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

Romain Birling & Antonin Faure

1 Introduction

Question 1.a) study the behavior of the model when epidemics are unmitigated

As shown on Figure 1, in the absence of any preventive measures, the number of susceptible people decreases rapidly within the first 10 weeks. This is primarily due to the high rate of transmission of the disease, which leads to a quick transition from susceptibility to infection and then to recovery or death. We can indeed see a sharp increase in the number of recovered individuals during the same period.

Moreover, if we focus on the total number of infected and dead individuals (Figure 2, we can see that the number of deaths grows rapidly until the 10th week, after which it plateaus at approximately 175,000. This indicates that an unmitigated spread leads to a significant number of deaths within a short period. The number of infections oscillates between 25,000 and 50,000 until the 10th week, after which it drops to zero. This pattern suggests that the virus has effectively spread throughout the population, leaving no one susceptible but many recovered or dead.

Looking at the data on a per-city basis on Figure 3, it's clear that the virus's spread and impact vary across different cities. Some cities see an early peak in infections and deaths, while others experience these peaks later in the time period. Despite these variations, all cities eventually reach a similar end-state with the deadly epidemic ending after week 15.

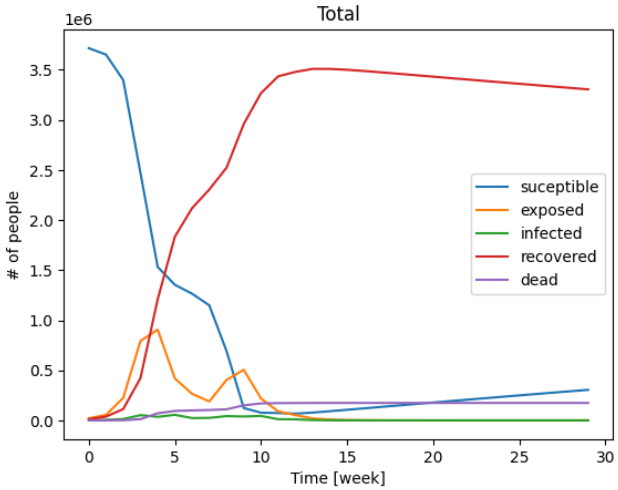


Figure 1: Population variables for unmitigated epidemic episode as function of time (week)

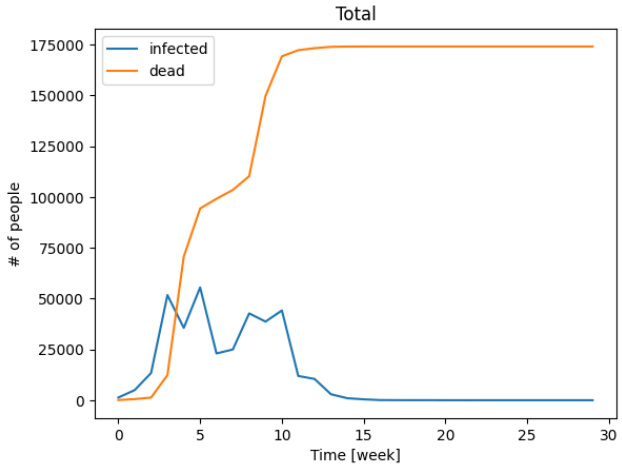


Figure 2: Total dead and infected for unmitigated epidemic episode as function of time (week)

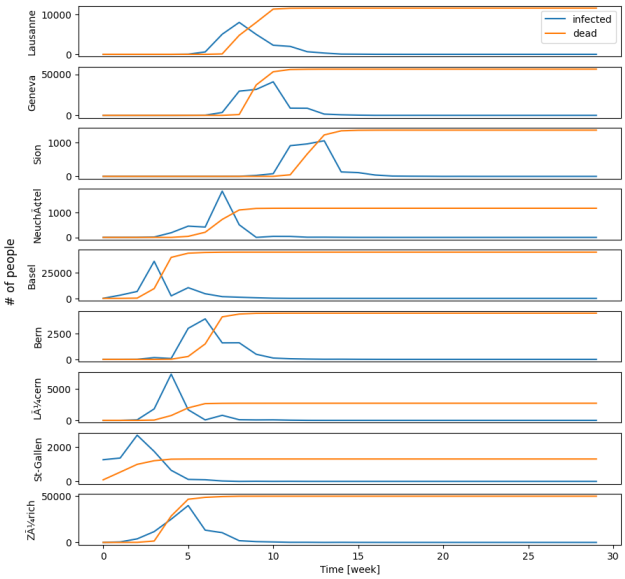


Figure 3: Dead and infected per city for unmitigated epidemic episode as function of time (week)

2 Professor Russo's Policy

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

As shown on Figure 6, with Russo's policy, the susceptible population is stable during each confinement period, implying that the spread of infection is being controlled during these times. The number of recovered individuals increases in plateaus, suggesting a slower rate of recovery during confinements, likely due to decreased infection spread. The death count also shows plateau patterns, indicating the policy's success in reducing fatalities.

On Figure 7 we can clearly see the number of infections spiking before the start of each confinement. This makes sense as the policy triggers a confinement when the number of infected cases surpasses a certain threshold (20,000). This shows that Russo's policy is responsive and is working as intended to restrict the spread during high infection periods. The death count's plateau behavior is a positive sign, reflecting the success of confinements in controlling the death toll. On the other hand, the Russo's policy cannot end the spread in less than 30 weeks.

The varied infection spikes in different cities shown on Figure 5 suggest that while the policy may effectively mitigate the disease spread in some regions, others might still experience heightened infections due to factors such as population density and timing of the first infection, with some cities never infected by the spread. The policy is somewhat effective on a national level but does not account for the heterogeneity in city-level outcomes.

The consistent and periodic confinements shown on Figure 4 reflect the policy's effectiveness in tracking and responding to the epidemic's evolution. Despite the periodic frequency of confinements, this strategy seems to manage the situation effectively, leading to plateaus in the death and recovery counts and successfully slowing down the disease's spread.

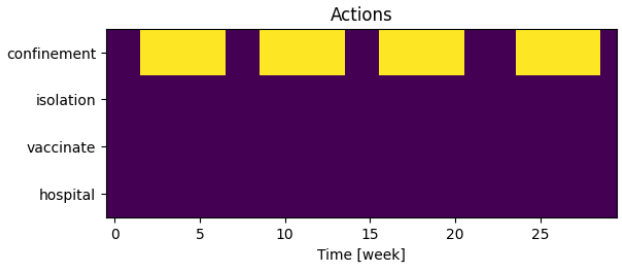


Figure 4: Actions for Pr. Russo's policy epidemic episode as function of time (week)

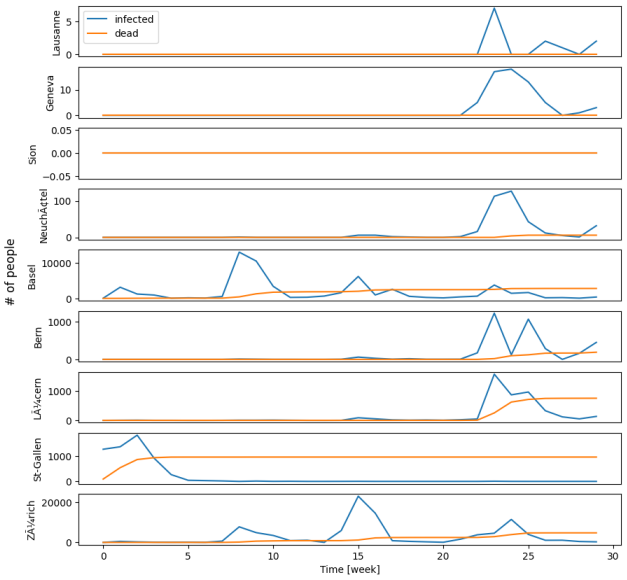


Figure 5: Dead and infected per city for Pr. Russo's policy epidemic episode as function of time (week)

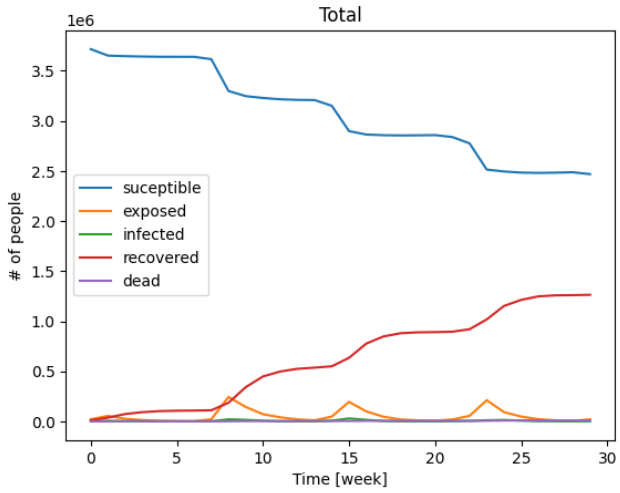


Figure 6: Population variables for Pr. Russo's policy epidemic episode as function of time (week)

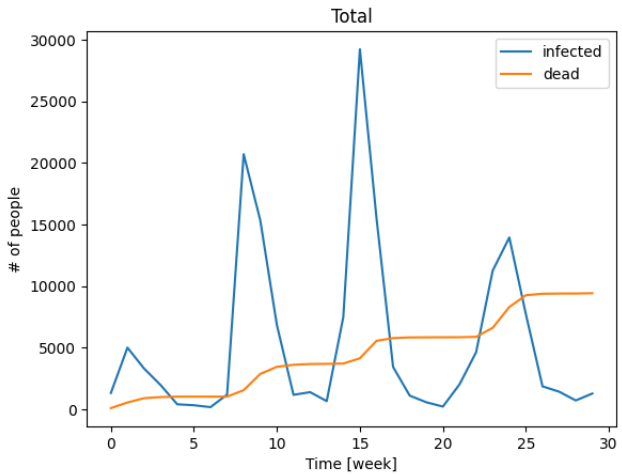


Figure 7: Total dead and infected for Pr. Russo's policy epidemic episode as function of time (week)

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

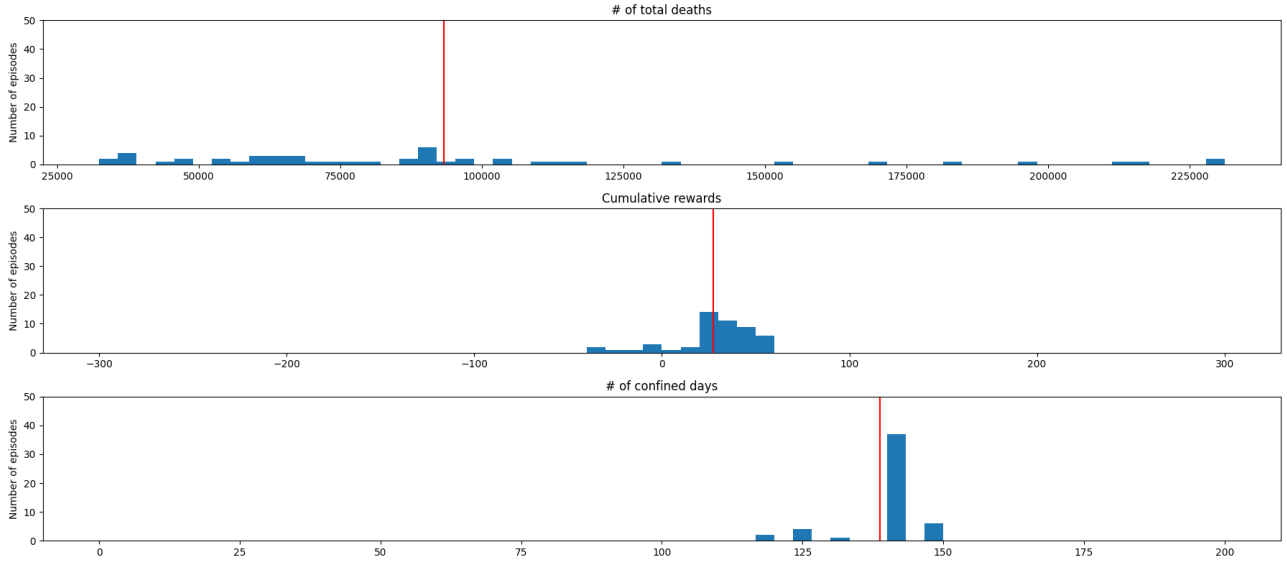


Figure 8: Evaluation of the Pr. Russo's Policy

3 A Deep Q-learning approach

3.1 Deep Q-Learning with a binary action space

Question 3.a) implementing Deep Q-Learning

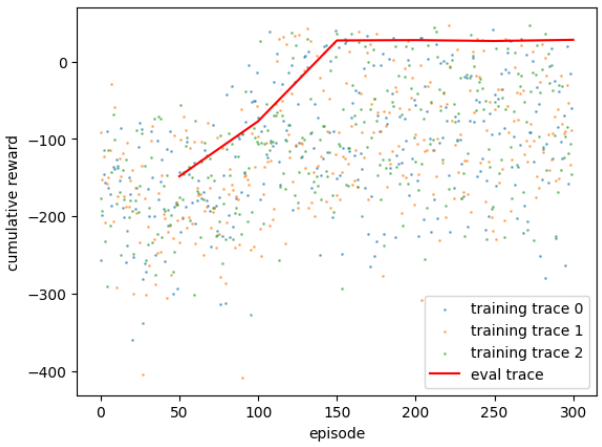


Figure 9: Training process of π_{DQN} with constant exploration

As seen in Figure 9, the evaluation trace, representing the average cumulative reward, increases until the 150 episode where it plateaus between 30-40. The training traces still spread uniformly between the eval trace and -300. Although the agent was able to learn a policy, the efficacy of that policy in mitigating the epidemic is debatable since Pr Russo's policy achieved the same average cumulative reward (around 40), as shown in Figure 8.

As seen in Figure 10, the learned policy characterizes itself by having different lengths confinements separated by short non-confinement periods. However, it doesn't seem to confine soon enough to prevent the disease's spread to other cities, resulting in sharp peaks of infected individuals. In the end the policy doesn't perform better than Russo's policy since it ends with more confined days while having about the same cumulative reward and number of deaths.

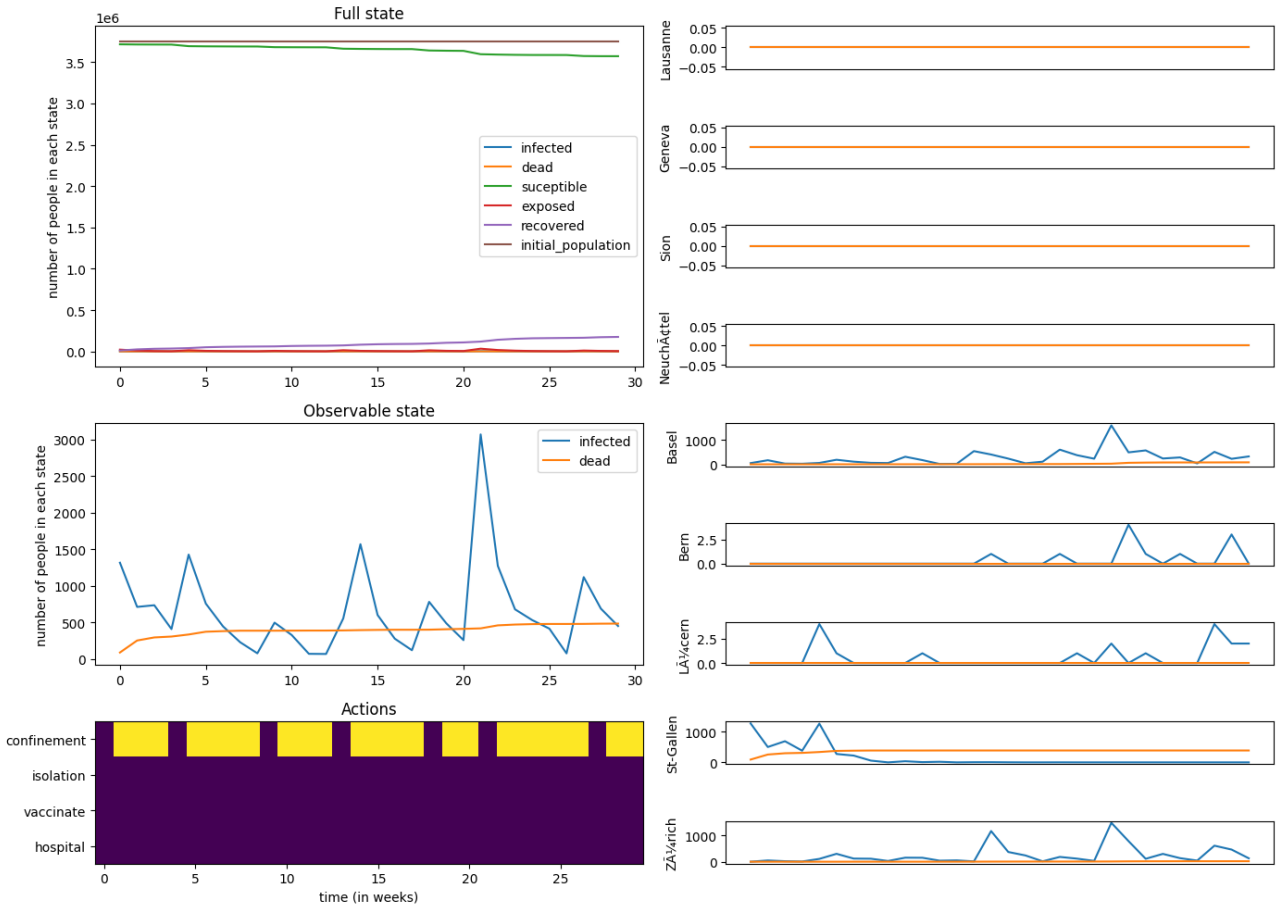


Figure 10: One episode of the best policy π^*_{DQN} with constant exploration

Question 3.b) decreasing exploration

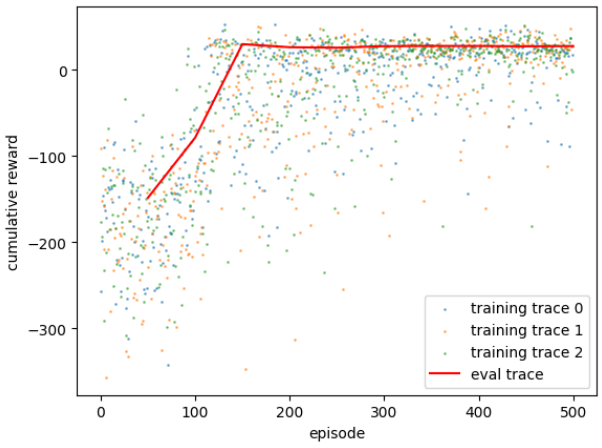


Figure 11: Training process of π_{DQN} with decreasing exploration

As seen in Figure 11, the evaluation trace, representing the average cumulative reward, increases until the 150 episode where it plateaus between 30-40. The training traces also tend to spread more around the eval trace.

As seen in Figure 12, the learned policy characterizes itself by having long confinements separated by short non-confinement periods with high infected spikes. However it doesn't end the epidemic in 30 weeks and does seem to confine too much. In the end the policy performs better than Russo's policy since it ends with about half the deaths.

It's hard to differentiate the two DQN policies since they have almost the same cumulative reward and confinement days. Nevertheless, if we had to choose one it'll be the one obtained with decreasing exploration (3.b) as shown in Figure 14 where we can see that the average number of death is lower than the one shown by policy 3.a) on Figure 13.

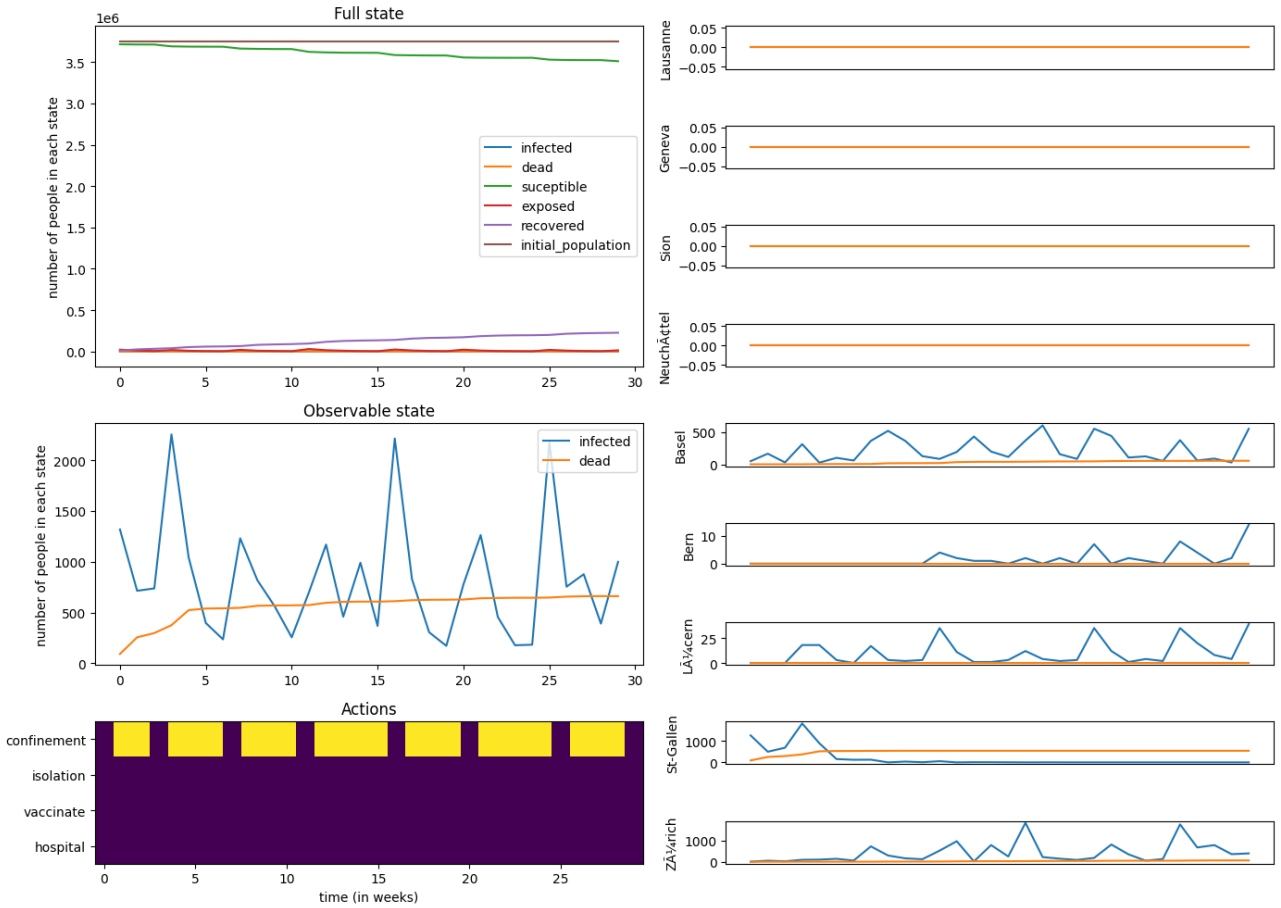


Figure 12: One episode of the best policy π^*_{DQN} with decreasing exploration

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

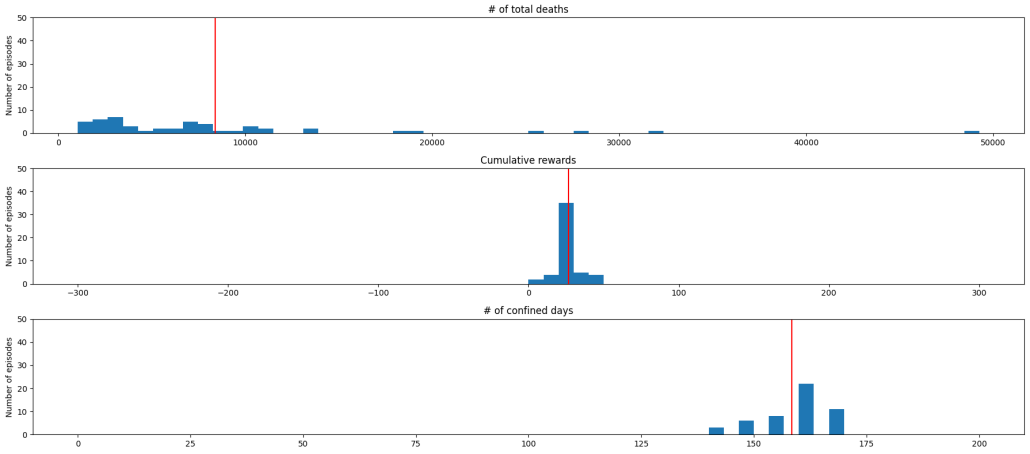


Figure 13: Evaluation of π^*_{DQN} without decreasing exploration

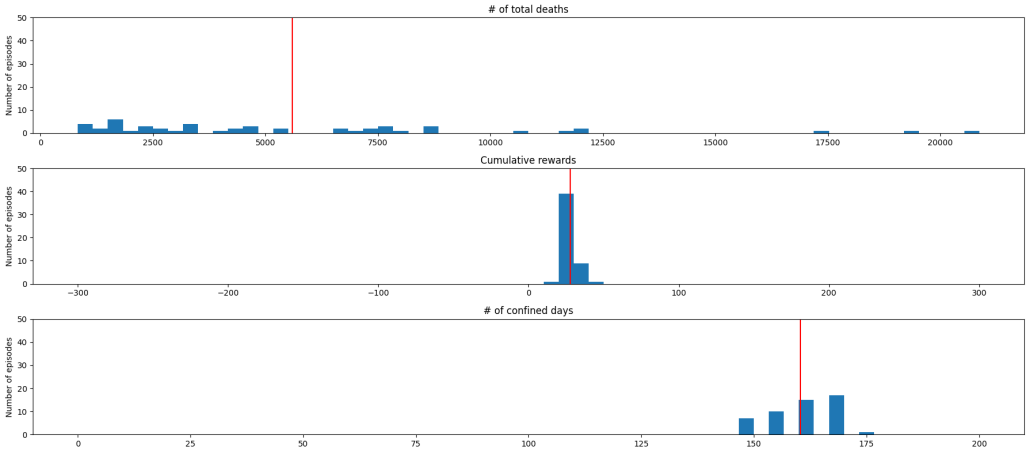


Figure 14: Evaluation of π^*_{DQN} with decreasing exploration

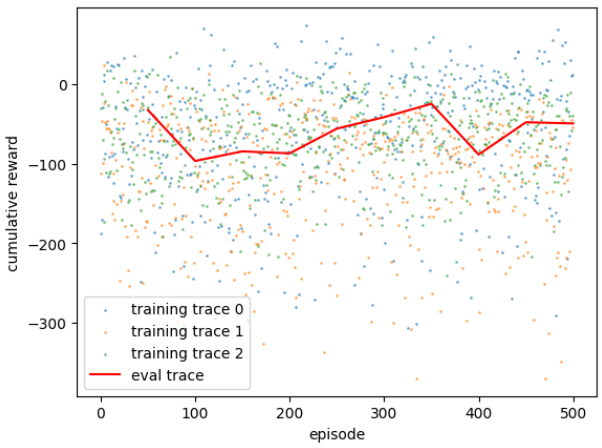
As said before, the best DQN policy is the one from 3.b) (with decreasing exploration) as shown in Figure 14. It exhibits better performance than Pr. Russo's strategy when evaluated on average confinement days and average death count. Despite having **higher average confinement days** (∼160 days compared to ∼140 days), it significantly **reduces the average death count** (∼5,500 compared to ∼90,000). While being potentially more disruptive in the short-term due to extended confinement periods, it crucially achieves a substantial reduction in fatalities, underlining its superior performance.

4 Dealing with a more complex action Space

Question 4.1.a) (Theory) Action space design

The toggle-action space proposed provides an elegant solution to handle more actions without expanding the direct action space dimensionality excessively. Instead of considering each possible state-action pair directly, we have a set of toggle actions that can modify the state of their respective actions. Thus we only need to estimate the Q-values for five toggle actions. This results in a more manageable action space, leading to a smaller, simpler network architecture that requires less computations, reducing the exploration time and variance during training.

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



As seen in Figure 15, the evaluation trace, representing the average cumulative reward doesn't look to increase so much, and is varying in a random way. Furthermore we can see that training traces don't converge to anything interesting.

As seen in Figure 16, the learned policy is the taking-no-action-policy. As discussed in 4.1.d network doesn't have a memory of past action. Since actions are dependent of each other, it converged to never toggling any action, which is probably better than randomly toggling any actions without knowledge of the previous ones. **It is not properly learning.**

Figure 15: Training process of π_{toggle}

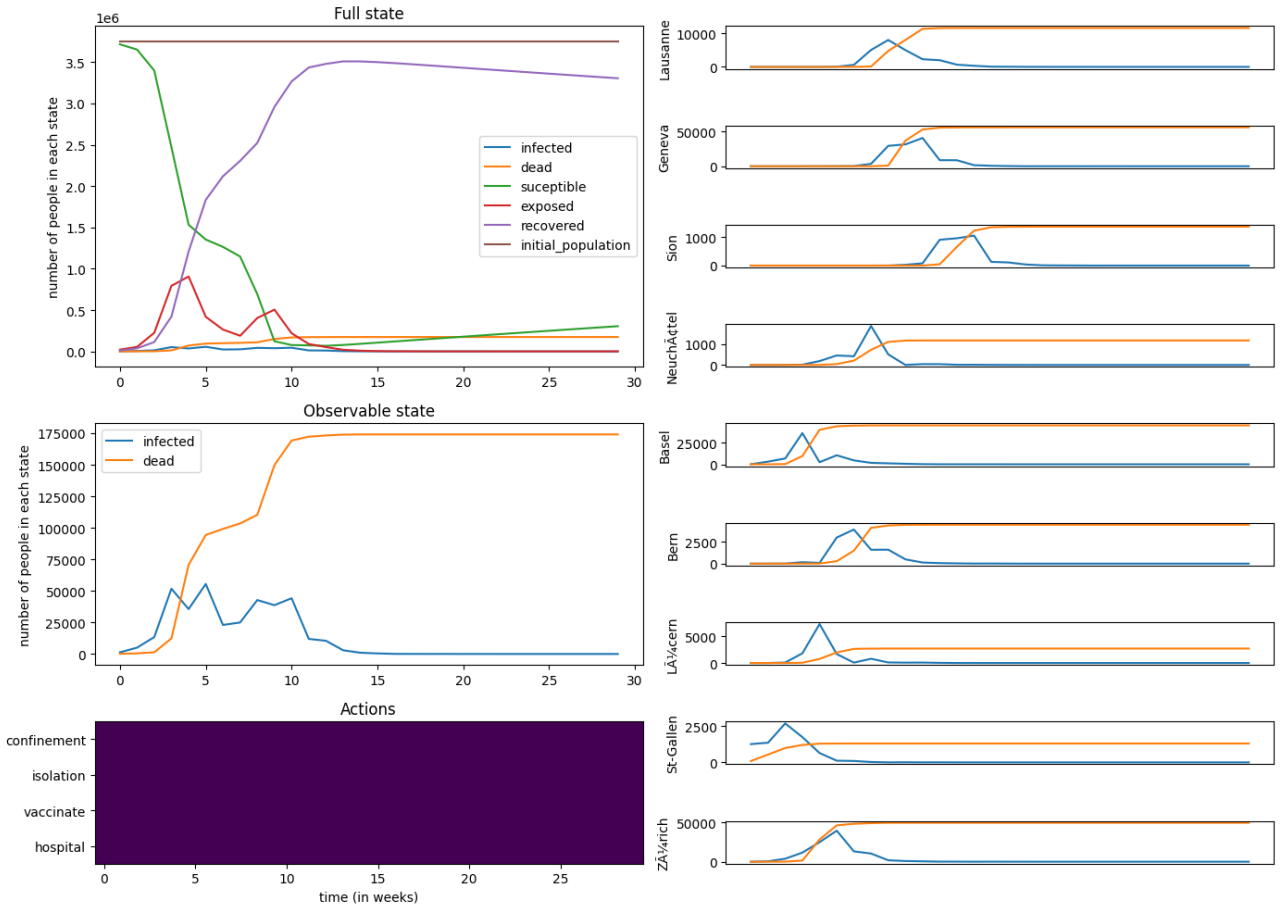
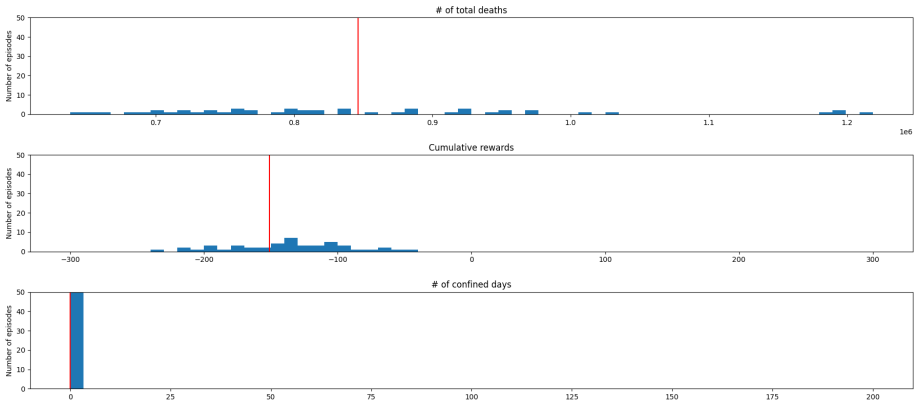


Figure 16: One episode of the best policy π^*_{toggle}

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



The toggle-action-policy performs much worse than the binary-action-policy because as we have seen, it is learning to not do anything at all. Its cumulative reward histogram is much lower.

Figure 17: Evaluation of π^*_{toggle}

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

It assumes that actions are **independent** of each other. This might not be valid if there are dependencies between actions. In addition, it doesn't have a **memory of past actions** which could lead to toggling actions on and off in consecutive time steps and may not be beneficial in some environments. Moreover, it assumes that actions have an **immediate effect** which is not always the case, for example toggling on/off a radiator doesn't affect instantly the temperature of a room.

4.1 Factorized Q-values, multi-action agent

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

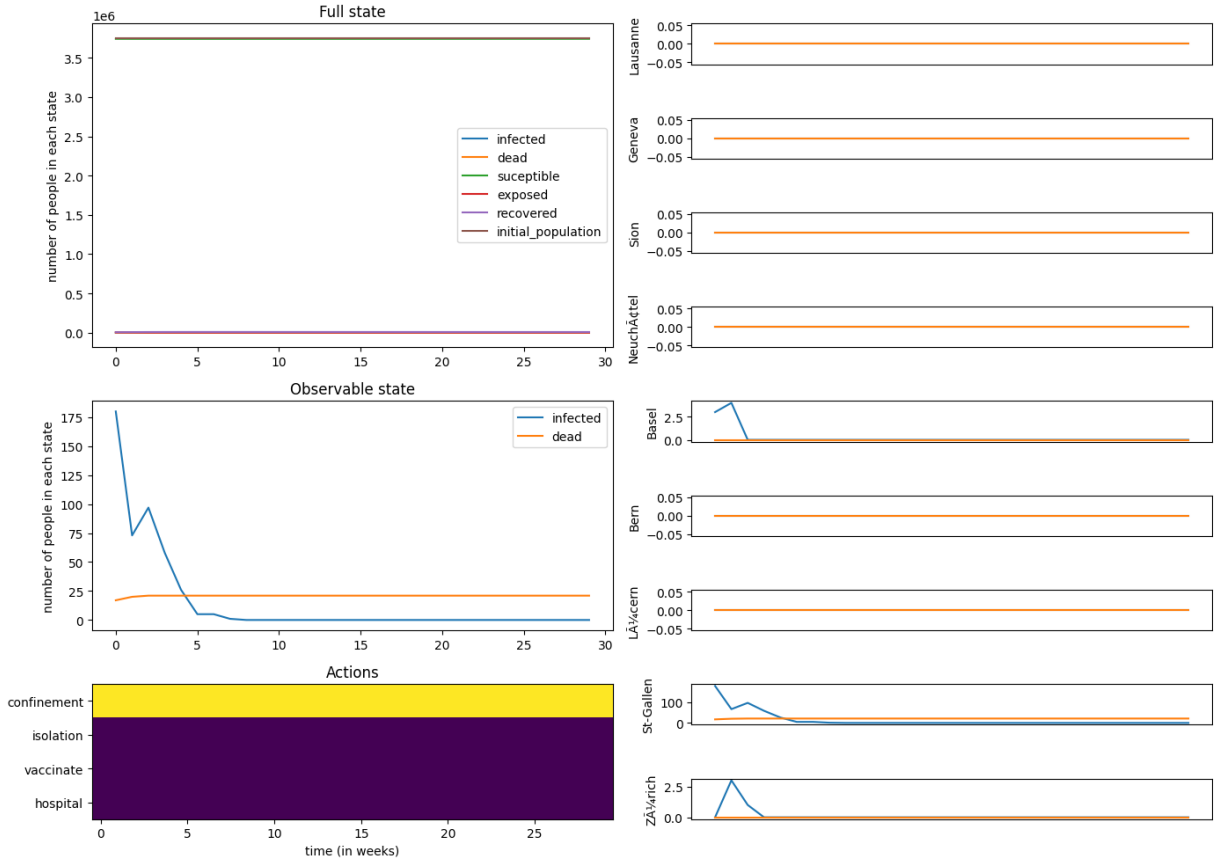


Figure 18: One episode of the best policy π^*_{Factor}

Figure 19: Training process of π_{Factor}

This agent doesn't successfully learn because it performs less good than binary-action-space policy having more freedom. For instance it looks like the useless actions distract the learning algorithm. **The policy is realistic** since it tends to predict to always confine people.

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

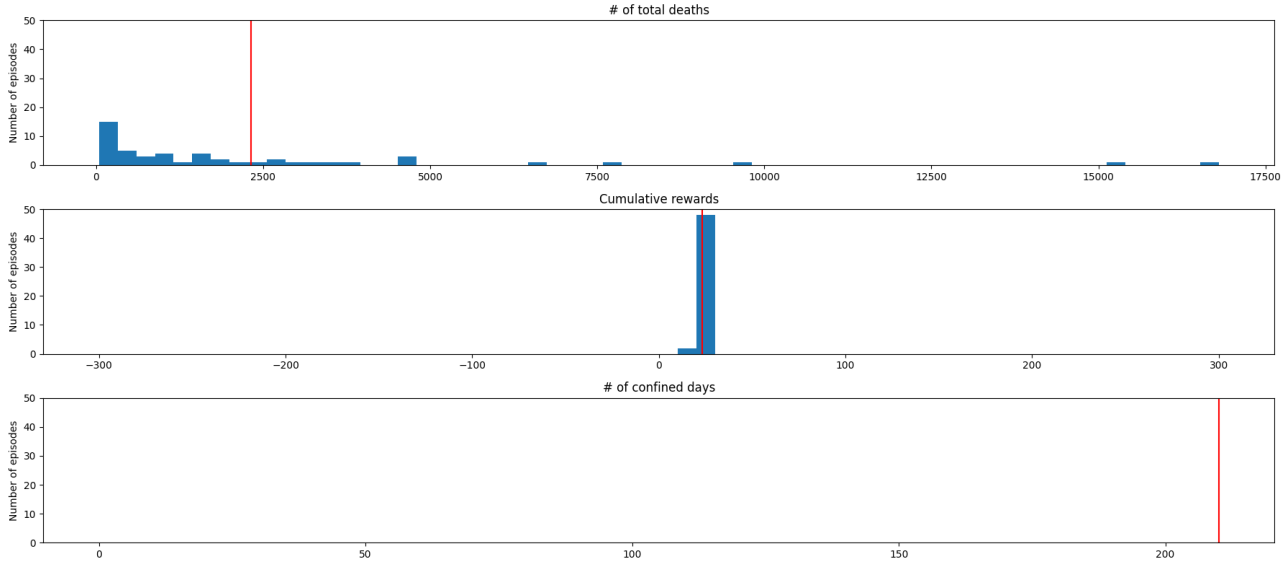


Figure 20: Evaluation of π^*_{Factor}

As shown in figure 20 **the factorized policy performs better than the toggled policy**. Its cumulated reward is slightly positive, while as shown in figure 17 the cumulated reward of the toggle action policy is slightly negative.

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

It assumes that actions are **independent** of each other and that there are no interaction effects between actions. In other words, the combined effect of two actions is assumed to be simply the sum of their individual effects. In addition, it assumes that all actions are of equal importance, as each action contributes independently to the Q-value. This might not be the case in many real-world problems, like for example in chess where the effectiveness of a move heavily depends on the sequence of prior moves.

5 Wrapping Up

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

- The single-action DQNs have an evaluation curve increasing then plateaus
- The toggle-action-space policy has an evaluation curve that doesn't seem to increase an randomly oscillate
- The factorized Q-values policy has a constant evaluation curve, which value highly depends on the seed.
- The Pr. Russo's policy doesn't learn thus it doesn't have an evaluation curve

The **single-action DQN** seem to perform best the learning process since it has an increasing cumulative reward curve.

Question 5.b) (Result analysis) Comparing policies

	avg$[N_{confinement}]$	avg$[N_{isolation}]$	avg$[N_{vaccination}]$	avg$[N_{hospital}]$	avg$[N_{deaths}]$	avg$[R_{cumulative}]$
π_{russo}	138.74	-	-	-	93250.46	27.328
π_{DQN}	160.3	-	-	-	5598.94	27.817
π_{toggle}	0	0	0	0	846481	-151.09
π_{factor}	210	0	0	0	2318.96	23.20

The Pr. Russo policy isn't the best in any criterion while other policies are the best performing in one of the criterion (thus extreme). The Pr. Russo policy can thus be interpreted as a good compromise between each criterion.

Question 5.c) (Interpretability) Q-values

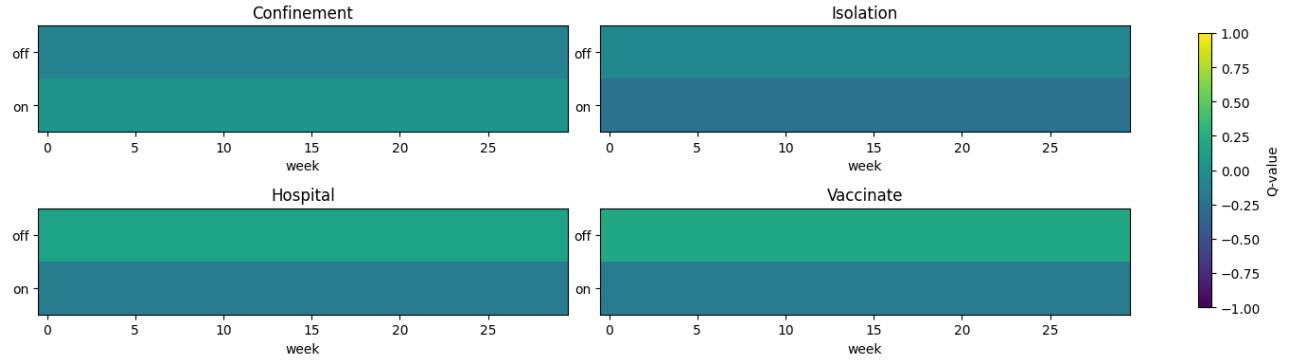


Figure 21: Q-values evolution heatmap for one episode of π^*_{Factor}

On Figure 21 we can see that for the π^*_{Factor} the Q-values are constant through time and only for "Confinement" do we have a "on" state with a higher q-value than the "off" state (thus being activated).

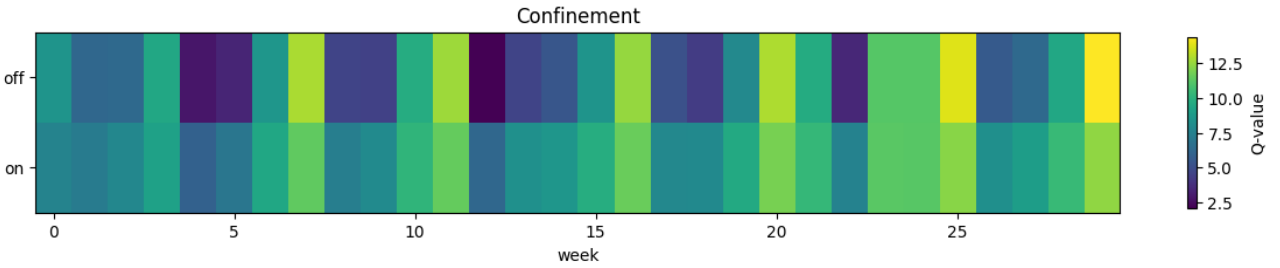


Figure 22: Q-values evolution heatmap for one episode of π^*_{DQN}

On the other hand for on Figure 22 we can see for the π^*_{DQN} policy that the Q-values varies a lot through time.

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

Adding more actions to an agent's action space **does not** automatically yield better rewards.

If the added actions are not relevant/beneficial given the specifics of the environment and task, they may not improve and even degrade performance. For example a useless action can distract the learning algorithm and make it harder to find the best policy by adding more complexity to the task.