Mini-project 1: Deep Q-learning for Epidemic Mitigation

In order to ensure the reproducibility of the code, we fixed all the seeds and sources of randomness. In addition, please notice that, when referring to the plots, the results shown in the histograms differ from those in the episodic evaluation since they are obtained from 50 episodes generated with different seeds. In order to provide results which are meaningfully comparable, the choice of seeds for each kind of plot is the same across the different agents and architectures.

1 Unmitigated epidemics

Question 1.a) Study the behaviour of the model when epidemics are unmitigated

In the initial part of the project, it was necessary to become familiar with the available environment, interpreting and studying its behaviour in the case where the epidemic spread was not mitigated by any countermeasure. This allowed us to have a baseline case against which to compare throughout the project, assessing the effectiveness of the actions implemented to face the epidemic spreading.

In this regard, we present below the requested plots, followed by some general considerations:

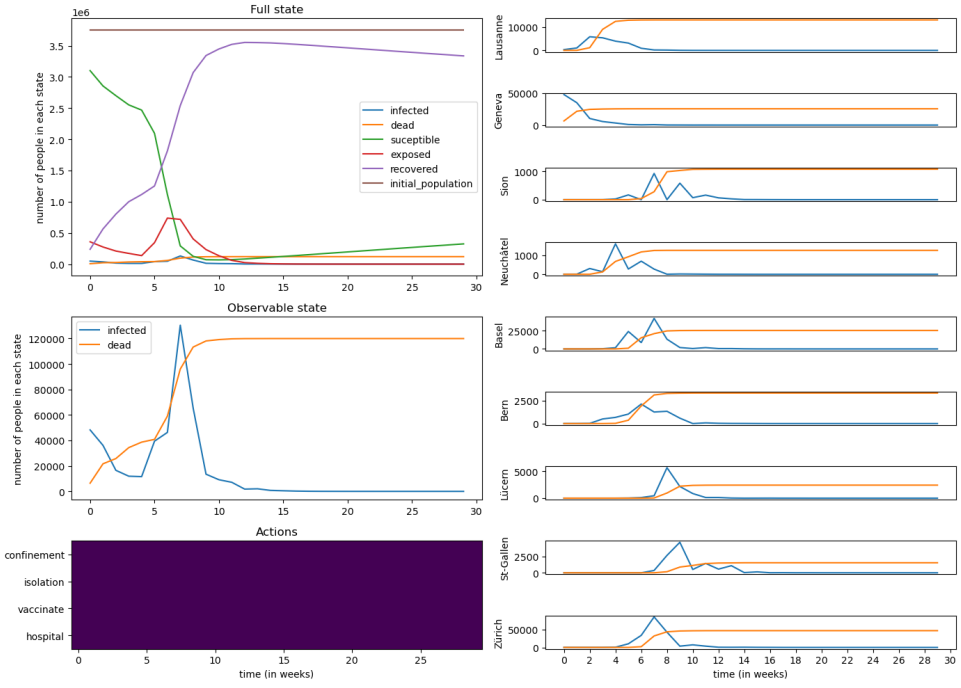

Figure 1: Unmitigated Scenario

The plots clearly show the physiological trend of the unmitigated epidemic:

- The susceptible population decreases dramatically as the number of recovered individuals increases, which is perfectly in line with our expectations.
- The increase in recovered individuals is an immediate consequence of the increase in exposed individuals.
- Similarly, the trend of the exposed population follows that of the infected population, although in considerably larger quantities.
- Regarding the specific trend of infected and deceased individuals, the peak of infections coincides temporally with the sudden increase in deaths.

2 Professor Russo's Policy

Question 2.a) Implement Pr. Russo's Policy

Below, we plot the trends of infected, deceased, susceptible, exposed, recovered individuals, and the initial population. Additionally, since this is the scenario where confinement is possible (unlike the unmitigated case), we will also plot the actual weeks of confinement within the 30 weeks considered.

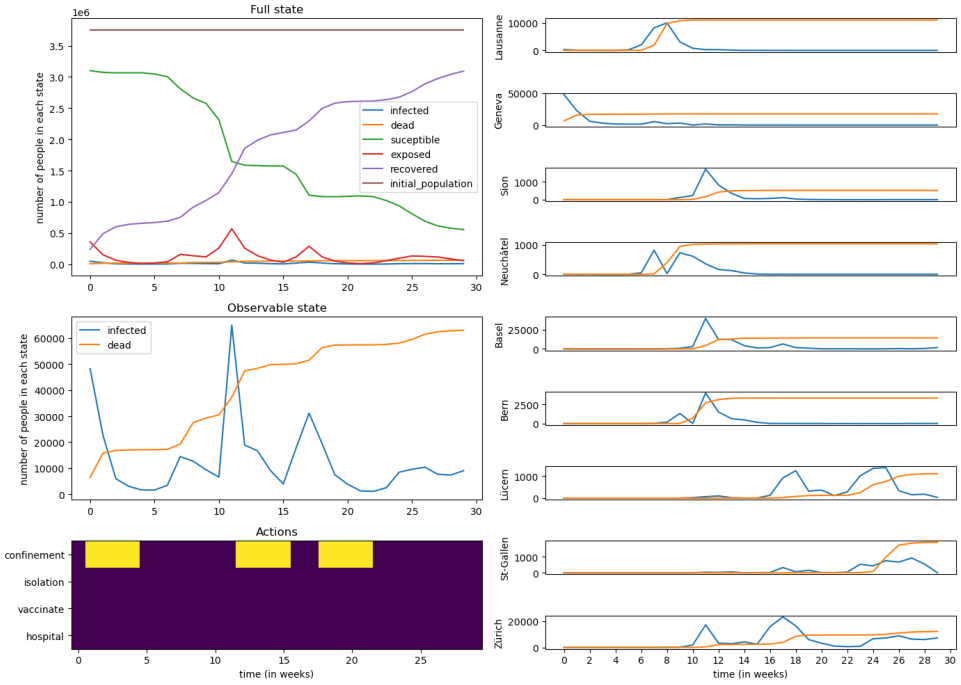

Figure 2: Episode generated by Pr. Russo's Policy

In light of these results and the unmitigated scenario, here are some observations:

- Firstly, it can be observed that the number of deaths and infected individuals has considerably decreased, indicating that the confinement measures have the desired effect.
- Additionally, it is interesting to note the "step trend" displayed by the recovered, susceptible, and deceased individuals. Plateaus are observed during the weeks of confinement, indicating that the situation stabilizes when this action is taken.

Question 2.b) Evaluate Pr. Russo's Policy

To evaluate Professor Russo's policy, we run 50 episodes and plot histograms (Figure 3) related to the total number of days of confinement, cumulative reward, and the total number of deaths.

Regarding the number of deaths, the average is around 60,000 units. The average cumulative reward is deeply negative (around -70), and the total number of days of confinement is around 98 (namely, more or less 14 weeks). For the exact numerical results, see Table 5.

3 A Deep Q-learning approach

In this phase of the project, we aim to improve Professor Russo's policy using a Deep Q-Learning approach. Regarding the network structure and the hyperparameters used, we strictly followed the guidelines provided in the project text.

Question 3.a) Implementing Deep Q-Learning

Results are presented in Figure 4. In the upper part of the plot, the training trace and evaluation trace of the three training processes are displayed separately, while the larger figure represents the average across the three training processes for both traces. It can then be inferred that the agent learns a meaningful policy as

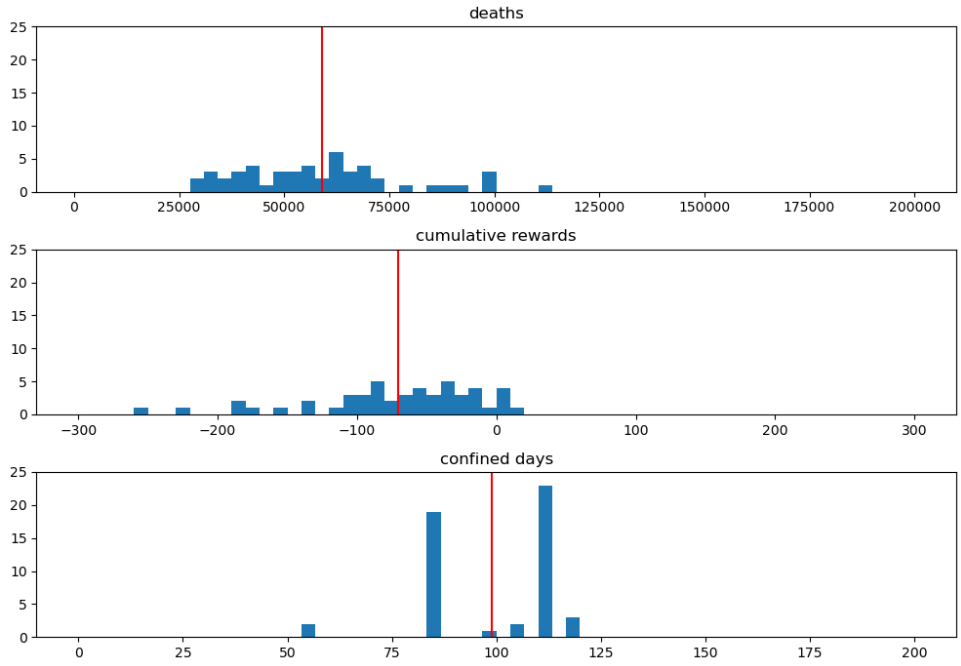

Figure 3: Professor Russo's policy - Histograms

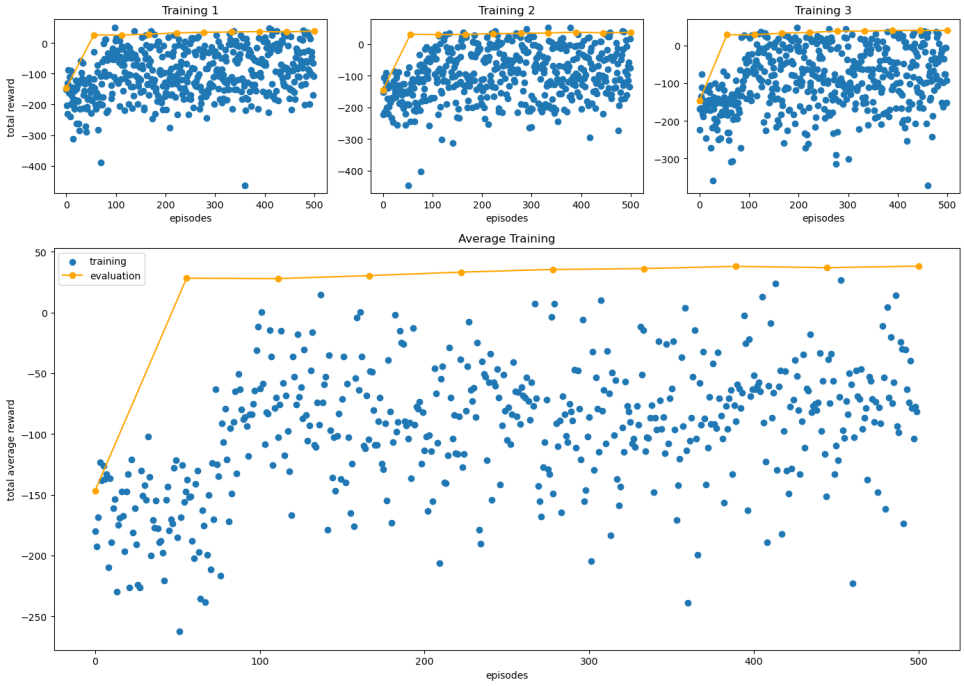

Figure 4: Training procedure for the binary action space

the evaluation trace increases with the number of training episodes. Specifically, it is observed that only 100 episodes are needed to achieve results that then remain relatively constant.

Once the model is trained, we can use it to simulate an episode of 30 weeks and plot the relevant quantities of interest (Figure 5) to compare the efficiency of the Deep Q-Network with respect to Russo's policy.

Again, the only action that can be taken is confinement. However, now the model is not subject to the constraint of confining the population in blocks of 4 weeks, making the DQN policy more flexible than Russo's. The presence of fewer constraints and a training model like a Deep Q-Network dramatically improves performances. Specifically, the number of deaths and infected individuals decreases significantly in all cities. Furthermore, the "step trend" mentioned earlier is still present but less pronounced, as the magnitudes are much lower. In practice, it appears that the ability to confine for multiple weeks from the beginning, even when the number of infected individuals is low, is beneficial, which is not possible in Russo's policy. Of course, the total number of confinement days is way larger than in Russo's policy, which is another reason why performances are better.

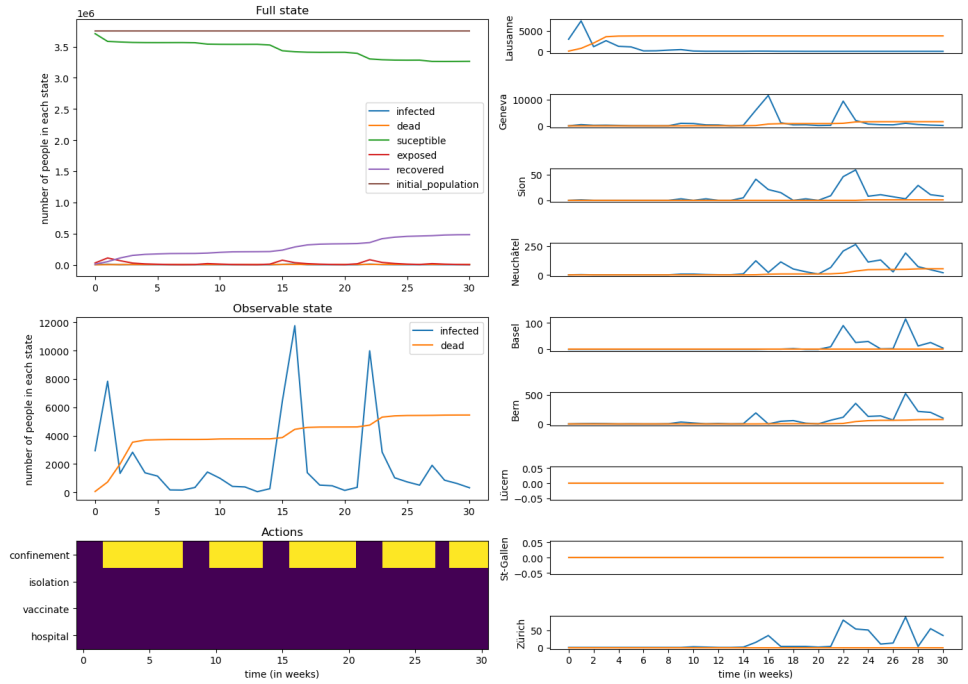

Figure 5: Episode generated by the best model for the DQN agent

Question 3.b) Decreasing exploration

Below there are the same plots for the case of a decreasing exploration coefficient ϵ. In Figure 6, it can be observed that the average total reward is slightly higher (although the difference is not excessively pronounced). More importantly, what stands out is how the learning trend of the training trace is much more stable. This is logical because as the training progresses, exploration of the environment is gradually reduced, favouring exploitation instead. In Figure 7 we observe that the weeks of confinement are exploited in longer blocks.

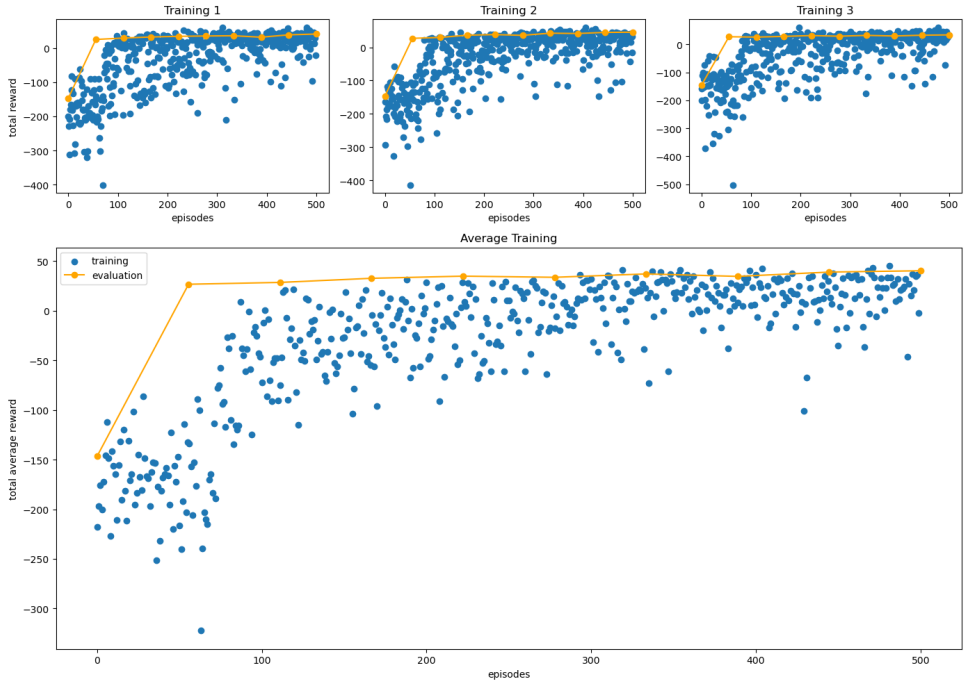

Figure 6: Training procedure for the binary action space and decreasing exploration coefficient

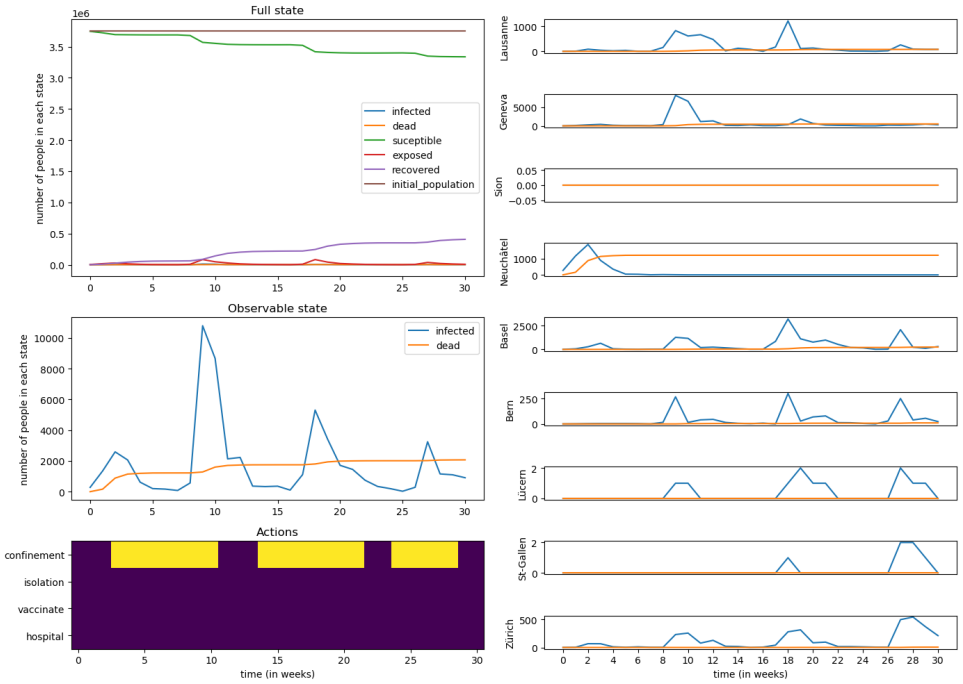

Figure 7: Episode generated by best model when using DQN agent with decreasing exploration

Question 3.c) Evaluate the best performing policy against Pr. Russo's policy

Finally, to determine if the use of Reinforcement Learning has indeed improved the performance in managing the epidemic, it is necessary to plot the histograms seen in the case of Russo's policy for the DQN policy as well (we consider the decreasing exploration agent, since it is the best agent between the two discussed in section 3.a and 3.b).

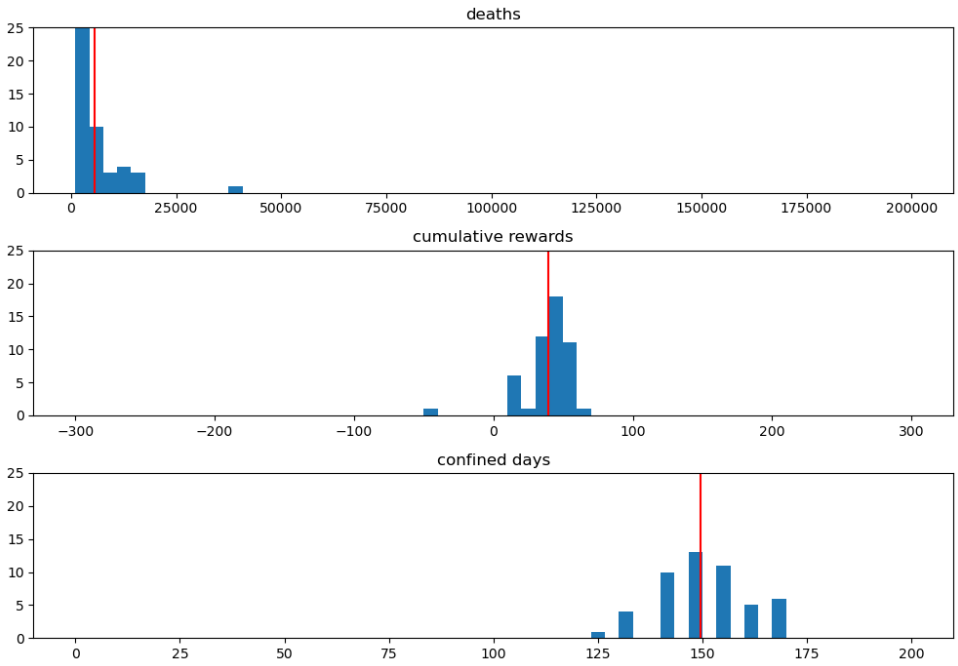

Figure 8: DQN Agent with decreasing exploration - Histograms

In Figure 8, it is evident that the average total number of deaths has significantly decreased compared to Russo's policy. Accordingly, the average cumulative reward has significantly increased, stabilizing at around 40. Finally, consistently with the increased effectiveness of the countermeasures, the average number of days of confinement has increased.

4 Dealing with a more complex action Space

In this section, we decide to work with a more complex action space, in order to progressively give the network more freedom when choosing the best action to take in different situations (states). In order to train the following architectures, we use the hyperparameters given in the project description.

4.1 Toggle-action-space multi-action agent

Question 4.1.a) Action space design

If we use the proposed toggle-action-space, we limit the choice of actions that we have in every state. Indeed, the output layer of the architecture will have 5 entries rather than 16 entries, which is the total amount of combinations we can obtain by taking or not 4 choices independently. This choice leads to a smaller number of parameters in the network, therefore making the training procedure faster.

Question 4.1.b) Toggle-action-space multi-action policy training

We now visualize the results obtained when training our agent under these conditions and assumptions in Figure 9. We can see that the evaluation trace is far below zero: the model seems to perform worse than the simpler

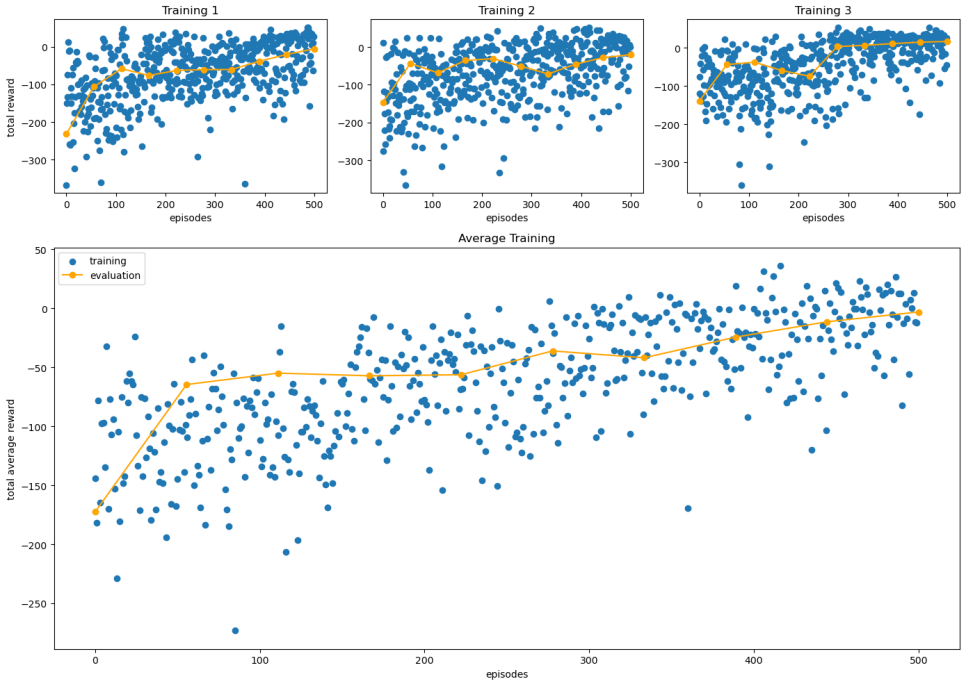

Figure 9: Training procedure for the toggle action space with decreasing eploration rate

architecture and approach employed in the previous section. We can say the agent is not learning effectively, which might be due to the fact that the learning rate (10^{-5}) and the number of epochs are not tuned and large enough to converge to a good set of parameters.

Moreover, if we analyze an episode in which the best model interacts with the environment (Figure 10), we can clearly see that the low death rate is entirely due to the strong confinement policy and the usage of hospital beds. Differently from Professor Russo's and previous policy, we now do not have a constraint on the actions and on the maximum number of weeks of confinement: we therefore incur in huge costs since confinement days are frequent and additional hospital beds are often exploited.

Question 4.1.c) Toggle-action-space multi-action policy evaluation

If we now compare the newly obtained policy to the binary action policy evaluated in task 3.c (Figure 11), we immediately observe that the death distribution has heavier tails, which suggests a worse policy used to deal with the epidemics. Moreover, while confinement days are still a key aspect of the adopted policy, the reward frequently assumes negative values, thus reflecting the negative behaviour observed during the training in Figure 9. Therefore, despite the additional freedom given to the agent, the performance is worse than before.

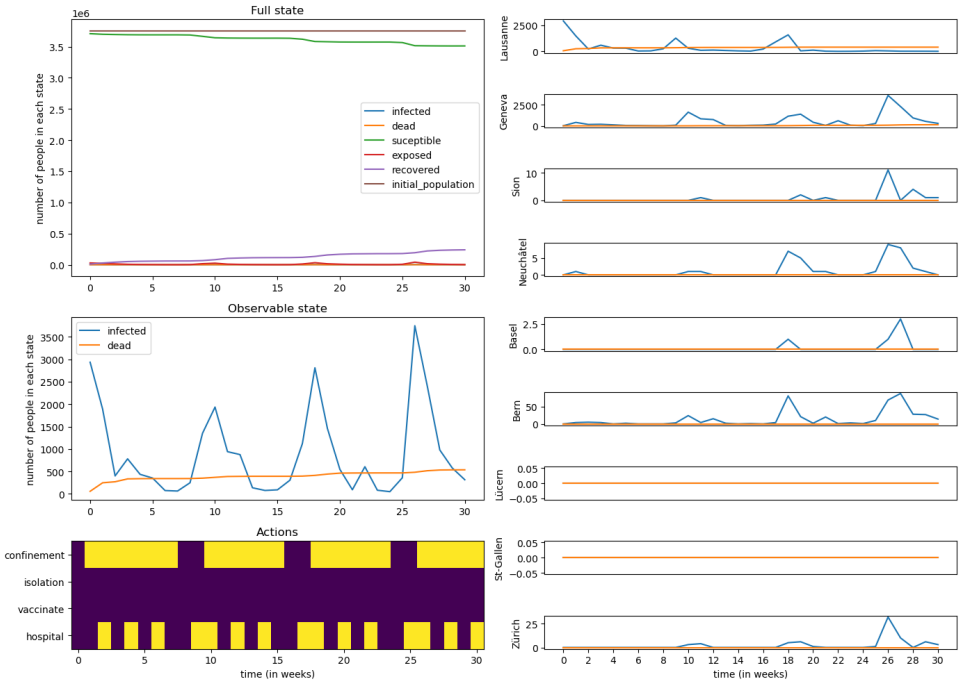

Figure 10: Episode generated by best model when using toggle-action-space

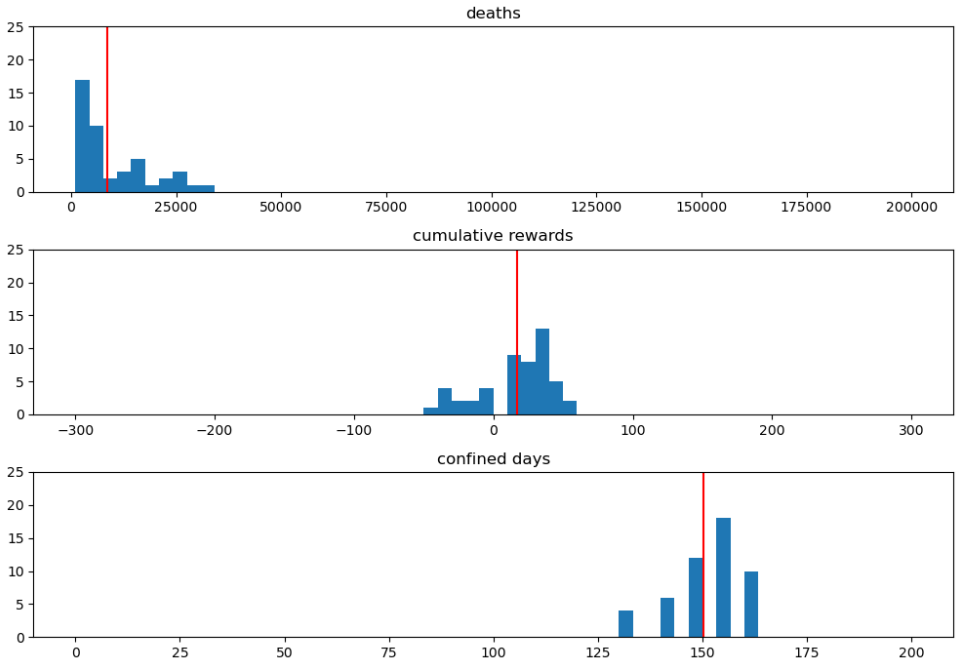

Figure 11: Toggle-Action space Agent - Histograms

Question 4.1.d) Question about toggled-action-space policy, what assumption does it make?

The use of such a technique on the action space implies that the agent cannot take an arbitrary action in the current state. As a matter of fact, the choice must depend on the action state at the end of the previous week. In general, if we consider each active action as a bit set to 1 (0 otherwise), we can say that the Hamming distance between two consecutive actions must be at most 1. Such a method cannot be used for action spaces in which taking two actions at the same time (e.g. Confinement and Hospital beds) is not equivalent to toggling them separately in two different moments. Moreover, the action space has to be discrete.

4.2 Factorized Q-values, multi-action agent

In this section, we decide to try an alternative approach, allowing the architecture to take any possible combination in our set of actions. This is achieved by using the factorized Q values approach.

Question 4.2.a) multi-action factorized Q-values policy training

From the result in Figure 12 we observe that the "factorized agent" performs better than the "toggle agent". This might be due to larger freedom when implementing actions. Nonetheless, the performance is still not optimal (the reward is slightly larger than 0 in the evaluation trace): a possible reason can be the not sufficiently large number of training epochs since the parameters of the new model are larger than those of the previous one. Moreover, if we look at the episode obtained deploying this agent (Figure 13), we immediately notice the high number of deaths and the large amount of time in which confinement is preferred over other choices: the policy seems to be unrealistic since such a rigid confinement procedure would be difficult to apply in real-life scenarios.

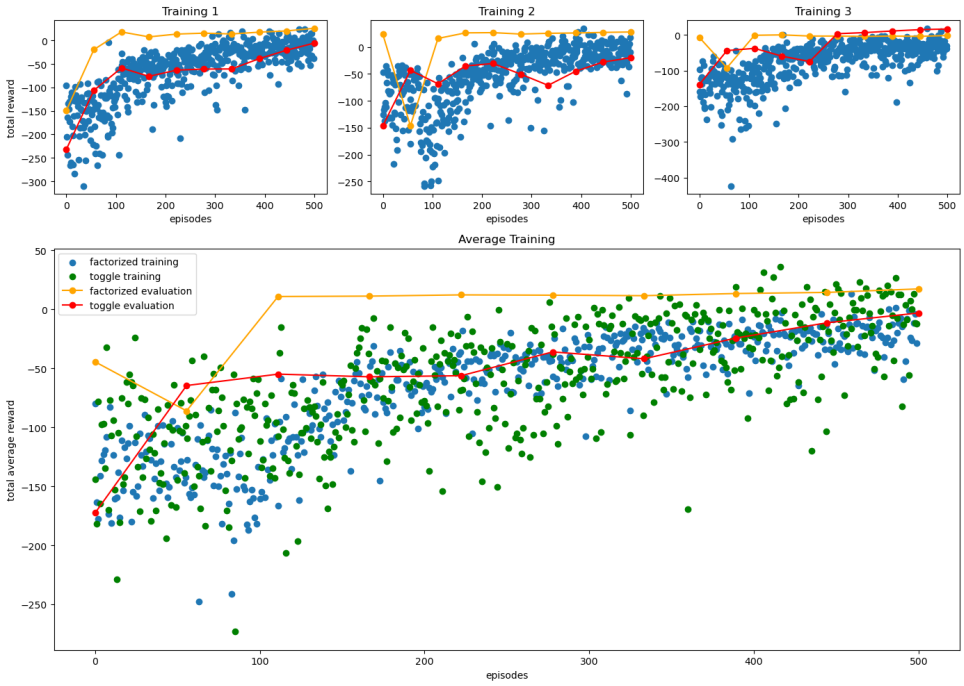

Figure 12: Training and evaluation trace comparison between factorized agent and toggle agent

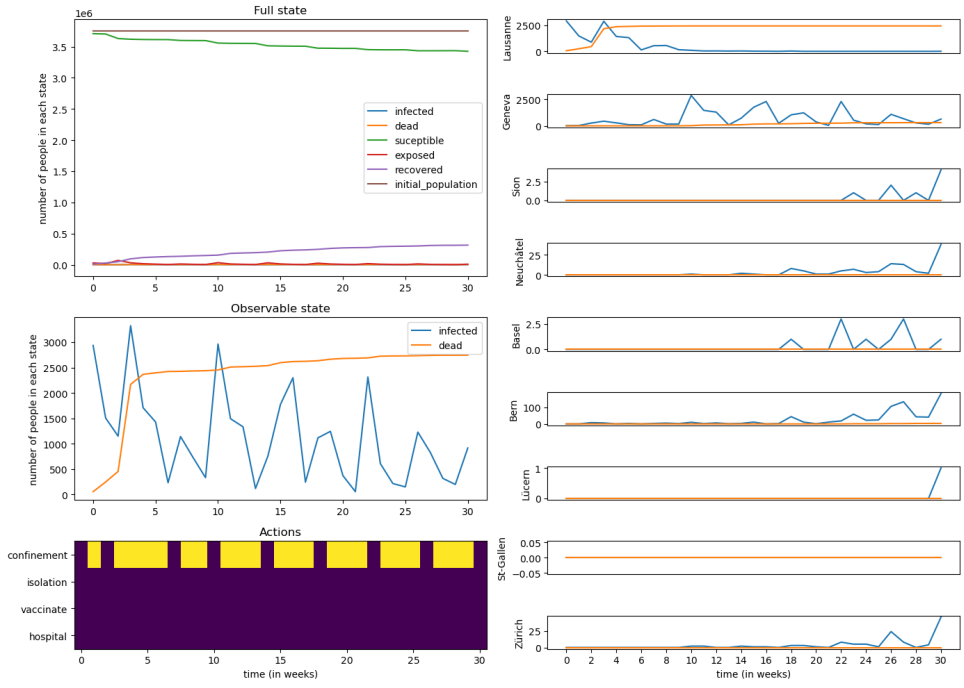

Figure 13: Episode generated by best model when using factorized Q values

Question 4.2.b) multi-action factorized Q-values policy evaluation

To better compare the "factorized agent" and the "toggle agent", we now look at the histograms obtained at the end of the 50 episodes evaluation procedure in Figure 14. As suggested by the training and evaluation traces, the number of deaths is lower than in the toggle-action space case (not observed in Figure 13 because of an unfortunate choice of seed). Moreover, the reward tends to be higher, while in the toggle case negative values were very frequent probably because of a larger number of dead people. For what regards confinement days, the mean in the factorized Q values case is significantly higher and the tails are heavy towards very high values.

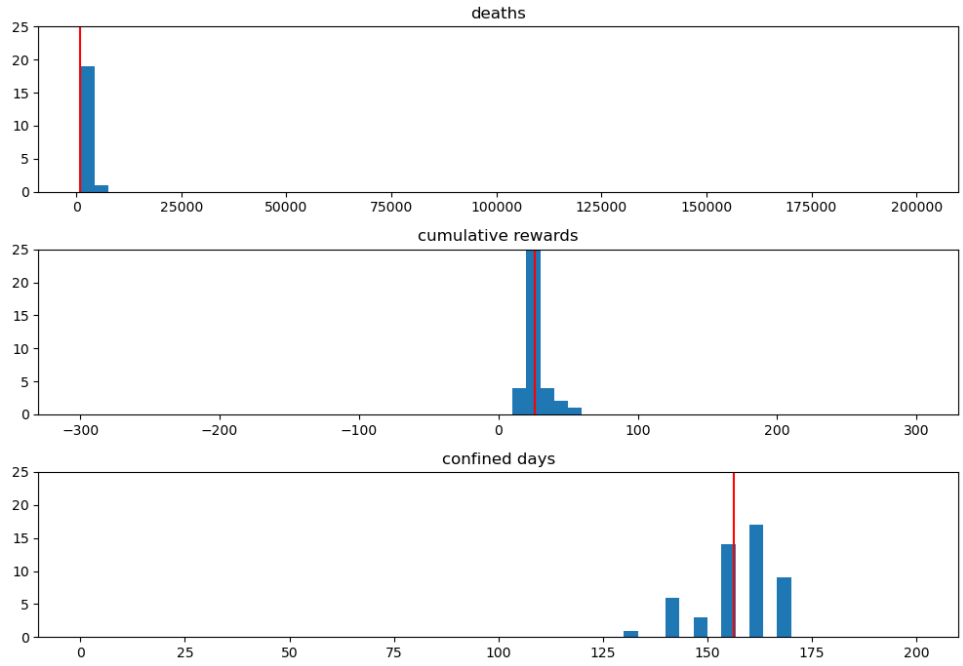

Figure 14: Factorized Q values - Histograms

Question 4.2.c) Factorized-Q-values, what assumption does it make?

This technique assumes that the actions can be taken independently, i.e. the choice of a specific action does not put any constraint on any other choice. This is the reason why, in order to choose the best combination to take, we can consider the maximum over each row of the 4×2 tensor given as output by the agent neural network. This approach cannot be used when such an assumption does not hold. A simple example could be the one proposed in the previous section since toggling an action impedes taking another one.

5 Wrapping Up

Question 5.a) (Result analysis) Comparing the training behaviors

We begin by noting that the Russo Agent does not require a training procedure, as it does not contain any network within the agent class.

Regarding the single-action DQN Agent, we observe that its performance in terms of training and evaluation traces is the best, reaching rewards ranging from 30 to 40. Although the Toggle Agent and Factorized Agent offer greater decision-making freedom for the agent, significantly worse results are observed. We suppose that these results are due to suboptimal tuning of hyperparameters and an insufficiently long training procedure to adequately train the network.

Analyzing the behaviour of a single episode of Russo with the singleDQN, we observe that a less restrictive constraint on the consecutive number of weeks of confinement results in a lower number of deaths both at the national level and in individual cities.

Question 5.b) (Result analysis) Comparing policies

From the summary table of the results, several conclusions can be drawn:

- Firstly, it can be observed that the Toggle policy is not optimal according to any of the considered metrics.

- Regarding the total days of confinement, the best policy is the one proposed by Professor Russo. This result was expected since it is the only policy that explicitly expresses constraints on the weeks of confinement.
- No days of isolation or vaccinations are ever implemented, probably because the costs are too high compared to the actual short-term impact on deaths.
- In terms of the addition of hospital beds, the preferable policy is the Factorized one, as it combines optimally a reduced number of additional hospitalizations with confinement days.
- The Factorized policy also yields the best results in terms of the total number of deaths. This result highlights the effectiveness of a winning combination of confinement days and hospital beds.
- Finally, the cumulative reward is the highest for the SingleDQN policy. It also indicates that the reward is not necessarily proportional to the number of actions implemented.

	Russo	DQN	Toggle	Factorized
Total Confined days	**98.98**	149.52	150.36	156.38
Total Isolation days	0	0	0	0
Total Vaccination days	0	0	0	0
Total Additional Hospital beds	0	0	91.7	**5.88**
Total Deaths	59053.52	5535.46	8548.4	**986.66**
Cumulative Reward	-70.62	**39.59**	16.91	26.04

Question 5.c) (Interpretability) Q-values

We notice that in the heatmap referring to DQN approach, (Figure 15 left), 'Do Confinement' is the action which assume the highest values (lighter colours), thus being chosen more frequently. This agrees with the results shown in task 5.b and Figure 7. Moreover, it is interesting to notice that, whenever 'Not Confinement' is chosen, its colour gets darker soon, as suggesting that without confinement the situation becomes worse.

For what regards the factorized Q values approach, it is important to notice that the single entries of the heatmap do not constitute the Q values. As a matter of fact, since an action is defined as the combination of 4 choices (confinement, isolation, hospital beds and vaccination) set either to True or False, the sum of 4 entries has to be considered as the way of computing the Q value for a specific action (more precisely, for each choice we can only take one value referring to the Do or Not operation). Even though the heatmap is not too interpretable to understand the Q values, it is still very informative to understand the behaviour imposed by the policy. As we can observe in Figure 15, "Do Confinement" tends to have higher values than "Not Confinement", thus being usually preferred. On the other hand, the "Do Not" operation is more common for the other choices. This behaviour is well illustrated in Figure 13, where the only chosen action is "Confinement".

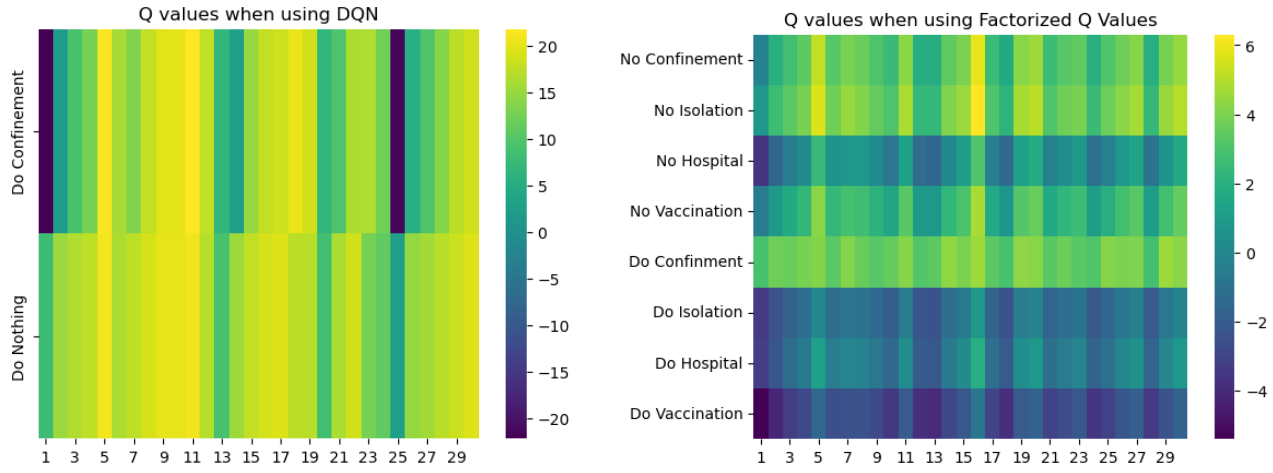

Figure 15: Heatmaps: Q values corresponding to the possible actions across all 30 weeks

Question 5.d) Is cumulative reward an increasing function of the number of actions?

From the results obtained throughout the whole project, we can state that the cumulative reward is not an increasing function of the number of actions. This is due to the fact that cumulative reward takes into account also the costs of countermeasures employed to face the epidemic. An increasing number of actions would for sure result in a decreasing number of deaths. However, it cannot be taken for granted that a lower number of total deaths would imply a higher cumulative reward, as it is about economic costs as well.