Expert Systems With Applications 249 (2024) 123686

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Exploring the Pareto front of multi-objective COVID-19 mitigation policies using reinforcement learning

Mathieu Reymond [a,*], Conor F. Hayes [b], Lander Willem [c], Roxana Rădulescu [a], Steven Abrams [c], Diederik M. Roijers [a], Enda Howley [b], Patrick Mannion [b], Niel Hens [d], Ann Nowé [a], Pieter Libin [a]

[a] Vrije Universiteit Brussel, Brussels, Belgium
[b] National University of Ireland Galway, Galway, Ireland
[c] University of Antwerp, Antwerp, Belgium
[d] Hasselt University, Hasselt, Belgium

ARTICLE INFO

Keywords:
Multi-objective reinforcement learning
Epidemic control
COVID-19 epidemic models

ABSTRACT

Infectious disease outbreaks can have a disruptive impact on public health and societal processes. As decision-making in the context of epidemic mitigation is multi-dimensional hence complex, reinforcement learning in combination with complex epidemic models provides a methodology to design refined prevention strategies. Current research focuses on optimizing policies with respect to a single objective, such as the pathogen's attack rate. However, as the mitigation of epidemics involves distinct, and possibly conflicting, criteria (i.a., mortality, morbidity, economic cost, well-being), a multi-objective decision approach is warranted to obtain balanced policies. To enhance future decision-making, we propose a deep multi-objective reinforcement learning approach by building upon a state-of-the-art algorithm called Pareto Conditioned Networks (PCN) to obtain a set of solutions for distinct outcomes of the decision problem. We consider different deconfinement strategies after the first Belgian lockdown within the COVID-19 pandemic and aim to minimize both COVID-19 cases (i.e., infections and hospitalizations) and the societal burden induced by the mitigation measures. As such, we connected a multi-objective Markov decision process with a stochastic compartment model designed to approximate the Belgian COVID-19 waves and explore reactive strategies. As these social mitigation measures are implemented in a continuous action space that modulates the contact matrix of the age-structured epidemic model, we extend PCN to this setting. We evaluate the solution set that PCN returns, and observe that it explored the whole range of possible social restrictions, leading to high-quality trade-offs, as it captured the problem dynamics. In this work, we demonstrate that multi-objective reinforcement learning adds value to epidemiological modeling and provides essential insights to balance mitigation policies.

1. Introduction

Infectious disease outbreaks represent a major challenge (Miranda et al., 2022). To this end, understanding the complex dynamics that underlie these epidemics is essential. Epidemiological transmission models allow us to capture and understand such dynamics and facilitate the study of prevention strategies through simulation. However, developing efficient mitigation strategies remains a challenging process, given the non-linear and complex nature of epidemics. To address these challenges, reinforcement learning provides a methodology to automatically learn mitigation strategies in combination with complex epidemic models (Libin, Moonens, et al., 2021). Previous research focused on optimizing policies with respect to a single objective, such as the pathogen's attack rate, while the mitigation of epidemics is a problem that inherently covers distinct and possibly conflicting criteria (i.a., prevalence, mental health, cost). Therefore, optimizing on a single objective requires that these distinct criteria are somehow aggregated into a single metric. Manually designing such metrics is time-consuming, costly and error-prone, as this non-intuitive process requires repetitive and tedious tuning to achieve the desired behavior (Roijers, Vamplew, Whiteson, & Dazeley, 2013). Moreover, taking a single objective approach reduces the explainability of the learned solution, as we cannot compare the learned behavior with alternatives (Hayes et al., 2021).

This challenging process can be circumvented by taking an explicitly multi-objective approach that aims to learn the different trade-offs

[*] Corresponding author.
E-mail addresses: mathieu.reymond@vub.be (M. Reymond), c.hayes13@nuigalway.ie (C.F. Hayes), lander.willem@uantwerpen.be (L. Willem), roxana@ai.vub.ac.be (R. Rădulescu), steven.abrams@uantwerpen.be (S. Abrams), diederik.roijers@vub.be (D.M. Roijers), enda.howley@nuigalway.ie (E. Howley), patrick.mannion@nuigalway.ie (P. Mannion), niel.hens@uhasselt.be (N. Hens), ann.nowe@ai.vub.ac.be (A. Nowé), pieter.libin@vub.be (P. Libin).

https://doi.org/10.1016/j.eswa.2024.123686
Received 9 July 2023; Received in revised form 12 February 2024; Accepted 13 March 2024
Available online 24 March 2024

regarding the considered criteria. By assuming that a decision maker will always prefer solutions for which at least one objective improves, it is possible to learn a set of optimal solutions referred to as the *Pareto front* (Hayes et al., 2021). This enables decision makers to review each solution on the Pareto front before making a decision, thereby being aware of the trade-offs that each solution implies.

In this work, we investigate the use of *multi-objective reinforcement learning* (MORL) to learn a set of solutions that approximate the Pareto front of multi-objective epidemic mitigation strategies. We consider the first wave of the Belgian COVID-19 epidemic, which was mitigated by a strict lockdown (Willem et al., 2021). When the incidence of confirmed cases was steadily decreasing, epidemiological experts were tasked to investigate deconfinement strategies, to reduce the severe social contact and mobility restrictions. Here, we consider an epidemiological model that was constructed to describe the Belgian COVID-19 epidemic and was fitted to hospitalization incidence data and serial sero-prevalence data (Abrams et al., 2021). This model constitutes a stochastic discrete-time age-structured compartmental model that simulates mitigation strategies by varying social distancing parameters concerning school, work and leisure contacts. Based on this model, we contribute a novel multi-objective epidemiological reinforcement learning environment (Multi-Objective Belgian COVID environment, MOBelCov), in the form of a multi-objective Markov decision process (MOMDP) (Roijers et al., 2013). MOBelCov encapsulates the epidemiological model developed by Abrams et al. (2021) to implement state transitions, with an action space that combines a proportional reduction of school, work and leisure contacts at each time step, Furthermore, it defines a reward function based on two objectives: the attack rate (i.e., proportion of the population affected by the pathogen) and the social burden that is induced by the mitigation measures.

To learn and explore the trade-offs between the attack rate and social burden we use a state-of-the-art MORL approach based on Pareto Conditioned Networks (PCN) (Reymond, Eugenio, & Nowè, 2022). PCN uses a single neural network to learn the policies that belong to the Pareto front. As PCN is an algorithm designed for discrete action-spaces, we extend it towards continuous action-spaces to accommodate MOBelCov's action-space. With this continuous action variant of PCN, we explore the Pareto front of multi-objective COVID-19 mitigation policies. As PCN makes no assumptions about the shape of the coverage set, it is particularly well suited for the complex decision problem that we consider, for which the shape of the coverage set is not known a priori.

By evaluating the solution set of mitigation policies learned by PCN, we observe that PCN minimizes the social burden in scenarios where hospitalization rates are sufficiently low. Therefore, in this work we illustrate that multi-objective reinforcement learning can provide important insights concerning the trade-offs between complex mitigation polices in real-world epidemiological models.

In Section 2, we formally introduce multi-objective reinforcement learning and explain a series of commonly used metrics to evaluate MORL learning performance. In Section 3, we introduce the methods used in this work, which includes the epidemiological model, the MOMDP that encapsulates this model, a short description of the Pareto Conditional Networks (PCN) MORL algorithm, and the extensions that were required to PCN. In Section 4, we present our experimental results regarding our study of the COVID-19 exit strategies, in terms of the social burden and hospitalization rates. In Section 5, we present the literature that is related to our study. In Section 6, we interpret our experimental results, discuss the limitations of our work, and look into future research avenues.

2. Multi-objective reinforcement learning

Real-world decisions problems typically consider multiple and possibly conflicting objectives. Multi-objective reinforcement learning (MORL) can be used to find optimal solutions for sequential decision making problems with multiple objectives (Hayes et al., 2021). Multi-objective sequential problems are typically modeled as a multi-objective Markov decision process (MOMDP), i.e., a tuple, $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \boldsymbol{\mathcal{R}} \rangle$, where $\mathcal{S}$, $\mathcal{A}$ denote the state and action spaces respectively, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes a probabilistic transition function, γ is a discount factor determining the importance of future rewards and $\boldsymbol{\mathcal{R}} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^n$ is an n-dimensional vector-valued immediate reward function, where n corresponds to the number of objectives.

While MOMDPs define the actions an agent can take, it does not define the agent's behavior, i.e., which actions are taken in each state. Given this MOMDP, an agent follows a policy π, that expresses the probability to take action $a \in \mathcal{A}$ when in state $s \in \mathcal{S}$: $\mathcal{S} \times \mathcal{A} \to [0, 1]$. We measure the performance of π through the expected sum of discounted rewards (denoted the *Value* $\mathbf{V}^\pi$) it achieves from the initial state $\mathbf{s}_0$ until the end of the decision problem:

$$\mathbf{V}^\pi = \mathbb{E}\left[\sum_{t=0}^{h} \gamma^t \mathbf{r}_t |\ \pi, \mathbf{s}_0 \right], \tag{1}$$

where $\mathbf{r}_t$ corresponds to the multi-objective reward observed at time t, by following policy π, given by the reward function $\mathcal{R}(\mathbf{s}_t, a_t, \mathbf{s}_{t+1})$, where $\mathbf{s}_t$ is the current state, a_t corresponds to the action that was taken, and $\mathbf{s}_{t+1}$ signifies the state reached by taking action a_t.

On the one hand, for single-objective RL, where $n = 1$, the goal is to find the policy π^* that maximizes the V-value:

$$\pi^* = \arg\max_{\pi} V^\pi. \tag{2}$$

On the other hand, for MORL, $n > 1$ which leads to vectorial returns. In this case, there can be policies for which, without any additional information, it is impossible to know if one is better than the other. For example, it is impossible to decide which policy between π_1, π_2 is optimal if both policies lead to expected returns $\mathbf{V}^{\pi_1} = (0, 1), \mathbf{V}^{\pi_2} = (1, 0)$ respectively. We call these solutions *non-dominated*, i.e., solutions for which it is impossible to improve an objective without decreasing the value of another. The set that contains all the non-dominated solutions of the decision problem is called the *Pareto front* $\mathcal{F}$. Our goal is to find the set of policies that lead to all the V-values contained in the Pareto front $\Pi^* = \{\ \pi | \mathbf{V}^\pi \ \in \ \mathcal{F}\}$. In general, we call any set of V-values a *solution set*. When a solution set contains only non-dominated V-values, it is referred to as a *coverage set*. In the case that no π exists that has a $\mathbf{V}^\pi$ dominating any of the solutions in a coverage set, then this coverage set is the Pareto front.

2.1. Multi-objective metrics

Comparing the learned coverage sets produced by different algorithms is a non-trivial task, as one algorithm's output might dominate the other in some region of the objective-space, but be dominated in another. Intuitively, one would generally prefer the algorithm that covers a wider range of decision maker preferences.

A widely used metric in the literature is called the *hypervolume* (Zitzler, Thiele, Laumanns, Fonseca, & Da Fonseca, 2003). This metric evaluates the learned coverage set by computing its volume with respect to a fixed reference point. The reference point is taken as a lower bound on the achievable returns so that the volumes are always positive. By definition, the solutions contained in the Pareto front dominate all other possible solutions. Thus, no other solution can further increase the volume under the Pareto front. This means that the hypervolume is the highest for the Pareto front. One drawback of the hypervolume is that it can be difficult to interpret. For example, when working in high-dimensional objective-spaces, adding or removing a single point can drastically change hypervolume values, especially if the point lies close to an extremum of said space.

To counterpoise these shortcomings, we can consider an additional metric called the ϵ-indicator I_ϵ (Zitzler et al., 2003), which measures

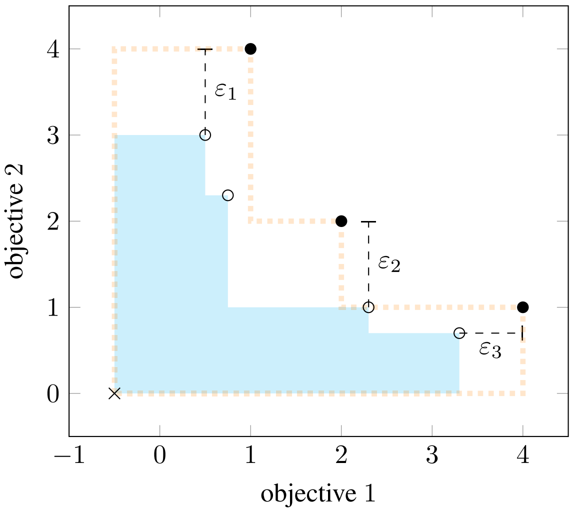

Fig. 1. Example of a Pareto front (black dots) and a coverage set (white dots) in a 2-objective environment. The hypervolume metric (in light blue) measures the volume of all dominated solutions with respect to a reference point (cross). The reference point is taken as a lower bound on the achievable returns, which is why it can differ from the origin. The ϵ metrics first compute the maximum distance between each point in the Pareto front and its closest point in the coverage set (ϵ_i). We can then take their maximum value to compute the I_ϵ metric, or their mean value to obtain the $I_{\epsilon-mean}$ metric of the coverage set.

how close a coverage set is to the Pareto front $\mathcal{F}$. Intuitively, I_ϵ shows that any solution of $\mathcal{F}$ is *at most* ϵ better with respect to each objective o than the closest solution of the evaluated coverage set:

$$I_\epsilon = \inf_{\epsilon \in \mathbb{R}} \{\forall\ \mathbf{V}^\pi \in \mathcal{F},\ \exists\ \mathbf{V}^{\pi'} \in \hat{\Pi}\ :\ \|\mathbf{V}^\pi - \mathbf{V}^{\pi'}\|_\infty \leq \epsilon\}, \tag{3}$$

where $\|.\|_\infty$, the L-infinity norm, is defined as the magnitude of the largest entry of a vector.

The main disadvantage of this metric is that we need the true Pareto front to compute it, which is unknown for our MOMDP. To still gain insights from our learned policies, we approximate the true Pareto front using the non-dominated policies across all runs.

We note that the ϵ-indicator metric is quite pessimistic, as it measures worst-case performance (Zintgraf, Kanters, Roijers, Oliehoek, & Beau, 2015), i.e., it will still report low performance as long as a single point of the Pareto front is incorrectly modeled, even if all the other points are covered. As such, we also use the $I_{\epsilon-mean}$ which measures the *average* ϵ distance of the solutions in $\mathcal{F}$ with respect to the evaluated coverage set (Reymond et al., 2022).

As an example, Fig. 1 shows a visual representation of the hypervolume and ϵ metrics in two dimensions.

3. Methods

In this section, we first introduce the compartment model that we use to simulate the epidemic (Section 3.1). Second, we explain the considered intervention strategies (Section 3.2). Third, we propose MOBelCov, a MORL environment that encompasses the epidemiological model and intervention strategies (Section 3.3). Finally we present the Pareto Conditioned Networks (PCN) algorithm (Reymond et al., 2022), a multi-objective reinforcement learning algorithm that will be used to approximate the Pareto front of the MOBelCov. We briefly present the original PCN algorithm and explain the methodological extensions we made with respect to continuous action spaces and stochastic transition functions, that were necessary for the MOBelCovenvironment.

Through the remainder of this manuscript, we use bold notation to denote variables with a vectorial value, while non-bold notation for its scalar counterpart.

3.1. Stochastic compartment model for SARS-CoV-2

To evaluate non-pharmaceutical interventions, we consider the compartmental model presented by Abrams et al. that was used to investigate exit strategies in Belgium after the first epidemic wave of SARS-CoV-2 (Abrams et al., 2021). This model concerns a discrete-time stochastic model, that considers an age-structured population. The model generalizes a standard SEIR model[1], extended to capture the different stages of disease spread and history that are associated with SARS-CoV-2 (i.e., pre-symptomatic, asymptomatic, symptomatic with mild symptoms and symptomatic with severe symptoms) and to represent the stages associated with severe disease, i.e., hospitalization, admission to the intensive care unit (ICU) and death.

A visual representation of the model is depicted in Fig. 2. It shows the different compartments, as well as the flow rates at which individuals move between compartments.

These flow rates are defined by a set of ordinary differential equations, which are outlined as follows:

$$\begin{aligned}
\frac{d\mathbf{S}(t)}{dt} &= -\lambda(t)\mathbf{S}(t), \\
\frac{d\mathbf{E}(t)}{dt} &= \lambda(t)\mathbf{S}(t) - \gamma\mathbf{E}(t), \\
\frac{d\mathbf{I}^{presym}(t)}{dt} &= \gamma\mathbf{E}(t) - \theta\mathbf{I}^{presym}(t), \\
\frac{d\mathbf{I}^{asym}(t)}{dt} &= \theta p\mathbf{I}^{presym}(t) - \delta_1\mathbf{I}^{asym}(t), \\
\frac{d\mathbf{I}^{mild}(t)}{dt} &= \theta(1-p)\mathbf{I}^{presym}(t) - \{\psi + \delta_2\}\mathbf{I}^{mild}(t), \\
\frac{d\mathbf{I}^{sev}(t)}{dt} &= \psi\mathbf{I}^{mild}(t) - \omega\mathbf{I}^{sev}(t), \\
\frac{d\mathbf{I}^{hosp}(t)}{dt} &= \phi_1\omega\mathbf{I}^{sev}(t) - \{\delta_3 + \tau_1\}\mathbf{I}^{hosp}(t), \\
\frac{d\mathbf{I}^{icu}(t)}{dt} &= (1-\phi_1)\omega\mathbf{I}^{sev}(t) - \{\delta_4 + \tau_2\}\mathbf{I}^{icu}(t), \\
\frac{d\mathbf{D}(t)}{dt} &= \tau_1\mathbf{I}^{hosp}(t) + \tau_2\mathbf{I}^{icu}(t), \\
\frac{d\mathbf{R}(t)}{dt} &= \delta_1\mathbf{I}^{asym}(t) + \delta_2\mathbf{I}^{mild}(t) + \delta_3\mathbf{I}^{hosp}(t) + \delta_4\mathbf{I}^{icu}(t)
\end{aligned}$$

In this set of ordinary differential equations, each state variable represents a vector over all age groups for a particular compartment at time t. For example, $\mathbf{S} = (S_1(t), S_2(t), \dots, S_k(t))^T$ is the vector representing the susceptible members of the population of each age group k at time t. Infection dynamics are governed by an age-specific force of infection λ:

$$\lambda(k,t) = \sum_{k'=1}^{K} \beta(k,k')I_{k'}(t), \tag{4}$$

where K is the total number of age groups, and $\beta(k,k')$ is the time-invariant transmission rate that encodes the average per capita rate at which an infectious individual in age group k makes an effective contact with a susceptible individual in age group k', per unit of time.

As we consider an age-structured population, we consider this extended SEIR structure for $K = 10$ age groups, i.e., $[0-10), [10-20), [20-30), [30-40), [40-50), [50-60), [60-70), [70-80), [80-90), [90, 100+)$. Contacts of the different age-groups, which impact the propagation rate of the epidemic, are modeled using social contact matrices (Willem et al., 2020). We define a social contact matrix for 6 different social environments: $\boldsymbol{C}_{\text{home}}, \boldsymbol{C}_{\text{work}}, \boldsymbol{C}_{\text{transport}}, \boldsymbol{C}_{\text{school}}, \boldsymbol{C}_{\text{leisure}}, \boldsymbol{C}_{\text{other}}$, for the home, work, transport, school, leisure, other environments respectively. The social contact matrix across all social environments is defined as:

$$\boldsymbol{C} = \boldsymbol{C}_{\text{home}} + \boldsymbol{C}_{\text{work}} + \boldsymbol{C}_{\text{transport}} + \boldsymbol{C}_{\text{school}} + \boldsymbol{C}_{\text{leisure}} + \boldsymbol{C}_{\text{other}} \tag{5}$$

[1] A standard SEIR model divides the population into four different states, i.e., susceptible, exposed, infectious and recovered individuals.

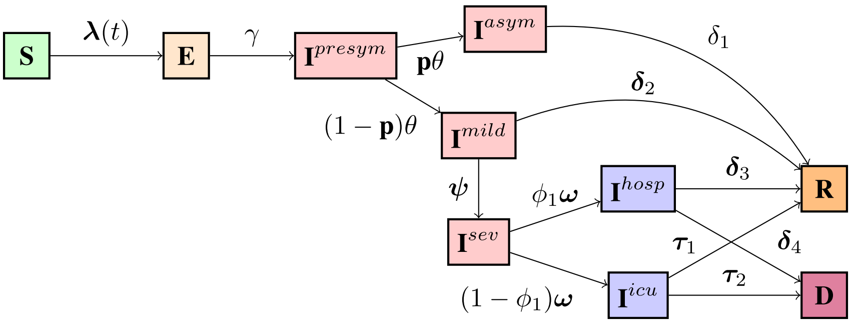

Fig. 2. Schematic diagram of the compartmental model for SARS-CoV-2 based on the work of Abrams et al. (2021), that is used to derive the MOMDP. The model consists of 10 compartments, listed here together with their abbreviations: susceptible (**S**), exposed (**E**), pre-symptomatic infection ($\mathbf{I}^{presym}$), asymptomatic infection ($\mathbf{I}^{asym}$), symptomatic infection with mild symptoms ($\mathbf{I}^{mild}$), symptomatic infection with severe symptoms ($\mathbf{I}^{sev}$), hospitalization ($\mathbf{I}^{hosp}$), admission to the ICU ($\mathbf{I}^{icu}$), death (**D**) and recovered (**R**).

Under the social contact hypothesis (Wallinga, Teunis, & Kretzschmar, 2006), we have that:

$$\beta(k,k') = q \cdot \boldsymbol{C}(k,k'), \tag{6}$$

where q is a proportionality factor.

Following Abrams et al. (2021), we rely on distinct social contact matrices for symptomatic and asymptomatic individuals, respectively $\boldsymbol{C}_s$ and $\boldsymbol{C}_a$. Therefore, we define the transmission rates for both symptomatic and asymptomatic individuals as follows:

$$\beta_s(k,k) = q_s \cdot C_s(k,k'), \tag{7}$$

and

$$\beta_a(k,k') = q_a \cdot C_a(k,k'). \tag{8}$$

The age-dependent force of infection can be defined as follows:

$$\lambda(t) = \boldsymbol{\beta}_a \times \{\mathbf{I}^{presym}(t) + \mathbf{I}^{asym}(t)\} + \boldsymbol{\beta}_s \times \{\mathbf{I}^{mild}(t) + \mathbf{I}^{sev}(t)\}, \tag{9}$$

where $\lambda(t) = (\lambda(1,t), \lambda(2,t), \dots, \lambda(K,t))$. For all further information about the different compartments and parameters we refer the reader to the work of Abrams et al. (2021).

Variability in social contact behavior, disease transmission, recovery and mortality impacts the epidemic outcome and is subject to chance. To evaluate intervention strategies that modulate infectious disease transmission and prevention, the use of stochastic epidemiological models is warranted (Abrams et al., 2021). Moreover, the effect of stochasticity is most pronounced when the number of infectious individuals is small or variability is high, for example studying the initial growth of an epidemic (Britton & Lindenstrand, 2009), or when implementing deconfinement strategies after strict lock-downs.

The spread of a virus in a population is a stochastic process, hence intervening by, for example, reducing social contacts or government interventions affects the further course of a stochastic outbreak. Therefore, to understand how interventions affect the spread of the disease we adopted a stochastic version of the deterministic model described above, which can capture the variability with and without interventions related to social contacts.

By formulating the set of differential equations defined above, as a chain-binomial, we can obtain stochastic trajectories from this model (Bailey, 1975). A chain-binomial model assumes a stochastic model where infected individuals are generated by some underlying probability distribution. The equations that underlie this model are specified in Section A of the Appendix.

For this stochastic model we consider a time interval $(t, t+h]$, where h is defined as the length between two consecutive time points. In this work we set $h = \frac{1}{240}$. We tuned this h to ensure that the average behavior of the binomial chain model matches the ODE model.

3.2. Intervention strategies

To model different types of non-pharmaceutical interventions, we follow the contact reduction scheme presented by Abrams et al. (2021). Firstly, to consider distinct exit scenarios, we modulate the social contact matrices to reflect a contact reduction in a particular age group.

We consider a contact reduction function that imposes a proportional reduction of work (including transport) p_w, school p_s and leisure p_l contacts, which is implemented as a linear combination of social contact matrices:

$$\hat{C}(p_w, p_s, p_l) = \boldsymbol{C}_{\text{home}} + p_w(\boldsymbol{C}_{\text{work}} + \boldsymbol{C}_{\text{transport}}) + p_s\boldsymbol{C}_{\text{school}} + p_l(\boldsymbol{C}_{\text{leisure}} + \boldsymbol{C}_{\text{other}}) \tag{10}$$

We denote $\hat{\boldsymbol{C}}_t$ the social contact matrix at timestep t, resulting from the reduction function $\hat{C}$. Secondly, we assume that compliance to the interventions is gradual and represent the transition from $\hat{\boldsymbol{C}}_t$ to $\hat{\boldsymbol{C}}_{t+1}$ using a logistic compliance function (see details in Sec. A.2 of the Appendix).

3.3. The MOBelCov environment

In order to apply multi-objective reinforcement learning, we construct the MOBelCov MOMDP based on the epidemiological model introduced in Section 3.1 and graphically depicted in Fig. 2. Moreover, we consider a finite-horizon setting where we simulate the compartment model for a fixed number of weeks.

Action-space: Our actions concern the installment of a social contact matrix with a particular reduction resulting from the reduction function $\hat{C}$ (see Section 3.2). To this end, we use the proportional reduction parameters p_w, p_s, p_l defined in Section 3.2. Thus, each $\mathbf{a} \in \mathcal{A}$ is a 3-dimensional continuous vector in $[0,1]^3$ (i.e., $\mathbf{a} = [p_w, p_s, p_l]$) which impacts the social contact matrix according to Eq. (10).

Transition function: The model defined by Abrams et al. (2021) utilizes a model transition probability M (see Sec. A of the Appendix for details on M), that progresses the epidemiological model in one timestep based on the currently installed social contact matrix $\hat{C}(p_w, p_s, p_l)$. We use this function as the transition function in MOBelCov.

In a classical MDP, executing an action $\mathbf{a}_t$ in any state $\mathbf{s}_t$ leads to a next state $\mathbf{s}_{t+1}$ according to the transition function $\mathcal{T}$. At every timestep t, the agent is free to choose the action to perform. In our case, this potentially results in a different restriction $[p_w, p_s, p_l]$ every week. However, we argue that in the context of mitigation policies, consistency is important and policies that impose changes too frequently will be hard to adhere to.

In order to obtain consistent mitigation policies, we introduce a *budget* regarding the number of times a policy can change its actions until the terminal state of the MOMDP is reached. To facilitate this, we maintain a budget for each of the actions. Concretely, when the action changes, i.e., if the social restriction proposed by the policy is different from the one that is currently in place, we reduce the budget for that action by one. We only allow action changes as long as there is budget left. Note that, since the actions are continuous values, we consider a change when the difference in action-value is greater than a delta (reported with the hyperparameter values in Sec. D.1). We note that we can mimic a no-limit budget setting by choosing a budget that corresponds to the horizon of the environment.

Finally, for each timestep t, our transition function $\mathcal{T}$ uses the model transition probability M to simulate the model for one week, using $\hat{\boldsymbol{C}}_t$ obtained from $\mathbf{a}_t$.

State-space: The state of the MOMDP concerns a 3-tuple. The first element, $\mathbf{s}_m$, directly corresponds to the aggregation of the state variables in the epidemiological model, i.e., a tuple,

$$\langle S_k, E_k, I_k^{presym}, I_k^{asym}, I_k^{mild}, I_k^{sev}, I_k^{hosp}, I_k^{icu}, H_k^{new}, D_k, R_k\rangle, \tag{11}$$

for each age group $k \in \{1, \dots, K\}$, where S encodes the members of the population who are susceptible to infection and E encodes the members of the population who have been exposed to COVID-19. Moreover, I^{presym}, I^{asym}, I^{mild}, I^{hosp}, I^{icu} denote the members of the population infected with COVID-19 and are, respectively, presymptomatic, asymptomatic, mildly symptomatic, hospitalized, or in the ICU. Finally, in addition to these compartments which define the transmission dynamics, we define a separate compartment H_k^{new} to keep track of the number of newly hospitalized individuals in age group k.

We parameterize the epidemiological model using the mean of the posteriors as specified by Abrams et al. (2021) (details in Sec. A.1 of the Appendix).

The second element of the tuple consists of the social contact matrix $\hat{\boldsymbol{C}}_t$ that is currently in place. The reason to incorporate it in the state–space is two-fold. First, Abrams et al. (2021) define a compliance function, simulating the time people need to get used to the new rules set in place. As such, during the simulated week, there is a gradual shift from the current $\hat{\boldsymbol{C}}_t$ to the new social contact matrix, $\hat{\boldsymbol{C}}_{t+1}$. Thus, we need to include the current $\hat{C}_t$, to establish a Markovian environment. Secondly, we require the current $\hat{\boldsymbol{C}}_t$ to determine whether the action changes the social restrictions in place, and thus consume part of the budget.

The third element of the tuple concerns the budget $\boldsymbol{b}$. We incorporate a distinct budget per action-dimension, so p_w, p_s and p_l each have their own budget, resulting in a vector $\boldsymbol{b} = [b_w, b_s, b_l]$. As such, it is possible that, at timestep t, the budget for one of the proportional reductions is reduced but not the others.

Therefore, we define a state in MOBelCov as follows:

$$\mathbf{s} = \mathbf{s}_m \cup \hat{C} \cup \boldsymbol{b} \tag{12}$$

Reward function: We define a vectorial reward function which considers multiple objectives: attack rate (i.e., infections or hospitalizations) and the social burden imposed by the interventions on the population.

The attack rate in terms of infections is defined as the difference in susceptibles from the current state to the next state. Since this is a cost that needs to be minimized, we defined the corresponding reward function as the negative attack rate (Libin, Moonens, et al., 2021):

$$\mathcal{R}_{\mathrm{ARI}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = -(\sum_{k=1}^{K} S_k(\mathbf{s}) - \sum_{k=1}^{K} S_k(\mathbf{s}')). \tag{13}$$

The reward function to reduce the attack rate in terms of hospitalizations is defined as the inverse of new hospitalizations:

$$\mathcal{R}_{\mathrm{ARH}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = -\sum_{k=1}^{K} H_k^{\mathrm{new}}(\mathbf{s}). \tag{14}$$

Finally, we use the contact reduction resulting from the intervention measures as a proxy for societal burden. To quantify this, we consider the original social contact matrix C and the installed social contact matrix $\hat{C}$ to compute the difference $\hat{C} - C$. The resulting difference matrix quantifies the social contact impairment. To determine the population-based loss, we apply the difference matrix to the population sizes of the respective age groups that are currently uninfected (i.e., susceptible and recovered individuals). Formally, we define the social burden reward function $\mathcal{R}_{\mathrm{SB}}$, as follows:

$$\mathcal{R}_{\mathrm{SB}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = \sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{C} - C)_{ij} S_i(\mathbf{s}) S_j(\mathbf{s}) + \sum_{i=1}^{K}\sum_{j=1}^{K}(\hat{C} - C)_{ij} R_i(\mathbf{s}) R_j(\mathbf{s}), \tag{15}$$

where $S_k(\mathbf{s})$ represents the number of susceptible individuals in age group k in state $\mathbf{s}$ and R_k represents the number of recovered individuals in age group k in state $\mathbf{s}$. In Section 4, we optimize PCN on two different variants for the multi-objective reward function: $[\mathcal{R}_{\mathrm{ARH}}, \mathcal{R}_{\mathrm{SB}}]$ and $[\mathcal{R}_{\mathrm{ARI}}, \mathcal{R}_{\mathrm{SB}}]$, to study the impact of these distinct attack rate quantities.

3.4 Pareto conditioned networks

In multi-objective optimization, the set of optimal policies can grow exponentially with the number of objectives. Thus, recovering them all is a computationally expensive process and requires an exhaustive exploration of the complete state space. To address this problem, we use Pareto Conditioned Networks (PCN), a method that uses a single neural network to encompass all non-dominated policies (Reymond et al., 2022). The key idea behind PCN is to use supervised learning techniques to improve the policy instead of resorting to temporal-difference learning. PCN uses a single neural network that takes a tuple $\langle \mathbf{s}, \hat{h}, \hat{\mathbf{R}}\rangle$ as input. $\hat{\mathbf{R}}$ represents the *desired return* of the decision maker, i.e., the return PCN should reach at the end of the episode. $\hat{h}$ denotes the *desired horizon* that expresses the number of timesteps that should be executed before reaching $\hat{\mathbf{R}}$. At execution time, both $\hat{h}$ and $\hat{\mathbf{R}}$ are chosen by the decision maker at the start of the episode. Then, at every timestep, the desired horizon is updated according to the perceived reward $\mathbf{r}_t$, $\hat{\mathbf{R}} \leftarrow \hat{\mathbf{R}} - \mathbf{r}_t$ and the desired horizon is decreased by one, $\hat{h} \leftarrow \hat{h} - 1$. More detail on the PCN algorithm can be found in Sec. B of the Appendix.

3.4.1 Training PCN for continuous actions

PCN trains the network as a classification problem, where each class represents a different action. Transitions $x = \langle \mathbf{s}_t, h_t, \mathbf{R}_t\rangle, y = a_t$ are sampled from the dataset, and the ground-truth output y is compared with the predicted output $\hat{y} = \pi(\mathbf{s}_t, h_t, \mathbf{R}_t)$. The predictor (i.e., the policy) is then updated using the cross-entropy loss function (Shore & Johnson, 1980):

$$H = -\sum_{a \in \mathcal{A}} y_a \log \pi(a|\mathbf{s}_t, h_t, \mathbf{R}_t) \tag{16}$$

where $y_a = 1$ if $a = a_t$ and $y_a = 0$ otherwise.

While the original PCN algorithm is designed for MOMDPs with discrete action-spaces, the problem we tackle is defined in terms of a continuous action-space. We thus extend PCN to the continuous action-space setting. We change the output of the neural network such that there is a single output value for each dimension of the action-space. Since the actions should be bound in the domain of possible actions ($[0, 1]$ in the case of MOBelCov, see Section 3.3), we apply a sigmoid non-linearity function on each output, as the output of the sigmoid function is bound in $[0, 1]$. Since the labeled dataset now uses continuous labels $y = \mathbf{a}_t$ instead of discrete ones, we have a regression problem instead of a classification problem. We thus use a Mean Squared Error (MSE) loss to update the policy:

$$MSE = \frac{1}{|\mathcal{A}|}\sum_{a \in \mathcal{A}} (\hat{y}_a - y_a)^2 \tag{17}$$

Learning the full set of Pareto-efficient policies Π^* requires that the policies $\pi^* \in \Pi^*$ are deterministic stationary policies. If stochastic policies are permitted, the set of optimal policies corresponds to the convex part of the Pareto front (Roijers et al., 2013). However, we argue that, in the context of mitigation policies, deterministic policies are required, as the population needs to be informed in advance of imposed measures. Thus, we use the output $\hat{\mathbf{y}}$, which is deterministic, as the action at execution time. However, PCN improves its policy through exploration, by continuously updating its dataset with better trajectories. Thus, at training time, we use a stochastic policy by adding random noise to the action (Lillicrap et al., 2015):

$$\mathbf{a}_t = \pi(\mathbf{s}_t, h_t, \mathbf{R}_t) + \eta s \text{ with } s \sim \mathcal{N}, \tag{18}$$

where $\mathcal{N}$ is the standard Normal distribution and η is a hyper-parameter defining the magnitude of noise to be added.

3.4.2 Coping with stochastic transitions in PCN

PCN trains its policy on a dataset that is collected by executing trajectories. It assumes that reenacting a transition from the dataset leads to the same episodic return. When the transition function $\mathcal{T}$ of the MOMDP is deterministic, the whole trajectory can be faithfully reenacted, which guarantees the same return. Combined with the fact that PCN's policy is deterministic at execution time, conditioning the policy on a target episodic return is equivalent to conditioning it on the V-value $\mathbf{V}$.

However, when $\mathcal{T}$ is stochastic this can no longer be guaranteed. To mitigate this, we add a limited amount of random noise to $\mathbf{R}_t$ when performing gradient descent, which reduces the risk of overfitting (Zur, Jiang, Pesce, & Drukker, 2009). Moreover, while the MOBelCov model is stochastic, the variation is entirely due to the sampling of the binomial distributions in the binomial-chain. While this variation accumulates over time, the time window we consider for each timestep (i.e., one week) is short enough that the accumulation remains bounded. Thus, the possible next-states resulting from a state–action pair are similar to each other. This allows PCN to compensate if $\mathbf{r}_t = \boldsymbol{\mathcal{R}}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$ is worse than expected.

Although we use a stochastic model to cope with the variability of the outcome of the outbreak, it is possible to deterministically evaluate the set of ordinary differential equations that define the model. We assess the validity of our approach by executing PCN on this deterministic variant, and observing similar performance as with the MOBelCov model. We show the results in Sec. C.1 of the Appendix. Related to this, Maillard, Mann, and Mannor (2014) define a "hardness" measure for MDPs and show that in practice many MDPs are far from the pathological, hard-to-solve MDPs that are analyzed in theory. In particular, deterministic MDPs show low hardness as the number of samples required to estimate the Value-function is much lower than MDPs with high stochasticity. The fact that PCN achieves similar performance on the stochastic and deterministic variant of the MOBelCov model indicates that they have a similar hardness measure. Thus, the stochasticity of the Binomial model is limited and has no long-term persecutions on the policy.

4 Results

Our goal is to use PCN to learn deconfinment strategies in the MOBelCov environment. We aim to learn policies that balance between the epidemiological objective of minimizing the attack rate (i.e., $\mathcal{R}_{\mathrm{ARH}}$ for hospitalization and $\mathcal{R}_{\mathrm{ARI}}$ for infection) and the social burden (i.e., $\mathcal{R}_{\mathrm{SB}}$) experienced by the population due to the implemented mitigation measures. Although infections and hospital admissions are correlated, the age-specific differences in transmission and disease burden might result in distinct optimal trade-offs, as the affected social environments are different. To this end, we consider two cases for the vectorial reward functions $[\mathcal{R}_{\mathrm{ARH}}, \mathcal{R}_{\mathrm{SB}}]$ and $[\mathcal{R}_{\mathrm{ARI}}, \mathcal{R}_{\mathrm{SB}}]$, to learn and analyze policies under different targets with respect to the considered attack rate.

To conduct this analysis, we apply our extension of PCN for continuous action-spaces on the MOBelCov model. As explained in Section 3.4.2, we extend PCN for environments with stochastic transitions.

As per Abrams et al. (2021), the simulation starts on the 1st of March 2020, by seeding a number of infections in the population. Two weeks later, on the 14th of March, the Belgian government initiated a strict lockdown. This is implemented by fixing the actions p_w, p_s, p_l to $0.2, 0, 0.1$ respectively. This lockdown ended on the 4th of May 2020, at which point the government decided on a multi-phase exit strategy to incrementally reduce teleworking, reopen schools and allow leisure activities, such as the gradual re-opening of bars and the cultural sector. It is from this day onward that PCN aims to learn policies for diverse exit strategies, compromising between the total number of daily hospitalizations and the total number of contacts lost as a proxy for social burden. The simulation lasts throughout the summer school holidays, from 01/07/2020 to 31/08/2020. Schools are closed during the school holidays, which is simulated by setting $p_s = 0$, regardless of the corresponding value outputted by the policy, i.e., during periods of school closure p_s is ignored.

We draw the analogy with the multi-phase exit strategy established by the Belgian government and the restriction on the number of action-changes imposed by the MOBelCov's budget $\boldsymbol{b}$. Indeed, on the 11th of May 2020, exactly one week after the end of the lockdown, stores and certain companies were allowed to reopen, under strict conditions. This corresponds to altering p_w in our MOMDP. One week later, on 18th of May, primary and secondary schools reopened for limited sized class-groups, and the cultural sector reopened partially. This is equivalent to increasing p_s and p_l. Further changes of restrictions occurred on the 8th of June, 1st, 9th and 25th of July. Thus, we argue that, with a limited budget, we can achieve realistic policies. In our experiments, we consider budgets of 2 to 5, as these closely relate to the number of changes that occurred until the end of the summer holidays of 2020. As an upper bound, we also consider a no-limit budget setting.

To evaluate the quality of the policies learned by PCN, we compare PCN to a baseline. This baseline consists of a set of 100 fixed policies, that iterate over all the possible social restriction levels, with values ranging between 0 and 1. Concretely, each policy uses a fixed proportional reduction $p_w = p_l = p_s = u, u \sim \mathcal{U}(0,1)$ throughout the episode. In other words, the fixed policies directly operate in a fine-grained manner on the whole contact reduction function $\hat{C}$. This allows us to obtain a strong baseline for potential exit strategies over the objective space. We note that while such fixed policies are a feasible approach, they do not scale well in terms of action and objective spaces and they will not be able to provide an adaptive restriction level, which is our aim using PCN. This baseline provides a reference to show the improvement of using dynamic learning methods. In this section, we focus on the policies learned by PCN. We confirm the quality of the learned policies in Sec. C.3 of the Appendix, by comparing PCN with Multi-Objective Natural Evolution Strategies (MONES) (Parisi, Pirotta, & Peters, 2017), another dynamic learning algorithm that searches for the set of Pareto-optimal policies.

All experiments are averaged over 10 runs. The initial choice for the number of runs was informed by our previous research experience with the PCN algorithm. To ensure that this number of experiments was sufficient, we assessed the variance of the Pareto front (Fig. 3) and the variance of the different evaluation metrics (Table 1). The hyper-parameters and the neural network architecture can be found in Sec. D.1 and Sec. D.2 of the Appendix, respectively.

4.1 Learned coverage set

We learn a coverage set (see Fig. 3) that ranges from imposing minimal restrictions to enforcing many restrictions. In Fig. 3, we display on the right the coverage set of the best-performing run in terms of hypervolume, for each budget setting. On the left, we show an interpolated average of the coverage sets learned by the different runs.

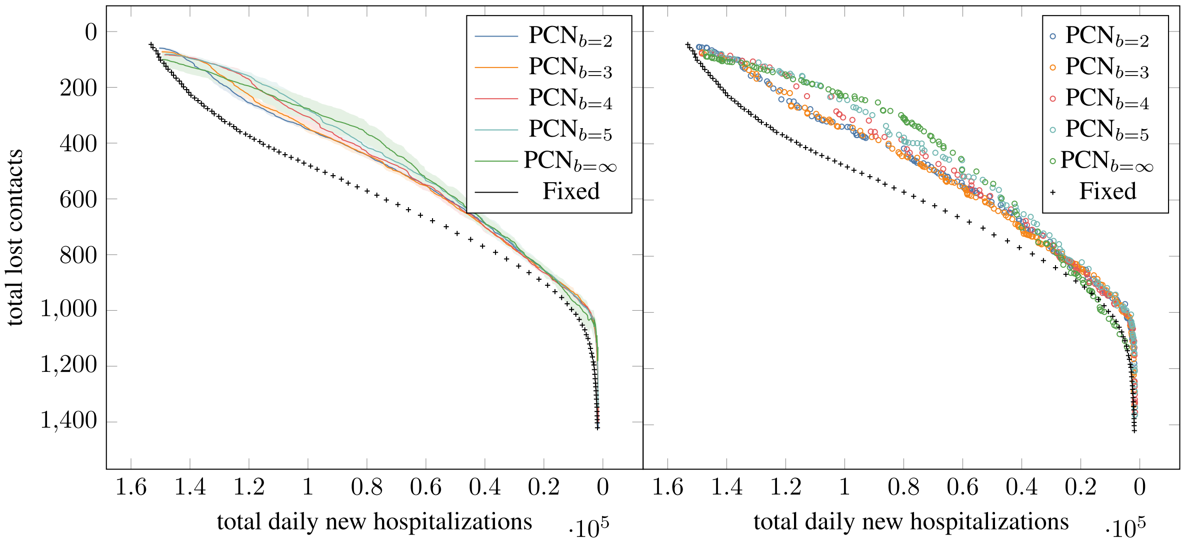

Fig. 3. The Pareto front of policies discovered by PCN using MOBelCov, showing the different compromises between the number of hospitalizations and the number of lost contacts. On the right, we show, for each budget setting (colored, subscript indicates budget) the coverage set learned by the best performing run. On the left, we show an interpolated average of the coverage sets learned by the different runs, with the shaded regions corresponding to the standard deviation. For comparison, the baseline is displayed on both plots (in black). As the budget increases, so does the size of the coverage set learnt by PCN. Changes are most noticeable in the less restrictive trade-offs in terms of social burden.

Table 1
Evaluation metrics for the coverage sets comparing hospitalizations with social burden. In general, an increase of budget results in a better coverage set. Training on infections (ARI) still provides a competitive coverage set in terms of hospitalizations. All PCN coverage sets outperform the baseline.

	$[\mathcal{R}_{\text{ARH}}, \mathcal{R}_{\text{SB}}]$			$[\mathcal{R}_{\text{ARI}}, \mathcal{R}_{\text{SB}}]$		
	Hypervolume	I_ϵ	$I_{\epsilon-mean}$	Hypervolume	I_ϵ	$I_{\epsilon-mean}$
$\text{PCN}_{b=2}$	158.370 ± 0.811	0.080 ± 0.011	0.033 ± 0.002	157.152 ± 1.023	0.087 ± 0.006	0.035 ± 0.002
$\text{PCN}_{b=3}$	158.721 ± 1.439	0.080 ± 0.012	0.032 ± 0.003	158.002 ± 2.081	0.084 ± 0.009	0.034 ± 0.005
$\text{PCN}_{b=4}$	160.642 ± 1.582	0.075 ± 0.007	0.028 ± 0.003	159.315 ± 2.601	0.088 ± 0.018	0.031 ± 0.006
$\text{PCN}_{b=5}$	163.104 ± 2.386	0.070 ± 0.023	0.022 ± 0.005	161.792 ± 2.464	0.075 ± 0.015	0.026 ± 0.005
Fixed	140.479 ± 0.000	0.139 ± 0.000	0.073 ± 0.000	140.479 ± 0.000	0.139 ± 0.000	0.073 ± 0.000
PCN	159.462 ± 7.713	0.264 ± 0.115	0.036 ± 0.020	159.852 ± 2.395	0.171 ± 0.093	0.032 ± 0.006

Regardless of the imposed budget, we notice that the coverage sets discovered by PCN almost completely dominate the coverage set of the baseline, demonstrating that there are better alternatives to the fixed policies. This is most evident in the compromising policies, where one has to carefully choose when to remove social restrictions while at the same time minimizing the impact on daily new hospitalizations. In these scenarios, PCN learns policies that drastically reduce the total number of new hospitalizations (e.g., more than 20000) for the same social burden. We analyze the executions of such policies in Fig. 4 (middle plot), that shows a flattened hospitalization curve, with a gradual increase of social freedom during the school holidays such that the curve of the epidemic is flattened and gradually decreases over time.

Interestingly, we notice that the most restrictive policy (i.e., the one that prioritizes hospitalizations over social burden, see Fig. 4, bottom plot) still starts to gradually increase p_w and p_l from the end of July onward. This is because by then, the epidemic has mostly faded out, and it is safe to reduce social restrictions. The timing of this reduction is important as reducing restrictions too soon can lead to a new wave. PCN learns the impact of its decisions over time, and correctly infers the timing at which restrictions can be safely lifted.

Finally, the top plot shows that, without imposing social restrictions, the number of hospitalizations peaks on the 15th of June. By the beginning of July, the epidemic has spread out over the majority of the population, and the number of admissions at the hospital has been reduced to a fraction of the number of hospitalizations at the peak. Thus, without social restrictions, we do not take advantage of the natural decrease of social contacts due to school holidays, as a significant proportion of the population has already been infected before the start of the holidays.

4.2 Impact of budgets on the coverage set

Fig. 3 demonstrates that the budget impacts the learned coverage set. In general, an increase of budget is associated with an increasingly better coverage set, as policies learned using a higher budget dominate the ones learned with a lower budget. This is to be expected, as a higher budget gives the agent more freedom to change its actions as the epidemic progresses.

Moreover, we observe that the difference is concentrated around the less restrictive policies in terms of social burden. We postulate that this region contains the most complex policies, as these try to maintain as much social freedom as possible, while containing the number of hospitalizations. In these cases, the timing of the actions coincide with the timing and duration of the peak of the epidemic, and a higher budget allows for more fine-grained control to manage this timing. To confirm this, we select, for each budget setting, the solution where the difference in performance is most noticeable, corresponding to the solutions with a total number of hospitalizations around 80000 (which is in the middle of the range of possible hospitalizations, as can be seen in Fig. 3). We plot the execution of the corresponding policies in Fig. 5 and analyze their impact in terms of social burden. First, we observe that the lower-budget policies are unable to reduce the social restrictions past the peak of the epidemic. In contrast, the setting with no budget restrictions meticulously controls the restrictions as the epidemic progresses, completely removing restrictions by the end of the

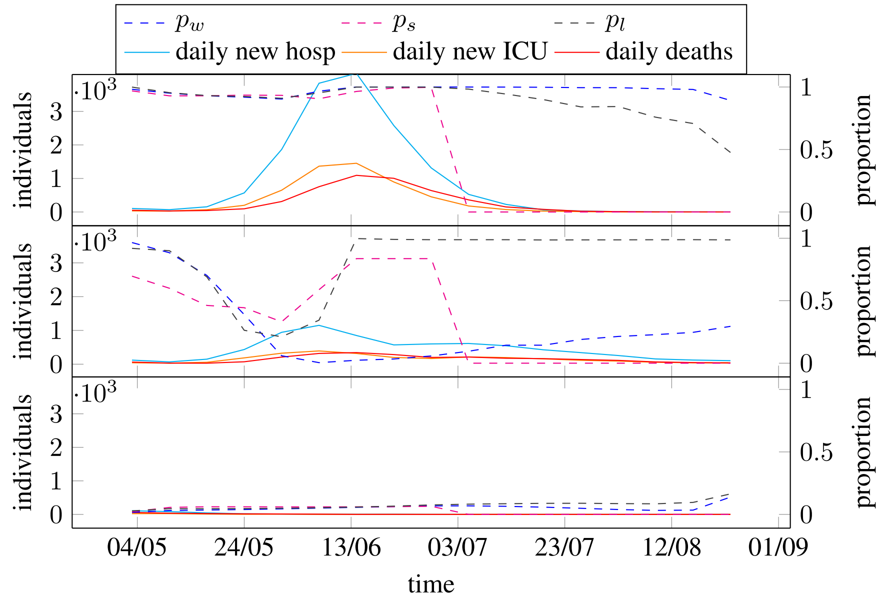

Fig. 4. Selection of policies learned by PCN$_{b=5}$, from least restrictive in terms of social burden (top) to most restrictive (bottom). The x-axis represents the time, starting from the end of the lockdown, on the 4th of May, until the end of the school holidays, on the 1st of September. Since the lockdown is simulated before the start of the exit strategy, the start-state differs for each episode (i.e., the hospital already contains infected individuals). The left y-axis represents the number of individuals affected by the epidemic. The full-lined plots represent the number of individuals admitted into the hospital, ICU and deceased between the last timestep and the current one. The plot showing the newly deceased individuals closely relates to the ICU admissions. The right y-axis represents the proportional reduction in effect, with 1 meaning a business-as-usual policy, and 0 meaning a complete suppression of social contacts. The dotted-lined plots represent the proportional reductions for the work, leisure and school environments. We note that the school reduction automatically goes to 0 at the start of the school holidays.

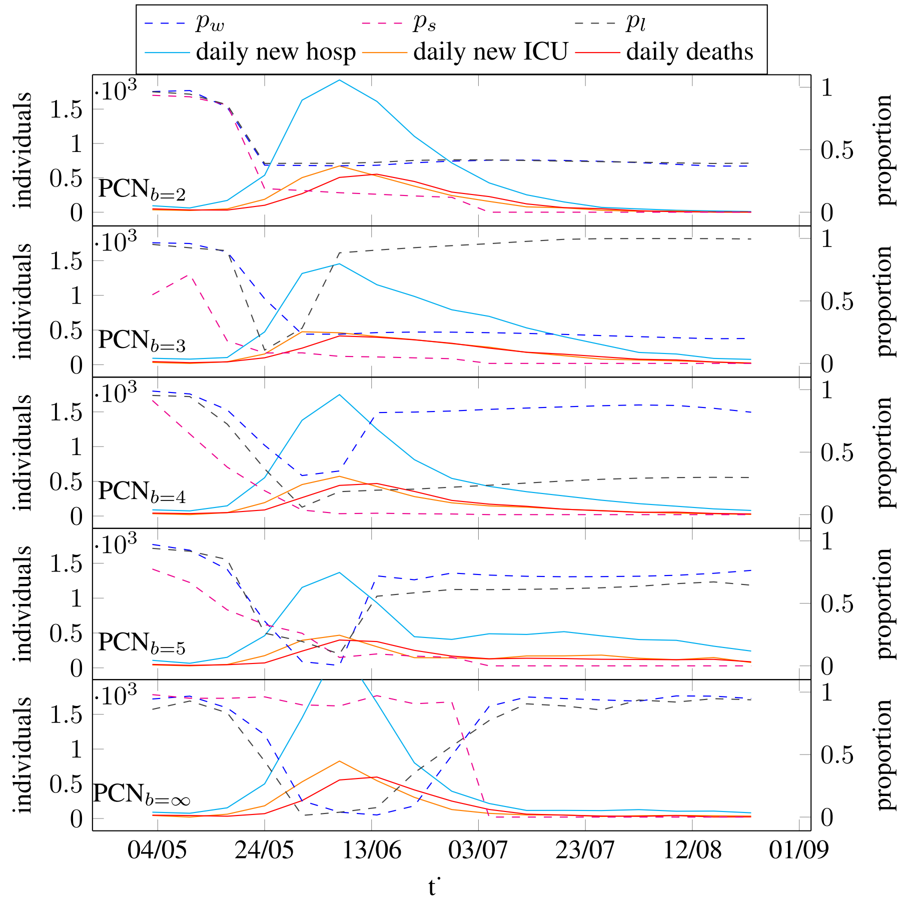

Fig. 5. Execution of the policies attaining a number of hospitalizations around 80000, for different budgets. From top to bottom we display the policy executions with budget 2, 3, 4, 5 and no-limit, respectively. We notice that the lower-budget policies are unable to reduce the social restrictions past the peak. The setting without budget restrictions finely controls the restrictions as the epidemic progresses, completely removing restrictions by the end of the wave. Finally, there is no consensus on which social environment to restrict most: certain policies provide similar restrictions for p_w and p_l, while others impose harsher restrictions on one social environment than the other.

wave. Second, we note that the policy with a budget of 5 resembles the execution of the one without restrictions. However, due to its budget, the policy is unable to progressively reduce the restrictions and instead resorts to a halfway compromise. Compared to this specific region of the coverage set, the difference in performance between different budget settings seems marginal around the extrema. At the extrema, the policies are less complex (e.g., business-as-usual, resulting in the same action executed throughout the episode) and are thus less impacted by the budget restrictions.

Finally, we observe that, while the extrema deliver similar trade-offs for any of the chosen budgets, these trade-offs differ for the setting without budget restrictions. Indeed, in this setting, PCN does not learn the most extreme policies with respect to restrictions, even though there are no constraints on the action-set. As explained in Sec. B.2, PCN searches for increasingly better solutions using a stochastic policy. Thus, at every timestep, the action can change compared to the previous one. Continuously outputting the same action (e.g., no social restrictions) becomes a complicated task. In comparison, for the settings with a limited budget, the action stays the same as the previous one once the budget has been spent. As such, it is easier to learn the most extreme policies. Thus, in the specific case where we have an unlimited budget, the freedom of action actually hinders PCN's search for certain regions of the reward-space.

4.3 Analysis of the coverage set's inflection point

The coverage set shown in Fig. 3 displays, on the far right of the plot, a sharp decrease in social contacts, for a marginal improvement on hospitalizations. We analyze the policies at the start of this decrease, when the social burden results in 1000 lost contacts. For each budget setting, we plot the policy executions that result in $\mathcal{R}_{\mathrm{SB}} = -1000$ in Fig. 6. On average, these policies result in a total of 3811 hospitalizations over the considered time-horizon (or 2.59% hospitalizations with respect to the business-as-usual policy). Regardless of the budget, these policies allow one of the 3 social environments to be open, while closing the 2 others. Except for a budget of 4 (which opens the working environment), each policy opens the leisure environment. On average, these policies enforce a reduction of 71.92% of the number of social contacts. Thus, allowing each citizen to see 28.08% of his social contacts by keeping the leisure environment open as per Fig. 6 results in a marginal number of hospitalizations.

4.4 Minimizing the hospitalizations versus minimizing the number of infections

Next, we assess the difference in coverage sets when optimizing on the number of hospitalizations $\mathcal{R}_{\mathrm{ARH}}$ versus the number of infections $\mathcal{R}_{\mathrm{ARI}}$. Although these reward functions have a different scale (there are more infected persons than hospitalized ones), our experiments show that infections and hospitalizations are tightly correlated. This is expected, as during the initial phase of the epidemic, limited immunity was present in the population (i.e., limited natural immunity and no vaccines), which induces a tight coupling between infection and hospitalization cases. This is confirmed in Table 1. In this table, we show the different performance metrics (hypervolume, I_ε, $I_{\varepsilon-mean}$) with respect to the objectives $[\mathcal{R}_{\mathrm{ARH}}, \mathcal{R}_{\mathrm{SB}}]$. The table is split in two parts. The left-side shows the different performance metrics, for PCN using $[\mathcal{R}_{\mathrm{ARH}}, \mathcal{R}_{\mathrm{SB}}]$ as optimization criteria. The right-side shows the same performance metrics, but with PCN using $[\mathcal{R}_{\mathrm{ARI}}, \mathcal{R}_{\mathrm{SB}}]$ as optimization criteria.

Even with the $\mathcal{R}_{\mathrm{ARI}}$, the increased budget shows an increase in hypervolume in terms of hospitalizations. Moreover, those hypervolumes are close to the ones trained on $\mathcal{R}_{\mathrm{ARH}}$. Combined with the plotted coverage sets (Fig. C.2, this indicates that their coverage sets are similar. However, regardless of the imposed budget, the hypervolumes are slightly worse. This is to be expected, since those experiments are not directly optimized on $\mathcal{R}_{\mathrm{ARH}}$. We draw a similar conclusion for I_ε: for budgets 2, 3 and 5, the difference between the worst-performing policy for the $\mathcal{R}_{\mathrm{ARI}}$ variant and the $\mathcal{R}_{\mathrm{ARH}}$ is less than 0.01, indicating less than 1% difference in return values between the two variants when comparing their worst-performing policy. As an exception, we notice that PCN without budget restrictions results in better performance across every metric for the $\mathcal{R}_{\mathrm{ARI}}$ variant. Still, due to the high standard deviation of the unlimited budget, $\mathcal{R}_{\mathrm{ARH}}$ setting, we do not consider this difference is meaningful. Thus, we could optimize on the attack rate of hospitalizations with $\mathcal{R}_{\mathrm{ARI}}$. As there is a 2-week delay for hospitalizations, this would facilitate learning policies to react to unexpected changes earlier than using $\mathcal{R}_{\mathrm{ARH}}$. This assumes that a good proxy to the actual number of infections was available (e.g., due to a scale up of PCR testing, as was the case after the first lockdown).

Table 2
Comparing the difference in the desired return provided to PCN and the actual return PCN obtained when executing its policy. We see that, regardless of the setting, the learned policy faithfully receives a return similar to its desired return.

	I_ε	$I_{\varepsilon-mean}$
$\mathrm{PCN}_{b=2}$	0.047 ± 0.020	0.009 ± 0.004
$\mathrm{PCN}_{b=3}$	0.048 ± 0.022	0.007 ± 0.002
$\mathrm{PCN}_{b=4}$	0.064 ± 0.018	0.011 ± 0.003
$\mathrm{PCN}_{b=5}$	0.058 ± 0.011	0.013 ± 0.003
PCN	0.035 ± 0.011	0.008 ± 0.004
Global average	0.050 ± 0.017	0.010 ± 0.004

Finally, we observe that, even though the obtained coverage sets (shown in Fig. C.2 of the Appendix)) are similar, the coverage set trained on $\mathbf{R}_{\mathrm{ARI}}$ is systematically dominated by the one when trained on $\mathbf{R}_{\mathrm{ARH}}$. While hospitalizations and infections are highly correlated, they differ in terms of age-groups. Older age-groups are more susceptible to be hospitalized after being infected, but they do not form the majority of the population. For trade-offs where infections and social burden need to be balanced, the proportional reductions target different social environments than for trade-offs balancing hospitalizations and social burden. For example, as the work environment is mostly comprised of individuals with a more robust immune system, reducing the social contact in this environment greatly affects the number of infections, but has a lesser impact on the number of hospitalizations.

Our experiments show, on average across all budgets, that policies with a similar social burden are 5.37% more permissive, 16.41% more permissive, 4.02% less permissive on p_w, p_s, p_l respectively, when trained on ARI than when trained on ARH. These numbers were computed as follows: at timestep t, the contact matrix depends on the proportional reduction p_w, p_s, p_l imposed by the policy π, i.e., $\hat{\boldsymbol{C}}_t^{\pi} = \hat{C}(p_w, p_s, p_l), \quad p_w, p_s, p_l \sim \pi(\mathbf{s}_t)$. These proportional reductions are different for policies trained on $\mathbf{R}_{\mathrm{ARI}}$ and policies trained on $\mathbf{R}_{\mathrm{ARH}}$. The difference $\Delta\hat{\boldsymbol{C}}_t = \hat{\boldsymbol{C}}_t^{\pi_{\mathbf{R}_{\mathrm{ARI}}}} - \hat{\boldsymbol{C}}_t^{\pi_{\mathbf{R}_{\mathrm{ARH}}}}$ shows which social contacts are more affected by $\pi_{\mathbf{R}_{\mathrm{ARH}}}$ than by $\pi_{\mathbf{R}_{\mathrm{ARI}}}$, the policy trained on $\mathbf{R}_{\mathrm{ARH}}$, $\mathbf{R}_{\mathrm{ARI}}$, respectively. Fig. 7 shows the average $\Delta\hat{\boldsymbol{C}}_t$, across all timesteps and all trained policies, i.e., $\mathbb{E}_{\pi_{\mathbf{R}_{\mathrm{ARI}}} \sim \Pi^*_{\mathbf{R}_{\mathrm{ARI}}}, \pi_{\mathbf{R}_{\mathrm{ARH}}} \sim \Pi^*_{\mathbf{R}_{\mathrm{ARH}}}} [1/T \sum_{t=0}^{T} \Delta\hat{\boldsymbol{C}}_t \mid \pi_{\mathbf{R}_{\mathrm{ARI}}}, \pi_{\mathbf{R}_{\mathrm{ARH}}}]$, where $\Pi^*_{\mathbf{R}_{\mathrm{ARI}}}$, $\Pi^*_{\mathbf{R}_{\mathrm{ARH}}}$ represents the Pareto front under $\mathbf{R}_{\mathrm{ARI}}$, $\mathbf{R}_{\mathrm{ARH}}$, respectively. Even though the proportional-reductions are not age-group specific, PCN learns the impact of the different social environments on the age-groups. The changes in p_w, p_s, p_l show that optimizing on infections is less restrictive regarding older age-groups (i.e., people aged 60 or more). In contrast, social restrictions are significantly higher for persons between 20 and 60 years old.

4.5 Robustness of policy executions

The dataset of trajectories that PCN is trained on is pruned over time to keep only the most relevant trajectories. The returns of these trajectories are used in Fig. 3 to visualize the learned coverage set. Each of these returns can be used as the desired return for policy execution.

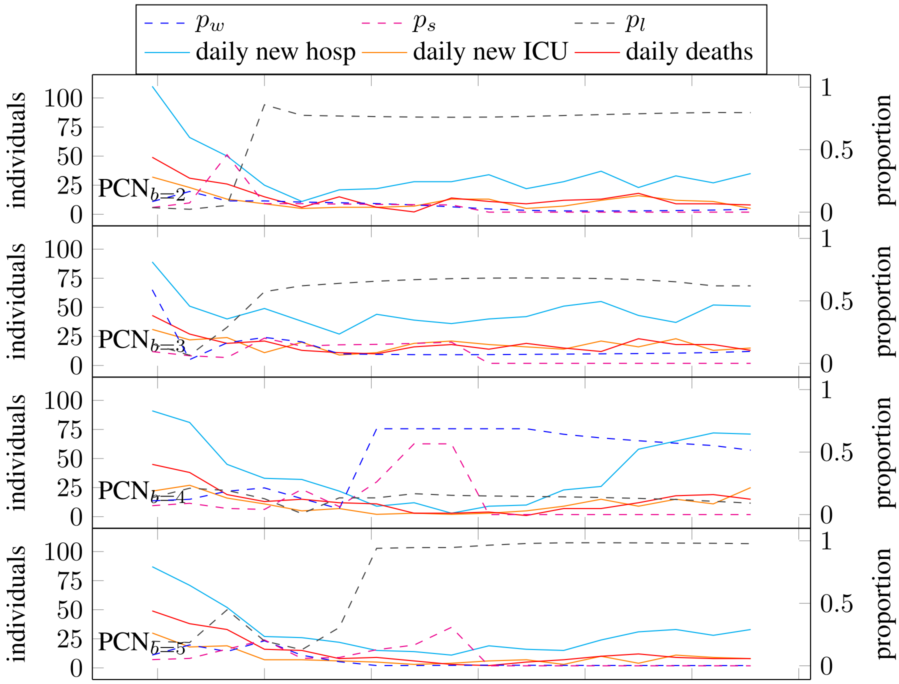

Fig. 6. Policy where $\mathcal{R}_{SB} = -1000$, for each budget setting. These policies focus on minimizing the number of hospitalizations, with a total of 3811 hospitalizations over the considered time-period, compared to the 150000 hospitalizations of business-as-usual policies. However, these policies still allow for a significant percentage of the pre-pandemic social contacts.

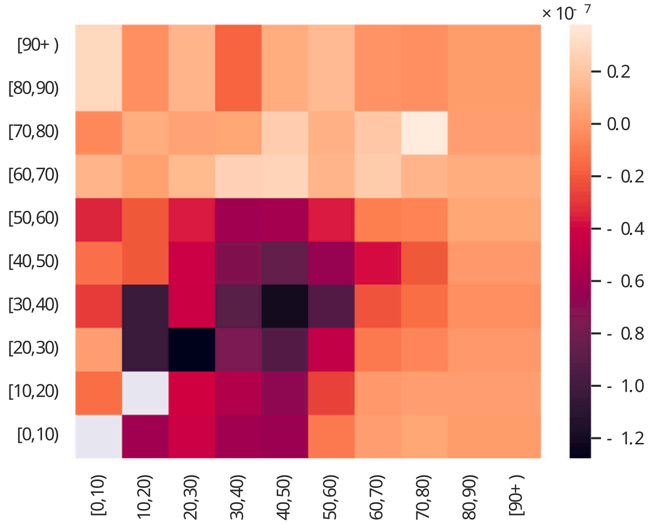

Fig. 7. We show a matrix that visualizes the expected difference between two contact matrices. To this end, we consider the social contact matrix $\hat{C}_t^{\pi} = \hat{C}(p_w, p_s, p_l), \quad p_w, p_s, p_l \sim \pi(s_t)$ resulting from the policy π. We compute the difference between the matrices resulting from policies optimized on $\mathbf{R}_{ARI}$ and $\mathbf{R}_{ARH}$. Thus, each cell represents, for two specific age-groups, how many more social interactions they have on average under a policy optimized on $\mathbf{R}_{ARH}$ than under a policy trained on $\mathbf{R}_{ARI}$. Note that matrix can be asymmetric, as interactions across age-groups can be asymmetric (e.g., 1 teacher interacts with a classroom of children). In general, the matrix shows that optimizing on infections is less restrictive on older age-groups (60+) than on younger ones.

We now assess the robustness of the executed policies, by comparing the return obtained after executing the policy with the corresponding target return. For each run, we execute the policy 10 times and compute the I_{ϵ} and $I_{\epsilon-mean}$ metrics with respect to the coverage set learned during the run. We show that the executed policies reliably obtain returns that are similar to the desired return used to condition PCN.

Results are shown in Table 2. The I_{ϵ} indicators shows that, regardless of the budget, the decision maker will lose at worst a 0.050 normalized return in any of the objectives. On average, it will lose 0.010 normalized returns, i.e., on average, the return obtained by executing a policy will either result in an additional 1441 hospitalizations than expected, or result in 12 additional social contacts lost. Moreover, we emphasize that the learned coverage set contains the non-dominated returns encountered over the whole training procedure. Since the MOBelCov model is stochastic, for multiple executions of the same policy, the executions kept in the coverage sets are the ones for which the samples from the binomial-chain resulted in a better progression of the epidemic than average. Thus, we expect our policy-executions to be close to the target selected from the coverage set, but not exactly on target. Based on this analysis, we conclude that the policies trained by PCN are robust and produce returns as close as possible from their chosen target.

5 Related work

Reinforcement learning (RL) has been used in conjunction with epidemiological models to learn policies to limit the spread of diseases and predict the effects of possible mitigation strategies (Libin, Moonens, et al., 2021; Libin et al., 2018; Probert et al., 2019).

RL and Deep RL have been used extensively as a decision making aid to reduce the spread of COVID-19. For example, to learn effective mitigation strategies (Ohi, Mridha, Monowar, Hamid, et al., 2020), to assess lockdown and travel restrictions (Kwak, Ling, & Hui, 2021) and to evaluate the limitation on the influx of asymptomatic travelers (Bastani et al., 2021).

Multi-objective methods have also been deployed to learn optimal strategies to mitigate the spread of COVID-19. Wan, Zhang, and Song (2021) implement a model-based multi-objective policy search method and demonstrate their method on COVID-19 data from China. Given that this method is model-based, a model of the transition function must be learned by sampling from the environment. The method proposed by Wan et al. (2021) only considers a discrete action space which limits the applicability of their algorithm. Wan et al. (2021) use linear weights to compute a set of Pareto optimal policies. However, methods which use linear weights can only learn policies on the convex-hull of the Pareto front (Vamplew, Yearwood, Dazeley, & Berry, 2008), therefore the full Pareto front cannot be learned. We note that the method proposed by Kompella et al. (2020) considers multiple objectives. However, the objectives are combined using a weighted sum with hand-tuned weights which are determined by the authors. The weighted sum is applied by the reward function and a single objective RL method is used to learn a single optimal policy. In contrast to previous work, our approach makes no assumptions regarding the scalarisation function of the user and is able to discover Pareto fronts of arbitrary shape.

In this work, due to the nature of the epidemiological decision problem, we have no prior knowledge on the expected shape of the coverage set. The fact that PCN does not make any assumptions on the shape of the Pareto front, motivates the use of PCN for our setting. In contrast, most other work in the MORL literature assumes that the preferences of the decision maker can be modeled as a weighted sum over the objectives (Abels, Roijers, Lenaerts, Nowé, & Steckelmacher, 2019; Alegre, Bazzan, & Da Silva, 2022; Alegre, Bazzan, Roijers, Nowé, & da Silva, 2023; Castelletti, Pianosi, & Restelli, 2012), in which case, the Pareto front is assumed to be convex (Roijers et al., 2013).

6 Discussion

Making decisions on how to maintain epidemic situations has important ethical implications with respect to public health and societal burden. In this regard, it is crucial to approach this decision making from a balanced perspective, to which end we argue that multi-objective decision making is essential. In this work, we establish a novel approach, i.e., an expert system, to study multi-faceted policies, and this approach shows great potential to study future epidemic mitigation policies. We are aware of the ethical implications that expert systems have on the decision process and we make the disclaimer that all results based on the expert system that we propose should be carefully interpreted by experts in the field of public health, and in a broader context that encompasses health economics, well-being and education. We note that the work in this manuscript was conducted by a interdisciplinary consortium that includes computer scientists and scientists with a background public health, epidemiology and bio-statistics, to allow for a balanced perspective regarding these disciplines.

In this work, we focus on the clinical outcomes of intervention strategies and use the reduced contacts as proxy for social burden. It is important to note that the definition of social burden in this work consists of a proxy that aggregates distinct aspects of the burden that a population experiences. Furthermore, we chose to focus on the burden of susceptible and recovered individuals, and did not include infected individuals in the social burden statistic. Overall, this is a reasonable assumption, however, one could argue that asymptomatic infected individuals also endure social burden. However, as the interactions of asymptomatic individuals may also induce new infections, and as such impact the other objective that is a derivative of the number of infected individuals, we choose not to include them in the social burden statistic. This could be extended into more formal health economic evaluations, by designing reward functions that explicitly consider distinct health economic principles. The COVID-19 pandemic demonstrates the broad impact of infectious diseases on sectors other than health care. This stresses the need to capture a societal and thus multi-objective perspective in the decision making process on public health and health care interventions. Our learned policies confirm this, showing that focusing solely on preventing hospitalizations admissions as much as possible, with the aim to keep these admissions very low, results in taking drastic measures – more than a thousand social interactions lost per person over the span of 4 months – that may have a long-lasting impact on the population. An important insight of our analyses is that policies that act fast, i.e., when the number of infections or hospitalizations are low, prove most effective, both with respect to the averted number of infections and the induced social burden. These observations confirm that the use of an infection barometer[2], that defines clear cutoffs regarding infections/hospitalizations and the rate at which they rise, and couples these thresholds to concrete actions, can be a useful instrument to mitigate an ongoing epidemic.

In MOBelCov, we consider the proportional contact reduction actions p_w, p_s, p_l. We acknowledge that these parameters are abstract in nature and reflect the average reduction to the respective contribution of work, school and leisure contacts. The reduction is thus proportional to all individuals, e.g., in this model, it is not possible to close some schools of a particular type completely and keep others open, and as such it concerns effective reduction. To translate these parameters to public health policy, contact reduction measures need to be combined to meet the desired contact reduction proportion. In the context of work contact reductions, certain jobs can be allowed on premise (e.g., supermarket staff, general practitioners and pharmacists), while other jobs can be restricted to be performed from home. In the context of school contacts, policy makers can choose to close certain types of schools (i.e., daycare vs primary vs secondary vs tertiary) or can decide to allow certain groups to attend live classes on particular days. In the context of leisure, policy makers can choose to close certain types of leisure, and keep others open (e.g., hospitality services vs sporting facilities).

Although we use an age-structured compartment model, with social contact matrices to model social interactions, this remains a model that evaluates the progression of the pandemic as an aggregated process over the population. Individual-based models enable more specific and targeted policies, such as contact tracing and household isolation, which potentially improve the epidemiological and social output, hence provide an interesting avenue for future work. However, due to the computational cost of simulating such models, and the number of interactions required by reinforcement learning in general, this remains a challenging problem, that will require fundamental research to improve the sample efficiency of multi-objective reinforcement learning algorithms.

While we are able to interpret and analyze the obtained policies and their corresponding trade-offs, as we can plot the Pareto front for two objectives, this approach cannot be used for problems with more objectives, which will be necessary to cover reward functions that cover distinct health economic principles. To facilitate this kind of research, new algorithms are necessary to enable reinforcement learning in many-objective contexts and to interpret the learnt policies.

In this work, PCN is able to cope with model stochasticity, as the stochasticity is limited due to the small time-window between timesteps, i.e., due to $h = \frac{1}{240}$. This results in more computations for the same time-period. For compartment models, this additional computation is negligible compared to the computation required for stochastic gradient descent used by PCN. However, in settings where the stochasticity is more pronounced, PCN is not suitable and further methodological extensions are warranted. This is for example the case for more complex models, such as individual-based models, that are for example necessary when designing policies to control the initial outbreak of an epidemic. The reason for that is that trajectories encountered at execution time might significantly deviate from the trajectories encountered at training time, on which PCN has been optimized. One way to mediate this issue would be to learn the transition probability function between two states. We can then measure the likelihood of a trajectory, and take it into account when training PCN. We are currently investigating this as ongoing research (Delgrange, Reymond, Nowé, & Pérez, 2023).

[2] https://motivationbarometer.com/en/

In this work we studied policies that aim to balance social burden and hospitalizations. Yet, the methodology that we propose shows promise to address a wide variety of public health challenges, such as balancing the number of lost schooldays with respect to the attack rate of infections in schools (Torneri et al., 2021), the efficacy versus burden of face masks for children (Esposito & Principi, 2020), contact tracing effort compared to the impact of such policies (Willem et al., 2021), the impact of antivirals on the epidemic while balancing the likelihood for resistance mutations to emerge (Torneri et al., 2020), to balance the efforts and insights of COVID-19 genomic surveillance (Chen et al., 2022), and to balance the cost of universal testing and its impact on an emerging epidemic (Libin, Willem, et al., 2021).

In MORL, there exist two possible optimization criteria: the Expected Scalarized Returns (ESR) and the Scalarized Expected Returns (SER). The difference between the two is where the preferences of the decision maker are applied: either on the episodic return (ESR), or on the expected return (SER). Methods that learn the whole Pareto front optimize for SER, as each solution on the Pareto front is expressed as an expected (vectorial) return. However, in the case of epidemiological outbreaks, the mitigation strategy will only be applied once (afterwards, the epidemic will be over). Thus, the appropriate optimization criterion to use is ESR. PCN learns a deterministic policy, and assumes that the environment has a deterministic transition function. Under those assumptions, optimizing SER is equivalent to optimizing ESR. When these assumptions hold, PCN thus optimizes policies under ESR. However, our environment and its transition function is stochastic and as such these assumptions do not hold exactly. We argue that, since PCN optimizes on episodic returns, and not on Q-values, the policies it learns are geared towards single executions, thus complying with ESR. Furthermore, although our environment is stochastic, this stochasticity is limited, and additional experiments displayed in Appendix Fig C.1 show that these policies are very similar to policies trained on a deterministic version of the environment. This indicates that the optimal policies for SER and ESR are similar.

To conclude, we show that multi-objective reinforcement learning provides decision maker with insightful and diverse alternatives on real-world problems. PCN automatically learns all Pareto-efficient trade-offs. Although extreme policies can be computed manually, the fact that PCN learns them shows that it explored the whole range of possible social restrictions, which led to many alternative trade-offs between these extreme policies. Moreover, the subtle differences in policies trained on the infection attack rate compared to the hospitalization attack rate show that the social interactions, rate of infection spreading and risk of severe sickness are well-captured by our learning algorithm, which indicates that the learned trade-offs are of high quality. Furthermore, we show that action budgets can act as a regularizer that facilitates learning realistic policies that can be easily conveyed to decision makers. Since the environment dynamics are well-captured, we can analyze the effect of different budget settings. Finally, we notice an inflection point on the right-side of the Pareto front, indicating that taking extreme measures (which can be computed manually) may not be necessary to root out the infection while minimizing the number of hospitalizations. In this work, we demonstrate that multi-objective reinforcement learning adds value to epidemiological modeling as it brings essential insights to balance mitigation policies and provides policy makers with a broader view on the decision space.

CRediT authorship contribution statement

Mathieu Reymond: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Visualization. **Conor F. Hayes:** Methodology, Software, Writing – review & editing. **Lander Willem:** Software, Validation, Investigation, Resources, Writing – review & editing. **Roxana Rădulescu:** Conceptualization, Methodology, Investigation, Writing – review & editing. **Steven Abrams:** Software, Resources, Writing – review & editing. **Diederik M. Roijers:** Supervision, Writing – review & editing. **Enda Howley:** Supervision, Writing – review & editing. **Patrick Mannion:** Supervision, Writing – review & editing. **Niel Hens:** Supervision, Writing – review & editing. **Ann Nowé:** Supervision, Writing – review & editing, Funding acquisition. **Pieter Libin:** Conceptualization, Investigation, Writing – original draft, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

C.F.H. is funded by the University of Galway Hardiman Scholarship, Belgium. This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program. This work also received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant number 101003688 – EpiPose project). P.J.K.L. gratefully acknowledges support from FWO via postdoctoral fellowship, Belgium 1242021N and the Research council of the Vrije Universiteit Brussel (OZR-VUB via grant number OZR3863BOF). N.H. acknowledges support from the Scientific Chair of Evidence-based Vaccinology under the umbrella of the Methusalem framework at the University of Antwerp. N.H. and A.N. acknowledge funding from the iBOF DESCARTES project (reference: iBOF-21-027). L.W. gratefully acknowledges support from FWO postdoctoral fellowship 1234620N. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. A.N. acknowledges the support by the FWO COVID-19 research project G0H0420N. P.L. and L.W. acknowledge support from FWO grant G059423N.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2024.123686.

References

Abels, A., Roijers, D. M., Lenaerts, T., Nowé, A., & Steckelmacher, D. (2019). Dynamic weights in multi-objective deep reinforcement learning. In *Proceedings of machine learning research*: *vol. 97*, *Proceedings of the 36th international conference on machine learning* (pp. 11–20). Long Beach, California, USA: PMLR.

Abrams, S., Wambua, J., Santermans, E., Willem, L., Kuylen, E., Coletti, P., et al. (2021). Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics*, *35*, Article 100449.

Alegre, L. N., Bazzan, A., & Da Silva, B. C. (2022). Optimistic linear support and successor features as a basis for optimal policy transfer. In *International conference on machine learning* (pp. 394–413). PMLR.

Alegre, L. N., Bazzan, A. L., Roijers, D. M., Nowé, A., & da Silva, B. C. (2023). Sample-efficient multi-objective learning via generalized policy improvement prioritization. arXiv preprint arXiv:2301.07784.

Bailey, N. T. (1975). The mathematical theory of infectious diseases and its applications. In *The mathematical theory of infectious diseases and its applications* (p. 413). 5a Crendon Street, High Wycombe, Bucks HP13 6LE: Charles Griffin & Company Ltd.

Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjicristodoulou, C., Lagiou, P., et al. (2021). Efficient and targeted COVID-19 border testing via RL. *Nature*, *599*(7883), 108–113.

Britton, T., & Lindenstrand, D. (2009). Epidemic modelling: aspects where stochasticity matters. *Mathematical Biosciences*, *222*(2), 109–116.

Castelletti, A., Pianosi, F., & Restelli, M. (2012). Tree-based fitted Q-iteration for multi-objective Markov decision problems. In *The 2012 international joint conference on neural networks* (pp. 1–8). IEEE.

Chen, Z., Azman, A. S., Chen, X., Zou, J., Tian, Y., Sun, R., et al. (2022). Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics*, *54*(4), 499–507.

Delgrange, F., Reymond, M., Nowé, A., & Pérez, G. A. (2023). WAE-PCN: Wasserstein-autoencoded Pareto conditioned networks. In *2023 adaptive and learning agents workshop at AAMAS* (pp. 1–7).

Esposito, S., & Principi, N. (2020). To mask or not to mask children to overcome COVID-19. *European Journal of Pediatrics*, *179*(8), 1267–1270.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., et al. (2021). A practical guide to multi-objective RL and planning. arXiv:2103.09568.

Kompella, V., Capobianco, R., Jong, S., Browne, J., Fox, S., Meyers, L., et al. (2020). Reinforcement learning for optimization of COVID-19 mitigation policies. arXiv preprint arXiv:2010.10560.

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), 1–15.

Libin, P. J. K., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P., et al. (2021). Deep reinforcement learning for large-scale epidemic control. In Y. Dong, G. Ifrim, D. Mladenić, C. Saunders, S. Van Hoecke (Eds.), *ECML* (pp. 155–170). Cham: Springer International Publishing.

Libin, P. J., Verstraeten, T., Roijers, D. M., Grujic, J., Theys, K., Lemey, P., et al. (2018). Bayesian best-arm identification for selecting influenza mitigation strategies. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 456–471). Cham: Springer.

Libin, P. J., Willem, L., Verstraeten, T., Torneri, A., Vanderlocht, J., & Hens, N. (2021). Assessing the feasibility and effectiveness of household-pooled universal testing to control COVID-19 epidemics. *PLoS Computational Biology*, *17*(3), Article e1008688.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Maillard, O.-A., Mann, T. A., & Mannor, S. (2014). How hard is my MDP? "The distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, *27*.

Miranda, M. N., Pingarilho, M., Pimentel, V., Torneri, A., Seabra, S. G., Libin, P. J., et al. (2022). A tale of three recent pandemics: Influenza, HIV and SARS-CoV-2. *Frontiers in Microbiology*, *13*.

Ohi, A. Q., Mridha, M., Monowar, M. M., Hamid, M., et al. (2020). Exploring optimal control of epidemic spread using RL. *Scientific Reports*, *10*(1), 1–19.

Parisi, S., Pirotta, M., & Peters, J. (2017). Manifold-based multi-objective policy search with sample reuse. *Neurocomputing*, *263*, 3–14.

Probert, W. J., Lakkur, S., Fonnesbeck, C. J., Shea, K., Runge, M. C., Tildesley, M. J., et al. (2019). Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies. *Philosophical Transactions of the Royal Society B*, *374*(1776), Article 20180277.

Reymond, M., Eugenio, B., & Nowè, A. (2022). Pareto conditioned networks. In *Proceedings of the 21st international conference on AAMAS (2022)*.

Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, *48*, 67–113.

Shore, J., & Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transaction on Information Theory*, *26*(1), 26–37.

Torneri, A., Libin, P., Vanderlocht, J., Vandamme, A.-M., Neyts, J., & Hens, N. (2020). A prospect on the use of antiviral drugs to control local outbreaks of COVID-19. *BMC Medicine*, *18*(1), 1–9.

Torneri, A., Willem, L., Colizza, V., Kremer, C., Meuris, C., Darcis, G., et al. (2021). Controlling SARS-CoV-2 in schools using repetitive testing strategies. (preprint).

Vamplew, P., Yearwood, J., Dazeley, R., & Berry, A. (2008). On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Australasian joint conference on artificial intelligence* (pp. 372–378). Springer.

Wallinga, J., Teunis, P., & Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, *164*(10), 936–944.

Wan, R., Zhang, X., & Song, R. (2021). Multi-objective model-based reinforcement learning for infectious disease control. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 1634–1644).

Willem, L., Abrams, S., Libin, P. J., Coletti, P., Kuylen, E., Petrof, O., et al. (2021). The impact of contact tracing and household bubbles on deconfinement strategies for COVID-19. *Nature Communications*, *12*(1), 1–9.

Willem, L., Van Hoang, T., Funk, S., Coletti, P., Beutels, P., & Hens, N. (2020). SOCRATES: an online tool leveraging a social contact data sharing initiative to assess mitigation strategies for COVID-19. *BMC Research Notes*, *13*(1), 1–8.

Zintgraf, L. M., Kanters, T. V., Roijers, D. M., Oliehoek, F., & Beau, P. (2015). Quality assessment of MORL algorithms: A utility-based approach. In *Benelearn 2015: proceedings of the 24th annual ML conference of Belgium and The Netherlands*.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., & Da Fonseca, V. G. (2003). Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, *7*(2), 117–132.

Zur, R. M., Jiang, Y., Pesce, L. L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical Physics*, *36*(10), 4810–4818.