MPr-1001

Studying the effect of different exploration strategies in Reinforcement Learning in the context of a multi-agent foraging system

Atthaluri shashank

5th Year Int. M.Sc. Mathematics, Department of Mathematics

UM-DAE Centre for Excellence in Basic Sciences

Electronic address:a.shashank@cbs.ac.in

Project Advisor: Prof. Sanjay Chandrasekharan

Homi Bhabha Centre for Science Education, Tata Institute of Fundamental Research

V. N. Purav Marg, Mankhurd, Mumbai 400088, India.

Electronic address:sanjay@hbcse.tifr.res.in

Abstract

Many organisms generate structures, such as pheromones, in their environment to reduce cognitive complexity. This project extended a multi-agent computational model (Chandrasekharan & Stewart, 2007) of how organisms could learn to generate such structures in the environment within their lifetime. The agents had some basic action and sensing capabilities, and they had to learn to put these elements together in ways that led to systematic generation of external markers. Agents used a reinforcement learning mechanism (Q learning) to achieve this objective. We studied different conditions under which such structures were created, including different exploration rates, as well as a new strategy, where the worst performing agents die and new agents were born. We report here the implementation, analysis and the results obtained. Our findings suggest that contrary to expectations, random actions were higher instead of systematic pheromone dropping and following actions. This suggests that random actions help agents reach places in the world where dropping and following pheromones would be useful. Our analysis also indicates that the q-table (policy function) in q-learning may not be working as a decision-support system, but as a condensed form of experience in a stochastic world filled with other agents.

Acknowledgements

This project could not have been completed without support and encouragement from these people. My greatest gratitude is extended to:

- My advisor, Prof. Sanjay Chandrasekharan, the best mentor I could have had. My academic pursuit would have been much harder without the plenty of freedom he gave me. Apart from the project, the course discussions have helped me in learning many things which will be very useful for my future work.

- The other one is Harshit Agarwal. I am grateful to him for teaching, guiding and supporting me in the the conceptualization, coding, and reporting of the project.

Contents

1 Introduction **4**

2 Background of Q-learning algorithm **6**
2.1 Exploration vs Exploitation and Exploration rate 7

3 System **8**
3.1 Agents 8
3.2 World 9

4 Experimentation **10**

5 Analysis **11**
5.1 Analysis of Each Iteration 11
5.2 Analysis across the experimental conditions 15

6 Results **18**

7 Conclusion **19**

8 References **20**

A Appendix **21**
A.1 Appendix-1 21
A.2 Appendix-2 23
A.3 Appendix-3 25
A.4 Appendix-4 27
A.5 Appendix-5 28

1 Introduction

"Good work is done the way ants do things: Little by little."

- Lafcadio Hearn

All animals face navigation problems, some of which involve orientation. To reduce the cognitive complexity (minimize search or inference) involved in this process, some organisms have developed the capability to add stable structures to the environment, for themselves or others or both. Wood mice (Apodemus sylvaticus) distribute small objects, such as leaves or twigs, which acts as visual markers that serve as points of reference while foraging. They do this way-marking even under laboratory conditions, using plastic disks. This way-marking diminishes the likelihood of losing interesting locations during foraging (Stopka & MacDonald, 2003). Red foxes (Vulpes vulpes) use urine to mark food caches they have emptied. This marking acts as a memory aid and helps them avoid unnecessary search (Henry, 1977, reported in Stopka & MacDonald, 2003). The male bowerbird builds colorful bowers (nest-like structures), which are used by females to make mating decisions (Zahavi & Zahavi, 1997). Most wasps, ants, and bees have specialized abdominal, foot, or head glands that they use to mark substrates.. Ants and other social insects use pheromones very successfully to mark trails connecting the nest to food sources (Holldobler and Wilson, 1990) and to recruit other ants to the foraging task. Many mammals mark their territories. In the immune system, antibodies help to fight microbes by recognising and marking the antigens as being foreign, and later on this marking helps macrophages to destroy those antigens.

Such "doping" of the world is observed mostly in organisms with smaller brains (such as insects and rodents) to lower the cognitive load. In general, animals with larger brains (such as horses, elephants, monkeys, etc.) do not exploit this strategy. However, humans are an exception to this. We exploit both internal and external traces simultaneously. Humans used an endless list of epistemic structures like: Markers, color-codes, page numbers, credit-ratings, stains, traces, badges, shelf-talkers, speed bugs, road signs, post-it notes, etc.

The pervasiveness of such restructuring of the world across species indicates that adding structures to the world (cognitive congeniality) is a basic adaptive strategy (Kirsh, 1996). From here onwards, we will call such stable organism-generated external structures that lower cognitive load epistemic structures (ESs). The term is derived from the distinction between epistemic and pragmatic action

developed by Kirsh and Maglio (1994).

Epistemic structures can be classified into three types, based on whom they are generated for. (examples of each in brackets).

- Structures generated for oneself (Cache marking, bookmarks).
- Structures generated for oneself and others (Pheromones, color codes)
- Structures generated exclusively for others (Warning smells, badges).

There are other ways to classify ESs (e.g. by function - structures for mating, foraging, etc.), but the above classification is more suited to the objective of our model.

In our model we consider non-human organism cases like ants, wood mice and red foxes.
Here we will make two reasonable assumptions:

- organisms sometimes generate random structures in the environment (pheromones, urine, leaf piles) as part of their everyday activity.
- organisms can track their physical or cognitive effort (i.e., they get 'tired'), and they have a built-in tendency to reduce tiredness.

During foraging and cache retrieval tasks we can observe a few random structures are being generated in the environment. These random structures help the organisms to ease the task (following pheromones reduces travel time, avoiding urine makes cache retrieval faster, avoiding leaf-piles reduces foraging effort). In other words, these structures help to shorten the path. Given the postulated bias to avoid tiredness, these paths get preference, and they are reinforced. Since more structure generation leads to more of these paths, structure generation behavior is also reinforced" (Chandrasekharan, 2007).

By using this framework we developed a computational model where we build 10 agents in a simple world. The agents learn systematically how to add the structures to the world. The foraging task consists of two phases: one leaving the home to search for the target and other to return back to home. Here the "target" can be thought of as a food source. By using this task the agents learn the forage by the shortest path. We perform this task using a simple method known as Q-learning (Watkins, 1989).

2 Background of Q-learning algorithm

"...what we want is a machine that can learn from experience."

- Alan Turing 1947.

Q-learning finds the optimal policy by learning the optimal Q-values for each (state, action) pair. In other words, Q-learning lets the agent use the environment's rewards to learn, over time, the best action to take in a given state.

The agents learn from the previous rewards. By the reference of previous reward to the reward received for the action in the current state, the Q-value is updated. The idea is to estimate the future rewards by the past experiences.

A Q-value for a particular (state, action) determines the "quality" of an action taken from that state. Better Q-values implies better chances of getting higher rewards.

We initialize the Q-values with zeros, and expose the agents to the environment where the agents receives different rewards by executing different actions. The Q-values are updated using the equation:

$$\text{Q(state,action)} \leftarrow Q(state, action) + \alpha(reward + \gamma \max_a Q_{new}(next\ state, all\ actions) - Q(state, action))$$

Where:

- α (learning rate) can simply be defined as how much you accept the new value vs the old value. In the above equation, we are taking the difference between new and old and then multiplying that value by the learning rate. The resulting value then gets added to our previous q-value which essentially moves it in the direction of our latest updated value.
- γ (discount factor) determines how much importance we want to give to future rewards. If the value of the discount factor is close to 1 the agent considers it as the long-term effective award, whereas, a discount factor is 0 makes our agent consider it as an immediate reward. Hence, It balances the immediate and future reward.
- Max is taking the maximum of the future reward and applying it to the reward for the current state.

Basically, we are learning the best action to take in the current state which will maximize the rewards in the future.

2.1 Exploration vs Exploitation and Exploration rate

Exploration is the act of exploring the environment to find out more information about the environment. Exploitation is the act of exploiting the information that is already known about the environment in order to maximize the return.

The goal of reinforcement learning is to maximize the expected return, so we might think that we want our agent to use exploitation all the time and don't bother about any exploration. But this strategy is not a good solution. Exploitation will be very ineffective if the agent has very limited or no information about the world. This information is gained via exploration.

Generally, an effective learning algorithm needs to balance exploration and exploitation. This project explores this balance between the two.

Exploration Rate in the context of Q-learning is the fraction of times that a random action would be chosen instead of resorting to the q-table which stores the policy function. This means that higher exploration rates would lead to agents taking actions which may not be preferred based on the agent's experience up-till then but might open up new parts of the search space for the agent.

3 System

"Our own body is in the world as the heart is in the organism... it forms with it a system."

- Maurice Merleau-Ponty

The system was the one used by Chandrasekaran & Stewart, 2007 with very minor modifications. The next two sections give a brief overview of the elements of the system.

3.1 Agents

Each agent in the system could detect the level of home and food pheromone in any cell. This constitutes the state of the agent. Each agent can perform the following actions:

- Drop Food pheromone
- Drop Home pheromone
- Follow Food pheromone
- Follow Home pheromone
- Move Randomly

One way to think about these actions is to consider the pheromone-following ability of ants. Common models of ant foraging (e.g. Bonabeau et al., 1999) postulate the automatic release of two pheromones: a home pheromone and a food pheromone. The ants go towards the home pheromone when they are searching for their home, and they go towards the food pheromone when foraging for food. This exactly matches these two actions in our agents. We had made a few assumptions which are assumed and performed in (Chandrasekharan, 2007) paper like episodic memory for the agents which has functionality of dropping pheromones of different types at a particular point in the world and able to recall data when in that particular context. Also a similar functionality is being added that when dropping new pheromone at the old dropped pheromone, the data of new ones makes only a small change to the amount of pheromones at that location. Dropping pheromone is an epistemic structure added by the agents to the world.

In our experiment, We ran 10 agents initially and we gave our agents a punishment (a reward of -1) whenever they perform an action, and a reward of 10 whenever they complete a trip(home to target, then target to home). Thus, to minimize their long-term punishment, they would need to travel from their home to the target and back as quickly as possible. Each agent uses the Q-learning algorithm independently without any communication in between them. Since the agents know nothing about the environment or the expected rewards for any state-action pair, all the Q-values are first initialized to zero. Over time, though, as the agents run several episodes(time steps), the Q-values produced for the state-action pairs that the agents experience will be used to update the Q-values. In the later episodes, the agents can look into this Q-values and base its next action on the highest Q-value for the current state.

3.2 World

Our world consisted of a 30×30 toroidal (doughnut-shaped) gridworld, with two 3×3 square patches which represents the agent's home and target as explained in Chandrasekharan & Stewart, 2007.

4 Experimentation

"The best way to show that a stick is crooked is not to argue about it or to spend time denouncing it, but to lay a straight stick alongside it"

- D.L. Moody

We developed multiple conditions to understand how changing exploration rates would affect the foraging behavior of the agents. We developed five different conditions and ran the model for each condition 20 times with 200,000 iterations.

Condition 1 is a replication of (Chandrasekharan, 2007) where the exploration rate starts from 10% and then drops to 0% over time.

Condition 2 is a modification of Condition 1 when the exploration rate is 10 percent throughout the simulation.

Condition 3 also is a modification of Condition 1 where the exploration rate starts from 10% and then drops to 1% over time.

In these conditions, all the agents have the same exploration rate at any given time in the simulation. Appendix-1 shows the decay of the exploration rates in conditions 1, 2 and 3.

Condition 4 is where after every 10,000 timesteps the agent which performed the worst (i.e. collected the least amount of food) in the past 10,000 timesteps dies and a new agent is added to one of the home cells. The new agent begins with empty q-table.

Condition 5 It's an adaptation of Condition 4 with a change that the new added agent inherits the memory (q-table) of the best performing agent in the last 10000 steps.

In conditions 4 and 5, the newly added agent starts at 10% exploration rate which reduces to zero with time. Agents which have been longer in the simulation will have lower exploration rates then their younger counterparts. Thus during the simulation, different agents will have different exploration rates.

5 Analysis

"When analytic thought, the knife, is applied to experience, something is always killed in the process."

- Robert M. Pirsig

5.1 Analysis of Each Iteration

We analyzed the data generated from the experiments from many different viewpoints. The data logged was: the position, direction, action, state and successful trip completion of each agent. This data was binned in 400 timestep sized blocks for the purpose of the analysis. We used the trip completion data to measure the performance of the agents in different experimental conditions as described below. We use Trial 11 from Condition 3 to illustrate the analysis done for each iteration.

Total actions and trips across all agents: Figure 1 shows one trial from Condition 3. The top part of the figure shows the distribution of the 5 actions. The lower part shows the number of successful trips completed by all the agents in the 400 time-step sized blocks. Sample graphs for other conditions is present in Appendix-2.

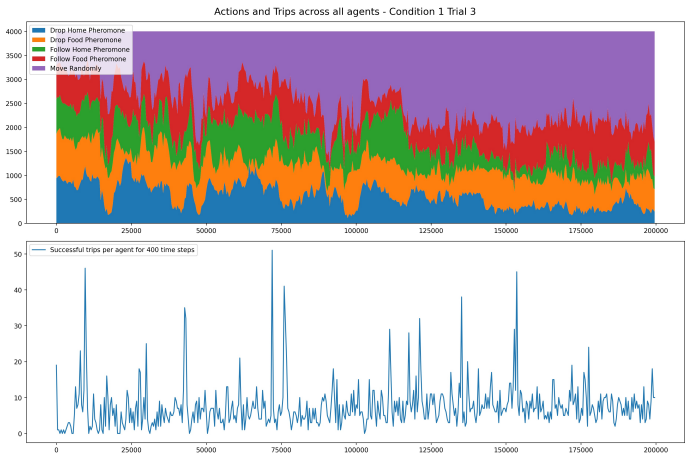

Figure 1

Fraction of each action across all agents: Figure 2 shows the distribution of actions summed over all agents. We observed the following pattern to hold for all

the experimental conditions:

Move Randomly > Follow Home Pheromone ≅ Follow Food Pheromone > Drop Home Pheromone ≅ Drop Food Pheromone.

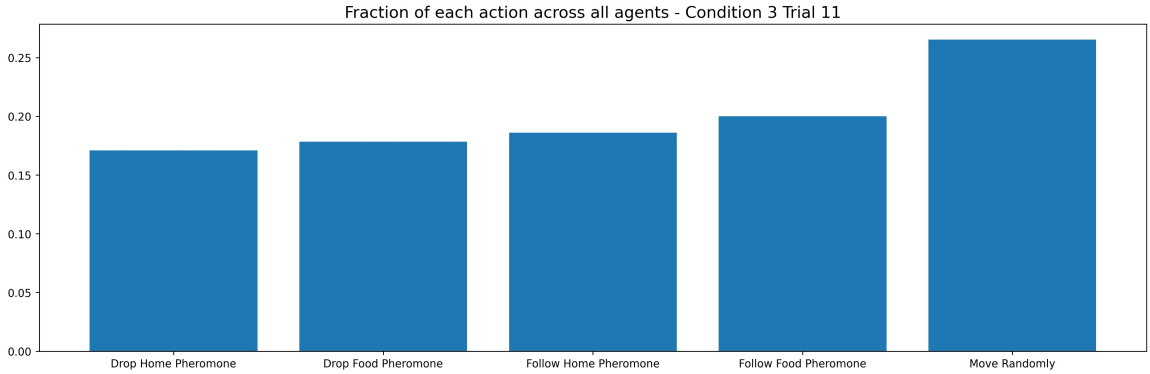

Figure 2

Histogram of successful trips: Figure 3 shows a typical distribution of successful trips. The distribution is skewed towards low values of trips counts with some outliers in higher trip counts. Because of this long tailed distribution, the median trip count is a better indicator of average performance than the mean values which are very sensitive to the presence of outliers.

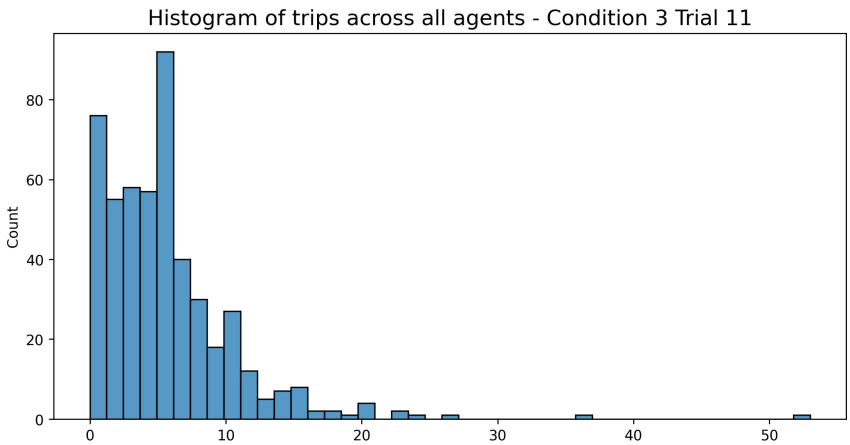

Figure 3

Number of trips across all agents: Figure 4 shows the number of trips successful trips complete by each agent.

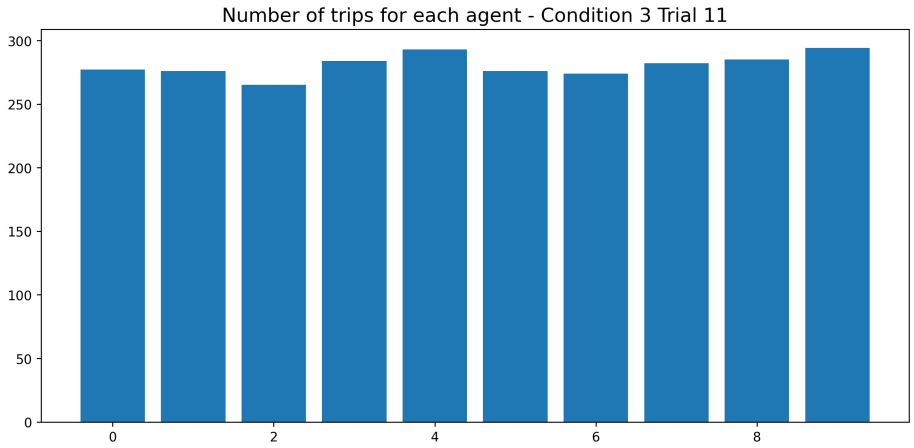

Figure 4

Number of states across all agents: In all the trials across different conditions, we found that the state (1,1) was encountered by the agents most as shown in figure 5. The primary reason for this is the dynamics of pheromones getting added via agents and its continuous evaporation and dispersion.

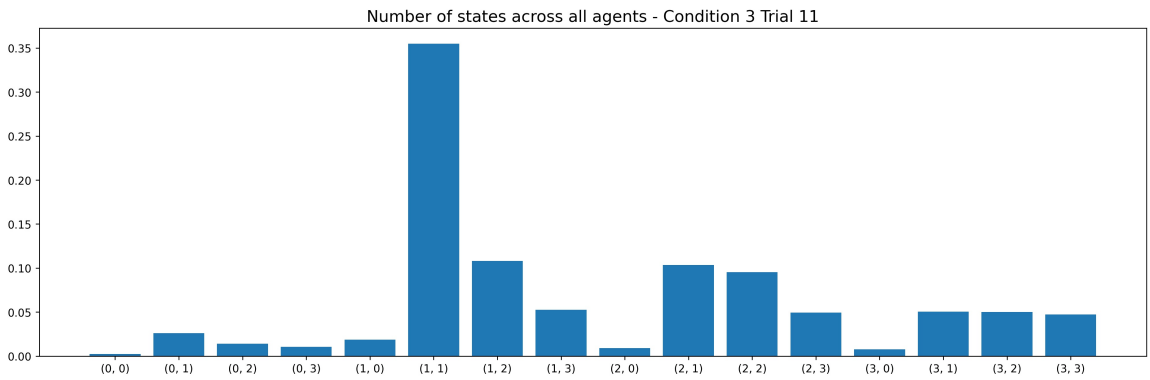

Figure 5

Pearson correlation test: Figure 6 shows the correlation matrix between the number of occurrences of each action and trips completed in the 400 step bins.

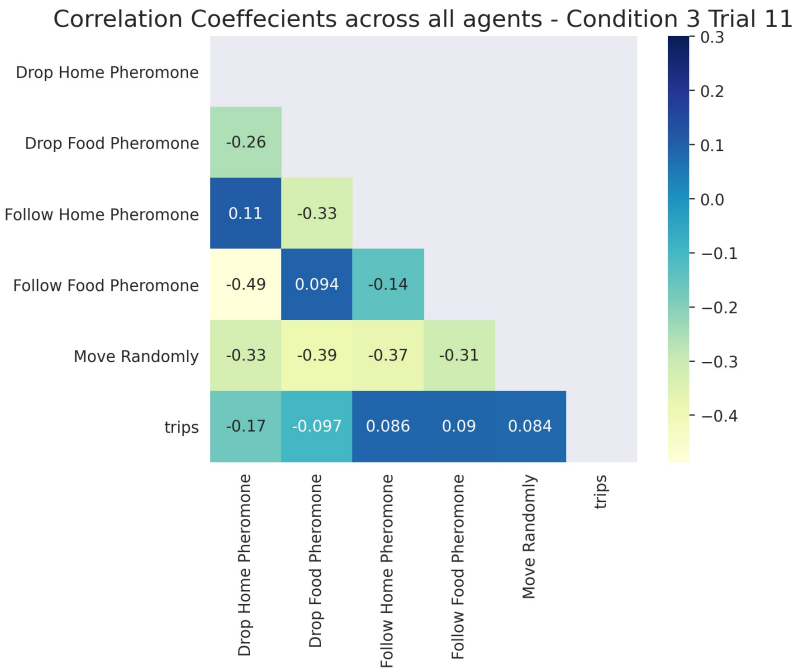

Figure 6

Action Transition Matrix: Figure 7 shows the transition matrix of counts of how many times an action occurred immediately after some other action. Random to Random transition was found to occur the most across all conditions.

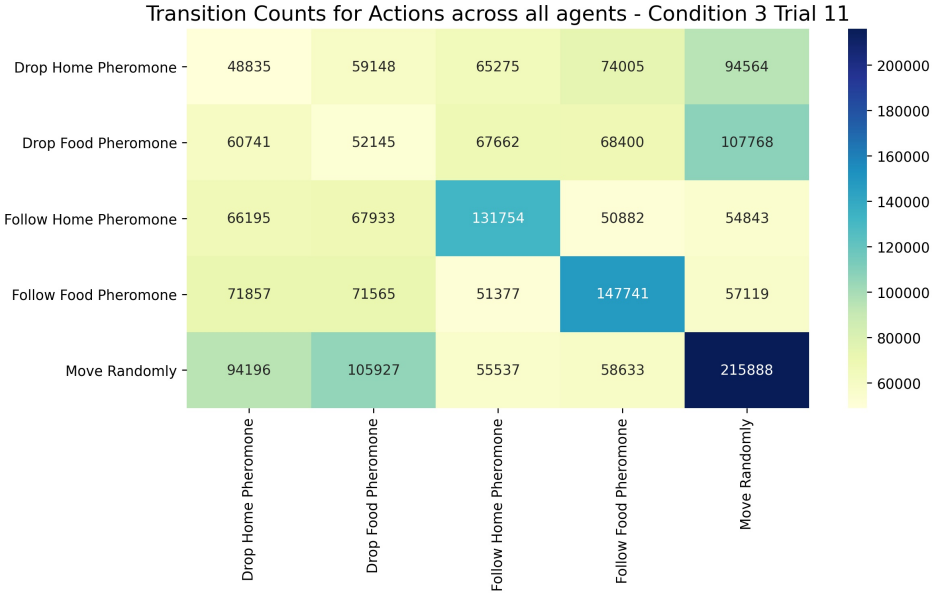

Figure 7

5.2 Analysis across the experimental conditions

Successful Trips: Figure 8 shows the performance of agents across all the experimental conditions. Mean Trips are much more susceptible to outlier values than Median Trips.

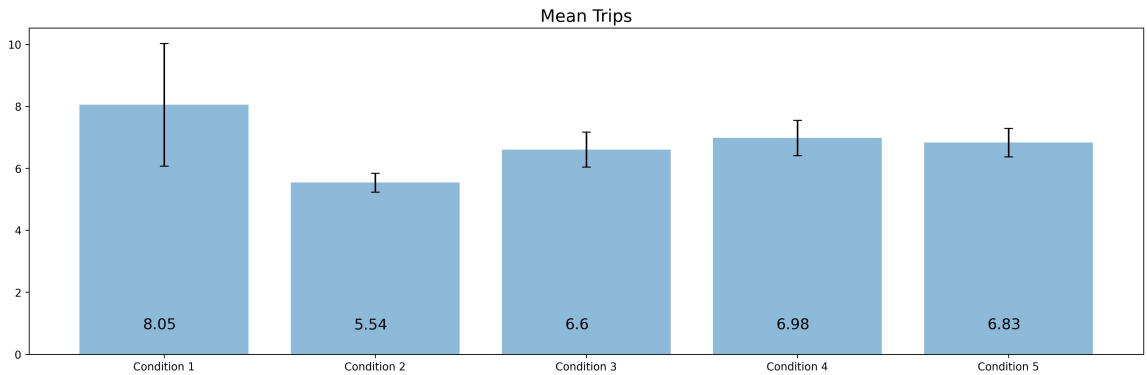

Figure 8

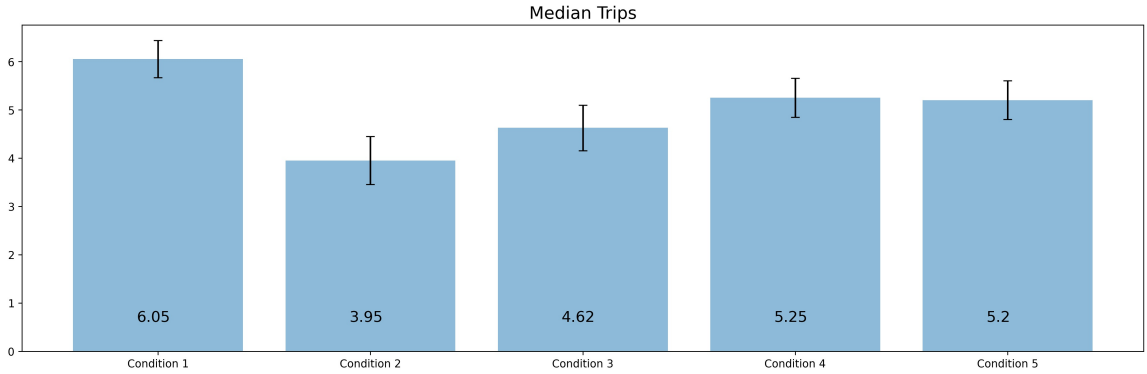

Figure 9

Skewness of trips: Skewness is a measure of asymmetry in the distribution. We used (Mean of all successful trips - Median of all successful trips) / Median of all successful trips as a measure to estimate it as shown in Figure 10.

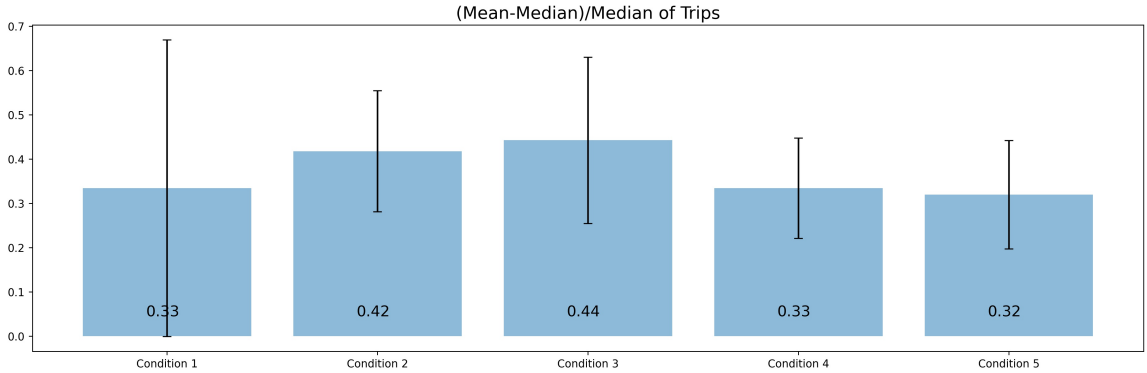

Figure 10

Distribution of Actions: Appendix-3 shows the distribution of actions for each condition. Condition 1,3,4 have similar distribution of actions with random action being significantly more than the other actions which have almost equal distribution. Across these conditions, the exploration rate for each agent reduces to zero with time. One hypothesis that could explain the presence of higher random action count relates to the fact that a large part of the world is not rewarding to explore. For cases where exploration in agents reduces to zero, the random action becomes favored because it provides the agents with the best chance of moving towards a more favorable part of the world.

Correlations: Appendix-4 shows the correlations between the binned count of various actions and trips across all trials of each condition. Some noteworthy obser-

vations are: there exists positive correlation between the occurrences of drop home pheromone and drop food pheromone actions.and between drop food pheromone and follow food pheromone actions across all conditions except condition 2. Condition 2 is the only condition where the exploration rate remains as high as 10% throughout the life of the agents. Another interesting observation is the presence of positive correlations between successful trips count and random action in Condition 1 and between successful trips and follow actions in other conditions.

6 Results

"As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality."

- Albert Einstein

Condition 1 is the baseline performed in Chandrasekharan, 2007. In Condition 1, we found that the random action is much higher than other actions and also the correlation between drop home to follow home (0.37) and drop food to follow food (0.33) is higher among all.

Condition 2 is modification of Condition 1, where the exploration rate is restricted to 10 percent always. Here we found equal distributions among all the actions with slight higher random action. By far Condition 2 is the worst in overall performance. There is positive correlation between follow (home and food) to trips (0.19) and trips to random (0.11).

Condition 3 is restricted with exploration rate in between 1 to 10 percent. Here we found random action being higher followed by and follow and drop actions. Even with less exploration rate compared to Condition 2 more random actions occur frequently. Also the correlation are: trips to follow home, follow food actions (0.16, 0.11) and follow food to drop food (0.12) and drop home to follow home (0.10).

Condition 4 is where after every 10000 steps the worst performing agent is killed and a new agent is being added to a home cell. Here we found the random action more and the correlation between trips to follow (0.06) and drop food to follow food (0.34) and drop home to follow home (0.29) which is higher compared to other actions. Overall performance is good.

Condition 5 is a modification of Condition 4 where the new agent added has the q-values of the best performing agent. Here we found the random action much higher compared to the other actions and the correlation between drop home to follow home (0.18) and drop food to follow food (0.21).

7 Conclusion

"The science of today is the technology of tomorrow."

- Edward Teller

Our findings suggest that contrary to expectations, random actions were higher instead of systematic pheromone dropping and following actions. This suggests that random actions help agents reach places in the world where dropping and following pheromones would be useful. Our analysis also indicates that the q-table (policy function) in q-learning may not be working as a decision-support system, but as a condensed form of experience in a stochastic world filled with other agents.

8 References

[1] Bonabeau E., Dorigo M. and Theraulaz G. (1999) Swarm intelligence: From natural to artificial systems. Santa Fe Institute studies in the sciences of complexity. New York: Oxford University Press.

[2] Chandrasekharan, S., & Stewart, T. C. (2007). The origin of epistemic structures and proto-representations. Adaptive Behavior, 15(3), 329-353.

[3] Hölldobler, B. & Wilson, E. O. (1990). The ants. Harvard University Press.

[4] Henry, J.D. (1977). The use of urine marking in the scavenging behaviour of the red fox (Vulpes vulpes). Behaviour, 62:82-105.

[5] Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. Cognitive Science, 18, 513-549.

[6] Kirsh, D. (1996). Adapting the environment instead of oneself. Adaptive Behavior, Vol 4, No. 3/4, 415-452.

[7] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[8] Stopka, P., & Macdonald, D. W. (2003). Way-marking behavior: An aid to spatial navigation in the wood mouse (Apodemus Sylvaticus). BMC Ecology, 3 (3).

[9] Watkins, C. (1989). Learning From Delayed Rewards, Doctoral dissertation, Department of Psychology, University of Cambridge, Cambridge, UK.

[10] Zahavi, A., & Zahavi, A. (1997). The Handicap Principle: A missing piece of Darwin's puzzle. Oxford: Oxford University Press.

A Appendix

A.1 Appendix-1

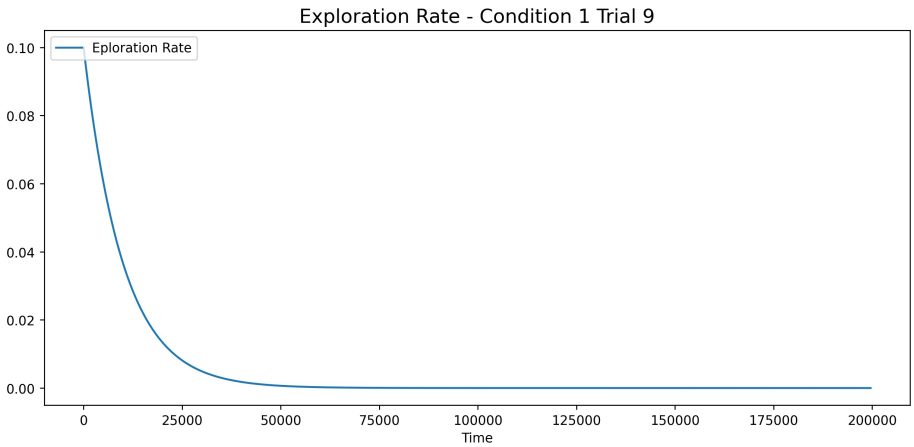

Figure 11

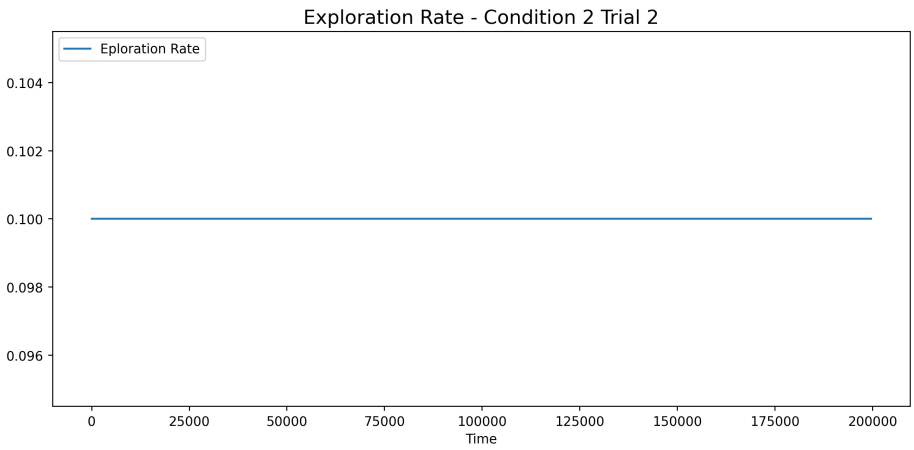

Figure 12

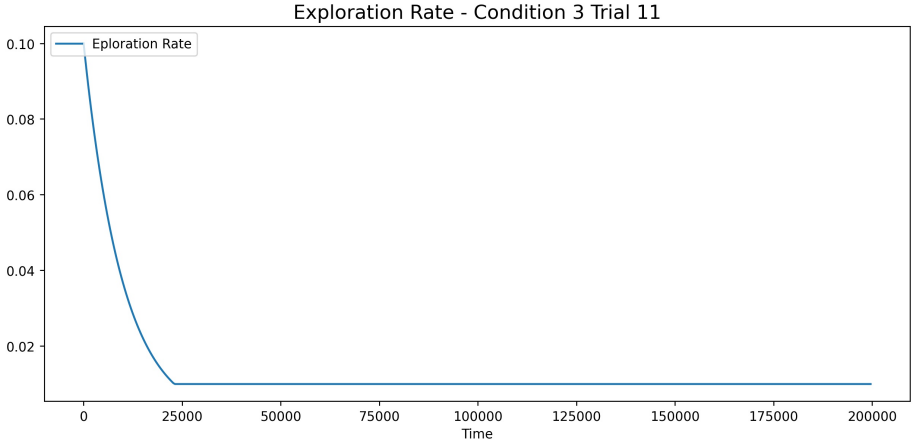
Exploration Rate - Condition 3 Trial 11
Eploration Rate
0.10
0.08
0.06
0.04
0.02
0
25000
50000
75000
100000
125000
150000
175000
200000
Time

Figure 13

A.2 Appendix-2

Figure 11, 12, 13, 14 represents the Total actions and trips of each condition.

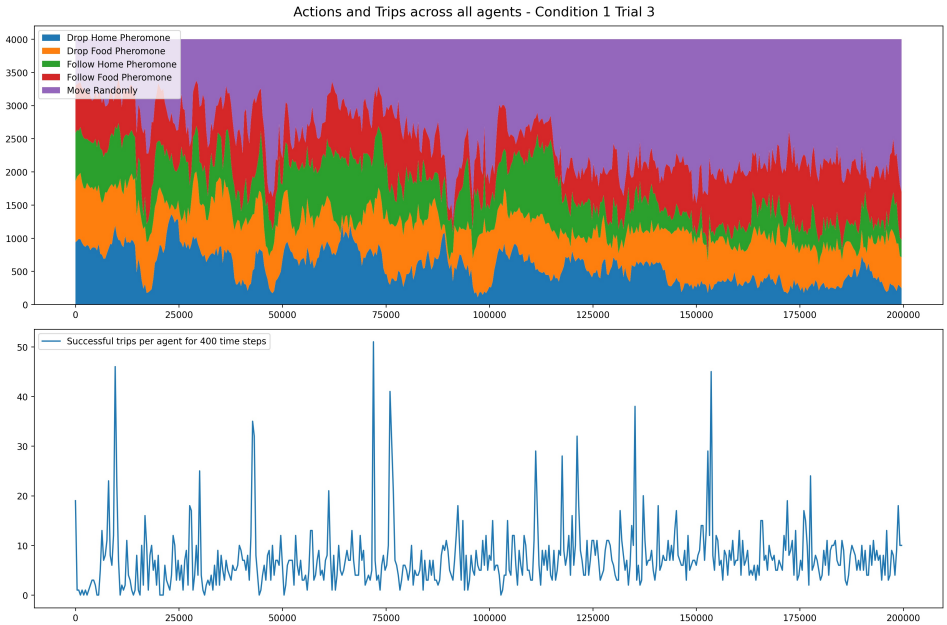

Figure 14

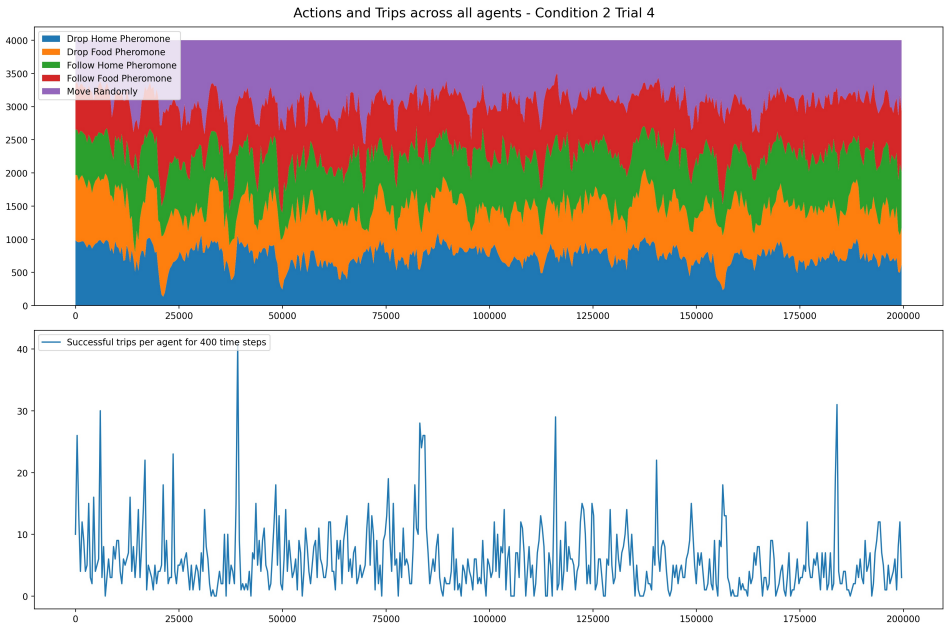

Figure 15

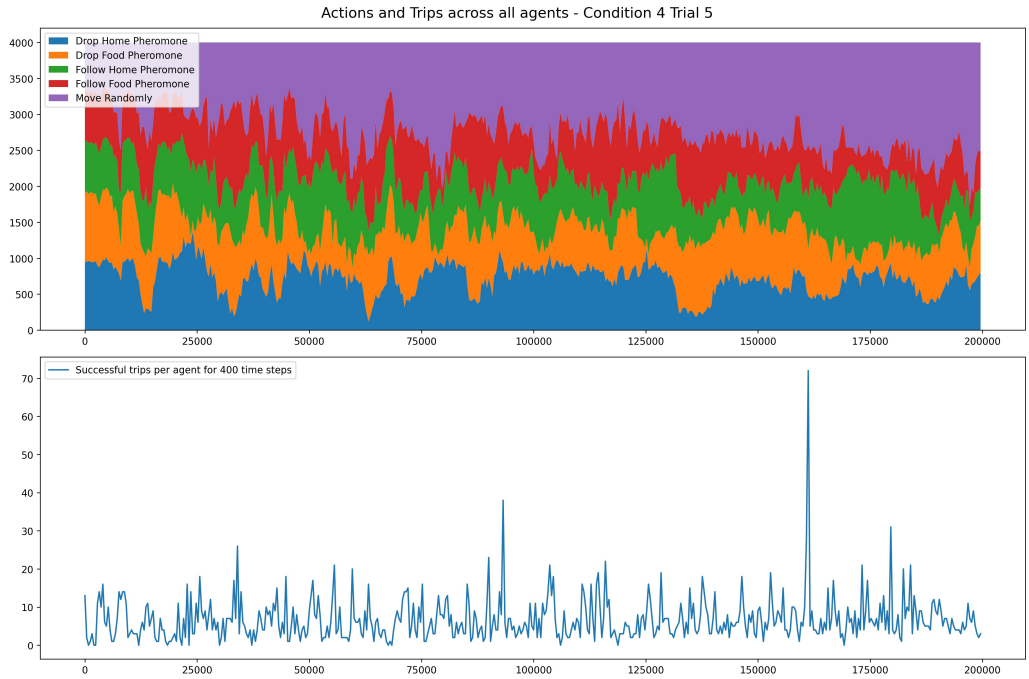

Figure 16

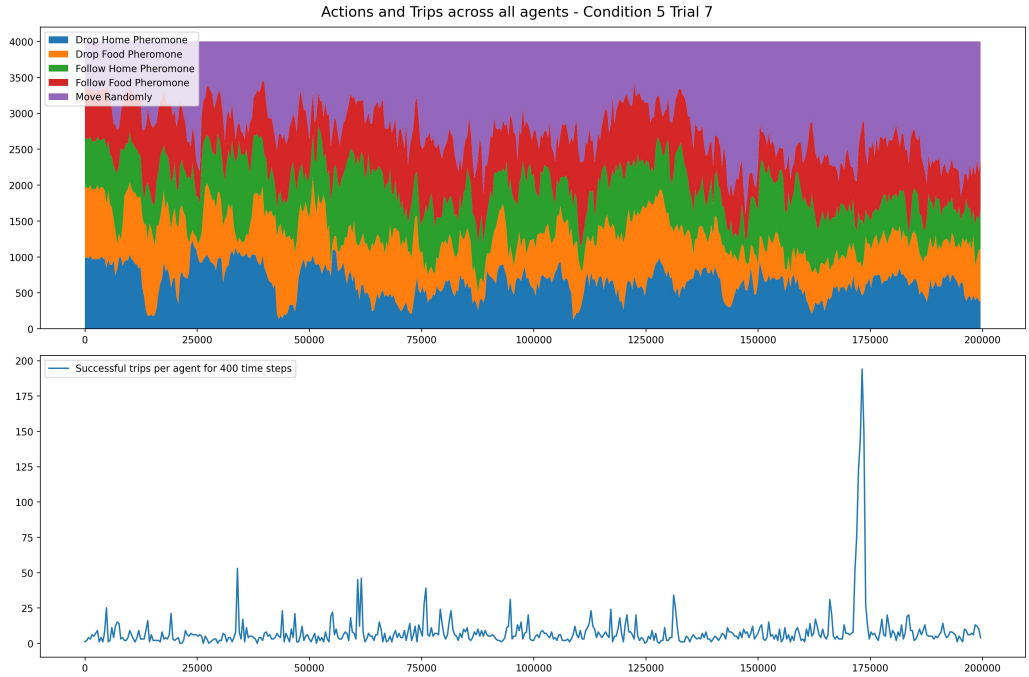

Figure 17

A.3 Appendix-3

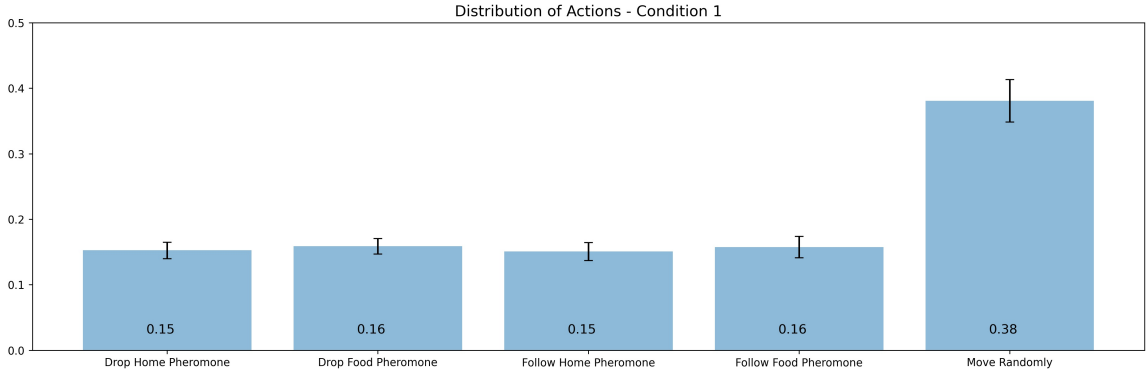

Figure 18

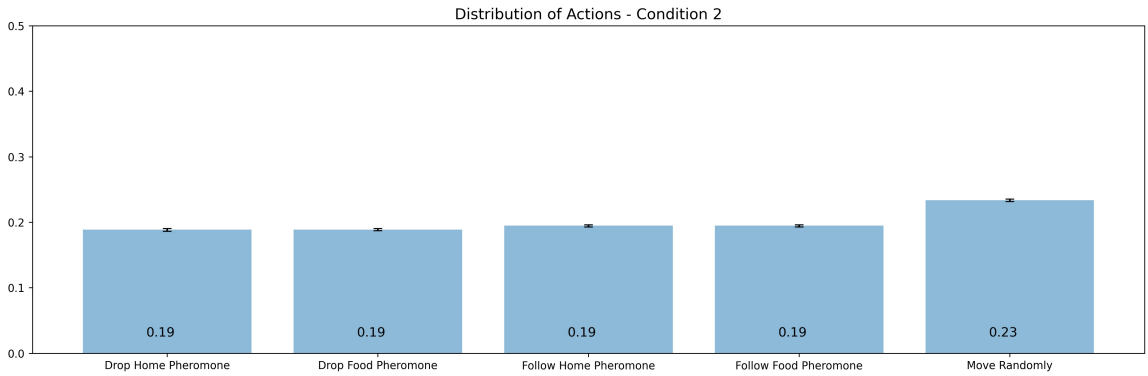

Figure 19

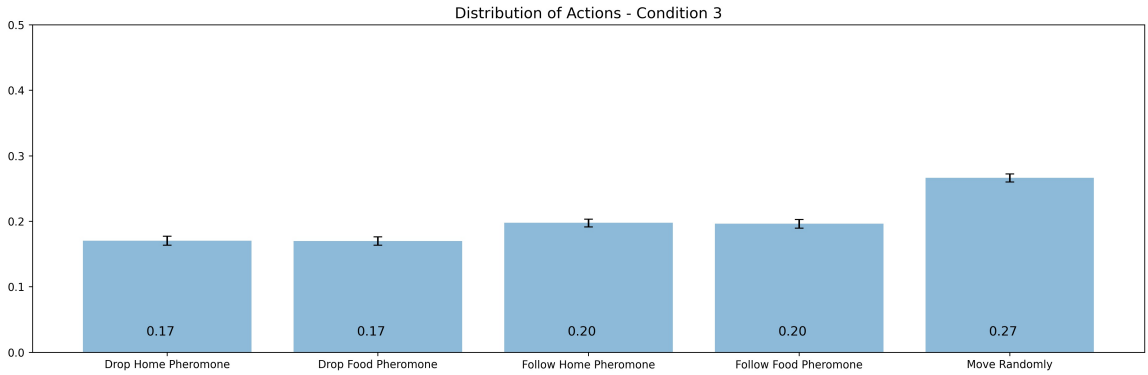

Figure 20

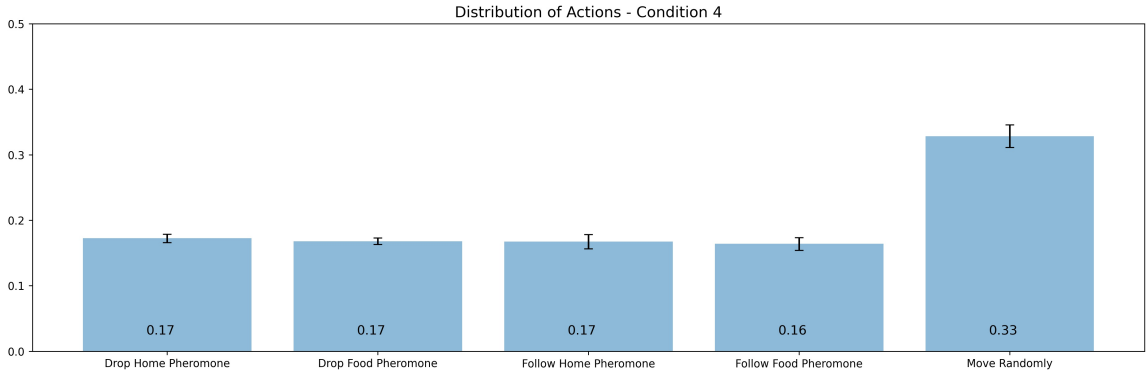

Figure 21

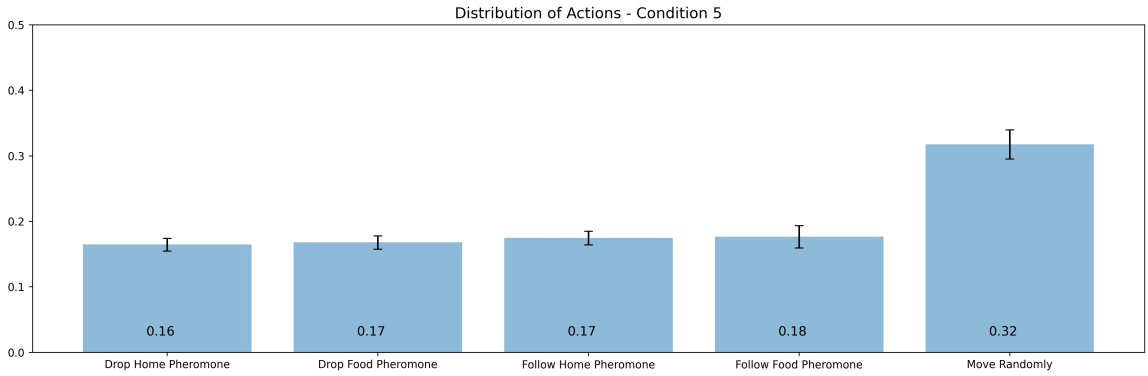

Figure 22

A.4 Appendix-4

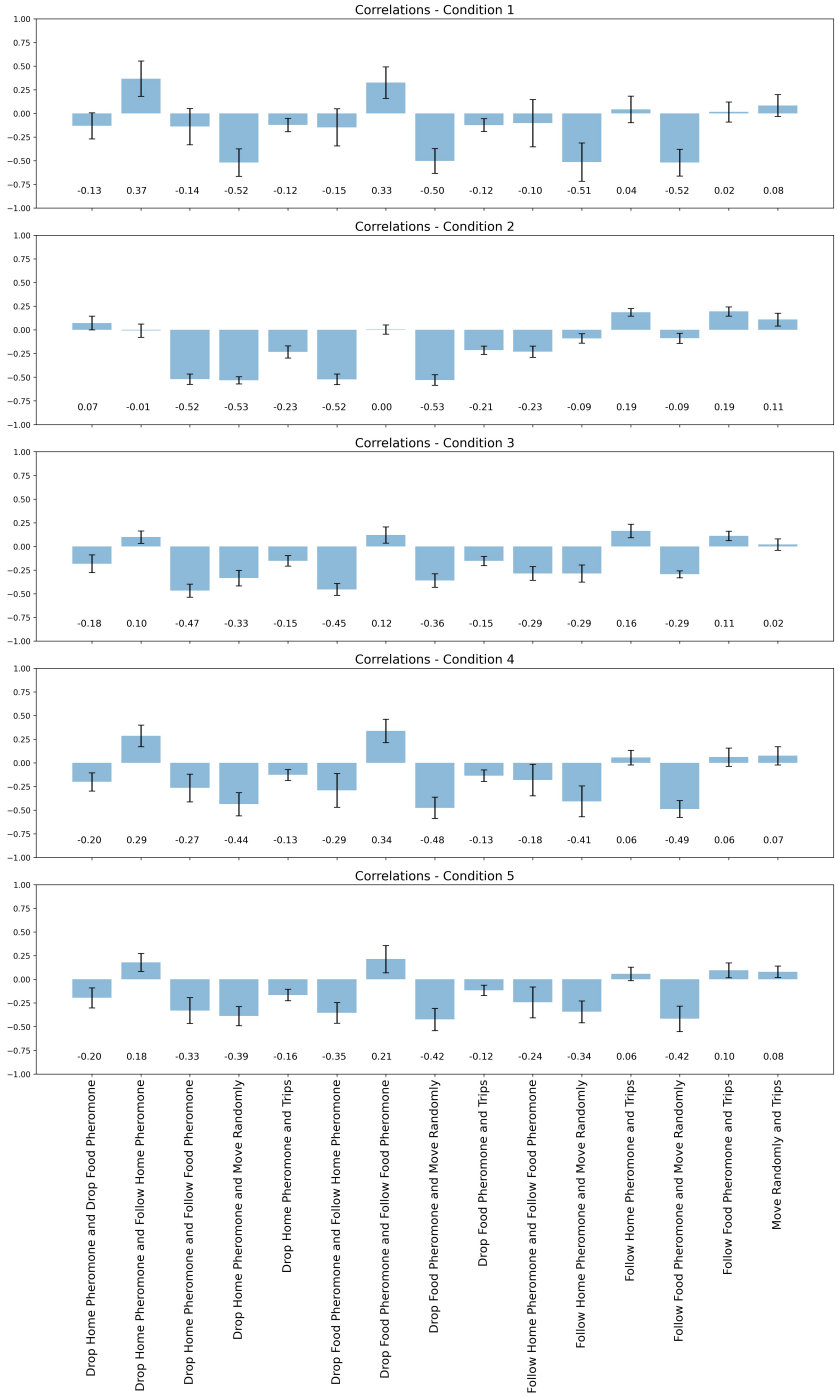

Figure 23

A.5 Appendix-5

State Transition Matrix: Figure 24 shows the transition matrix of counts of how many times an state occurred immediately after some other state. (1,1) to (1,1) transition was found to occur the most across all conditions.

Figure 24

State-Action Transition Matrix: Figure 25 shows the transition matrix of counts of how many times an action occurred in that certain state. (1,1) to Random transition was found to occur the most across all conditions.

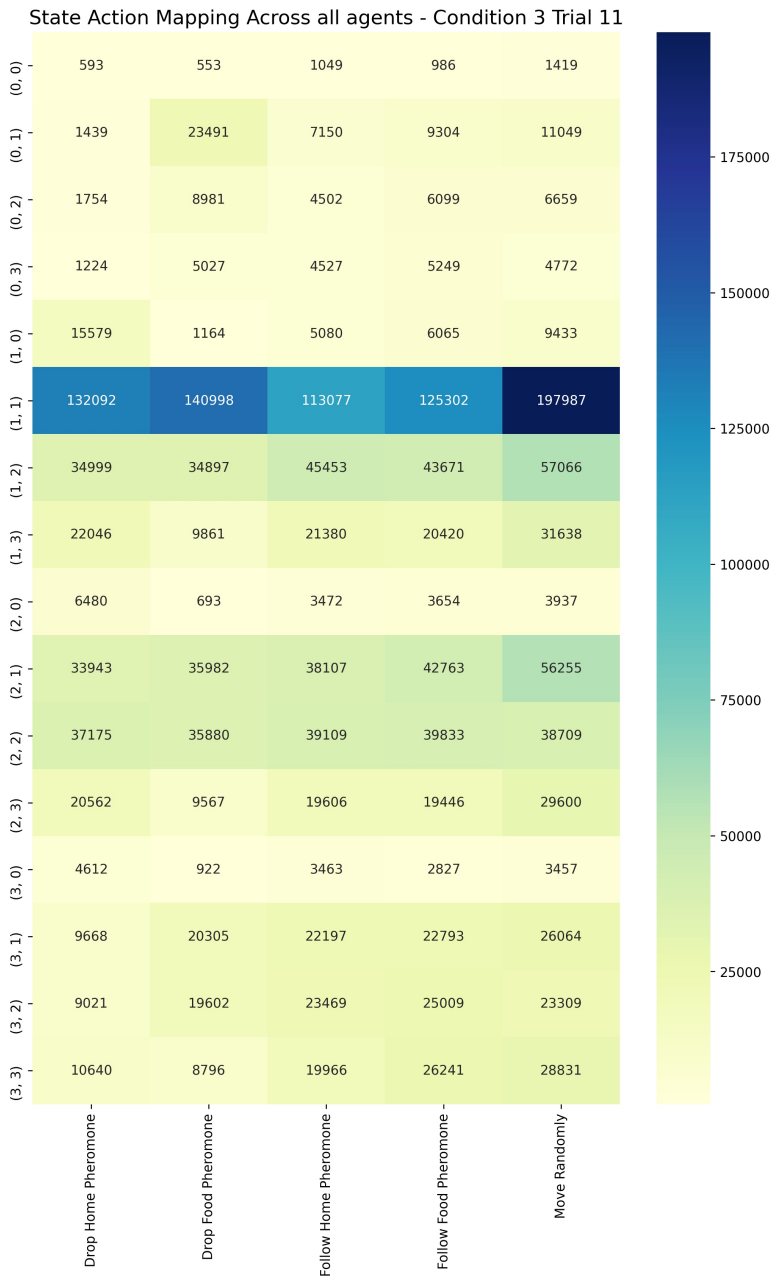

Figure 25