (ii) the relative advantage of each action, provided by the second branch (the "advantage" network). By aggregating these elements, the model can effectively yield the approximate Q-values required for Q-learning. This relationship is formally expressed in Equation (2).

Figure 2
Schematic structure of the dueling Q-network



$$\mathrm{Q(s,a;\alpha,\beta) = V(s;\theta,\beta) + (A(s,\alpha;\theta,\alpha) - max'_a \in |A| A(s,a';\theta,\alpha))} \quad (1)$$

The variables Q, V, s, a, and a′ in Equation (1) have the same consistent meaning as those in the mathematical notation in Figure 1. Furthermore, the letter denotes the advantage value. The parameter vector associated with the convolutional layer is shared between the Value and Advantage networks. The parameter set specific to the Advantage network and the State-Value function are denoted by α and β, respectively. Within the context of function approximation, the outputs of any network are expressed in terms of the parameters of the corresponding estimating network. This distinction ensures clarity when multiple approximators infer values for the same variable. The governing equation specifies that the Q-value indexed by (β,α,β) for a given state-action pair equals the estimated state value, representing the absolute utility of being in that state, as derived from the State-Value network. Further consideration of identifiability suggests that, in its simplest form, this relationship can be expressed as shown in Equation (3).

$$\mathrm{Q(s,a;\alpha,\beta) = V(s;\theta,\beta) + A(s,a;\theta,\alpha)} \quad (2)$$

A limitation of this straightforward formulation is that although the Q-value (action-value) can be calculated given a state–action pair (S,A), the inverse mapping does not hold. Specifically, it is not possible to uniquely recover the state and action values from a given Q-value, as shown in Equation (3). This issue is referred to as unidentifiability. To address this issue, Equation (4) introduces an improved variant of Equation (3), in which the last term is slightly modified. Although this adjustment involves subtracting a constant and may introduce a small numerical deviation, it does not affect the learning process, as the relative comparisons between action values remain unchanged. Moreover, this reformulation helps improve the stability during optimisation.

$$\mathrm{Q(s,a;\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|A|}\sum_a A(s,a;\theta,\alpha)\right)} \quad (3)$$

We provide the details of the D3QN algorithm in Table 1 below.

Table 1 presents a step-by-step implementation of the D3QN algorithm, which is designed to generate an optimal policy for determining the appropriate timing of lockdown measures and the imposition of travel restrictions on UK domestic movements.

3.3. Action space

This paper proposes the use of a discrete action space to determine the level of severity of lockdowns and travel restrictions. The recommended values for the three action outputs display variability within the predefined range, suggesting that this recommendation engine offers more flexibility and customisation than a continuous action space. See Table 2 below.

We define a 3×3 action space for implementing local lockdowns and travel restrictions. The lockdown policy is classified into three levels: Level 0 (L0) means no intervention, Level 1 (L1) imposes restrictions on public social gatherings, and Level 2 (L2) corresponds to a nationwide lockdown. Similarly, the travel policy included three scenarios: T0 indicates no action will be taken, T1 involves suspension of air travel, and T2 means the complete closure of all borders within the UK.

Table 1
D3QN algorithm for optimal policy implementation of COVID-19

```
1. Initialize:
    1) Priority experience replay buffer D with capacity N.
    2) Parameters for the initial Q-network (θ) and target network (θ⁻).
    3) Set ε (epsilon) for the ε-greedy policy.
2. For each episode (t = 1 to M):
    1) Reset the environment.
    2) Initialize the input state sequence $s_0$.
    3) Reset the UK COVID-19 burden (e.g., case numbers, hospital load).
3. For each time step (t = 1 to T):
    1) For each batch i in batch size $b_i$:
    a. Use the target network to compute the next state $s_{t+1}$.
    b. Select action $a_t$ for batch i with probability $p_i$.
    c. Store the transition ($s_{t+1}$, $a_t$) in the policy network.
    d. Determine the next state $s_{t+1}$ based on accumulated reward.
    e. Compute the immediate reward $r_{t,i}$, reflecting acceleration in COVID-19 cases (if episode ends).
    f. Update the Q-value using the immediate reward and discount factor γ.
    g. Update the loss function L(θ).
4. Update:
    1) Adjust sampling weights using precision error, selection weight, and ε.
    2) Update the replay buffer D with new transitions
    3) At each target update interval, copy weights from the policy network to the target network.
```

Table 2
Lockdown and travel restriction policy levels in the UK

Action Space	Acronym
No lockdown	L0
Restriction on social gathering in the UK	L1
Complete national lockdown	L2
No aircraft restrictions	T0
Aircraft cancellation	T1
Border closure	T2

3.4. Reward

Our rewards system is designed to discourage rapid escalation of infections and death rates while also boosting rapid case recovery in a 2:1:1 ratio. The argument for creating this reward system has two main goals. According to earlier research, it is crucial to emphasise the prevention of new infections as this is associated with a reduced death rate [3]. Furthermore, there is a significant temporal gap between the onset of a new infection case and the subsequent occurrence of recovery or death. Previous research has demonstrated that this temporal delay often leads to the postponement of governmental interventions, which can be seen as a potential aspect to consider in reinforcement learning [7]. According to a 2022 study by the University of Oxford, the UK's early approval and mass deployment of the Oxford-AstraZeneca vaccine, which began in December 2020, was a major success, significantly reducing hospitalisations and deaths in the UK and enabling society to return to normal sooner. Despite rising infections and warnings from the scientific community, the UK delayed its first lockdown in March 2020, allowing the virus to spread rapidly. A 2025 report from Queen's University Belfast highlighted that this delay contributed to one of the highest excess death rates in Europe [22].

Table 3 shows the general sign direction for the definition rationale of awards, based on death severity. The same logic applies to recovery, which will take the opposite sign.

Table 3 shows an example of how to design a reward sign indicating the seriousness of death. The same principle applies to death and recovery, with the latter displaying the opposite sign. If no action is taken, more deaths are likely due to the transmissibility of the disease. If the activities improve the situation, it may take a few days for the effects to become apparent. However, if the reward is favourable, the agent will be motivated.

Table 3
Example of reward signal design reflecting death severity in policy learning

Severity	Meaning	Reward sign
High	Increase in death cases without action taken	Negative
Moderate	Considerate decrease in death cases compared to previous date due to action taken	Positive
Low	Further reduction in death cases or no death due to certain action taken	Positive

As a result, we decided to impose greater penalties on the increased rate of new infections than on the increase in mortality or compensation for the higher rate of recovered cases. We created a system of positive incentive stability, which involved making no modifications and minimising the number of new infection cases. In contrast, we used negative rewards when the number of new infection cases rose despite the interventions performed. Positive rewards were not given for the absence of changes in new infection rates, as a lack of increase in new case rates is often suggestive of stability in the early phases of an outbreak [11]. Following this overarching approach, the reward function rt was constructed as described below. It is important to note that the relative weights of its components can be adjusted to suit specific objectives. In this study, the reward function was designed to capture the dual impact of COVID-19 on economic performance and quality of life. A value of 10 corresponds to a strong economy and high quality of life, whereas 0 represents the opposite. While both economic conditions and quality of life are influenced by numerous factors, for this research, these endpoints were assigned arbitrarily, with 0 representing "poor" and 10 representing "good." Thus, we have the following equations:

$$\mathrm{c_0 = c_f + \tfrac{1}{ld},\ c_0 = c_f + \tfrac{1}{lr},\ c_1 = c_f + \tfrac{1}{ld},\ c_1 = c_f + \tfrac{1}{lr}} \tag{4}$$

$$\mathrm{r_t = r_1^{rc} + r_t^{dt} + r_t^{cf}} \tag{5}$$

$$r_t^i = \begin{cases} -c_0 - c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^t > s_t^i\right) \text{ and } \left(s_t^i > 0\right) \\ -0.5xc_0 + c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^i = s_t^i\right) \text{ and } \left(s_t^i \neq 0\right) \\ \text{and } \left(\left(a_{t+1}^{\text{lockdown}} > 0\right) \text{ or } \left(a_{t+1}^{\text{travel\_ban}} > 0\right)\right) \quad \text{for } i \text{ in } rc \text{ and } dt \\ c_0 - c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^i < s_t^t\right) \\ -0.5xc_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{otherwise} \end{cases} \tag{6}$$

$$r_t^i = \begin{cases} c_0 + c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^t > s_t^i\right) \text{ and } \left(s_t^i > 0\right) \\ 0.5 \times c_0 + c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^t = s_t^i\right) \text{ and } \left(s_t^i \neq 0\right) \\ \text{and } \left(\left(a_{t+1}^{\text{lockdown}} > 0\right) \text{ or } \left(a_{t+1}^{\text{travel\_ban}} > 0\right)\right) \quad \text{for } i \text{ in } cf \\ -c_0 + c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{if } \left(s_{t+1}^t < s_t^i\right) \\ -c_1 \times \left(s_{t+1}^i - s_t^i\right) \quad \text{otherwise} \end{cases} \tag{7}$$

where

Ld: lockdown
tr: travel ban
Lq: quality of life = 0 or 10
econ: economy = 0 or 10
rc: recovered cases
cf: confirmed cases
dt: death cases
r_t: timestamp t for reward
s_t: timestamp t for state
c_0, c_1: constant values

The design initially sets the range of the quality-of-life and economic indices between 0 and 10 to provide an intuitive, normalised scale; however, we recognise that this choice introduces an element of arbitrariness. To mitigate potential bias, future work will calibrate these scales using empirical indicators such as gross domestic product contraction, unemployment rate, and quality-adjusted life-years loss. Likewise, the epidemiological rationale behind giving more weight to infections than deaths was that preventing new infections can indirectly prevent downstream fatalities. Nonetheless, the relative weights can be data driven in future implementations by regressing observed policy outcomes against measurable public-health and economic variables.

3.5. Model performance evaluation

The experimental evaluation compares the performance of the standard D3QN model with the proposed V-D D3QN algorithm under identical parameter settings. In this setup, the discount factor (γ) was fixed at 0.99, with a maximum of 100 episodes. The maximum average reward over 100 steps was limited to 10,000. Both the policy and value networks employed a learning rate of 0.0001. The replay buffer size was set to 10,000, and the batch size was 32. Additionally, the soft update parameter (τ) was configured to 0.005. Figure 3(b) illustrates the enhanced D3QN architecture, referred to as V-D D3QN, which serves as a baseline for policy comparison against the standard D3QN implementation.

The moving average rewards during the training and evaluation phases are shown in Figures 3(a) and 3(b), respectively. During evaluation, performance stabilises rapidly within the range of 3,000 to 4,000 in the initial episodes and subsequently exhibits a gradual upward trend. These results demonstrate that the proposed model achieves strong and consistent performance across episodes.

3.6. Validation against epidemiological indicators

To evaluate the epidemiological plausibility of the proposed policies, the timing and intensity of the interventions recommended by the model were compared with publicly available reproduction-number (R_t) estimates from the UK Health Security Agency and the simulated Susceptible-Exposed-Infectious-Recovered (SEIR) trajectories. The agent's recommendation for an early lockdown implementation in late January 2020 coincided with periods where R_t was above 1.5, while subsequent easing of restrictions occurred when the R_t fell below 1.0. This qualitative alignment suggests that the learned policy captured biologically meaningful dynamics. Although a full quantitative coupling to SEIR equations was beyond the present scope, the comparison supports the epidemiological plausibility of the learned strategies.

3.7. Sensitivity and uncertainty analysis

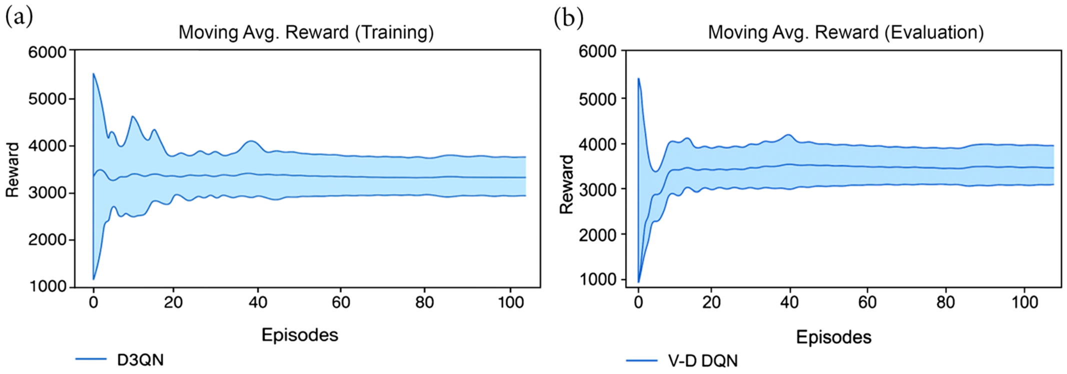
To evaluate the robustness of the D3QN model and assess sensitivity to hyperparameter variations, several controlled experiments were performed by perturbing the core learning parameters. Specifically, the learning rate (α), discount factor (γ), and replay buffer size (B) were adjusted within ±20% of their baseline values (α = 0.0001, γ = 0.99, B = 10,000). The resulting change in the average cumulative reward over 50 evaluation episodes is summarised in Table 4.

The analysis shows that model performance remains stable under moderate parameter variations, with cumulative reward changes less than ±3%. This suggests that the D3QN configuration used in this study is robust to small perturbations in key hyperparameters. Nevertheless, full uncertainty quantification is recommended for future work. Probabilistic methods such as Monte Carlo dropout or bootstrapped DQN ensembles can be implemented to generate confidence intervals around policy estimates, further strengthening interpretability and reliability.

Table 4
Sensitivity of model performance to key hyperparameter variations

Parameter	Baseline Value	Variation	Mean Cumulative Reward	% Change vs Baseline
Learning rate (α)	0.0001	20%	3985	2.80%
Learning rate (α)	0.0001	−20%	3840	−0.9%
Discount factor (γ)	0.99	0.05	3920	−1.5%
Discount factor (γ)	0.99	−0.05	4010	3.10%
Replay buffer size (B)	10,000	25%	3960	1.90%
Replay buffer size (B)	10,000	−25%	3825	−1.3%

Figure 3
Baseline policy comparison for D3QN: (a) Training phase, (b) Evaluation phase



4. Experimental Results

4.1. Data

Our research analysed data from the UK from January 21, 2020 (the date of the WHO's initial report on COVID-19) to March 2021. Data on the index case date (the date of the first confirmed patient), confirmed infections, recoveries, and deaths were obtained from reliable sources, including the Johns Hopkins COVID-19 Data Repository, reports from the Centers for Disease Control and Prevention, and case reports from the WHO [16, 21]. In addition, details on the date and severity of local lockdown measures, as well as international travel restrictions, were provided. After linear interpolation from the date of the index case in the UK, data were collected by averaging values over three consecutive days. This method was intended to eliminate potential biases caused by delayed reporting and changes in viral activity. Instead of collecting data daily, a three-day interval was chosen. This approach was adopted due to the need for time-sensitive information at each time stamp, as well as to reduce potential biases caused by delayed reporting and variable viral testing capability on weekends.

4.2. Result

In the results section, the term "Agent Model" refers to the proposed reinforcement learning framework, whereas "Public Health" denotes the officially implemented policy model. Experimental analyses were conducted using empirical COVID-19 data from the UK, covering the period from January 2020 to January 2021, with the first confirmed case reported on January 31, 2020. To enhance the agent's ability to identify optimal policies, a well-structured incentive function was developed. The primary objective was to mitigate the impact of the pandemic by dynamically adjusting the severity of lockdown measures and mobility restrictions daily. From a behavioural standpoint, individuals naturally aim to minimise high infection and mortality rates because of their negative societal consequences. Conversely, targeted interventions must be incentivised to effectively reduce infection and mortality rates to predefined thresholds.

The reward structure within the environment comprises two distinct components: mortality and recovery, each associated with specific incentives and penalties that collectively define the overall reward function. This study proposes a discrete action space to determine the severity levels of lockdowns and travel restrictions instead of the continuous action space proposed by [16] in their DDGP implementation. The continuous approach is less reliable, especially when addressing epidemic termination scenarios.

Figures 4(a), 4(b), and 4(c) depict the evolution of lockdown intensity, travel restrictions, and the combined severity of these interventions during the first three months of the COVID-19 pandemic in the UK.

Figure 4

Lockdown, travel ban, and total intensity for the first 3 months: (a) Lockdown for the first 3 months, (b) Travel ban for the first 3 months, (c) Total intensity for the first 3 months

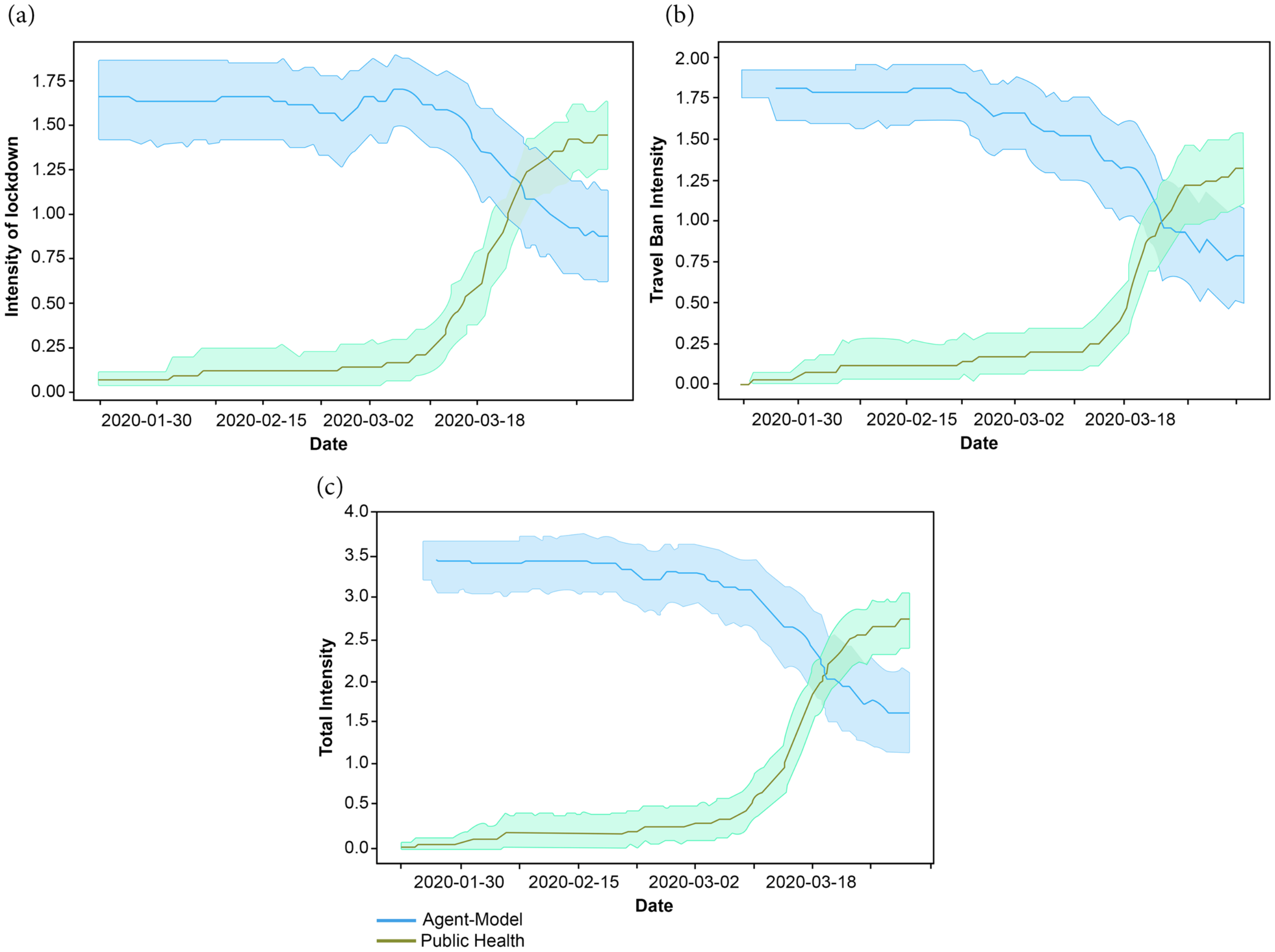


Table 5 shows how the intensity of travel bans and lockdowns, and their combined effect changed over four key dates during the first three months of the pandemic. We can see that travel restrictions were tightened and eased at different times, possibly reflecting shifts in the government's response to the spread of the virus. Lockdown measures also varied, becoming stricter or more relaxed depending on the situation. When we look at the total intensity, which combines both types of interventions, it tends to be highest during periods of strong public health action and lower when measures were relaxed. Overall, the table gives a clear picture of how policies were adjusted over time to respond to the changing nature of the pandemic.

The proposed dueling network was trained, and its performance was analysed using data from three distinct phases of the outbreak: the first three months, the entire duration, and the most recent three months relative to the dataset. In Figures 5(a), 5(b), and 5(c), the agent was trained exclusively on data from the first three months. Initially, the agent recommended strict regulation of both domestic and international policies. However, the number of recommended interventions decreased markedly beginning in mid-March. By late March, the key policy measures were implemented.

Figures 5(a), 5(b), and 5(c) depict the lockdown, travel restriction, and total intensity of both travel restriction and lockdown in the last three months of the COVID-19 pandemic in the UK.

Table 6 shows how the intensity of travel bans and lockdowns, and their combined effect changed over four key dates during the last three months. We can see that travel restrictions tightened compared

Table 5
Comparison of combined intensity measures between the proposed model and the existing model for the first 3 months

	Travel Ban Intensity		Lockdown Intensity		Total Intensity	
Date	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**
2020-01-30	1.78	0.00	1.50	0.00	3.50	0.00
2020-02-15	1.75	0.20	1.50	0.50	3.50	1.00
2020-03-02	1.50	0.50	1.25	1.00	3.00	2.00
2020-03-18	1.00	1.00	1.00	1.25	1.00	3.00

Figure 5
Lockdown, travel ban, and total intensity for the last 3 months: (a) Lockdown for the last 3 months, (b) Travel ban for the last 3 months, (c) Total intensity for the last 3 months

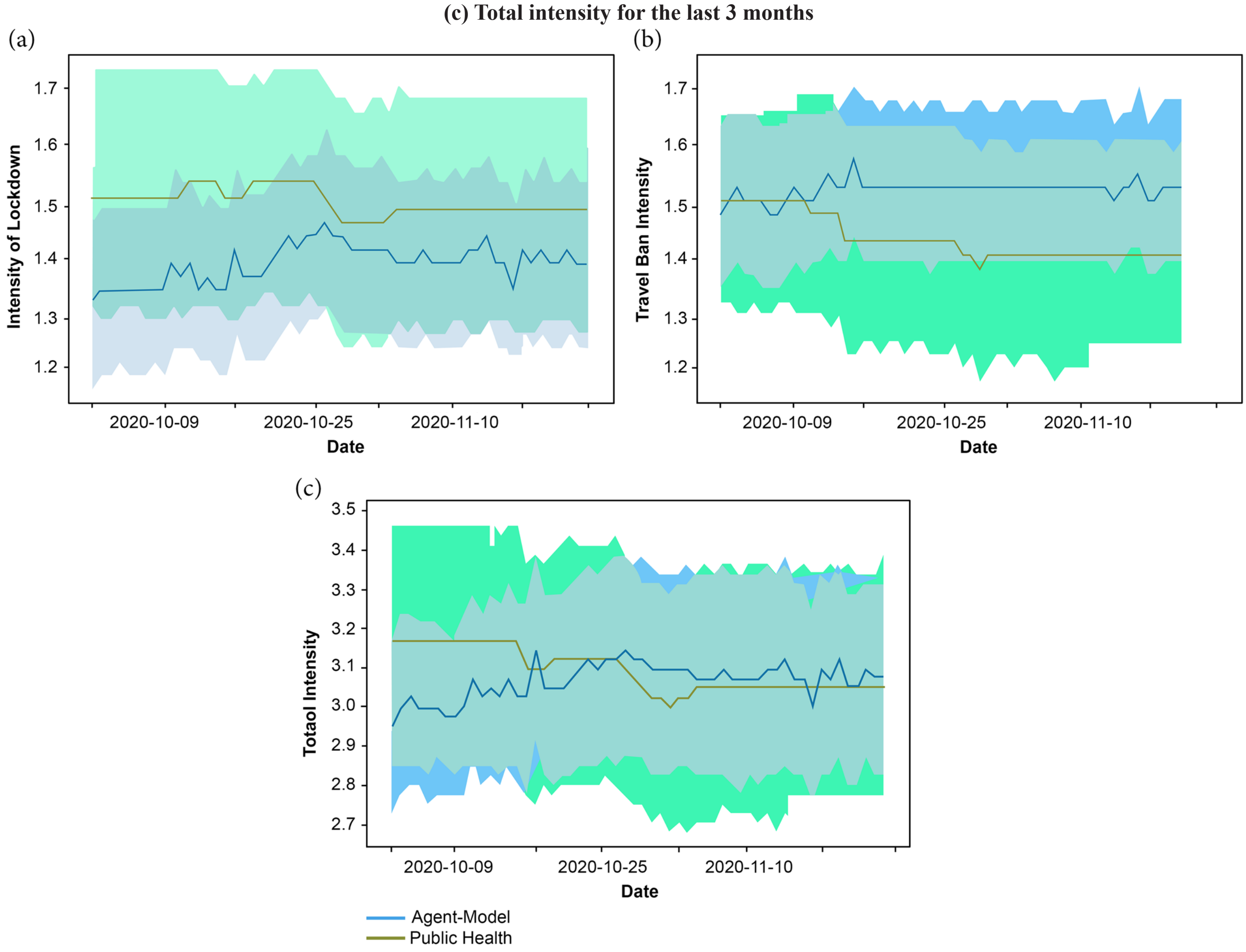


Table 6
Comparison of combined intensity measures between the proposed model and the existing model for the last 3 months

	Travel Ban Intensity		Lockdown Intensity		Total Intensity	
Date	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**
2020-10-09	1.35	1.52	1.63	1.65	2.97	3.18
2020-10-25	1.45	1.55	1.68	1.58	3.15	3.15
2020-11-10	1.40	1.48	1.68	1.55	3.10	3.08

to the government response, and eased at different times. Lockdown measures also varied, becoming stricter or more relaxed depending on the situation. When we look at the total intensity, which combines both types of interventions, it tends to be the same for both the model and government intervention. Overall, the table gives a clear picture of how policies were adjusted over time to respond to the changing nature of the pandemic.

During the subsequent period, the policies generated by the proposed model in the final three months exhibited substantial convergence with the officially implemented public health strategies, as depicted in Figures 5(a), 5(b), and 5(c). Specifically, the recommended lockdown policy demonstrated a slightly lower level of intensity compared to the corresponding public health measures, while the official travel restriction policy maintained a marginally higher level of severity. Figures 6(a), 6(b), and 6(c) further depict the temporal evolution of lockdown intensity and travel restrictions, and the combined severity of both interventions throughout the observation period.

Table 7 summarises the variation in the intensity of travel bans and lockdowns, and their combined effect across four critical dates during the observation period. Although the individual measures exhibit slight offsets, the overall strategy remains largely consistent over time. This table provides a clear representation of how policy interventions were progressively adjusted in response to the evolving dynamics of the pandemic.

The introduction of these regulations was informed by the agent's assessment of epidemic progression. As illustrated in Figure 6(c), the agent consistently recommended implementing lockdowns or travel bans at Level 1, substantially earlier than the official adoption by public health authorities or the government. Specifically, the agent advised initiating at least minimal restrictions in late January, despite

Figure 6
Travel ban, lockdown, and total intensity for the overtime period: (a) Travel ban for the overtime period, (b) Lockdown for the overtime period, (c) Total intensity for the overtime period

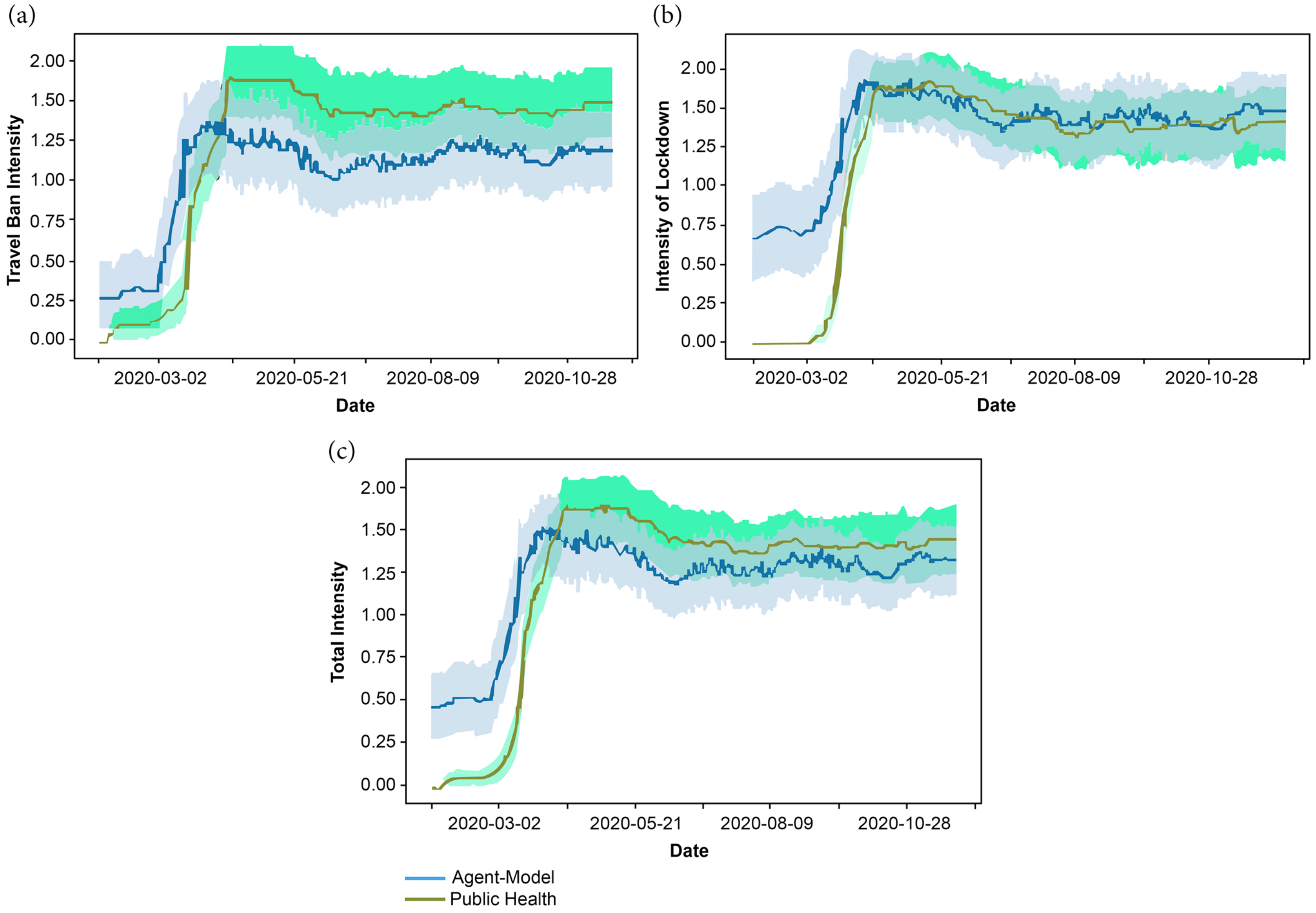


Table 7
Comparison of combined intensity measures between the proposed model and the existing model for the overtime period

	Travel Ban Intensity		Lockdown Intensity		Total Intensity	
Date	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**	**Proposed Model**	**Existing Model**
2020-03-02	0.26	0.00	0.61	0.00	1.00	0.00
2020-05-21	1.25	1.68	1.42	1.41	2.57	3.30
2020-08-09	1.19	1.48	1.39	1.38	2.50	2.52
2020-10-28	1.16	1.49	1.40	1.39	2.55	2.57

the first confirmed case occurring in late March. In contrast, while public health experts advocated for immediate interventions, the agent suggested delaying policy implementation in the absence of exponential case growth, reflecting a more conservative approach to early stage containment.

4.3. Discussion

This study introduces a novel approach for training an agent to determine the optimal timing and severity of lockdown measures and travel bans in the UK. The proposed approach employs DRL and uses a dataset from the Johns Hopkins COVID-19 Data Repository, which includes both worldwide and UK COVID-19 epidemiology data, with a focus on UK data. To identify the most effective course of action for specific states over time, we conducted a temporal analysis of policy implementation across varying levels of crisis intensity. The current investigation employed DRL approaches, focusing on incentive structures and discrete state spaces. In contrast to the measures that were eventually imposed, our algorithm primarily indicated that public health officials take a more lenient approach to lockdown measures and travel bans in the context of the COVID-19 pandemic. Our agent advocated implementing a lockdown during the early stages of the outbreak, when the situation was more serious. Nonetheless, the agent ultimately complied with public health policies. Furthermore, it was in line with recommendations from public health officials to ease the stringency of confinement measures in the final phases of the pandemic. Early implementation of COVID-19 mitigation strategies must be carefully considered, considering the possible economic, social, and health implications. After comparing the first and official implementations, the agent suggested travel limitations as a minimum measure. Contrary to the agent's projections on travel, public health officials have implemented harsher travel restrictions.

The algorithm and findings indicate that the agent does not favour prolonged, high-intensity lockdowns or travel restrictions for public health management in the UK. Figures 4(a), 4(b), and 4(c) support this analysis. Figures 5(a), 5(b), and 5(c) show that the agent initially recommended stringent regulatory measures but gradually relaxed these measures starting in mid-March. These changes in policy intensity are consistent with the algorithm's training on data from the first three months. Figures 6(a), 6(b), and 6(c) illustrate that agents have reached an agreement on the decisions made by public health, which is consistent with the policies proposed throughout the three-month pandemic data analysis in Figures 4(a), 4(b), and 4(c). Inconsistencies were found in travel restrictions. Compared to public health authorities, travel restrictions were more liberal in the second half of the study period. However, the travel restrictions imposed by the agent were slightly stricter than those proposed by the government. This was primarily due to the agent's reliance on information from three months earlier. Implementing the optimal policy recommended by our agent provides greater benefits than adopting a cautious, risk-averse approach from the outset. Such a strategy helps avoid incurring unnecessary costs associated with premature or overly stringent interventions.

The DRL strategy for determining optimal lockdown and travel restrictions may fail due to inconsistent and delayed data. In this approach, instead of using daily case data, a three-day average was used to smooth fluctuations. While this reduces noise, it introduces lag, causing the agent to react too slowly to sudden outbreaks or improvements. Additionally, real-world data may be incomplete or biased, further impairing learning. These limitations can lead to suboptimal or mistimed policy decisions. To address this, integrating real-time data correction techniques, uncertainty-aware models, and hybrid epidemiological-DRL frameworks can enhance responsiveness and robustness in policy optimisation.

4.3.1. Model generalisation and transferability

Although this study was trained solely on the UK COVID-19 dataset, the underlying framework is extensible to other regions and future outbreaks. National differences in demographics, healthcare capacity, and behavioural response can significantly alter pandemic trajectories, potentially limiting direct generalisation. Nevertheless, the reinforcement learning architecture can be adapted using transfer learning or fine-tuning, where the agent initialised on UK data is retrained on new regional data to learn localised response dynamics. In future iterations, federated reinforcement learning can enable decentralised agents trained on data from multiple countries to share policy parameters without exposing sensitive health information. This approach would strengthen the global applicability and robustness of the learned policy strategies beyond a single-country context.

4.3.2. Behavioural and mobility factors

The present framework focuses on epidemiological variables on confirmed cases, recoveries, and deaths, while omitting behavioural and mobility determinants that substantially influence disease spread. Real-world interventions are modulated by public compliance, population movement, and intensity of social interaction. Incorporating such exogenous data, for example, from Google Mobility Reports or Apple Mobility Trends, can allow the agent to capture behavioural feedback loops and improve policy responsiveness. Additionally, explicit modelling of public adherence levels can refine reward attribution, producing more realistic simulations of policy impact.

4.3.3. Dynamic reward adaptation

Another limitation of the current framework is that the reward structure remains static over time. In real pandemic settings, policy priorities evolve with vaccination rollout, the emergence of viral variants, and changing public compliance. A future adaptive reward system can incorporate time-varying weights that emphasise different objectives such as vaccination coverage or variant transmissibility, thereby allowing the agent to learn phase-specific strategies. Implementing non-stationary or meta-reinforcement learning mechanisms would enable the reward function to evolve alongside the pandemic context.

4.3.4. Comparison with existing AI-based policy optimisation model

The proposed dueling DQN framework complements earlier reinforcement learning approaches such as DDPG and Proximal Policy Optimization. While continuous-action algorithms such as DDPG offer fine-grained control, they can suffer from instability and overestimation in rapidly changing epidemic environments. The discrete-action D3QN adopted here provides a more interpretable and stable alternative for policy discretisation, aligning naturally with categorical government interventions (e.g., partial or full lockdown). Future comparative experiments can systematically benchmark these models using identical epidemiological datasets.

4.3.5. Interpretability and policy transparency

Trust in algorithm-based decision-making is essential for adoption in policymaking. Although the present model functions as a black-box optimiser, future versions can integrate interpretable-AI tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to attribute each policy recommendation to specific epidemiological drivers. Visualising action-value saliency maps across time can further help policymakers understand why the agent prefers a certain level of restriction, thereby improving transparency and acceptability of AI-driven policy guidance.

5. Conclusion

In summary, the proposed DRL framework demonstrates the feasibility of data-driven policy optimisation for epidemic control. While the present analysis focused on the UK, the architecture is inherently generalisable and can be extended to other regions through transfer learning or multi-country federated reinforcement schemes. Future work will incorporate dynamic rewards, mobility and behavioural variables, and uncertainty quantification to improve realism and reliability. By coupling the DRL agent with epidemiological simulators and interpretable analytics, this framework can evolve into a practical decision-support system for pandemic preparedness and real-time outbreak management.

Ethical Statement

This study did not require formal ethical approval, as per the University of Salford's research ethics policy, because it is a computational modeling study based exclusively on publicly available, aggregated, and fully anonymized secondary datasets. No human participants were recruited, no interventions were performed, and no identifiable personal data were collected or analyzed. According to the standard UK research governance and institutional ethics guidelines, research that rely solely on secondary public data and simulation methods are exempt from formal IRB/ethics committee review.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Johns Hopkins COVID-19 Data Repository at https://github.com/CSSEGISandData/COVID-19, World Health Organization at https://www.who.int/emergencies/diseases/novel-coronavirus-2019, Our World In Data COVID-19 pandemic data at https://ourworldindata.org/coronavirus, and Wikipedia at https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic.

Author Contribution Statement

Raphael Ibraimoh: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Project administration. **Mohammed Saraee:** Validation, Writing – review & editing, Supervision, Project administration. **Kaveh Kiani:** Writing – review & editing, Supervision, Project administration. **Danial Saraee:** Writing – review & editing.

References

[1] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

[2] Moy, N., Antonini, M., Kyhlstedt, M., Fiorentini, G., & Paolucci, F. (2023). Standardising policy and technology responses in the immediate aftermath of a pandemic: A comparative and conceptual framework. *Health Research Policy and Systems*, *21*(1), 10. https://doi.org/10.1186/s12961-022-00951-x

[3] Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., ... & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, *5*(4), 529–538. https://doi.org/10.1038/s41562-021-01079-8

[4] Al Awaidy, S. T., Ghazy, R. M., Mahomed, O., & Wesonga, R. (2023). The impact of non-pharmaceutical interventions (NPIs) on communicable diseases. *Frontiers in Public Health*, *11*, 1270709. https://doi.org/10.3389/fpubh.2023.1270709

[5] Rietveld, J., Hobson, T., Mani, L., Avin, S., & Sundaram, L. (2024). The UK's pandemic preparedness and early response to the COVID-19 pandemic. *Global Public Health*, *19*(1), 2415499. https://doi.org/10.1080/17441692.2024.2415499

[6] Bærøe, K., Árnason, V., Jansen, M., Yamin, A. E., Ruano, A. L., & Davis, A. P. (2025). Pandemic and crisis Preparedness and Response: Conceptualizing cultural, social and Political Drivers of trustworthiness and collective action. *Public Health Ethics*, *18*(2), phaf004. https://doi.org/10.1093/phe/phaf004

[7] Weets, C. M., Carlson, C. J., Robertson, H., Toole, K., McGivern, L., Graeden, E., & Katz, R. (2025). The WHO disease outbreak news during the Covid-19 pandemic. *PLOS Global Public Health*, *5*(1), e0004025. https://doi.org/10.1371/journal.pgph.0004025

[8] Cooper, A., Lewis, R., Gal, M., Joseph-Williams, N., Greenwell, J., Watkins, A., ... & Edwards, A. (2024). Informing evidence-based policy during the COVID-19 pandemic and recovery period: Learning from a national evidence centre. *Global Health Research and Policy*, *9*(1), 18. https://doi.org/10.1186/s41256-024-00354-1

[9] Lee, J. M., Jansen, R., Sanderson, K. E., Guerra, F., Keller-Olaman, S., Murti, M., ... & Khan, Y. (2023). Public health emergency preparedness for infectious disease emergencies: A scoping review of recent evidence. *BMC Public Health*, *23*(1), 420. https://doi.org/10.1186/s12889-023-15313-7

[10] Fisher, A., Roberts, A., McKinlay, A. R., Fancourt, D., & Burton, A. (2021). The impact of the COVID-19 pandemic on mental health and well-being of people living with a long-term physical health condition: A qualitative study. *BMC Public Health*, *21*(1), 1801. https://doi.org/10.1186/s12889-021-11751-3

[11] Khadilkar, H., Ganu, T., & Seetharam, D. P. (2020). Optimising lockdown policies for epidemic control using reinforcement learning: An AI-driven control approach compatible with existing disease and network models. *Transactions of the Indian National*

Academy of Engineering, *5*(2), 129–132. https://doi.org/10.1007/s41403-020-00129-3

[12] Salamanca-Buentello, F., Katz, R., Silva, D. S., Upshur, R. E., & Smith, M. J. (2024). Research ethics review during the COVID-19 pandemic: An international study. *PLOS One*, *19*(4), e0292512. https://doi.org/10.1371/journal.pone.0292512

[13] Bahakel, H., Waghmare, A., & Madan, R. P. (2024). Impact of respiratory viral infections in transplant recipients. *Journal of the Pediatric Infectious Diseases Society*, 13, S39–S48. https://doi.org/10.1093/jpids/piad094

[14] Tian, H., Liu, Y., Li, Y., Wu, C. H., Chen, B., Kraemer, M. U., ... & Dye, C. (2020). An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, *368*(6491), 638–642. https://doi.org/10.1126/science.abb6105

[15] Dunn, K., Hurwitz, H. H., Toledo, J. P., Schwaber, M. J., Chu, M., Chou, R., ... & Baller, A. (2024). Summary of WHO infection prevention and control guideline for covid-19: Striving for evidence based practice in infection prevention and control. *BMJ*, 385. https://doi.org/10.1136/bmj.q645

[16] Kwak, G. H., Ling, L., & Hui, P. (2021). Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. *PLOS One*, *16*(5), e0251550. https://doi.org/10.1371/journal.pone.0251550

[17] Oh, J., Farquhar, G., Kemaev, I., Calian, D. A., Hessel, M., Zintgraf, L., ... & Silver, D. (2025). Discovering state-of-the-art reinforcement learning algorithms. *Nature*, 1–2. https://doi.org/10.1038/s41586-025-09761-x

[18] Liu, K. (2022). Optimal control policy on COVID-19: An empirical study on lockdown and travel restriction measures using reinforcement learning. *International Journal of High School Research*, *4*(3), 60–68. https://doi.org/10.36838/v4i3.11

[19] McDermid, P., Craig, A., Sheel, M., Blazek, K., Talty, S., & Seale, H. (2022). Examining the psychological and financial impact of travel restrictions on citizens and permanent residents stranded abroad during the COVID-19 pandemic: International cross-sectional study. *BMJ Open*, *12*(5), e059922. https://doi.org/10.1136/bmjopen-2021-059922

[20] Zhu, J., Wu, F., & Zhao, J. (2021). An overview of the action space for deep reinforcement learning. In *Proceedings of the 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, 1–10. https://doi.org/10.1145/3508546.3508598

[21] Al-Hamadani, M. N., Fadhel, M. A., Alzubaidi, L., & Harangi, B. (2024). Reinforcement learning algorithms and applications in healthcare and robotics: A comprehensive and systematic review. *Sensors*, *24*(8), 2461. https://doi.org/10.3390/s24082461

[22] Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, *368*(6489), 395–400. https://doi.org/10.1126/science.aba9757

How to Cite: Ibraimoh, R., Saraee, M., Kiani, K., & Saraee, D. (2025). Optimal Policy Strategy for Pandemic Outbreak Control: A Deep Reinforcement Approach. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52026822