**Question 1. (Exercise 6.2)**

TD update would be better on average than a Monte Carlo update, if we have lots of experience driving home from work and then we move to a new building and a new routine before entering the highway. Since we still enter the highway at the same place, the states and estimation after entering the highway would be the same, thus the value function of the two scenarios should be similar. This way if we take an initial guess that is close to the original scenario (starting from old building), it is more likely that we can get our result faster.

Also, MC method can only get the return after one episode is done, sometimes we cannot afford to wait if an episode is too long or the task is continuous without the present of an episode. TD method does not require a complete episode to generate a return, it only needs one time step to get that, so it is often faster than the MC method. In this scenario a TD method is preferred.

In original scenario, the above applies as well, if the initial guess of the value function is close to the real value function.

**Question 2. (Exercise 6.3)**

**Question 3. (Exercise 8.2)**

For the blocking experiment, the original path does not lead to the goal anymore. Dyna-Q+ performs better than Dyna-Q in the first phase because it explores more often with the presence of a bonus exploration, which ensures that Dyna-Q+ finds the shortest path faster than Dyna-Q, since in the early phase of path-finding process, more exploration is preferred.

After blocking the shortest path on the left (second phase), Dyna-Q agent will have to try the original path for many times until it realized that the previous path was blocked, whereas Dyna-Q+ always check the tiles that have not been tried for a while, it is easier for Dyna-Q+ to reach the goal through a new path with its bonus exploration. Also, once the original path was blocked, the agent has to find a new path from the beginning, which is the same as solving a different maze, so Dyna-Q+ will be faster, for the reason suggested above.

For the shortcut experiment, the original policy can still lead to the goal. The reason that Dyna-Q+ performs faster than Dyna-Q in the first phase is the same as in the blocking experiment.

In the second phase, there is a great possibility that the Dyna-Q agent cannot find the new optimal path, because it can still reach the goal through the original path, so it may not be able to find the new optimal path with limited ability of exploration. Whereas the Dyna-Q+ keeps track for each state-action pair of how many time steps have elapsed since the pair was last tried in a real interaction with the environment, so it can notice the change in environment right away and find the shortest path and develop a new optimal policy.