Section C – Final report

Manwel Borg – IT-MSD-6.3B

This study employed quantitative and qualitative data collection methods. Quantitative data included the statistical graphs generated by ML-Agents. These were used to evaluate the training runs for both approaches. The agents' performance during inference cannot be measured quantitatively. For this reason, the prototypes were personally evaluated; they were also sent to three participants who are familiar with the RTS genre. Data was collected from the participants through qualitative interviews. As stated by Creswell (2014), this method was used with the intention of eliciting views and opinions from the participants. They were requested to experiment with the prototype and asked open-ended questions about it.

Quantitative data:

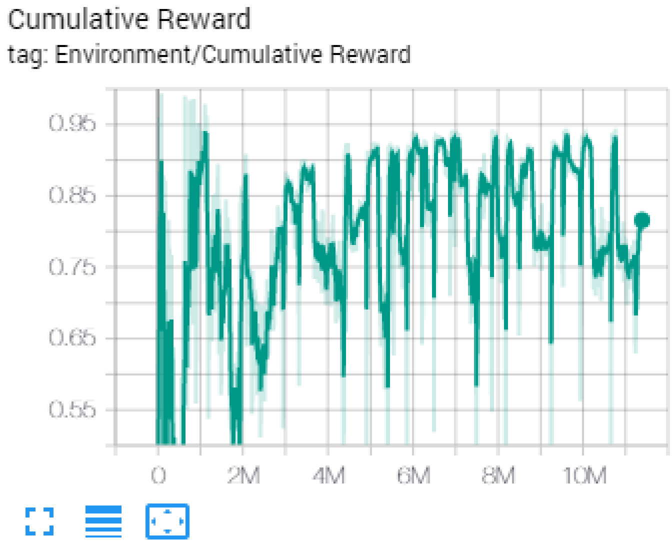

Figure 1: Cumulative reward (generalised reinforcement learning agents)

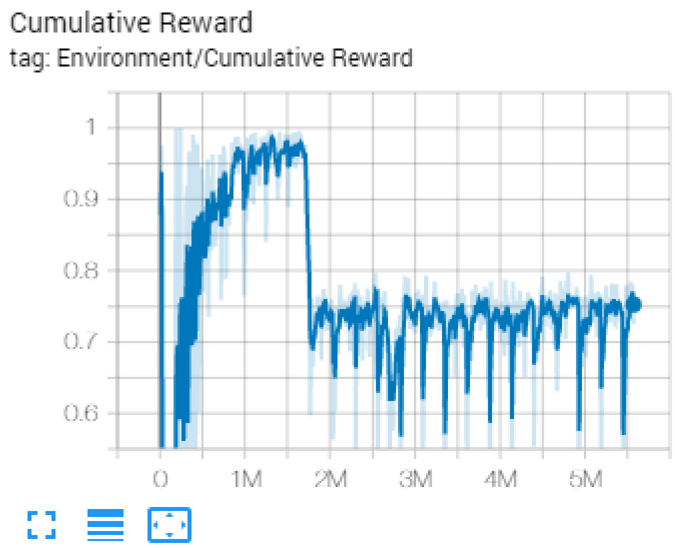

Figure 2: Cumulative reward (curriculum learning)

In Figures 3 – 7, the green graph is the first approach (generalised reinforcement learning agents); blue is the second approach (curriculum learning).

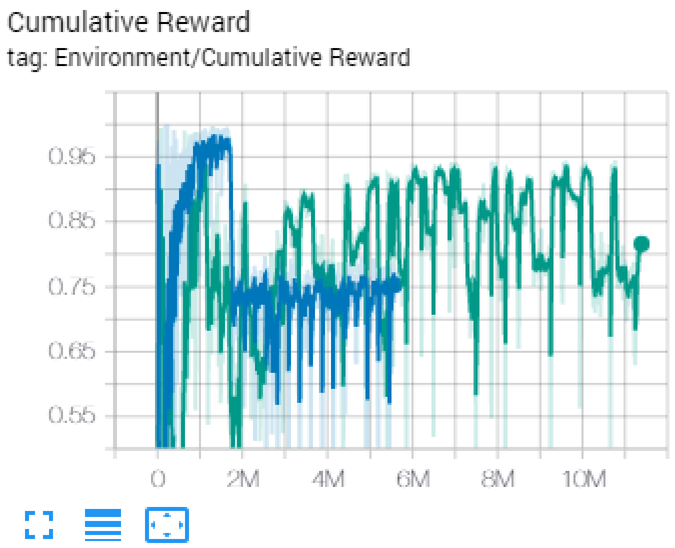

Figure 3: Cumulative reward for both approaches

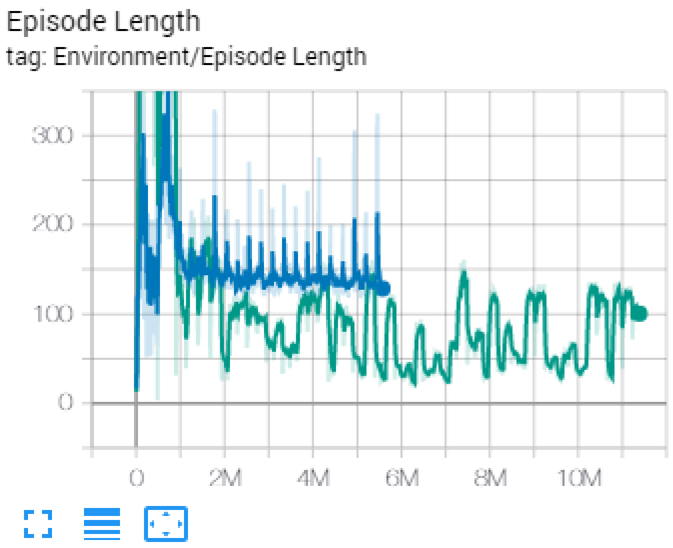

Figure 4: Episode length for both approaches

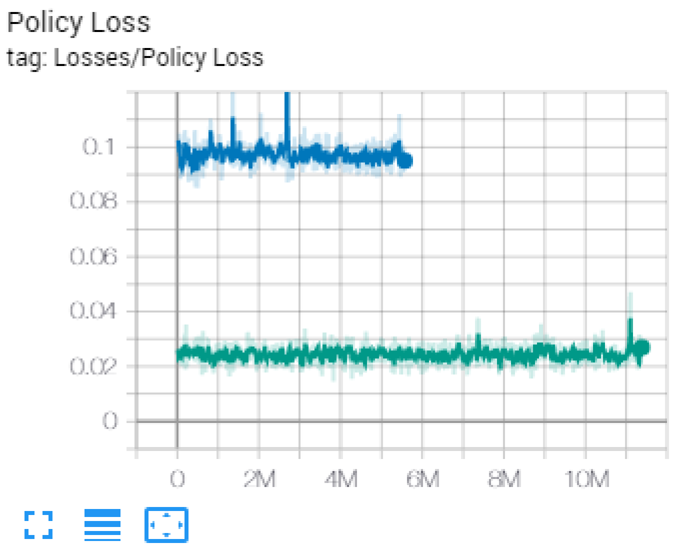

Figure 5: Policy loss for both approaches

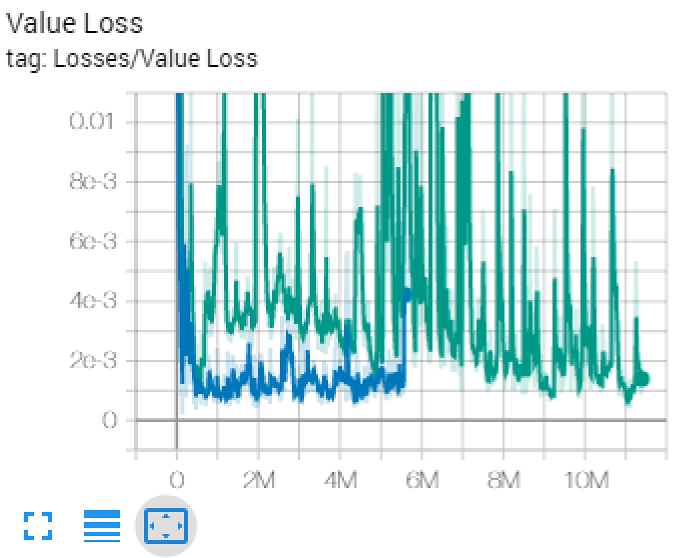

Figure 6: Value loss for both approaches

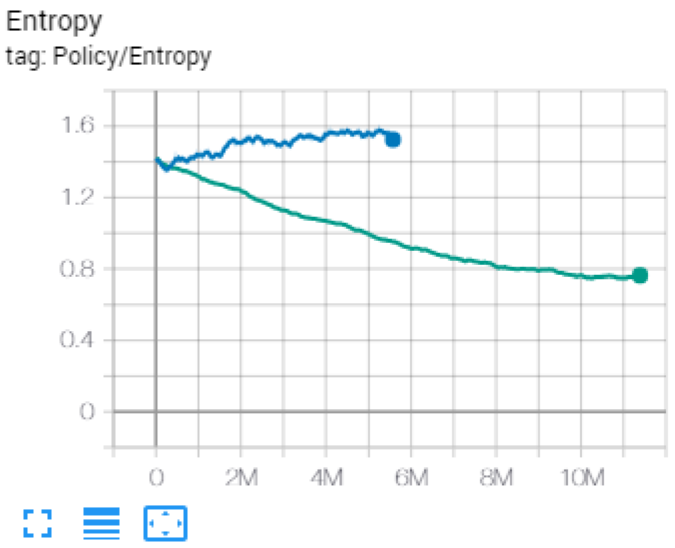

Figure 7: Entropy for both approaches

The above graphs were all explained in detail in the dissertation. From these results, it could be deducted that the training has been more stable in the first approach.

These results cannot be compared to results from other studies due to the difference in approaches. In the study by Vadim, Tatyana and Oksana (2018), for example, the agent was able to achieve a cumulative reward of 100 after 50,000 steps, but the methods with which it was trained and given rewards differed from the ones used in this study.

Training differences:

In their study, the agent was trained via imitation learning and reinforcement learning simultaneously. With this approach, the agent has shown to require significantly less training time compared to when reinforcement learning was used by itself. There are two ways with which imitation learning can be used in conjunction with reinforcement learning in ML-Agents; Behavioural Cloning (BC) and Generative Adversarial Imitation Learning (GAIL). Although it is not explicitly stated, it is evident that Vadim, Tatyana and Oksana (2018) used BC for their approach since ML-Agents did not offer the possibility to train agents via GAIL at the time. As stated in the ML-Agents documentation, BC is ideal when there exist demonstrations for nearly all the states that can be experienced by the agent. This is because it cannot generalise past the examples shown in the demonstrations, thus being unsuitable for this study. The ML-Agents documentation states that GAIL, on the other hand, works well when there is a limited number of demonstrations. For this reason, during the prototype development stage, an attempt was made to train the agents via Generative Adversarial Imitation Learning (GAIL). Several demonstrations were recorded in an environment consisting of a 20 by 20 plane and several obstacles. The agents were then trained with these demonstrations, but no difference was observed in the learning rate.

Reward differences:

For their approach, Vadim, Tatyana and Oksana (2018) rewarded their agent for multiple tasks per step. In this study, the agent's reward was set to 1, rather than added. As stated in the ML-Agents documentation, when there are multiple additions to the reward for a single agent decision, the rewards are summed together to evaluate how good the previous decision was. It is for this reason that the agent in their study was able to achieve a cumulative reward of 100. When the reward is set, all previous rewards are overridden, so the maximum the cumulative reward can be is the number it is set to. In this study, the agent had its reward set to 1 whenever it reached its target because that was the only task it was rewarded for.

The approach used in the study by Youssef et al. (2019) is comparable to that used by Vadim, Tatyana and Oksana (2018).

Qualitative data:

Below are two (one for each approach) of the ten test cases that are in the dissertation, followed by their analysis:

All tests were performed on a plane of 20 by 20 or smaller.

Test case no.	01
Approach	Generalised reinforcement learning agents
Description	A single agent moved from one point to another.
Expected result	If the target is within training coordinates, the agent takes the shortest path to it while avoiding any obstacles in the way.
Actual result	The agent would behave erratically if its target is far away from it; it would move in the opposite direction, fall off the map, or take an unnecessarily long path to the target. This has shown to occur primarily from the target being placed very far away from the agent, such as from bottom left to top right on a 20 by 20 plane, for example, but has shown to be hit-or-miss, meaning that it has happened as well with smaller maps. Otherwise, the agent moves normally while managing to take the shortest path and avoid obstacles. It was observed that sometimes, the agent would collide with an obstacle and then correct its behaviour shortly after, rather than avoid it altogether. It was also observed that the agent

	would occasionally behave erratically when there is a cluster of objects between it and its target; it would act as if there is a wall that is preventing it from moving further.

Test case no.	06
Approach	Curriculum learning
Description	A single agent moved from one point to another.
Expected result	If the target is within training coordinates, the agent takes the shortest path to it while avoiding any obstacles in the way.
Actual result	Unlike in test case no. 1, the agent's performance did not seem to be impacted if its target is very far away from it. In some cases, an agent would miss its target and then correct its behaviour shortly after.

During the development stage, testing was initially performed on a plane whose size was set so that it could only grow to a maximum of 10 by 10 during training, as opposed to 20 by 20. In that case, the agents have shown to perform outstandingly, but a size of 10 by 10 was considered too low to be viable for an RTS game, so it was doubled to allow for at least two players to be able to play comfortably in a hypothetical scenario. As a result, a decrease in training stability was observed. It is believed that this can be attributed to the fact that the agents receive a small penalty per action but are only given a reward upon reaching their targets. As previously mentioned in the training results section, a larger map increases the likelihood that an agent would have to travel over longer distances to reach its targets. This means that it would have to earn a substantial number of small penalties until it is given a reward.

It is believed that increasing the maximum map size has directly contributed to the erratic behaviour that was observed in test case no. 1, as this behaviour was not recorded when the agents were trained in a map that could expand to a maximum of 10 by 10. Although the agents should have been able to perform well in a map of up to 20 by 20 in size, their performance seems to have degraded even in maps that are smaller than 20 by 20. Strangely, this behaviour did not occur when the agents were trained via curriculum learning. The exact reason for this could not be pinpointed. It is believed that it is because the agents spent a longer time training in the 20 by 20 version of the map, as unlike in the generalized learning approach, the size of their map did not change periodically throughout the entirety of the training.

The test cases have shown that in both approaches, the agents' behaviour lacked in consistency. It was observed that the agents from the first approach were unable to adapt to their environment due to the behaviour they have shown when moved to positions that are far away from them. This behaviour was not observed in agents from the second approach. Moreover, at times, the agents from both approaches were unable to pass through clusters of obstacles despite there being enough space for them to do so.

A difference in movement between approaches was also observed. The agents from the first approach have shown that their movement is much more refined; those from the second approach tended to move roughly. In an RTS game, the units would ideally move smoothly, as the zig zags in movement lengthens the path.

This has shown that the agents from both approaches struggled to adapt to their environment as their performance was not satisfactory in all parts of the map.

1-on-1 interviews:

All figures in this section are screencaps from the interview recordings.

For the first approach, the following observations have been made by the participants as highlighted in the appendix:

- The agents behave erratically when their targets are placed very far away from them, but it has been observed that this does not always happen. An example of this can be seen in Fig. 8. The agents were sent the position of the cursor; they were able to move to it without problems.

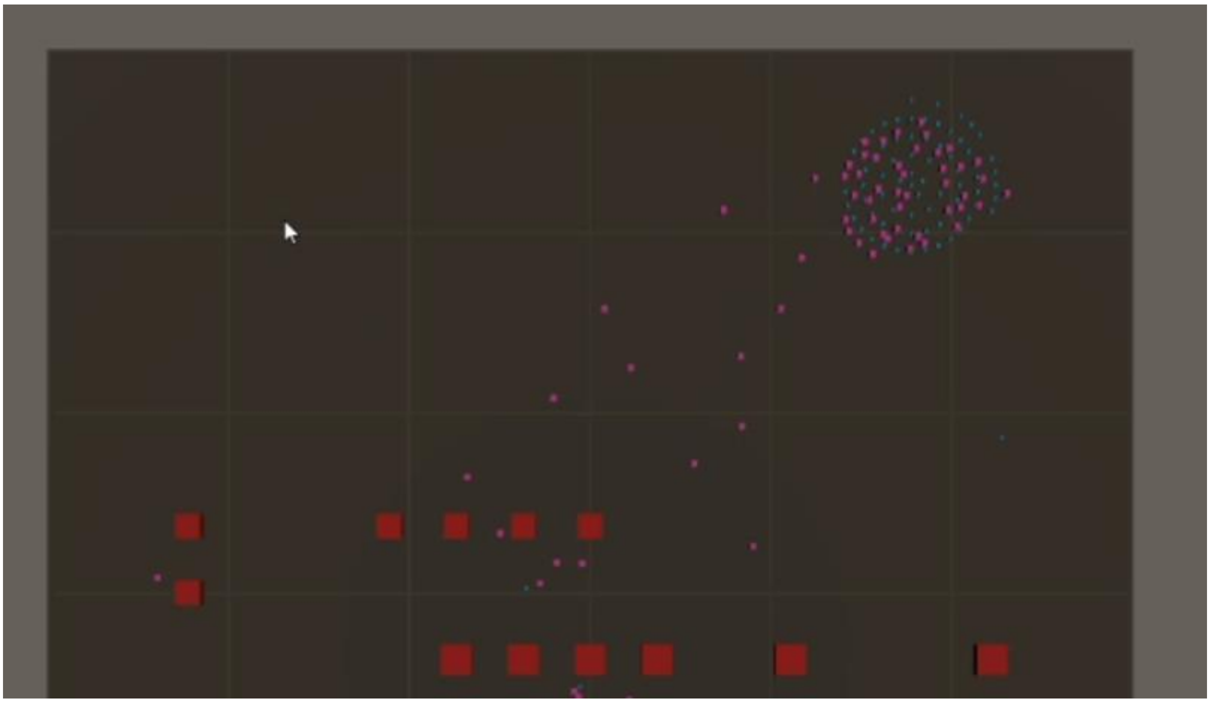

Figure 8: Sending the agents far from their current position

- Single unit movement was okay for the most part.

- When the agents are moved in large groups, some of the agents were surrounded by those that had managed to stop at their targets as shown in Fig. 9. As such, they either kept moving in circles or pushed past stationary agents to reach their targets.

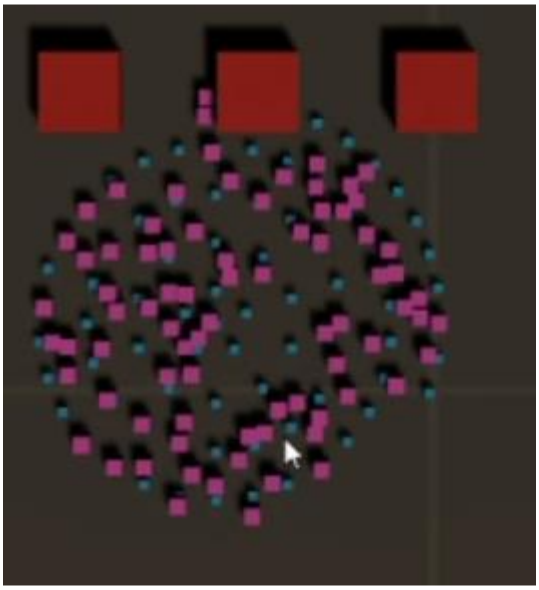

Figure 9: The remaining agents getting surrounded by the others

- When the agents are faced with obstacles, they sometimes bump into them and then correct their behaviour immediately after. However, they were generally good at avoiding obstacles.

- The agents have shown to be good at avoiding each other, albeit some pushing occurred.

- The agents moved smoothly; their movement was free from abrupt stops or zigzags.

- When moved in groups, some of the agents freeze; the other move normally.

- The agents do not always take the optimum path when sent to a location. An example can be seen in Fig. 10, in which although the agents were sent to the right, they chose to first move down and then to their paths, rather than going right altogether. In Fig. 10, the obstacles could also have negatively impacted the agents.

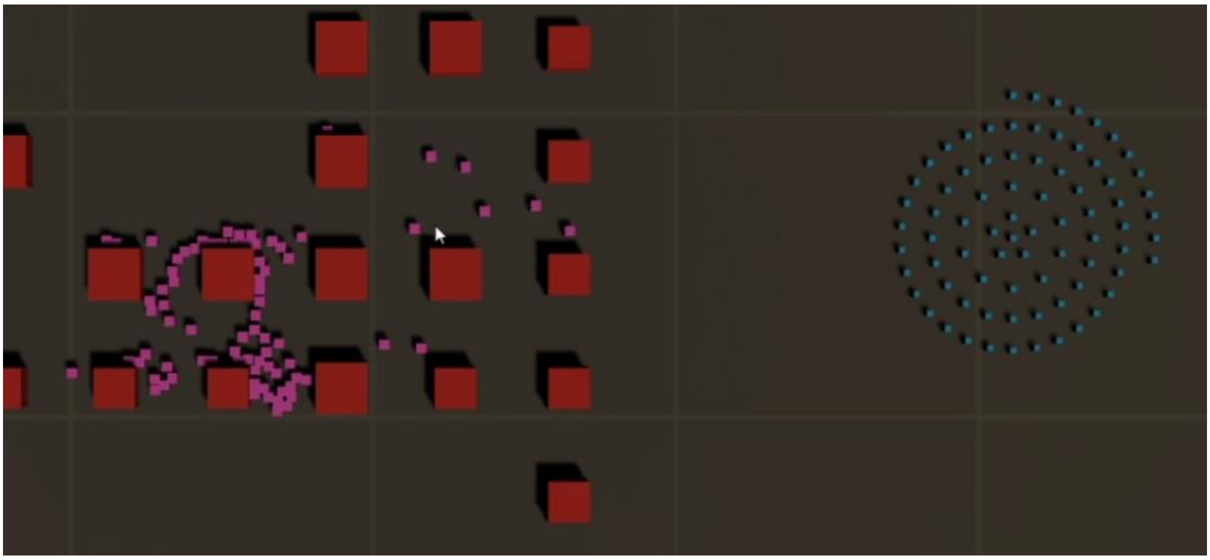

Figure 10: The agents taking a longer path than necessary

- In some cases, the agents take long to pass through clusters of obstacles as if there is a wall as shown in the bottom right of Fig. 11.

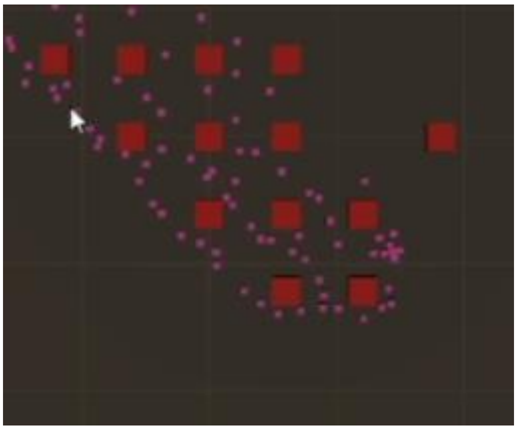

Figure 11: Some of the agents reacting poorly to obstacles

- This approach is not production ready; several improvements must be made to it before it can be used in a finished game.

For the second approach, the following observations have been made by the participants as highlighted in the appendix:

- When they are sent to targets that are far away from them, the agents do not exhibit the same behaviour as in the first approach. However, their ability to stop at their targets was worse as shown in Fig. 12; but this did not happen everywhere in the map. In Fig. 13, for example, the agents were sent near the centre of the map and were able to reach their targets.

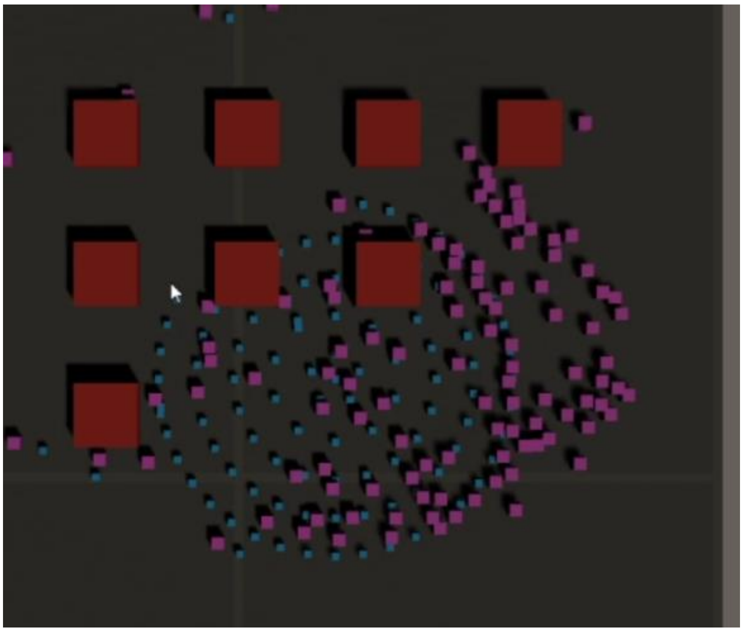

Figure 12: The agents attempting to reach their targets

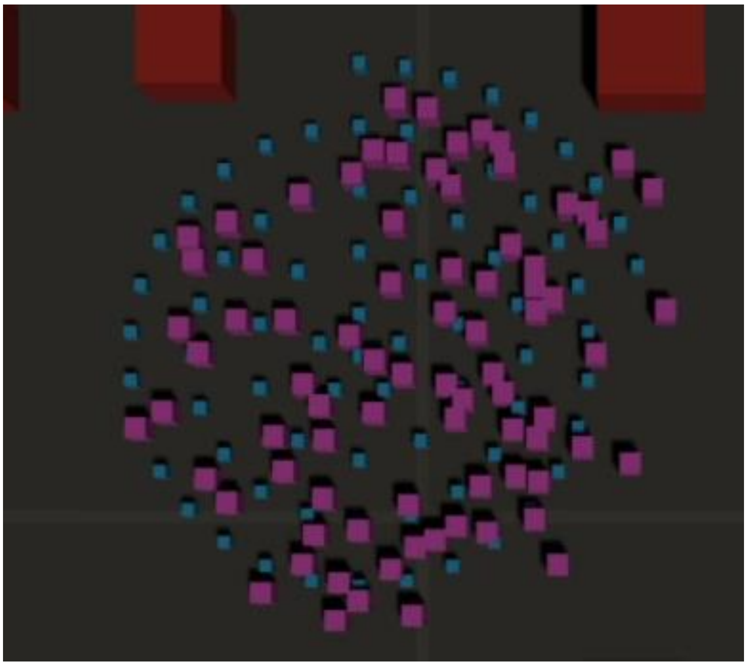

Figure 13: The agents reach their targets without much problems

- Single unit movement was also good for the most part.

- The agents in this approach behaved similarly to those in the first approach when faced with obstacles or other agents.

- Group movement was not bad; the agents in this approach scattered more as they moved.

- The agents' movement was less smooth; they have shown to zigzag when moving.

- Like in the first approach, the agents do not always take the shortest path to their targets.

- The agents frequently get stuck trying to "reach" targets that do not belong to them.

- The agents' performance seems to have been the worst near the edges of the map.

- The agents react poorly to clusters of obstacles; this was observed as well in agents from the first approach.

The observations that have been made by the participants are comparable to those from the test cases. It has been observed by the participants that both approaches lacked in consistency; their performance, for example, would be acceptable in some parts of the map but declines in others. Two of the participants have stated that they preferred the first approach over the second one. The participants have observed the following significant differences between approaches:

- Movement: The agents from the first approach have shown that their movement is refined whereas the others have shown to move roughly, their movement consisted of sudden zigzags.

- Group movement: In the second approach, group movement has not been bad per se, but the agents have shown to be unable to arrive to their targets in groups in most cases.

- Behaviour in large maps: In the second approach, the map size did not seem to have severely affected the agents' ability to move to their targets when they are sent to a position that is far away from them.

- Behaviour near the edges of the map: The agents from the second approach have shown a tendency to behave erratically when sent to a target that is close to the edge of map.

- Reaction to targets: In the second approach, the agents would frequently attempt to reach targets that do not belong to them.

These results also cannot be compared to results from other studies, as there is no research that addresses the implementation of reinforcement learning pathfinding in RTS games or includes in-depth post-training observations such as the ones in this study.