Deep Q-Learning for Epidemic Mitigation

Lucas Reymond[a], Pranav Krishna[a]

[a]*EPFL, Chemin du Barrage, Lausanne, Lausanne, 1015, Switzerland*

Keywords: DQN, Reinforcement Learning, Artificial Neural Networks, Deep Learning



Question 1

We see the effect of an unmitigated pandemic on Figure 1. For each city, we see that the number of deaths lags the number of infected people by roughly 1-2 weeks and that the pandemic hits various cities at different times, confounding the lead-lag relationship in the aggregate.

The cities directly connected to Zurich are impacted first, while the cities which are not, are impacted after two weeks. This is due to the structure of the graph, wherein the pandemic has to pass through Bern or Neuchâtel before it affects Lausanne and Geneva. This is also shown in the graph for the number of susceptible and exposed people, which also appear to have change-points when the pandemic starts to hit the French speaking cities.

In the temporal domain, we see that there are a lot of deaths(around 120,000) in the first 10 weeks of the pandemic, and almost no new deaths after. Similarly, we see that after 10 weeks, 3.5 million people recovered from the disease and have immunity. This may be enough to stop the spreading of the pandemic. Furthermore, the number of infected people is also very high in the first 10 weeks but does not rise above 10,000 after that period.

Lastly, we remark that as a proportion of the population, very few people (80,000 max) are infected at any particular instant of time, whereas during the entire pandemic, we see that the number of recovered people (almost the whole population) rises to be of the same order as the total initial population.

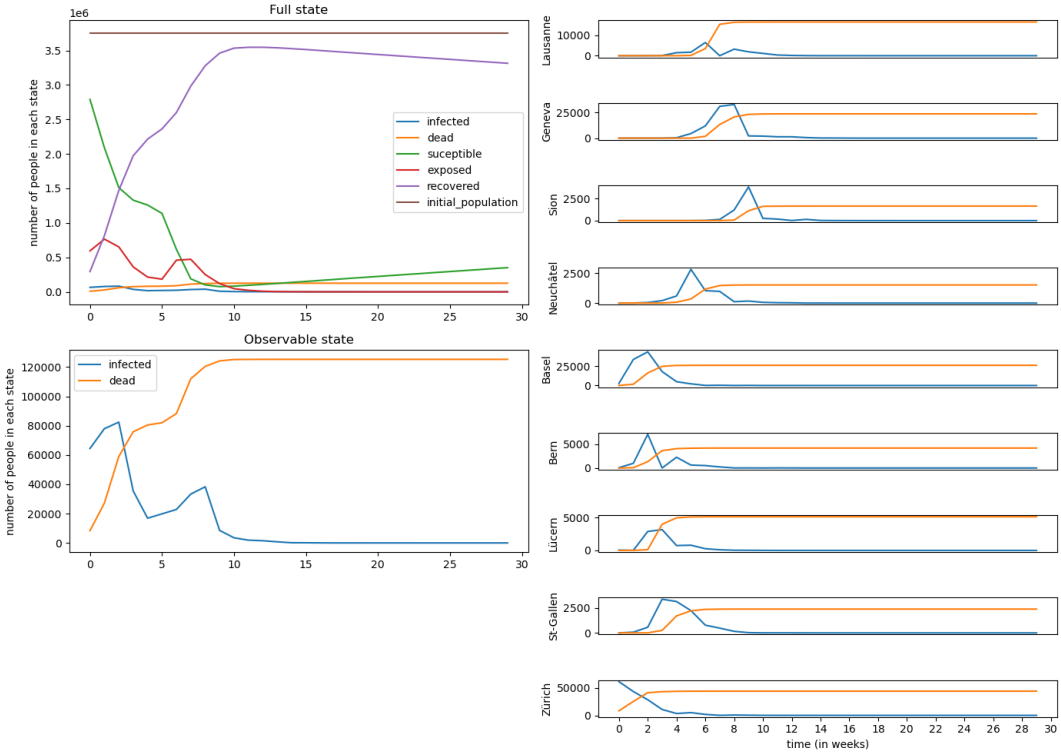


Figure 1: Plots for Question 1.a

Question 2

Question 2.a

In this section, we study the effect of Prof. Russo's policy. Clearly, we have lengthened the duration of the pandemic, which had nearly run its course in 10 weeks in the unmitigated case.

Most importantly, we see on Figure 2 that this policy halves the number of deaths to around 60,000 when compared with the unmitigated pandemic. In the graphs for deaths and infected people, we see clearly, the periodicity of the increase and the corresponding decrease before and after the confinements. We remark that in this policy, we cannot take any action in the first week, which is shown by the relatively high number of infected people in the initial week of the pandemic, which are mostly in Zurich. Once we confine, we see that the

pandemic dynamics follow a roughly periodic structure, with a sharp increase when the confinement is removed and a slower decline in the intervening four weeks before the next spike. This is also visible in the susceptible, exposed and recovered parts of the population.

The order in which the pandemic hits the cities is broadly the same, with the pandemic starting in Zurich, hitting the German speaking part, then the French speaking part and lastly, Sion.

We also remark that while the lead-lag relationship in the infected and dead number of people still holds for each city, it is also visible in the cumulative number of dead and infected people, with the largest number of deaths usually in the week just after the one when the confinement is implemented. Lastly, we see that after one more cycle of Prof. Russo's policy, we expect the number of susceptible and recovered people to become roughly the same as the unmitigated pandemic with roughly half the number of deaths.

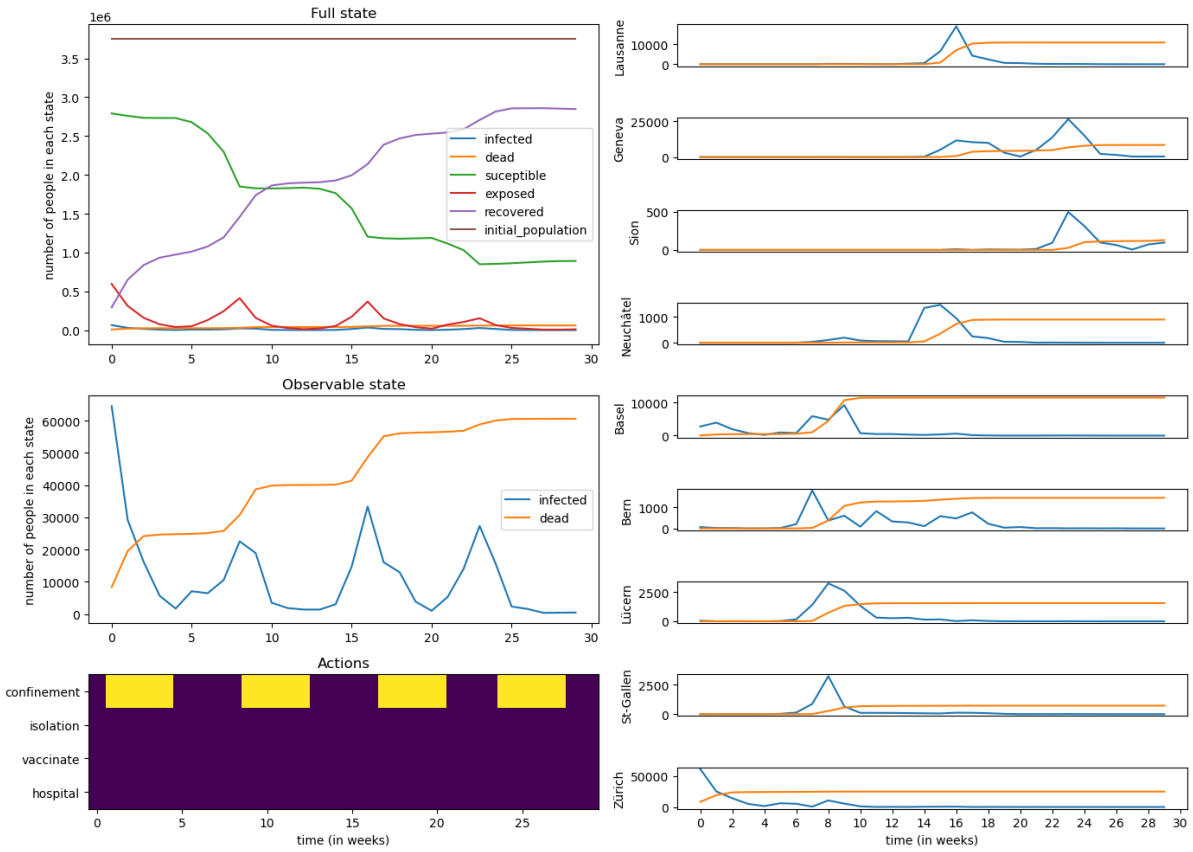


Figure 2: Plots for Question 2.a

Question 2.b

Figure 3 required histogram as mentioned in the problem statement. We will compare it with other policies in Question 5 part a.

Question 3

In this section, we will study the effect of a policy trained using Deep Q-Learning, for a binary action space, i.e. choosing to confine or not to confine. We use the approach outlined in (1).

Question 3.a

We see in Figure 4, a consistent pattern in the training and evaluation traces for all the three training instances. They are at around -120 in the beginning and rise above zero in the third evaluation. This is because we require at least $\frac{2048}{30} \sim 68.2$ experiences for replay, and this is completed only after 69 episodes. The agent learn a meaningful policy.

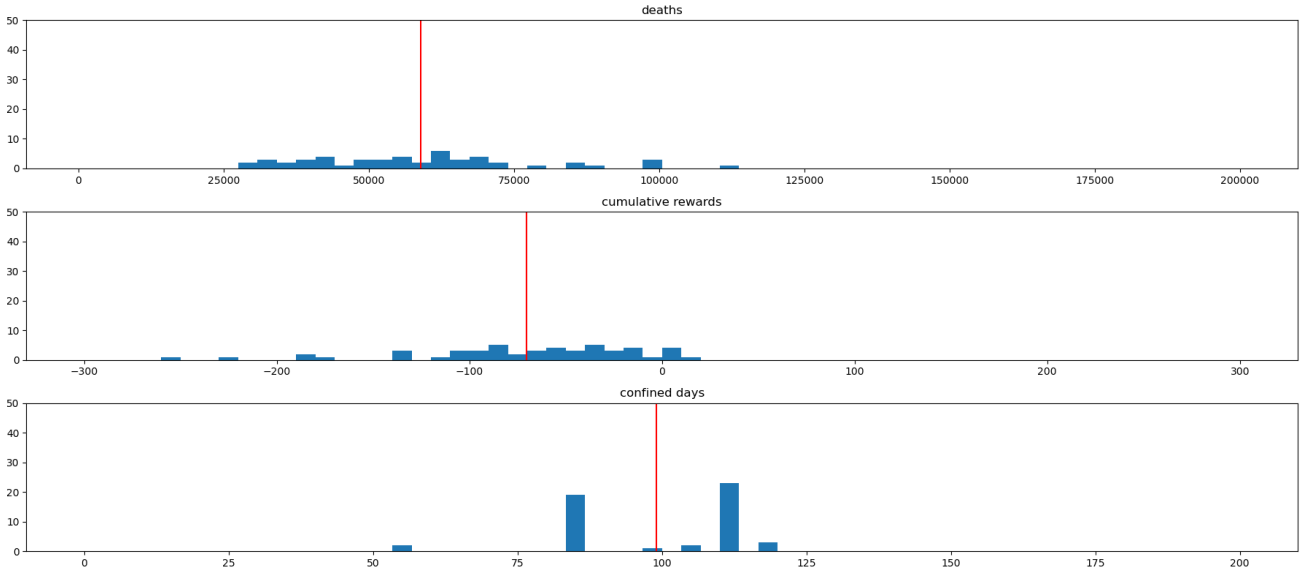


Figure 3: Histograms for Question 2.b

On Figure 5, we see that the pandemic is still not finished as there are still a lot of susceptible people and it has barely started in the French speaking part of Switzerland.

Clearly, as the pandemic has not run its course, we cannot directly compare the total number of dead or infected people with Russo policy. We try to interpret the policy of the agent as follows. We see that while the policy looks similar to Russo policy in that we run a few weeks of confinement repeatedly, the agent is much more aggressive towards confining. This is especially visible in week 30 where the agent decides to confine even when there are less than 2500 infected people.

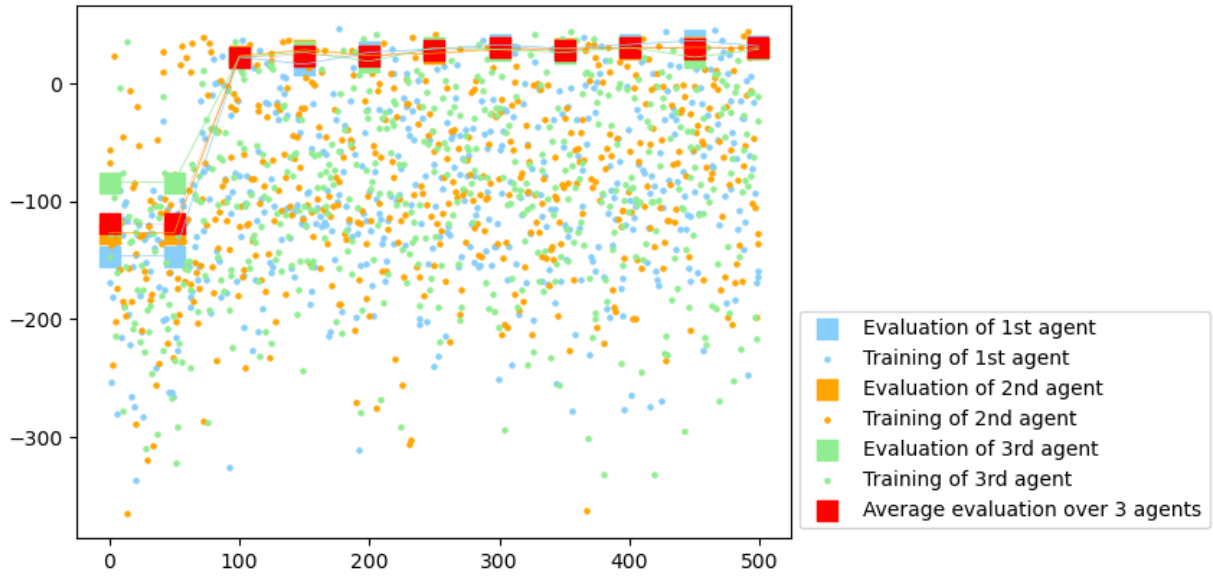


Figure 4: Training and Evaluation traces for Question 3.a

Question 3.b

We see clearly in Figure 6 that the effect of decreasing exploration is visible in the training trace. Compared to Question 3.a, we explore less in the later part of the training and thus, have lesser instances of a negative reward, which is better. Besides that, we see that both policies have explored the loss landscape adequately and have a comparable evaluation trace. To determine which one is the best we computed the average cumulative

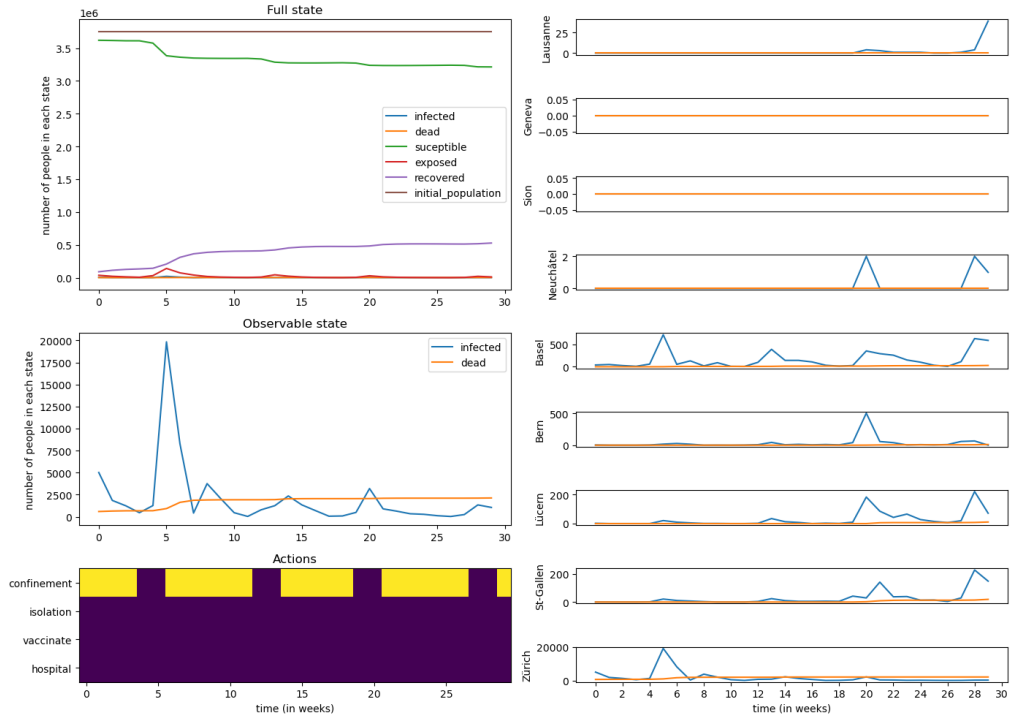


Figure 5: Plots for Question 3.a

rewards and obtained that the one for 3.a is 27.5 while the one for 3.b is 34.7. So the policy with decreasing exploration rate gets better results in terms of average cumulative reward.

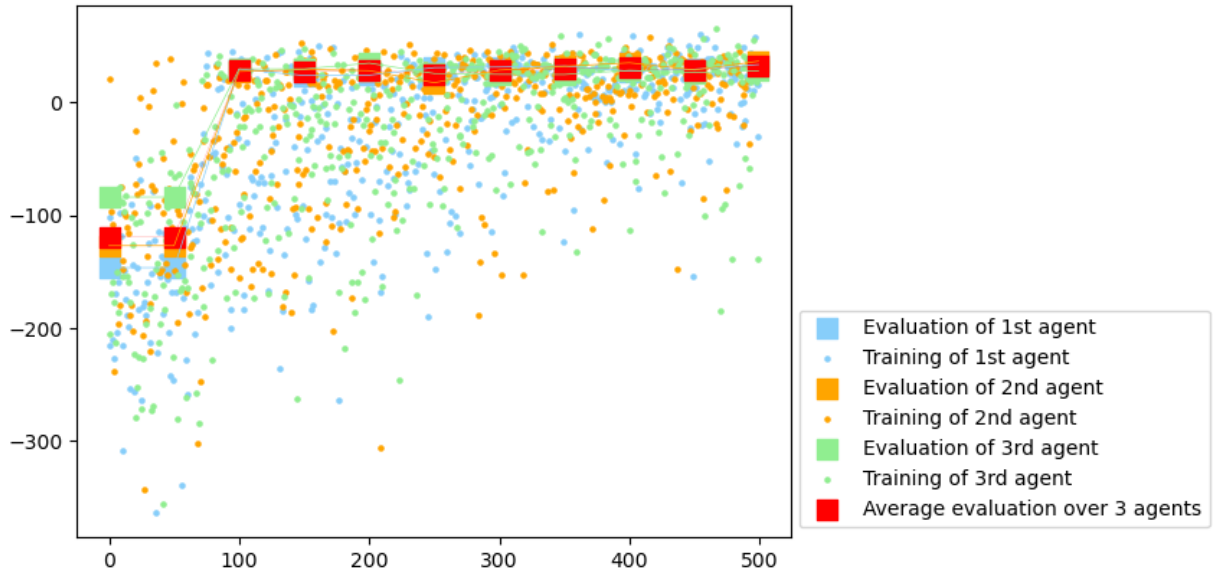


Figure 6: Training and Evaluation traces for Question 3.b

Question 3.c

We see that the DQN policies clearly outperform Russo-policy in term of the reward function. The average cumulative reward for Russo policy is -70 while the average cumulative reward for the DQN policy with decreasing exploration rate is 34. A similar comparison holds in terms of average number of deaths as the Russo policy has a higher average number of deaths (58923) compared to the DQN based policy with decreasing exploration rate (2856). However, we remark that the DQN policy has a larger average number of days in confinement than Russo policy (176 and 99 respectively), which means that the Russo policy is preferable in this regard.

Question 4

In this section, we will try to augment the action space to make a more realistic response to pandemics by allowing for isolation, vaccination and an increase in hospital beds as possible actions.

Question 4.1.a

In Q-learning, we need estimators for the function Q(s, a) for each possible state-action pair. However, when a toggle-action-space is not used, the number of actions per state is quite large. Without considering any specific constraints, we have 16 potential actions for each state, as each possible configuration for each of the four measures is required to define an action. This high number of actions leads to computational intensity, making it impractical to directly compute and store Q-values for all action-state pairs.

To address this, we use the toggle-action-space approach by augmenting the state-space with additional information. By including the previous actions taken as part of the state, we can take advantage of a clever perspective on the action space. We effectively reduce the number of actions (5 instead of 16) required to capture the complexity of the problem. This modification allows us to condense the action space, making it more manageable and computationally efficient.

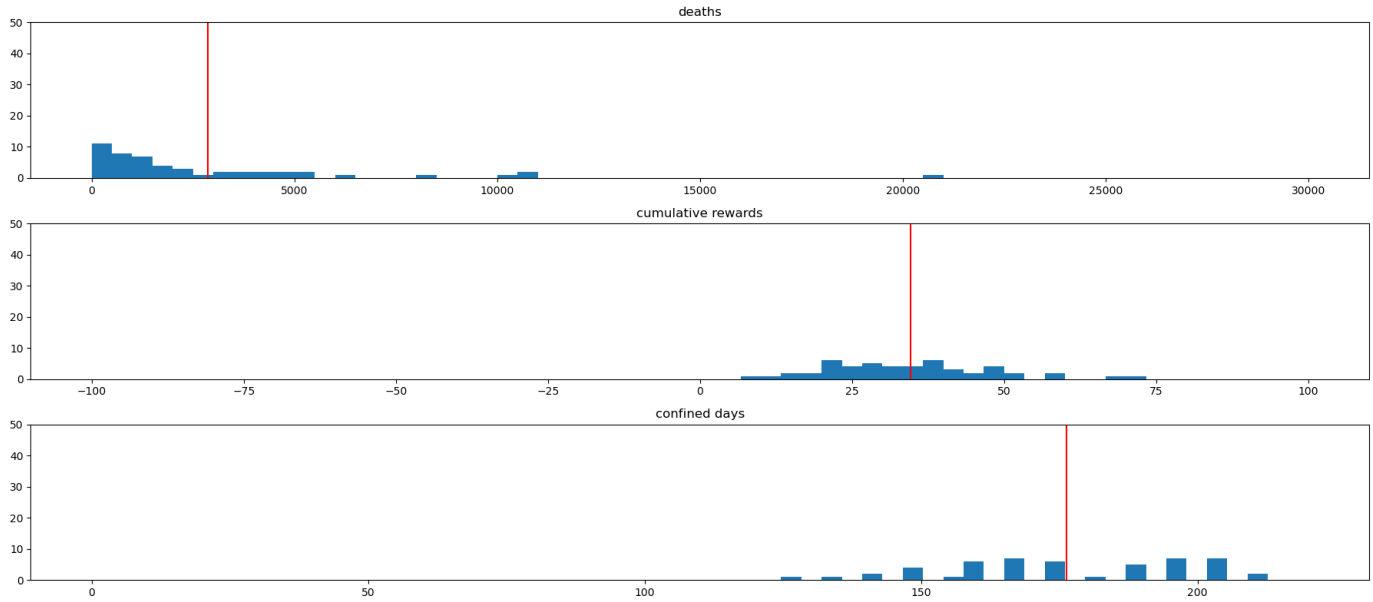


Figure 7: Plots for Question 3.c

Besides this, it is also possible that the shared representation of the state-and pandemic measure status can also lead to the ability of the network to be able to capture interactions between the measures. Recall that the reward does not contain any interaction terms, which makes it unlikely that this particular perspective is helpful.

Lastly, we observe that the toggle-action-space simplifies the network architecture as we only have five output neurons, while retaining the complexity of the possible combinations of actions and their interactions. On a related note, we also see that the toggle policy simplifies the training as we do not have as many weights which need to be optimised.

Question 4.1.b

We see in Figure 8, a consistent rise in the reward for all the three training processes. Due to a small prescribed learning rate (10^{-5}), we expect that more episodes of training may be beneficial. We see that the training is slightly unstable, as we see large variation in the evaluation traces. This is due to the fact that we have a larger number of weights in the neural network and all of them must be tuned suitably.

We see in Figure 10 that the agent does not vaccinate or isolate in the 30 weeks and tires to only toggle confinement and adding beds in hospitals. It is not possible to interpret the policy more than this because we see that the agent does not confine in Week 28 even though there are a large number of infected people.

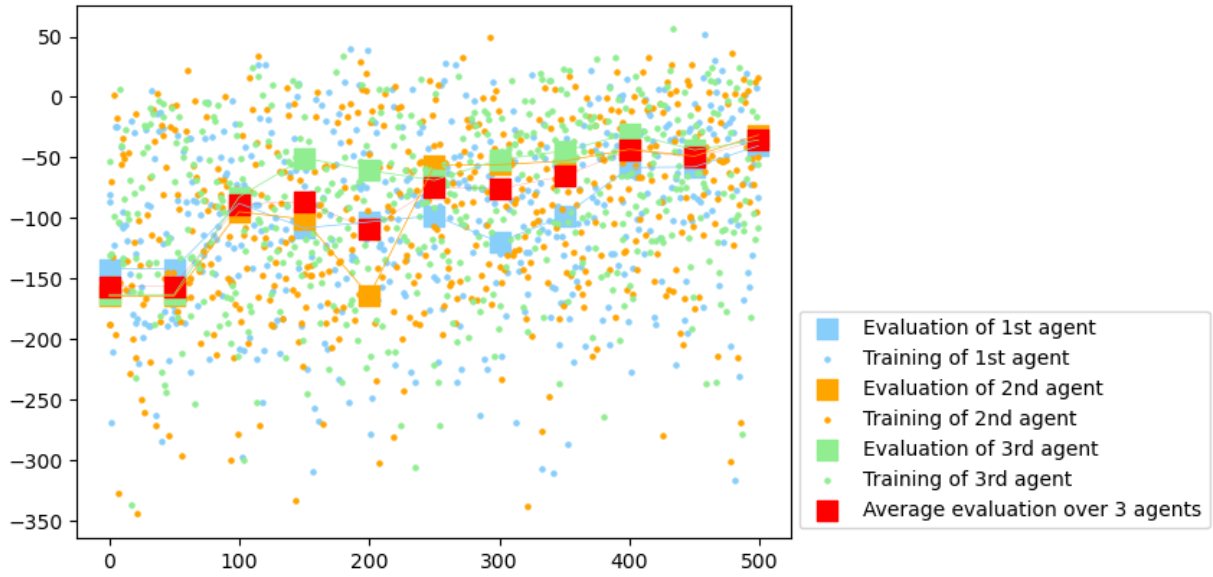


Figure 8: Training and Evaluation traces for Question 4.1.b

Question 4.1.c

We see in Figure 9 that the average cumulative reward for this policy is very bad (around -30) compared to the binary action policy(around 34). We also observe that the number of deaths is larger (around 30,000) than the binary action-pace policy, but there is slightly less average number of confinement days (around 150).

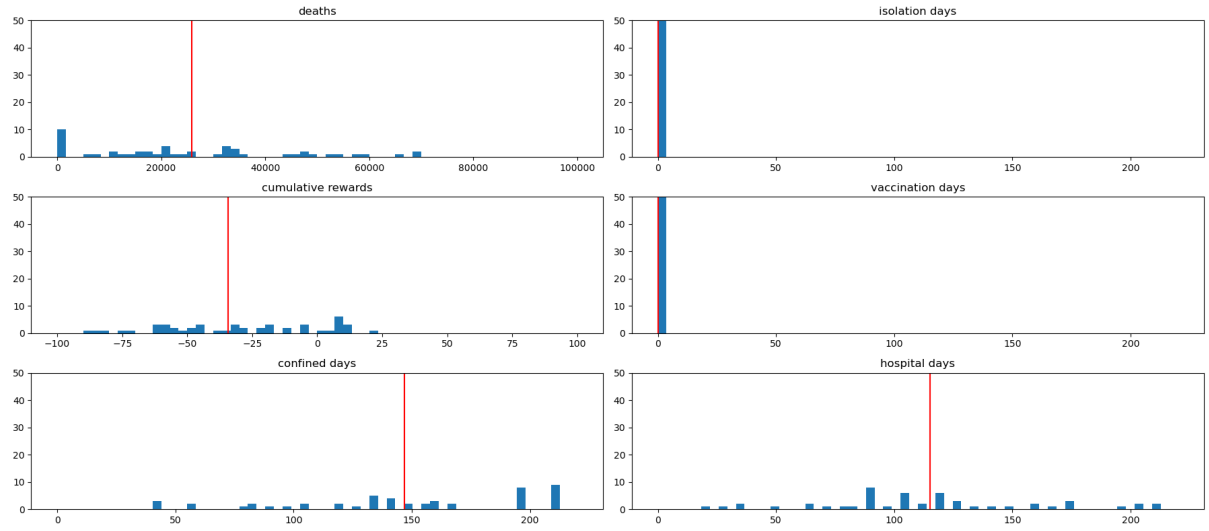


Figure 9: Plots for Question 4.1.c

Question 4.1.d

While we have simplified the architecture of the network, we must also point out that the toggle-action-space uses the fact that the actions are binary. In real life, it is possible to have actions which are not necessarily binary(or discrete) and may possibly be continuous. For example, another action which was possible would be the maximum number of people allowed to work or stay in one room. Such an action space would not be adequately modelled by a toggle-action-space.

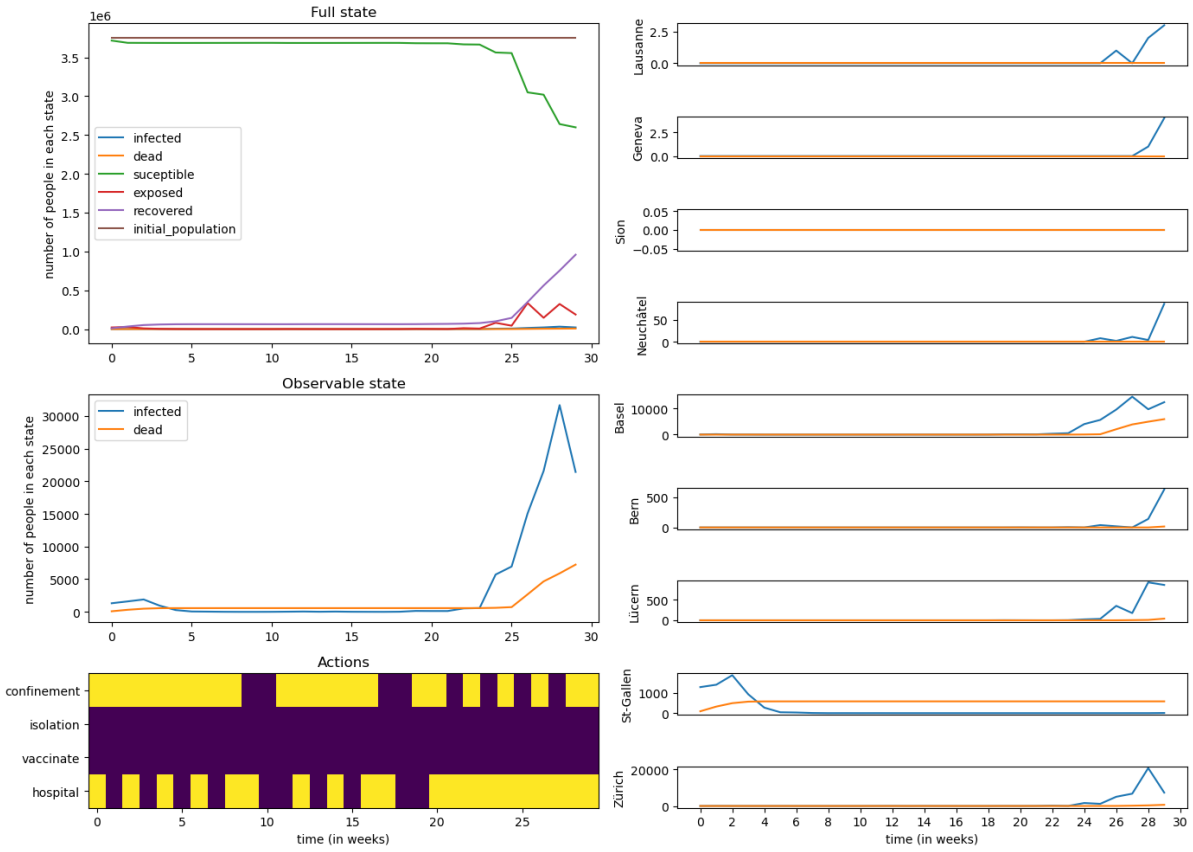


Figure 10: Plots for Question 4.1.b

Question 4.2.a

We see in Figure 11 that the training and evaluation traces are similar, but they do not have a consistent behaviour. There is a very large variation, especially in the evaluation trace for the third agent. We also observe that the training traces rise a little when the training starts, but then fall down after 300 episodes. This suggests a lack of convergence and we cannot say that the agent has learnt successfully. We think is is due to a large prescribed learning rate.

We see in Figure 12 that this policy always takes all possible actions all the time and thus, is not very efficient. We caution the reader about the previous plot. One must note that the bottom left plot shows that no action is ever taken, but in reality each of the 4 actions is taken for all 30 weeks. A display problem shows the opposite.

Such a policy is not realistic, at least in Switzerland.

Question 4.2.b

We see in Figure 13 that this policy is slightly better than the Toggle policy. The average cumulative reward is still negative but slightly higher (around -21), but there are fewer number of deaths (around 200). However, the number of days spent in confinement is higher (210 days, so the whole time).

Question 4.2.c

In this part, we operated under the main assumption that the four basic actions (confinement, isolation, vaccination, and adding beds in hospitals) are independent of each other. This means that each action can be implemented without affecting or being influenced by the others. An example of action space where we could not do that is if we introduce the action "close all the hospitals" because this new action would then not be independent of "add beds in hospitals". Another example is that there can be possible interactions between the actions, for example, implementing isolation and confinement at the same time might have interaction effects, which ends up improving the Q value for the action compared to just the sum of the factorised Q values associated to confinement and isolation.

Question 5

We discuss the results of our experiments in this section.

Question 5.a

The evaluation traces of Questions 3.a and 3.b (i.e. DQN policy) are fast to converge and almost reach their best policy after 100 episodes (soon after the buffer is full and we can optimise). The evaluation traces are very similar for both. The training traces however are slightly different: the ones associated with a decreasing exploration rate are less scattered.

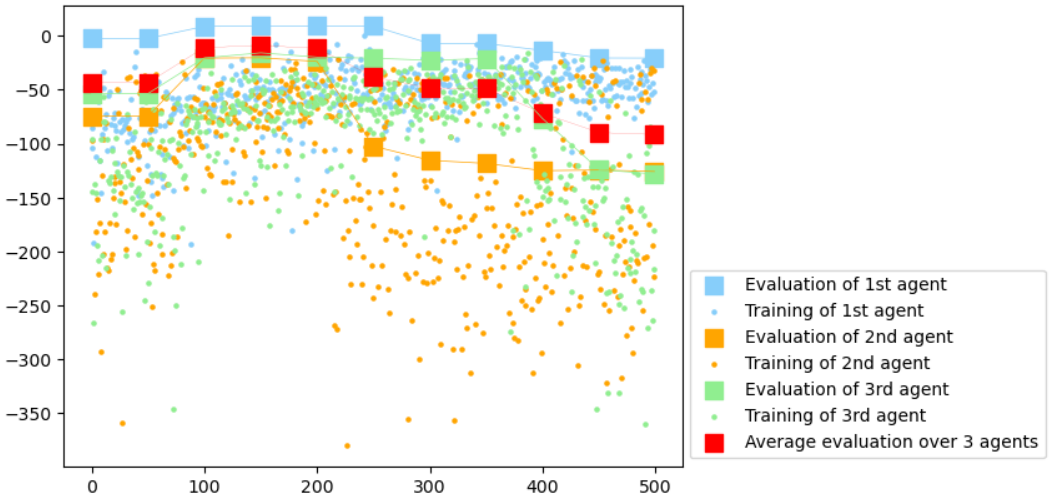


Figure 11: Training and Evaluation traces for Question 4.2.a

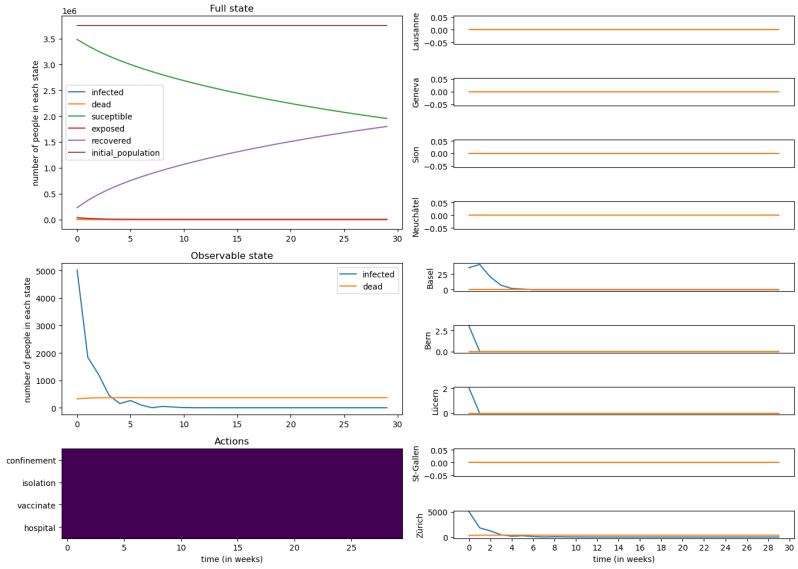


Figure 12: Plots for Question 4.2.a

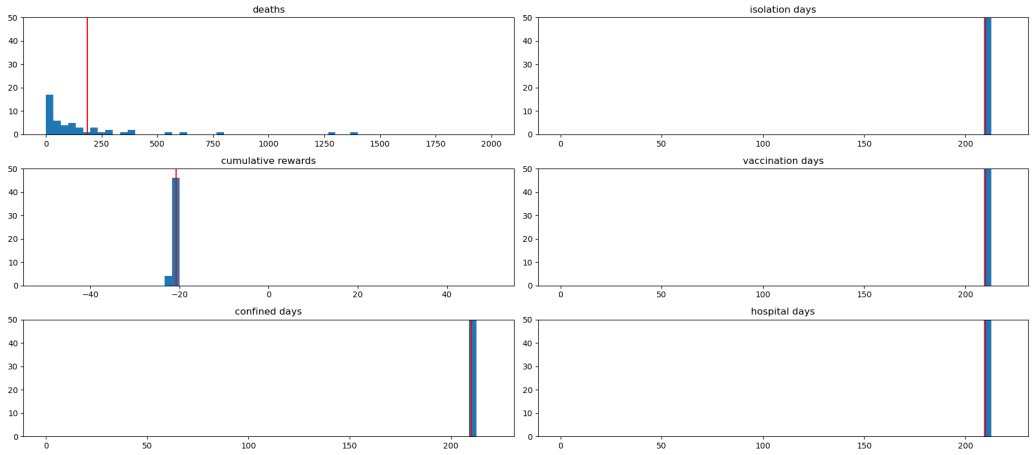


Figure 13: Plots for Question 4.2.b

The traces of the non-binary action spaces policies are not as good. The evaluation traces of the Toggle policy have a slow convergence and don't seem to have finish to converge because we don't see the reward stabilise. Figure **??** in the Appendix shows the traces with a learning rate equal to 10^{-3} which appear more stable. Similarly, the evaluation traces of the Factor policy starts to increase but then, decrease after 250 episodes. This suggests that the learning rate is too high. Figure 15 shows those traces with a smaller learning rate equal to 10^{-6} which appear to have converged.

Lastly, we remark that the Russo policy is sub-optimal because it waits that there are a lot of infections to start a confinement. The DQN policy tries to learn when it is useful to do a confinement. The Toggle policy uses only the actions confinement and adding hospital beds even though all actions are theoretically possible. The Factor policy applies all the possible actions all the time but then succeed the have a low number of deaths.

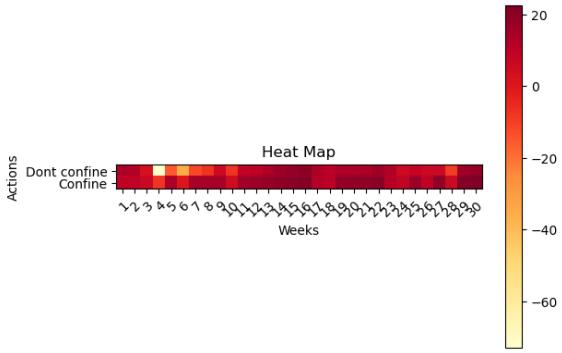
To determine what is the best policy, we should first determine what means the 'best'. If it is the fewer number of deaths then the Factor policy is the best. If it is the cumulative reward because if captures both the cost of taking an action and the cost of deaths, then the DQN policy is the best. In the next question we determine which policy is the best for the six possible metrics.

Question 5.b

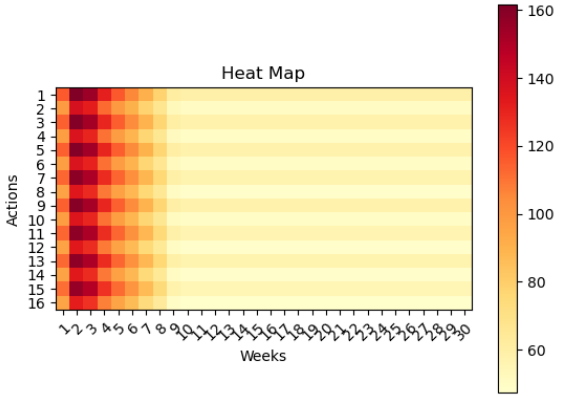
Based on the table 1, let's discuss the best performing policy with respect to each metric while excluding policies that do not have access to an action for the relevant metric:
N_confinement: The best performing policy is Russo with a value of 99.
N_isolation: The best performing policy is Toggle with a value of 0. The worst is Factor with a value of 210.
N_vaccination: The best performing policy is Toggle with a value of 0. The worst is Factor with a value of 210.
N_hospital: The best performing policy is Toggle with a value of 115. The worst is still Factor with a value of 210.
N_deaths: The best performing policy is Factor with a value of 186. The worst is Russo with a value of 58923.
R_cumulative: The best performing policy is DQN with a value of 35. The worst is Russo again with a value of -70.

Table 1: Policy Comparison

Policy	N_confinement	N_isolation	N_vaccination	N_hospital	N_deaths	R_cumulative
Russo	98.98	NA	NA	NA	58922.90	-70.29
DQN	176.26	NA	NA	NA	2856.28	34.66
Toggle	147.0	0	0	115.08	25901.96	-34.37
Factor	210.0	210.0	210.0	210.0	186.08	-20.80
Best Policy	Russo	Toggle	Toggle	Toggle	Factor	DQN



(a) Heat-map of Q valuesfor the DQN policy



(b) Heat-map of Q values for the Factor policy where explanation of actions is in the appendix

Figure 14: Heat-map of Q values

Question 5.c

Figure 14a and 14b show a heat-map of the evolution of all the Q-values as a function of time for the DQN and Factor policy. We can see a clear description of each action in Table 2.

The Q-values for the DQN policy are smaller than for the Factor policy. For the DQN policy, the Q-values are somewhat the same over the 30 weeks and for both actions: 'confine' and 'not confine'. For the Factor policy, the Q-values are somewhat the same for all actions but decrease over time and almost don't change after 10th week. We also remark that given the history of the episode, the Q-values are consistent with taking the first action all the time.

Question 5.d

In theory, we would expect the cumulative reward to increase as the policy has more freedom and flexibility to explore different actions. This is because a wider range of actions allows for greater adaptability and potentially better performance. However, in practice, we observed that the actual learning process is different.

One factor contributing to this is the sub-optimal learning rates employed. These learning rates determine how quickly the neural network adapts its weights and biases based on the observed rewards. If the learning rate is not appropriately calibrated, the network may struggle to converge to an optimal policy as seen in Figure 11.

Surprisingly, despite being the most restrictive in terms of available actions, the Deep Q-Network (DQN) policy consistently outperforms other policies in terms of average cumulative reward. It demonstrates that a well-designed and carefully optimised algorithm can overcome the limitations imposed by a smaller set of available actions.

These findings highlight the importance of algorithmic choices and optimisation strategies in reinforcement learning. It emphasises that a policy's performance cannot solely rely on the number of available actions, but also on the effectiveness of the learning algorithms employed.

Number	Confinement	Isolate	Vaccinate	Hospital
1	Confinement	Isolate	Vaccinate	Hospital
2	No confinement	Isolate	Vaccinate	Hospital
3	Confinement	No isolate	Vaccinate	Hospital
4	No confinement	No isolate	Vaccinate	Hospital
5	Confinement	Isolate	No vaccinate	Hospital
6	No confinement	Isolate	No vaccinate	Hospital
7	Confinement	No isolate	No vaccinate	Hospital
8	No confinement	No isolate	No vaccinate	Hospital
9	Confinement	Isolate	Vaccinate	No hospital
10	No confinement	Isolate	Vaccinate	No hospital
11	Confinement	No isolate	Vaccinate	No hospital
12	No confinement	No isolate	Vaccinate	No hospital
13	Confinement	Isolate	No vaccinate	No hospital
14	No confinement	Isolate	No vaccinate	No hospital
15	Confinement	No isolate	No vaccinate	No hospital
16	No confinement	No isolate	No vaccinate	No hospital

Table 2: Table 2: Explanation of the actions for heat-map for Question 4 part 2

Appendix

0.1. Improved Learning Rates

In this section, we detail the experiments we tried out with different learning rates for the cases when the neural networks did not converge optimally.

We recall that the training for 4.1 appeared to be incomplete even after 500 rounds. Thus, we tried to run training with the larger learning rate 10^{-3} and observed a clear improvement in the training and evaluation traces. We see a clear and consistent behaviour for all the three times we trained the agents. We recall that the training for 4.2 appeared to be very unstable. Thus, we tried to run training with the smaller learning rate 10^{-6} and observed a clear improvement in the training and evaluation traces. We see a clear and consistent behaviour for all the three times we trained the agents.

0.2. Explanation of Actions

The actions are as follows -

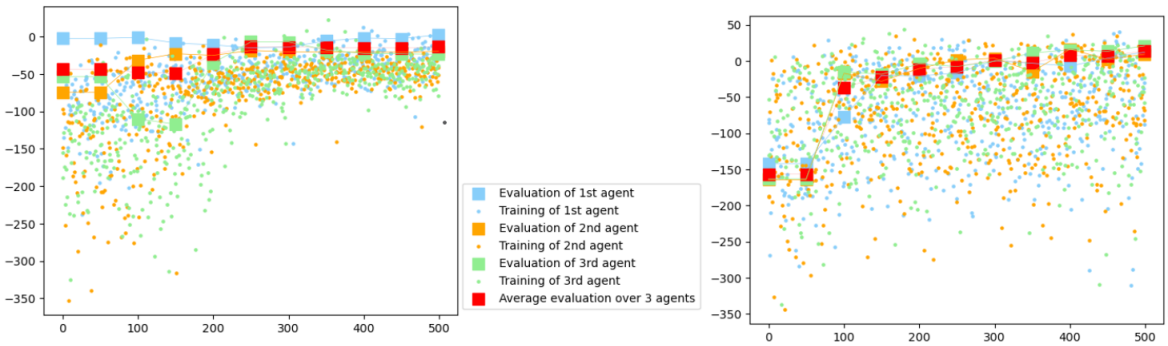


Figure 15: Left:Training and Evaluation Traces for Question 4.1 with learning rate 10^{-3} and Right:Training and Evaluation Traces for Question 4.2 with learning rate 10^{-6}

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.