Deep reinforcement learning for large-scale epidemic control

Pieter J.K. Libin
Vrije Universiteit Brussel
Brussels, Belgium
pieter.libin@vub.be

Arno Moonens
Vrije Universiteit Brussel
Brussels, Belgium
arno.moonens@vub.be

Timothy Verstraeten
Vrije Universiteit Brussel
Brussels, Belgium
tiverstr@vub.be

Fabian Perez-Sanjines
Vrije Universiteit Brussel
Brussels, Belgium
fperezsa@vub.be

Niel Hens
Hasselt University
Hasselt, Belgium
niel.hens@uhasselt.be

Philippe Lemey
KU Leuven
Leuven, Belgium
philippe.lemey@kuleuven.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

ABSTRACT

Epidemics of infectious diseases are an important threat to public health and global economies. Yet, the development of prevention strategies remains a challenging process. For this reason, we investigate a deep reinforcement learning approach to automatically learn prevention strategies in an epidemiological model, in the context of pandemic influenza. To this end, we construct a new epidemiological meta-population model, with 379 patches, that balances between model complexity and computational efficiency such that the use of reinforcement learning techniques becomes attainable. First, we set up a ground truth such that we can evaluate the performance of the "Proximal Policy Optimization" algorithm to learn in a single district of this epidemiological model. Next, we consider a larger scale problem, by conducting an experiment where we aim to learn a joint policy to control the districts in a community of 11 tightly coupled districts, for which no ground truth can be established. This experiment shows that deep reinforcement learning can be used to learn mitigation policies in complex epidemiological models with a large state space. Moreover, through this experiment, we demonstrate that there can be an advantage to consider collaboration between districts when designing prevention strategies.

KEYWORDS

multi-agent system, epidemic control, pandemic influenza, deep reinforcement learning

1 INTRODUCTION

Epidemics of infectious diseases are an important threat to public health and global economies. The most efficient way to combat epidemics is through prevention. To develop prevention strategies and to implement them as efficiently as possible, a good understanding of the complex dynamics that underlie these epidemics is essential. To properly understand these dynamics, and to study emergency scenarios, epidemiological models are necessary. Such models enable us to make predictions and to study the effect of prevention strategies in simulation. The development of prevention strategies, which need to fulfil distinct criteria (i.a., prevalence, mortality, morbidity, cost), remains a challenging process. For this reason, we investigate a deep reinforcement learning (RL) approach to automatically learn prevention strategies in an epidemiological model. The use of model-free deep reinforcement learning is particularly interesting, as it allows us to set up a learning environment in a complex epidemiological setting (i.e., large state space and non-linear dependencies) while imposing few assumptions on the policies to be learned. In this work, we conduct our experiments in the context of pandemic influenza, where we aim to learn optimal school closure policies to mitigate the epidemic.

Pandemic preparedness is important, as influenza pandemics have made many victims in the (recent) past [35] and the ongoing COVID-19 epidemic is yet another reminder of this fact [51]. Contrary to seasonal influenza epidemics, an influenza pandemic is caused by a newly emerging virus strain that can become pandemic by spreading rapidly among naive human hosts (i.e., human hosts with no prior immunity) worldwide [35]. This means that at the start of the pandemic no vaccine will be available and it will take several months before vaccine production can commence [45]. For this reason, learning optimal strategies of non-therapeutic intervention measures, such as school closure policies, is of great importance to mitigate pandemics [32].

To meet this objective, we consider a reinforcement learning approach. However, as the state-of-the-art of reinforcement learning techniques require many interactions with the environment in order to converge, our first contribution entails a realistic epidemiological model that still has a favourable computational performance.

Specifically, we construct a meta-population model that consists out of a set of 379 interconnected patches, where each patch corresponds to an administrative region in Great Britain and is internally represented by an age-structured stochastic compartmental model. To conduct our experiments, we establish a Markov Decision Process with a state space that directly corresponds to our epidemiological model, an action space that allows us to open and close schools on a weekly basis, a transition function that follows the epidemiological model's dynamics, and a reward function that is targeted to the objective of reducing the attack rate (i.e., the proportion of the population that was infected). In this work, we will use "Proximal Policy Optimization" (PPO) [39] to learn the school closure policies.

First, we set up an experiment in an epidemiological model that covers a single administrative district. This setting enables us to specify a ground truth that allows us to empirically assess the performance of the policies learned by PPO. In this analysis, we consider different values for the basic reproductive number R_0 (Definition 1.1) and the population composition (i.e., proportion of adults, children, elderly, adolescents) of the district. Both parameters induce a significant change of the epidemic model's dynamics.

Definition 1.1 (Basic reproductive number). The basic reproductive number, R_0, is the number of infections that is, on average, generated by one single infected individual that is placed in an otherwise fully susceptible population.

Through these experiments, we demonstrate the potential of deep reinforcement learning algorithms to learn policies in the context of complex epidemiological models, opening the prospect to learn in even more complex stochastic models with large action spaces. In this regard, we consider a large scale setting where we examine whether there is an advantage to consider the collaboration between districts when designing school closure policies. We conduct an experiment in our epidemiological model with 379 districts and attempt to learn a joint policy to control the districts in the Cornwall-Devon community, a set of 11 tightly coupled districts. To this end, we assign an agent to each of the 11 districts of the Cornwall-Devon community and use a reinforcement learning approach to learn a joint policy. We compare this joint policy to a non-collaborative policy (i.e., aggregated independent learners).

2 RELATED WORK

The closing of schools is an effective way to limit the spread of an influenza pandemic [32]. For this reason, the use of school closures as a mitigation strategy has been explored in variety of modelling studies [4, 5, 8, 10, 16, 19–21, 34], of which the work by Germann et al. [16] is the most recent and comprehensive study.

The concept to learn dynamic policies by formulating the decision problem as a Markov decision process (MDP) was first introduced in [47]. The proposed technique was used to investigate dynamic tuberculosis case-finding policies in HIV/tuberculosis coepidemics [48]. Later, the technique was extended towards a method to include cost-effectiveness in the analysis [49], and applied to investigate mitigation policies (i.e., school closures and vaccines) in the context of pandemic influenza in a simplified epidemiological model. On the one hand, the work presented in [47, 49] uses a policy iteration algorithm to solve the MDP. On the other hand, the use of on-line reinforcement learning techniques (e.g., TD-learning, policy gradient) has only been explored to a limited extent[1], and motivated us to do the work presented in this manuscript. Note that the "Deep Q-networks" algorithm was recently used to investigate culling and vaccination in farms in a simple individual-based model to delay the spread of viruses in a cattle population [36]. However, to our best knowledge, the work presented in this manuscript is the first attempt to use deep reinforcement learning algorithms directly on a complex meta-population model. Furthermore, we experimentally validate the performance of these algorithms using a ground truth, in a variety of model settings (i.e., different census compositions and different R_0's). This is the first validation of this kind and it demonstrates the potential of on-line deep reinforcement learning techniques in the context of epidemic decision making. Finally, we present a novel approach to investigate how intervention policies can be improved by enabling collaboration between different geographic districts, by formulating the setting as a multi-agent problem, and by solving it using deep multi-agent reinforcement learning algorithms.

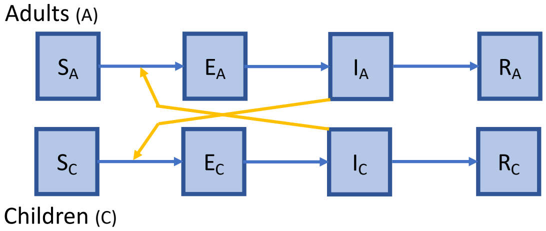

Figure 1: We depict an age-structured SEIR model that considers two age groups (i.e., adults and children). This model consists out of two SEIR models, one for each age group, that are connected to represent mixing between the age groups (yellow arrows). Note that it is also possible to mix within the age groups. Note that we use two age groups in this figure to provide a clear visualization of the model. In our actual model, we consider four different age groups.

3 EPIDEMIOLOGICAL MODEL

We construct a meta-population model that consists out of 379 patches, where each patch represents one administrative region in Great Britain. Great Britain consists out of three countries with the following administrative regions: 325 districts in England, 22 unitary authorities in Wales and 32 council areas in Scotland. Each patch consists out of a stochastic age-structured compartmental model, which we describe in sub-section 3.1, and the different patches are connected via a mobility model, as detailed in sub-section 3.2. In sub-section 3.3 we discuss how we validate and calibrate the model. We analyse the model's computational complexity and discuss the model's performance in the Supplementary Information.

3.1 INTRA-PATCH MODEL

We consider a stochastic SEIR compartmental model from which we sample trajectories. We first describe the model in terms of ordinary differential equations (i.e., a deterministic representation) that we than transform to stochastic differential equations [1] to make a stochastic evaluation possible. An SEIR model divides the population in a susceptible, exposed, infected and recovered compartment, and is commonly used to model influenza epidemics [10]. More specifically, we consider an age-structured SEIR model (see Figure 1 for a visualization) with a set of n disjoint age groups [10, 14]. This model is formally described by this system of ordinary differential equations (ODEs), defined for each age group i:

$$\begin{aligned}\frac{\mathrm{d}S_i}{\mathrm{d}t} &= -\phi_i(t)S_i(t)\\ \frac{\mathrm{d}E_i}{\mathrm{d}t} &= \phi_i(t)S_i(t) - \zeta E_i(t)\\ \frac{\mathrm{d}I_i}{\mathrm{d}t} &= \zeta E_i(t) - \gamma I_i(t)\\ \frac{\mathrm{d}R_i}{\mathrm{d}t} &= \gamma I_i(t).\end{aligned} \tag{1}$$

[1] The recent perspective report by [50] reached the same conclusion.

Every susceptible individual in age group i is subject to an age-specific and time-dependent force of infection:

$$\phi_i(t) = \sum_{j=1}^{n} \beta M_{ij}(t) \frac{I_j(t)}{N_j(t)}, \tag{2}$$

which depends on:

- The probability of transmission β when a contact occurs.
- The time-dependent contact matrix M, where $M_{ij}(t)$ is the average frequency of contacts that an individual in age group i has with an individual in age group j [15].
- The frequency that contacts with infected individuals (in age group j) occur: $I_j(t)/N_j(t)$

Once exposed, individuals move to the infected state according to the latency rate ζ. Individuals recover from infection (i.e., get better or die) at a recovery rate γ.

We omit vital dynamics (i.e., births and deaths that are not caused by the epidemic) in this SEIR model, as the epidemic's time scale is short and we therefore assume that births and deaths will have a limited influence on the epidemic process [41]. Therefore, at any time:

$$N_i(t) = S_i(t) + E_i(t) + I_i(t) + R_i(t), \tag{3}$$

where the total population size N_i corresponds to age-specific census data. Our model considers 4 age groups: children (0-4 years), adolescents (5-18 years), adults (19-64 years) and elderly (65 years and older).

Note that the contact frequency $M_{ij}(t)$ is time-dependent, in order to model school closures, i.e., a different contact matrix is used for school term and school holiday. Following [10], we consider *conversational contacts*, i.e., contacts for which physical touch is not required. As we aim to model the effectiveness of school closure interventions, we use the United Kingdom contact matrices presented in [10], which encodes a contact matrix for both school term and school holiday. These contact matrices are the result of an internet-based social contact survey completed by a cohort of participants [10]. The contact matrices encode for the same age groups as mentioned before: children, adolescents, adults and elderly.

We defined the SEIR model in terms of a system of ordinary differential equations which implies a deterministic evaluation of the system. However, for predictions, stochastic models are preferred, as they to account for stochastic variation and allow us to quantify uncertainty [25]. In order to sample trajectories from this set of differential equations, we transform the system of ordinary differential equations (ODEs) to a system of stochastic differential equations (SDEs), using the transformation procedure presented by Allen et al. [1]. This procedure introduces stochastic noise to the system by adding a Wiener process to each transition in the ODE. We evaluate the SDE at discrete time steps using the Euler-Maruyama approximation method [1].

Each compartmental model is representative of one of the administrative districts and as such the compartmental model is parametrised with the census data of the respective district, i.e., population counts stratified by age groups. We use the 2011 United Kingdom census data made available by NOMIS (https://www.nomisweb.co.uk). We present more details on the census data in the Supplementary Information[2].

3.2 BETWEEN-PATCH MODEL

Our model, that is comprised of a set of connected SEIR patches, is inspired by the recent BBC pandemic model [27]. The BBC pandemic model was in its turn motivated by the model presented in [17].

At each time step, our model checks whether a patch p becomes infected. This is modulated by the patch's force of infection, which combines the potential of the infected patches in the system, weighted by a mobility model, that represents the commuting of adults between the different patches:

$$\mathring{\phi}_p(t) = \sum_{p' \in \mathcal{P}} \mathcal{M}_{p'p} \cdot \beta \cdot \left(S_p^{\mathrm{A}}(t)\right)^{\mu} \cdot \mathcal{I}_{p'}(t), \tag{4}$$

where $\mathcal{P}$ is the set of patches in the model, $\mathcal{M}_{p'p}$ is the mobility flux between patch p' and p, β is the probability of transmission on a contact, $S_p^{\mathrm{A}}(t)$ is the susceptible population of adults in patch p at time t and its contribution is modulated by parameter μ (range in $[0, 1]$), and $\mathcal{I}_{p'}(t)$ is the infectious potential of patch p' at time t. We define this infectious potential as,

$$\mathcal{I}_{p'}(t) = I_{p'}^{\mathrm{A}}(t) \cdot M_{\mathrm{AA}}, \tag{5}$$

where $I_{p'}^{\mathrm{A}}(t)$ is the size at time t of the infectious adult population in patch p' and M_{AA} is the average number of contacts between adults, as specified in the contact matrix (see sub-section 3.1) This infectious potential corresponds to infectious adult individuals that commute from district p' to district p.

$\mathcal{M}$ is a matrix based on the mobility dataset provided by NOMIS[3]. This dataset describes the amount of commuting between the districts in Great Britain.

In general, this between-patch model is constructed from first principles i.e., census data, a mobility model, the number of infected individuals and the transmission potential of the virus. However, for the parameter μ that modulates the contribution of the susceptibles in the receptive patch (while it is commonly used in literature [11, 17, 26]) no such intuition is readily available. Therefore, this parameter is typically fitted to match the properties of the epidemic that is under investigation [11, 17, 26]. We will calibrate this parameter such that it can be used for a range of R_0 values, as detailed in the next sub-section.

Given this time-dependent force of infection, we model the event that a patch becomes infected with a non-homogeneous Poisson process [44]. As the process' intensity depends on how the model (i.e., the set of all patches) evolves, we cannot sample the time at which a patch becomes infected a priori. Therefore, we determine this time of infection using the time scale transformation algorithm [6]. Details about this procedure can be found in Supplementary Information. Following Klepac et al. [27], we assume that a patch will become infected only once.

By using this time scale transformation algorithm and evaluating the stochastic differential equation at discrete time steps, we

[2] http://plibin-vub.github.io/epidemics-rl/supplement.pdf

[3] We use the NOMIS WU03UK dataset that was released in 2011.

produced a model with favourable performance, i.e., we can run about 2 simulation runs per second on a MacBook Pro.

3.3 CALIBRATION AND VALIDATION

Our objective is to construct a model that is representative for contemporary Great Britain with respect to population census and mobility trends. This model will be used to study school closure intervention strategies for future influenza pandemics. While in many studies [11, 17, 26] a model is created specifically to fit one epidemic case, we aim for a model that is robust with respect to different epidemic parameters, most importantly R_0, the basic reproduction number.

To validate our model according to these goals, we conduct two experiments. In the first experiment, we compare our patch model to an SEIR compartmental model that uses the same contact matrix and age structure, but with homogeneous spatial mixing (i.e., no spatial structure). While we do not expect our model to behave exactly like the compartmental model, as the patches and the mobility network that connects them induces a different dynamic, we do observe similar trends with respect to the epidemic curve and peak day. This experiment also enables us to calibrate the μ parameter. We present a detailed description of this analysis and report the results in Supplementary Information. In the second experiment we show that our model is able to reproduce the trends that were observed during the 2009 influenza pandemic, commonly known as the swine-origin influenza pandemic (A(H1N1)v2009), that originated in Mexico. The 2009 influenza pandemic in Great Britain is an interesting case to validate our model for two reasons. Firstly, the pandemic occurred quite recently and thus our model's census and mobility scheme are a good fit, as both the datasets on which we base our census and mobility model were released in 2011. Secondly, due to the time when the virus entered Great Britain, the summer holiday started 11 weeks after the emergence of the epidemic. The timing of the holidays had a severe impact on the progress of the epidemic and resulted in a epidemic curve with two peaks. This characteristic epidemic curve enables us to demonstrate the predictive power of our age-structured contact model with support for school closures. In Figure 2, we show a set of model realisations in conjunction with the symptomatic case data, which shows that we were able to closely match the epidemic trends observed during the British pandemic in 2009 (details on this case study in the Supplementary Information). Note that our model reports the number of infections while the British Health Protection Agency only recorded symptomatic cases. Therefore we scale the epidemic curve with a factor of $\frac{1}{4}$. This large number of asymptomatic cases produced by our model is in line with earlier serological surveys [33] and with previous modelling studies [28].

4 LEARNING ENVIRONMENT

In order to apply reinforcement learning, we construct an MDP based on the epidemiological model that we introduced in the previous section. This epidemiological model consists out of patches that correspond to administrative regions.

We have an agent for each patch that we attempt to control, and for each agent we have an action space $\mathcal{A} = \{\text{open}, \text{close}\}$ that allows us to open and close schools for one week. Each agent has a predefined budget $\mathscr{b}$ of school closure actions it can execute. Once this budget is depleted, executing a close action will default to executing an open action. We refer to the remaining budget at time t as $\mathscr{b}^{(t)}$.

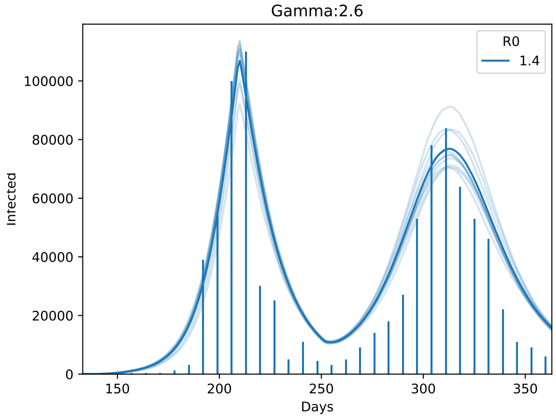

Figure 2: We show that our model (blue epidemic curves) is able to match the trends observed in the British pandemic of 2009 (the vertical bars represent the number of infected individuals that was recorded during the epidemic). We show 10 stochastic trajectories.

For each patch, we consider a state space that combines the state of the SEIR model and the remaining budget of school closures $\mathscr{b}_p^{(t)}$. For the SEIR model, we have 16 state variables (i.e., $\mathbb{R}^{16}$), as we have an SEIR model (4 state variables) for each of the four age groups. The remaining school closure budget is encoded as an integer, resulting in a combined state space of 17 variables. We refer to the state space of one patch p, that thus combines the epidemiological states and the budget, as $\mathcal{S}_p$. The state space of the MDP $\mathcal{S}$ corresponds to the aggregation of the state space of each patch that we attempt to control:

$$\bigtimes_{p \in \mathcal{P}^c} \mathcal{S}_p, \tag{6}$$

where $\mathcal{P}^c$ is the set of patches that we control.

The transition probability function $T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ evolves the state space to the next week in the epidemic, taking into account the school closure actions that were chosen, using the epidemiological dynamics as defined in the previous section.

To reduce the attack rate, we consider an immediate reward function that quantifies the number of susceptible individuals lost at time step t:

$$R_{\text{AR}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = S(\mathbf{s}) - S(\mathbf{s}'), \tag{7}$$

where $S(.)$ is the function that determines the total number of susceptible individuals given the state of the epidemiological model.

For PPO, we use both a policy and value network. The policy network accepts the state of the epidemiological model as input (details in Section 4) and the output of the network contains 1 unit, which is passed through a sigmoid activation function. This output thus represents the probability of keeping the schools open in the district. Every hidden layer in the PPO network uses the hyperbolic tangent activation function. The value network has the same architecture as the policy network, with the exception that the output is not passed through an activation function. We will refer to this settings throughout this work as the single-district PPO agent.

PPO's hyper-parameters are tuned (hyper-parameter values in Supplementary Information) on a single-district (i.e., the Greenwich district) learning environment with $R_0 = 1.8$. To this end, we performed a hyper-parameter sweep using Latin hypercube sampling ($n = 1000$) [40].

We conduct two kinds of experiments: in the context of a single district and in the context of the Great Britain model that combines all 379 districts. We consider two values for the reproductive number, i.e., $R_0 = \{1.8, 2.4\}$, to investigate the effect of distinct reproductive numbers. $R_0 = 1.8$ represents an epidemic with moderate transmission potential [12] and $R_0 = 2.4$ represents an epidemic with high transmission potential [31]. We investigate the effect of different school closure budgets, i.e., $\mathscr{b} = \{2, 4, 6\}$ weeks. The epidemic is simulated for a fixed number of weeks, chosen beforehand, to ensure that the epidemic fades out after its peak. Following Baguelin et al. [2], we use a latent period of one day ($\zeta = \frac{1}{1}$) and an infectious period of 1.8 days ($\gamma = \frac{1}{1.8}$).

5 COMPARE PPO TO THE GROUND TRUTH

We now establish a ground truth for different population compositions, i.e., the proportion of the different age groups in a population. We will use this ground truth to empirically validate that PPO converges to the appropriate policy.

To establish this ground truth[4], first consider that when we deal with a single district, we can approach the 'average' behaviour of the model by removing the stochastic terms from the differential equations. Hence, for a particular parameter configuration (i.e., district, R_0, γ, ζ), the model will always produce the same epidemic curve. This means that the state space of this deterministic epidemic model directly corresponds to the time of the epidemic. For an epidemic that spans w weeks, we can formulate a school closure policy as a binary number with w digits, where the digit at position i signifies whether schools should be open (1) or closed (0) during the i-th week. For short-lived epidemics, such as influenza epidemics, we can enumerate these policies and evaluate them in our model (i.e., using exhaustive policy search). Note that, in the epidemiological models that we consider, the epidemic spans no more than 25 weeks, and thus exhaustive search is possible.

In this analysis, we consider different values for the basic reproductive number R_0 and the population composition of the district, both parameters that induce a significant change of the epidemic model's dynamics. To this end, we select 10 districts that are representative of the population heterogeneity in Great Britain: one district that is representative for the average of this census distribution and a set of nine districts that is representative for the diversity in this census distribution. Details on this selection procedure can be found in the Supplementary Information.

To evaluate PPO with respect to the ground truth, we repeat the experiment for which we established a ground truth (i.e., $R_0 \in \{1.8, 2.4\}$, 10 districts and $\mathscr{b} \in \{2, 4, 6\}$) and learn a policy using PPO, in the stochastic epidemic model. For each experimental setting (i.e., the combination of a district, an R_0 value, and a school closure budget $\mathscr{b}$), we run PPO 5 times (5 trials), to asses the variance of the learning curve. Each PPO trial is run for $5 \cdot 10^5$ time steps. We show the learning curves for the district that is representative for the average of the census distribution (i.e., the Barnsley district in England), with $R_0 = 2.4$ in Figure 3, for the other settings we report similar learning curves in the Supplementary Information.

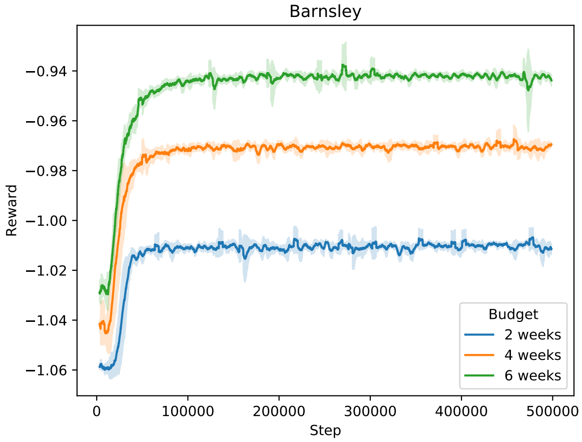

Figure 3: PPO learning curves for the Barnsley district with $R_0 = 2.4$ for the three school closure budgets $\mathscr{b} = \{2, 4, 6\}$.

To compare each of the learned policies to its ground truth (one for each district), we take the learned policy and apply it 1000 times in the stochastic model, which results in a distribution over model outcomes (i.e., attack rate improvement: the difference between the attack rate produced by the model and the baseline when no schools are closed). We then compare this distribution to the attack rate improvement that was recorded for the ground truth. We show these results, for the setting with a school closure budget of 6 weeks and $R_0 = 2.4$, in Figure 4, and for the other settings in Supplementary Information. These results show that PPO learns a policy that matches the ground truth for all districts and combinations of R_0 and $\mathscr{b}$.

Note that for these experiments, we use the same hyper-parameters for PPO that were introduced in Section 4. This demonstrates that, for different values of R_0 and for different census compositions (which induce a significant change in dynamics in the epidemic model) these hyper-parameters work well. This indicates that these hyper-parameters are adequate to be used for different variations of the model.

In this section, we compare a proxy to the ground truth (that has been found through an exhaustive policy search) to a policy learned by PPO, a deep reinforcement learning algorithm. This allows us to empirically validate that PPO converges to the optimal policy. This experimental validation is important, as it demonstrates the potential of deep reinforcement learning algorithms to learn policies in the context of complex epidemiological models. This indicates that it is possible to learn in even more complex stochastic models

[4] Note that this is a proxy to the ground truth, as we use a deterministic version of the model.

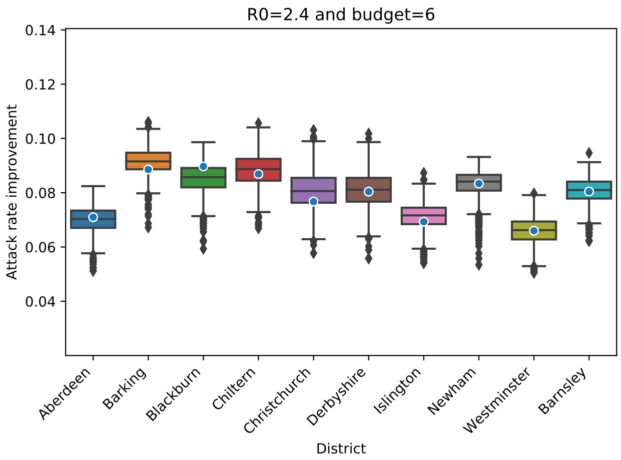

Figure 4: We compare the PPO results to the ground truth for $R_0 = 2.4$ and $b = 6$. Per district, we show a box plot that denotes the outcome distribution that was obtained by simulating the policy learned by PPO 1000 times. On top of this box plot, we show the ground truth, as a blue dot.

with large action spaces, for which it is impossible to compute a proxy to the ground truth. In Section 7, we investigate such a setting, where we aim to learn a joint policy for a set of agents, using deep multi-agent reinforcement learning.

6 FINDING COMMUNITIES

To investigate the collaborative nature of school closure policies, we apply deep multi-agent reinforcement learning algorithms. In our model, we have 379 agents, one for each district, as agents represent the district for which they can control school closure. As the current state-of-the-art of deep multi-agent reinforcement learning algorithms is limited to deal with about 10 agents [22], we thus need to partition our model into smaller groups of agents, such that deep multi-agent reinforcement learning algorithms become feasible.

To this end, we consider the mobility matrix $\mathcal{M}$ and define a directed commute graph for $\mathcal{M}_{ij} \geq 0$ (Definition 6.1).

Definition 6.1 (Commute graph). For a commuting matrix $\mathcal{M}$ that describes the mobility flux between a set of districts $\mathcal{D}$, we define a commute graph,

$$G_c = \langle V_c, A_c \rangle, \tag{8}$$

where V_c is the set of vertices, with a vertex for each of the districts in $\mathcal{D}$, and A_c is the adjacency matrix that specifies the vertices that are connected:

$$(A_c)_{ij} = \begin{cases} 1, & \mathcal{M}_{ij} > 0 \\ 0, & \mathcal{M}_{ij} = 0 \end{cases} \tag{9}$$

Each pair of connected vertices i and j has a weight $\mathcal{M}_{ij}$.

To detect communities in the commute graph, we used the Leiden algorithm [42], an algorithm that searches for communities that maximize the network modularity [29]. We found a partition of which we demonstrated the robustness (p-value ≤ 0.001) using a bootstrapping approach presented by Radivojević and Grujić [37]. Furthermore, by rendering this partition on top of the map of Great Britain, as is shown in Figure 5, we show that the districts belonging to the same community are close to each other geographically, as we would expect. Moreover, when we overlay the NUTS-2 administrative regions[5] on the partitioning (Figure 5), we observe that our partitioning scheme mostly overlaps with the NUTS-2 regions, which indicates that the Leiden algorithm produces a sensible partitioning.

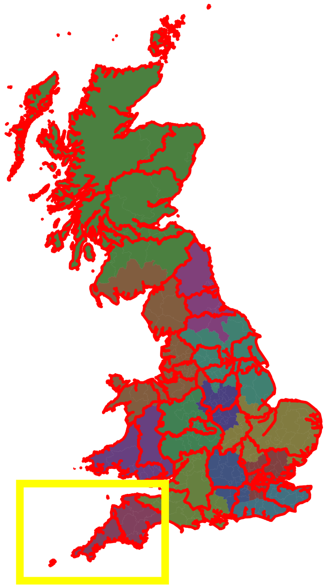

Figure 5: We show the communities, that resulted from applying the Leiden algorithm, on the map of Great Britain. We show all administrative districts colour-coded by the community they belong to and the add the borders of the NUTS-2 administrative regions on top of this map. We annotate the Cornwall-Devon community with a yellow rectangle.

We conduct our multi-agent reinforcement learning experiments in the community with 11 districts, to which we will refer as the Cornwall-Devon community (see Figure 5), as it is comprised of the Cornwall and Devon NUTS-2 regions. In Section 8 we will discuss possible ways to deal with larger communities.

7 MULTI-DISTRICT RL

We now examine whether there is an advantage to consider the collaboration between districts when designing school closure policies. We conduct an experiment in our epidemiological model with 379 districts, and attempt to learn a joint policy to control the districts in the Cornwall-Devon community. To this end, we assign an agent to each of the 11 districts of the Cornwall-Devon community, and use a reinforcement learning approach to learn a joint policy.

[5] NUTS (Nomenclature of Territorial Units for Statistics) is a geocode standard constructed by Eurostat to reference the subdivisions of European countries. NUTS-2 is the second level and corresponds to basic regions for the application of regional policies.

We compare this joint policy to a non-collaborative policy (i.e., aggregated independent learners).

We remind the reader, that we refer to the state space of one patch p as $\mathcal{S}_p$, as detailed in Section 4. The state space of the MDP $\mathcal{S}$ corresponds to the aggregation of the state space of each patch that we attempt to control:

$$\mathcal{S} = \mathop{\times}_{p \in \mathcal{P}^c} \mathcal{S}_p, \tag{10}$$

where $\mathcal{P}^c$ is the set of patches we attempt to control. In this experiment, $\mathcal{P}^c$ corresponds to the 11 districts in the Cornwall-Devon community.

In order to learn a joint policy, we need to consider an action space that combines the actions for each district $p \in \mathcal{P}^c$ that we attempt to control. This results in a joint action space with a size that is exponential with respect to the number of agents. To approach this problem, we use a PPO super-agent that controls multiple districts simultaneously, to learn a joint policy. To this end, we use a custom policy network that gets as input the combined model state of each district $p \in \mathcal{P}^c$ (Equation 10), and as a result, the input layer has $17 \cdot |\mathcal{P}^c|$ input units. In contrast to the single-district PPO, that was introduced in Section 4, the output layer of the policy network of this agent has a unit for each district that we attempt to control. Again, each output unit is passed through a sigmoid activation function, and hence corresponds to the probability of closing the schools in the associated district. Similar to the single-district PPO, each hidden layer uses the hyperbolic tangent activation function. The value network has the same architecture for the input layers and hidden layers, but only has a single output unit that represents the value for the given state. We will refer to this agent as *multi-district PPO*.

We conduct experiments for $R_0 = 1.8$ (i.e., moderate transmission potential) and $R_0 = 2.4$ (i.e., high transmission potential), and we consider a school closure budget of 6 weeks, i.e., $b = 6$. We run multi-district PPO 5 times, to assess the variance of the learning signal, for $5 \cdot 10^6$ time steps, and we show the learning curves in Figure 6. These learning curves demonstrate a stable and steady learning process, for $R_0 = 1.8$ the reward curve is still increasing, while for $R_0 = 2.4$ the reward curve indicates that the learning process has converged.

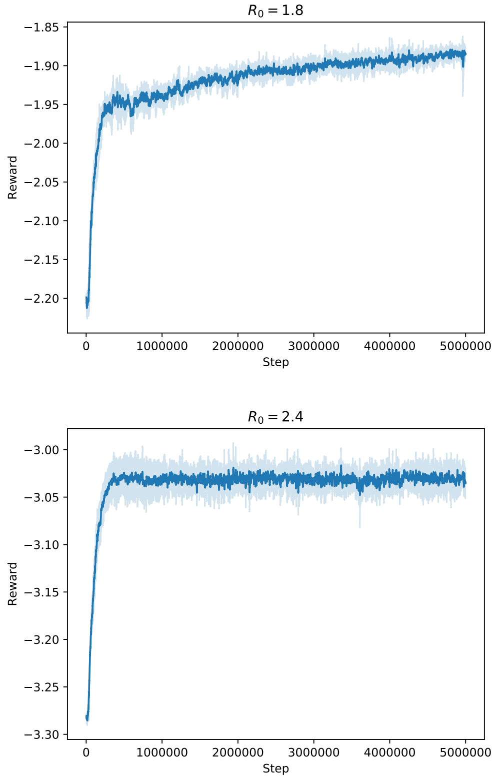

Figure 6: We show the reward curves for multi-district PPO for $R_0 = 1.8$ (top panel) and $R_0 = 2.4$ (bottom panel). The reward curves are visualized using a rolling window of 100 steps. The shaded area shows the standard deviation of the reward signal, over 5 multi-district PPO runs.

To investigate whether these *joint policies* provide a collaborative advantage, we compare it to the aggregation of single district policies, to which we will refer as the *aggregated policy*. To construct this aggregated policy, we learn a distinct school closure policy for each of the 11 districts in the Cornwall-Devon community, using PPO, following the same procedure as in Section 5. To evaluate this aggregated policy, we execute the distinct policies simultaneously. For the districts that are not controlled (both for the joint and aggregated policy) we keep the schools open for all time steps. For both $R_0 = 1.8$ and $R_0 = 2.4$, respectively, we simulate the joint and the aggregated policy 1000 times, and we show the attack rate improvement distribution in Figure 7. These results corroborate that there is a collaborative advantage when devising school closures policies, for both $R_0 = 1.8$ and $R_0 = 2.4$. However, the improvement is most significant for $R_0 = 1.8$. We conjecture that this difference is due to the fact that there is less flexibility when the transmission potential of the epidemic is higher, since there is less time to act. Although, we observe an improvement when a joint policy is learned, it remains challenging to interpret deep multi-agent policies, and we discuss in Section 8 possible directions for future work with respect to explainable multi-agent reinforcement learning.

In this analysis, where we have a limited number of actions per agent, the use of multi-district PPO proved to be successful. However, the use of more advanced multi-agent reinforcement learning methods is warranted when a more complex action space is considered. For this reason, we also investigated the recently introduced QMIX [38] algorithm. We searched for hyper-parameters to optimize QMIX's performance, but the resulting learning curve proved to be quite unstable (shown in Supplementary Information). Next to QMIX, there are other algorithms (e.g., Counterfactual multi-agent policy gradients [13], Actor-Attention-Critic for Multi-Agent Reinforcement Learning [23] and deep coordination graphs [3]) of interest to epidemiological decision making. In particular, we discuss the attention-based multi-agent reinforcement learning algorithms (e.g., Actor-Attention-Critic for Multi-Agent Reinforcement Learning [23]) as a direction for future work in Section 8.

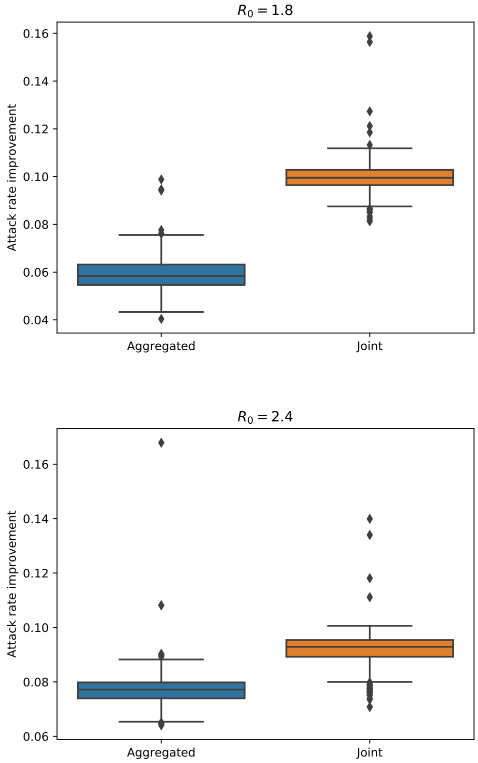

Figure 7: We compare the simulation results of the aggregated policy (blue) and the joint policy (orange) for $R_0 = 1.8$ (top panel) and $R_0 = 2.4$ (bottom panel). For both distributions (i.e., aggregated versus joint) , we show a box plot that denotes the outcome distribution that was obtained by simulating the respective policy 1000 times.

We conducted our experiments in the setting of school closures, and our findings are of direct relevance with respect to the mitigation of pandemic influenza. Furthermore, our novel approach to investigate the collaborative nature of prevention strategies as a multi-agent reinforcement learning problem, can be applied to other epidemiological settings, as we discuss in Section 8.

8 DISCUSSION

We demonstrate the potential of deep reinforcement learning in the context of epidemiological decision making by conducting experiments that show that PPO converges to the optimal policy. Next, we investigate and show that there is a collaborative advantage when devising school closures policies, by formulating this hypothesis as a multi-agent problem.

The work conducted in this manuscript indicates that there is the potential to use reinforcement learning in the context of complex stochastic epidemiological models. For future work, it would be interesting to investigate how well these algorithms scale to even larger state and/or action spaces. To increase the scalability, a possible research direction is the use of graph convolutional neural networks instead of multi-layer perceptron networks [9].

Another important concern is to scale these reinforcement learning methods to epidemiological models with a greater computational burden. In this work, we construct a custom model where we attempt to balance between model complexity and computational efficiency. However, constructing such models is cumbersome and time-consuming, and the resulting model is specifically tailored to address one particular research question (in our case the evaluation of school closure policies). An alternative to such custom models is the use of individual-based models, as such models can be easily configured to approach a variety of research scenarios. However, the computational burden that is associated with individual-based models complicates the use of reinforcement learning methods. To this end, it would be interesting to devise methods to automatically learn a surrogate model from the individual-based model, such that the reinforcement learning agent can learn in this computationally leaner surrogate model [46].

While we show that deep reinforcement learning algorithms can be used to learn optimal mitigation strategies, the interpretation of such policies remains challenging [18]. This is especially the case for the multi-district setting we considered, where state and time do not match, and the infection onset of the patches is highly stochastic. To this end, further research into explainable reinforcement learning, both in a single-agent and multi-agent setting, is warranted. An interesting direction for further research is the use of Soft Decision Trees (i.e., a hybrid model that combines decision trees and simple neural networks) as a surrogate for the deep RL policy that was learned, as presented by Coppens et al. [7].

Furthermore, in order to address problems with a larger state/action space and to scale to a larger number of agents, the use of attention-based multi-agent reinforcement learning algorithms could be explored [23, 43]. Based on this mechanism, algorithms can be applied on a graph of agents, which is either assumed [24] or learned [30].

9 ACKNOWLEDGMENTS

Pieter Libin and Timothy Verstraeten were supported by a PhD grant of the FWO (Fonds Wetenschappelijk Onderzoek - Vlaanderen). We thank Jelena Grujic (Vrije Universteit Brussel) for her comments on the network analysis. We thank the reviewers for their insightful comments and suggestions.

REFERENCES

[1] Edward J Allen, Linda JS Allen, Armando Arciniega, and Priscilla E Greenwood. 2008. Construction of equivalent stochastic differential equation models. *Stochastic analysis and applications* 26, 2 (2008), 274–297.
[2] Marc Baguelin, Albert Jan Van Hoek, Mark Jit, Stefan Flasche, Peter J White, and W John Edmunds. 2010. Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine* 28, 12 (2010), 2370–2384.
[3] Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. 2019. Deep Coordination Graphs. *arXiv preprint arXiv:1910.00091* (2019).
[4] Shawn T Brown, Julie HY Tai, Rachel R Bailey, Philip C Cooley, William D Wheaton, Margaret A Potter, Ronald E Voorhees, Megan LeJeune, John J Grefenstette, Donald S Burke, et al. 2011. Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost?: a computational simulation of Pennsylvania. *BMC public health* 11, 1 (2011), 353.
[5] Constanze Ciavarella, Laura Fumanelli, Stefano Merler, Ciro Cattuto, and Marco Ajelli. 2016. School closure policies at municipality level for mitigating influenza spread: a model-based evaluation. *BMC infectious diseases* 16, 1 (2016), 576.
[6] Erhan Cinlar. 2013. *Introduction to stochastic processes*. Courier Corporation.

[7] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, and Ann Nowé. 2019. Distilling Deep Reinforcement Learning Policies in Soft Decision Trees. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*. Macau, 1–6.

[8] Giancarlo De Luca, Kim Van Kerckhove, Pietro Coletti, Chiara Poletto, Nathalie Bossuyt, Niel Hens, and Vittoria Colizza. 2018. The impact of regular school closure on seasonal influenza epidemics: a data-driven spatial transmission model for Belgium. *BMC infectious diseases* 18, 1 (2018), 29.

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*. 3844–3852.

[10] Ken TD Eames, Natasha L Tilston, Ellen Brooks-Pollock, and W John Edmunds. 2012. Measured dynamic social contact patterns explain the spread of H1N1v influenza. *PLoS computational biology* 8, 3 (2012), e1002425.

[11] Rosalind M Eggo, Simon Cauchemez, and Neil M Ferguson. 2010. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *Journal of the Royal Society Interface* 8, 55 (2010), 233–243.

[12] Neil M Ferguson, Derek AT Cummings, Christophe Fraser, James C Cajka, Philip C Cooley, and Donald S Burke. 2006. Strategies for mitigating an influenza pandemic. *Nature* 442, 7101 (2006), 448.

[13] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*.

[14] Laura Fumanelli, Marco Ajelli, Piero Manfredi, Alessandro Vespignani, and Stefano Merler. 2012. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS computational biology* 8, 9 (2012), e1002673.

[15] Isaac Chun-Hai Fung, Manoj Gambhir, John W Glasser, Hongjiang Gao, Michael L Washington, Amra Uzicanin, and Martin I Meltzer. 2015. Modeling the effect of school closures in a pandemic scenario: exploring two different contact matrices. *Clinical Infectious Diseases* 60, suppl_1 (2015), S58–S63.

[16] Timothy C Germann, Hongjiang Gao, Manoj Gambhir, Andrew Plummer, Matthew Biggerstaff, Carrie Reed, and Amra Uzicanin. 2019. School dismissal as a pandemic influenza response: When, where and for how long? *Epidemics* (2019), 100348.

[17] Julia R Gog, Sébastien Ballesteros, Cécile Viboud, Lone Simonsen, Ottar N Bjornstad, Jeffrey Shaman, Dennis L Chao, Farid Khan, and Bryan T Grenfell. 2014. Spatial transmission of 2009 pandemic influenza in the US. *PLoS computational biology* 10, 6 (2014), e1003635.

[18] David Gunning and David W Aha. 2019. DARPA's Explainable Artificial Intelligence Program. *AI Magazine* 40, 2 (2019), 44–58.

[19] Michael J Haber, Davis K Shay, Xiaohong M Davis, Rajan Patel, Xiaoping Jin, Eric Weintraub, Evan Orenstein, and William W Thompson. 2007. Effectiveness of interventions to reduce contact rates during a simulated influenza pandemic. *Emerging infectious diseases* 13, 4 (2007), 581.

[20] Nilimesh Halder, Joel K Kelso, and George J Milne. 2010. Developing guidelines for school closure interventions to be used during a future influenza pandemic. *BMC infectious diseases* 10, 1 (2010), 221.

[21] M Elizabeth Halloran, Neil M Ferguson, Stephen Eubank, Ira M Longini, Derek A T Cummings, Bryan Lewis, Shufu Xu, Christophe Fraser, Anil Vullikanti, Timothy C Germann, and Others. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences* 105, 12 (2008), 4639–4644.

[22] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* (2019), 1–48.

[23] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*.

[24] Jiechuan Jiang, Chen Dun, and Zongqing Lu. 2018. Graph Convolutional Reinforcement Learning for Multi-Agent Cooperation. *arXiv preprint arXiv:1810.09202* (2018).

[25] Aaron A King, Matthieu Domenech de Cellès, Felicia MG Magpantay, and Pejman Rohani. 2015. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences* 282, 1806 (2015), 20150347.

[26] Stephen M Kissler, Julia R Gog, Cécile Viboud, Vivek Charu, Ottar N Bjørnstad, Lone Simonsen, and Bryan T Grenfell. 2019. Geographic transmission hubs of the 2009 influenza pandemic in the United States. *Epidemics* 26 (2019), 86–94.

[27] Petra Klepac, Stephen Kissler, and Julia Gog. 2018. Contagion! The BBC Four Pandemic–The model behind the documentary. *Epidemics* (2018).

[28] Ruben J Kubiak and Angela R McLean. 2012. Why was the 2009 influenza pandemic in England so small? *PloS one* 7, 2 (2012), e30223.

[29] Elizabeth A Leicht and Mark EJ Newman. 2008. Community structure in directed networks. *Physical review letters* 100, 11 (2008), 118703.

[30] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. 2019. Multi-Agent Game Abstraction via Graph Attention Neural Network. *arXiv preprint arXiv:1810.09202* (2019). arXiv:cs.AI/1911.10715

[31] Ira M Longini, Azhar Nizam, Shufu Xu, Kumnuan Ungchusak, Wanna Hanshaoworakul, Derek AT Cummings, and M Elizabeth Halloran. 2005. Containing pandemic influenza at the source. *Science* 309, 5737 (2005), 1083–1087.

[32] Howard Markel, Harvey B Lipman, J Alexander Navarro, Alexandra Sloan, Joseph R Michalsen, Alexandra Minna Stern, and Martin S Cetron. 2007. Nonpharmaceutical interventions implemented by US cities during the 1918-1919 influenza pandemic. *Jama* 298, 6 (2007), 644–654.

[33] Elizabeth Miller, Katja Hoschler, Pia Hardelid, Elaine Stanford, Nick Andrews, and Maria Zambon. 2010. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *The Lancet* 375, 9720 (2010), 1100–1108.

[34] George J Milne, Nilimesh Halder, and Joel K Kelso. 2013. The cost effectiveness of pandemic influenza interventions: a pandemic severity based analysis. *PloS one* 8, 4 (2013).

[35] Catharine Paules and Kanta Subbarao. 2017. Influenza. *The Lancet* (2017), 697–708.

[36] William JM Probert, Sandya Lakkur, Christopher J Fonnesbeck, Katriona Shea, Michael C Runge, Michael J Tildesley, and Matthew J Ferrari. 2019. Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies. *Philosophical Transactions of the Royal Society B* 374, 1776 (2019), 20180277.

[37] M Radivojević and J Grujić. 2017. Community structure of copper supply networks in the prehistoric Balkans: An independent evaluation of the archaeological record from the 7th to the 4th millennium BC. *Journal of Complex Networks* 6, 1 (2017), 106–124.

[38] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 4295–4304.

[39] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[40] Michael Stein. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* 29, 2 (1987), 143–151.

[41] S Towers and Z Feng. 2012. Social contact patterns and control strategies for influenza in the elderly. *Mathematical biosciences* 240, 2 (2012), 241–249.

[42] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* 9 (2019).

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. [n.d.]. Attention is all you need. In *Advances in Neural Information Processing Systems 30*.

[44] Lin Wang and Joseph T Wu. 2018. Characterizing the dynamics underlying global spread of epidemics. *Nature communications* 9, 1 (2018), 218.

[45] Richard J Webby and Robert G Webster. 2003. Are we ready for pandemic influenza? *Science* 302, 5650 (2003), 1519–1522.

[46] Lander Willem, Sean Stijven, Ekaterina Vladislavleva, Jan Broeckhove, Philippe Beutels, and Niel Hens. 2014. Active Learning to Understand Infectious Disease Models and Improve Policy Making. *PLoS Comput Biol* 10, 4 (2014), e1003563.

[47] Reza Yaesoubi and Ted Cohen. 2011. Dynamic health policies for controlling the spread of emerging infections: influenza as an example. *PloS one* 6, 9 (2011).

[48] Reza Yaesoubi and Ted Cohen. 2013. Identifying dynamic tuberculosis case-finding policies for HIV/TB coepidemics. *Proceedings of the National Academy of Sciences* 110, 23 (2013), 9457–9462.

[49] Reza Yaesoubi and Ted Cohen. 2016. Identifying cost-effective dynamic policies to control epidemics. *Statistics in medicine* 35, 28 (2016), 5189–5209.

[50] Andrea Yanez, Conor Hayes, and Frank Glavin. [n.d.]. Towards the Control of Epidemic Spread: Designing Reinforcement Learning Environments. ([n. d.]).

[51] Na Zhu, Dingyu Zhang, Wenling Wang, Xinwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).