TME – Navigation Strategies

The code for the different gating strategies is included in the 'strategyGating.py' file.

The data for the random persist strategy was recorded using 1 random seed for 40 trials, while the data for the q-learning strategy was recorded using 10 random seeds (0 to 9) for 40 trials each.

The trial duration data for the random persist strategy are under 'log_randomPersist', and the trial duration and q-values data for the q-learning strategy are under 'log_qlearning'.

Step 2: Random Persist Trial Durations

The trial durations for the random persist strategy will be used in later parts as a 'control' for the Q-learning strategy.

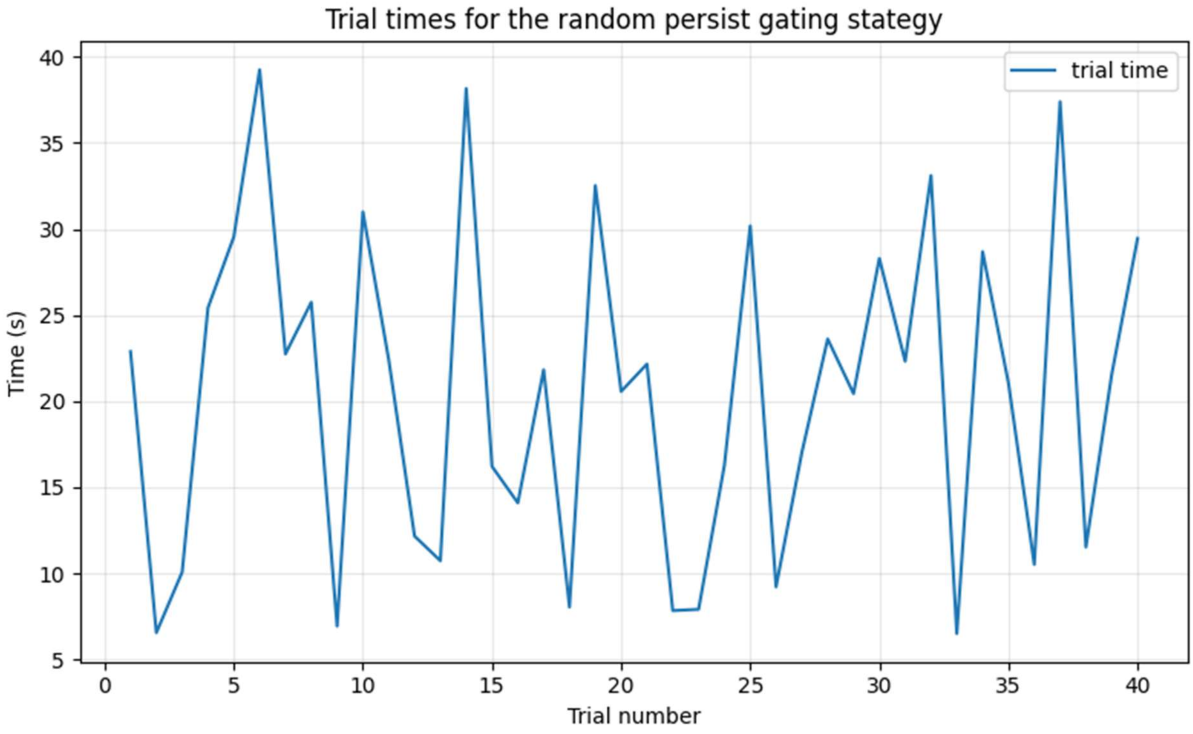

Figure 1 Trial times for the random persist gating strategy. *Trial times were recorded over 40 trials, and through one run.*

Table 1: Random Persist Quartile Trial Durations	
Median	21.691389560699463
First quartile	11.328941583633423
Third quartile	28.386833667755127

Steps 4 & 6: Q-values

As a quick reminder, the higher the Q-value for an action in a particular state, the more beneficial that action is expected to be, so:

- If Q(S, 0) > Q(S, 1), the robot will favor the wallFollower strategy when in state S.
- If Q(S, 1) > Q(S, 0), the robot will favor the radarGuidance strategy when in state S.

The Q-values of the desired states are summarized in the table below:

Table 2: Q-values

State	Action == 0 (wallFollower)	Action == 1 (radarGuidance)
‘00002’	0.36028811	0.41152139
‘00072’	0.36222706	0.41556849
‘00000’	0.99941692	1.10317686
‘00070’	1.10976456	1.03668722
‘11101’	-0.88327389	-1.39716876
‘11171’	-1.09773849	-1.28663101

In states ‘00002’ and ‘00072’, the first three digits indicate that no walls are detected, while the robot is far. Both strategies yield moderate positive Q-values for both strategies, but radarGuidance has slightly higher values than wallFollower (0.36 vs. 0.41 and 0.36 vs. 0.42). This indicates a marginal preference for radar guidance in open spaces, which aligns with expectations since the radar guidance strategy helps orient the robot toward the target, which is optimal when there are no walls around.

In states ‘00000’ and ‘00070’, the only difference with the two prior states is that the robot is near. Consequently, both strategies achieve higher Q-values compared to the previous states since the robot is closer to reaching the goal. In state ‘00000’, radarGuidance performs slightly better (1.00 vs. 1.10), as expected as no walls are recorded to be near. In contrast, in state ‘00070’, the wallFollower strategy has a slightly higher Q-value than radarGuidance (1.10 vs. 1.04), which could be due to the fact that the obstacles are near the robot, so the Q-values could be influenced by the future reward associated to taking a turn towards the obstacle walls. This is also motivated by the high value associated to wallFollower for the ‘00000’.

In states ‘11101’ and ‘11171’, three walls are detected. We can see that all four Q-values are negative, reflecting penalties associated with the robot getting stuck facing the walls. We can also see that wallFollower performs better than radarGuidance in both states, with less negative values (-0.88 vs -1.40, and -1.09 vs. -1.29). That is expected in such positions as the wallFollowing strategy should get the robot away from the walls while radar guidance would get it stuck.

We can see a general pattern that corresponds to expectations given the problem: radarGuidance has higher Q-values in open areas and when the robot is near the target, while wallFollower has higher scores when the robot is facing walls, or even is near walls.

Steps 5 & 7: Q-Learning Trial Durations

Figure 2 shows the time each of the 40 trials took for each seed used (0 to 9). Unexpectedly, we can see that the values oscillate between 10 and 200 seconds (with one 350 seconds outlier), while the random persist trial times only ranged from 10 to 30 seconds. To see this patten more clearly, we averaged the curves and superposed the Q-learning and random persist results.

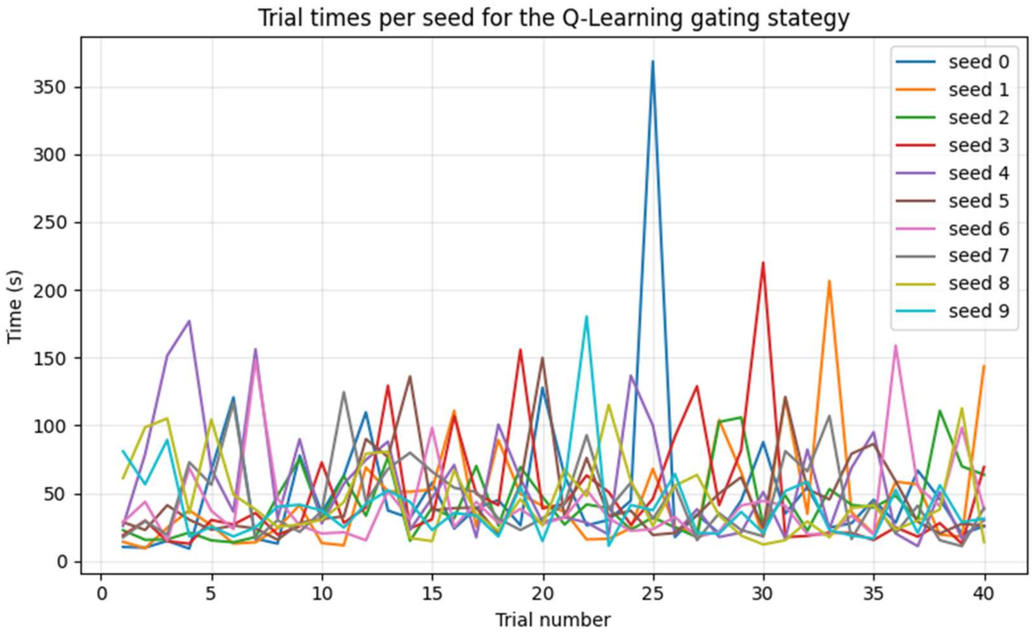

Figure 2 Trial times per seed for the Q-Learning gating strategy. *Trial times were recorded over 40 trials and the curves 'seed i' correspond to the curve for seed == i.*

Figure 3 shows that the Q-learning performance is clearly poorer than the random persist. Although Q-learning is not necessarily the optimal solution for the problem, it is expected to perform at least better than random choice. Therefore, I believe the long trial durations for Q-learning could be due to the machine running too many scripts in parallel, as I ran 5 seeds in parallel, and then 5 other seeds in parallel. Since the trials were measured through duration of the trial and not number of steps until the target is reached, the computer slow down thus could have made it seem like Q-learning has worse performance, while measuring using steps might have had opposite results. I won't be able to take the measures again as I do not have access to the PPTI computers anymore.

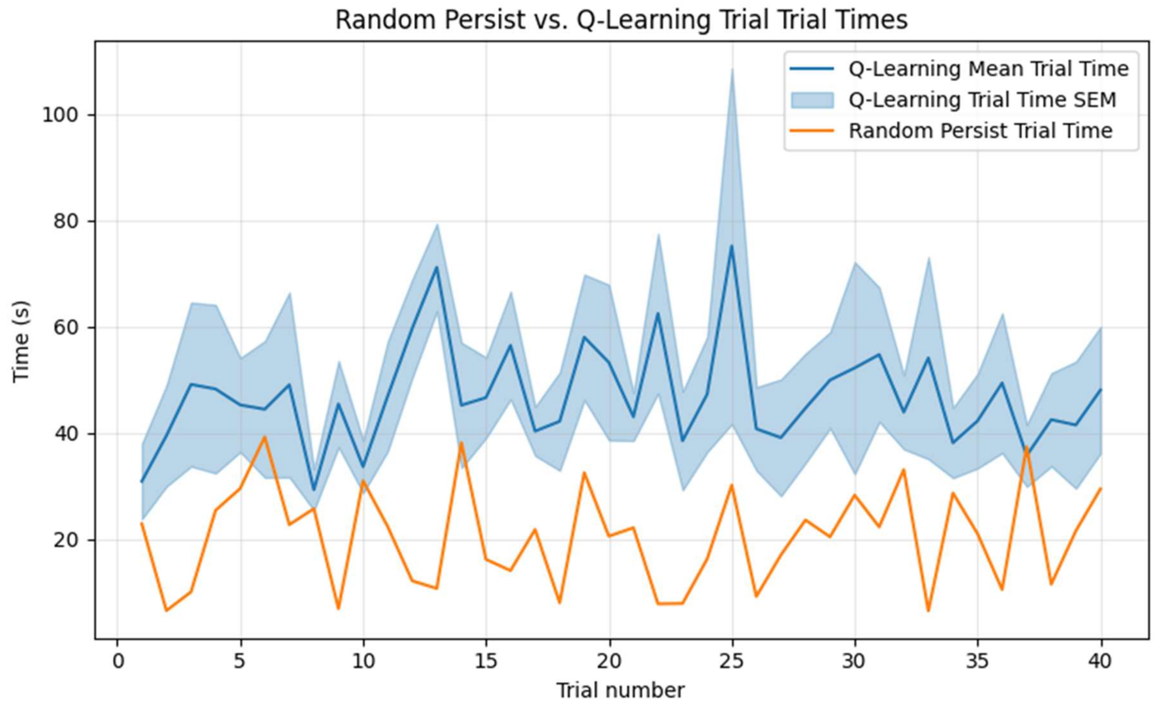

Figure 3 Random Persist vs. Q-Learning trial times. *The trial times for Q-learning were recorded over 40 trials and repeated 10 times with different random seeds.. Unexpectedly, the Q-Learning method shows longer trial times than the random persist one. This could be explained by overuse of computational resources, or by Q-learning not being optimal to use for the task at hand.*

This pattern also shows in the median, 25th, and 75th quartile measures, where we expected Q-learning times to be smaller than random persist times, but ended up having all Q-learning measures being higher than all random persist measures. The last 10 trials of Q-learning also should have taken less time than the first 10 trials of Q-learning, as the Q-table should contain values that have learned from experience in the last 10 values, but instead we have a higher median (41.7 > 35.1) and 75th quartile for the last 10 trials (49.1 > 47.6).

Table 3: Quartile Trial Durations for both strategies	
Random Persist Trial Durations	
Median	21.691389560699463
First quartile	11.328941583633423
Third quartile	28.386833667755127
Q-Learning Trial Durations (First 10 Trials)	
Median	44.867052137851715
First quartile	35.11361702084541
Third quartile	47.568386846780776
Q-Learning Trial Durations (Last 10 Trials)	
Median	43.212720370292665
First quartile	41.68889576792717
Third quartile	49.064426028728484