NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

BALANCING EXPLORATION AND EXPLOITATION IN AGENT LEARNING

by

Ozkan Ozcan

September 2011

Thesis Advisor:	Christian J. Darken
Second Reader:	Jonathan K. Alt

This thesis was done at the MOVES Institute
Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** September 2011	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE Balancing Exploration and Exploitation in Agent Learning	**5. FUNDING NUMBERS**
6. AUTHOR(S) Ozkan Ozcan	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number _____N.A_____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

Controlling the ratio of exploration and exploitation in agent learning in dynamic environments is a continuing challenge in applying agent-learning techniques. Methods to control this ratio in a manner that mimics human behavior are required for use in the representation of human behavior in simulations, where the goal is to constrain agent-learning mechanisms in a manner similar to that observed in human cognition.

The Cultural Geography (CG) model, under development in TRAC Monterey, is an agent-based social simulation. It simulates a wide variety of situations and scenarios so that a dynamic ratio between exploration and exploitation makes the decisions more sensible. As part of an attempt to improve the model, this thesis investigates enhancements to the exploration-exploitation balance by using different techniques. The work includes design of experiments with a range of factors in multiple environments and statistical analysis related to these experiments. As a main finding from this research, for small environments and for short runs techniques based on subjective utility give better results, while for long runs techniques based on time obtain higher utilities than other techniques. In more complex and bigger environments, a combined technique performed better in long runs.

14. SUBJECT TERMS Cultural Geography, Social Simulations, Reinforcement Learning, Agent-Based Modeling, Exploration and Exploitation.	**15. NUMBER OF PAGES** 81
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

BALANCING EXPLORATION AND EXPLOITATION IN AGENT LEARNING

Ozkan Ozcan
1st Lieutenant, Turkish Air Force
B.S., Turkish Air Force Academy, 2005

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN
MODELING, VIRTUAL ENVIRONMENTS, AND SIMULATION (MOVES)

from the

NAVAL POSTGRADUATE SCHOOL
September 2011

Author:	Ozkan Ozcan
Approved by:	Christian J. Darken Thesis Advisor
	Jonathan K. Alt Second Reader
	Mathias Kölsch Chair, MOVES Academic Committee
	Peter J. Denning Chair, Computer Science Academic Committee

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Controlling the ratio of exploration and exploitation in agent learning in dynamic environments is a continuing challenge in applying agent-learning techniques. Methods to control this ratio in a manner that mimics human behavior are required for use in the representation of human behavior in simulations, where the goal is to constrain agent-learning mechanisms in a manner similar to that observed in human cognition.

The Cultural Geography (CG) model, under development in TRAC Monterey, is an agent-based social simulation. It simulates a wide variety of situations and scenarios so that a dynamic ratio between exploration and exploitation makes the decisions more sensible. As part of an attempt to improve the model, this thesis investigates enhancements to the exploration-exploitation balance by using different techniques. The work includes design of experiments with a range of factors in multiple environments and statistical analysis related to these experiments. As a main finding from this research, for small environments and for short runs techniques based on subjective utility give better results, while for long runs techniques based on time obtain higher utilities than other techniques. In more complex and bigger environments, a combined technique performed better in long runs.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

CG	Cultural Geography
FOP	Frequency of Optimal Pulls
GW	Grid World
IW	Irregular Warfare
MAS	Multi Agent Simulation
RPT	Regret Per Trial
TRAC	TRADOC Analysis Center
TRADOC	Training and Doctrine Command

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

First, I want to thank to my country, Turkey, and the Turkish Air Force for giving me this great opportunity of being a Master's student at Naval Postgraduate School (NPS). It was a great honor for me to study at NPS.

During the progress of this study, I had lots of support from many people. Without their support, I would not be able to accomplish this study. First, I want to thank Dr. Chris Darken, my thesis advisor. His knowledge formed the skeleton of this study. Lieutenant Colonel Jon Alt was the one who changed this skeleton into life. I was encouraged to finish this work by his enthusiasm. TRAC Monterey provided enormous support to me, especially Major Francisco Baez and Mr. Harold Yamauchi. They both assisted greatly in the construction of the model scenarios for Cultural Geography model.

Besides educational support, I needed moral support to reinforce my enthusiasm to finish such a study. My wife, Selda, provided this support. I always feel lucky to be married to such a great person. I owe all my success to my wife and my daughter.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Balancing the ratio of exploration and exploitation is an important problem in reinforcement learning. The agent must develop a strategy for successfully interacting with its environment. The agent can choose to explore its environment and try new actions in search of better ones to be adopted in the future, or exploit already tested actions and further reinforce successful ones. A strategy of pure exploration or pure exploitation will not typically yield best results (Sutton & Barto, 1998).

In agent learning, agents must find a balance between exploration and exploitation to obtain the best value. This ratio can be related to the estimated life span of the agent. To increase performance in the short term, an agent's action selection can become greedier, while to increase performance in the long term its action selection policy can become more exploratory. The development of techniques to adjust dynamically the ratio between exploration and exploitation has the potential to expand the application of reinforcement learning as an action selection mechanism.

A. GOALS OF STUDY

In this thesis, three new approaches are examined to investigate the exploration and exploitation problem in reinforcement learning from different perspectives. In each approach, the aim is to make an agent's action selection dynamic to produce more sensible results. Focusing on the correlation between the simulation's scenario lengths, sizes of environments, and expected utility of agent produces a relationship that helps the user design the simulation based on his needs. Experiments are conducted to answer the research questions identified below.

B. RESEARCH QUESTIONS

The two questions this research will address are:

- What techniques can be used to control the level of exploration and exploitation within cognitive agents?
- How do changes in individual agent behaviors affect the macro-level results of a test-bed simulation scenario within the Cultural Geography (CG) model?

C. BENEFITS OF STUDY

This research demonstrates the application of three novel approaches to controlling the level of exploration and exploitation in reinforcement learning. These approaches support TRAC Monterey in the development of an appropriate methodology in the CG model for developing more realistic behaviors.

In the CG model, agents select their actions from a small set of predefined action choices. At this point, the level of exploration and exploitation is defined manually to make the agent greedy or explorative (Alt et al., 2009). Although this implementation is working, because the ratio is static it is not known if the given ratio best fits the model. In the CG model, there are different situations and scenarios so that a dynamic ratio makes the decisions more realistic. “Understanding the behavioral response of the civilian population in irregular warfare operations presents a major challenge area to the joint modeling and simulation community where there is a clear need for the development of models, methods, and tools to address civilian behavior response" (Alt et al., 2009). As part of an attempt to improve the model, this research seeks to enhance the exploration and exploitation balance by using different techniques. The research examines and compares techniques to address the research questions stated above and to define the effects of different techniques in the CG model. Finally, the work examines ways to make the ratio dynamic in different conditions in simulation and analyzes identified techniques.

The scope of this study is not limited to the CG model. The focus of this thesis is to treat one of the main reinforcement learning problems by using different techniques and to generalize it into other situations. Techniques were evaluated in three different environments: the two armed bandit example, the grid world example and the Cultural Geography model. Future work will apply these algorithms to more complex environments, with the intended application representing human behavior within modeling and simulation.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND

A. CULTURAL GEOGRAPHY MODEL

In this era, operations in stability, security, transition, and reconstruction (SSTR) require a cultural understanding of the population in which they are conducted. Knowing military capabilities of an enemy is not enough anymore. Eliminating the enemy in irregular warfare is based on understanding the civilian population and cultural terrain. In the modeling and simulation community, cognitive social simulations identify this complex problem as human behavior modeling.

Cognitive social simulation provides a framework for complex dynamic social systems. Cognitive social simulations combine the use of cognitive architectures with agent-based simulation methods. They are distinguished by the level of sophistication of the agent's cognitive processes and the existence of a formal structure for the agent's cognitive processes that is based on what is known regarding human information processing. The use of agent-based models and cognitive social simulations in the examination of issues related to Irregular Warfare, such as resource allocation, infrastructure improvement, and the impact of information operations, have not been extensively explored (Alt et al., 2009).

One example of cognitive social simulation is the Cultural Geography model, under development by the U.S. Army Training and Doctrine Command (TRADOC) Analysis Center, Monterey (TRAC-Monterey). The Cultural Geography (CG) model is a government-owned, open-source agent-based model designed to address the behavioral response of civilian populations in conflict environments. The CG model is a multi-agent simulation (MAS) designed to represent social behaviors of a population. Agents within the CG model select their action according to a constant number, which defines the ratio between exploration and exploitation. To enhance the functionality of agent action selection and to obtain more realistic results with better utilities, this constant is changed to a dynamic parameter that depends on time and utility employing different techniques.

B. REINFORCEMENT LEARNING

Reinforcement learning provides a mechanism to facilitate agent action selection across multiple domains. The basic elements of reinforcement learning are a policy, a reward function, a value function, and an optional model of the environment. These elements allow the agent to map situations to actions based on feedback from the environment. Model-free methods, such as Q-learning, provide robust methods across environments. In this research, time based, utility based, and a combination of both reinforcement learning methods are used (Russell & Norvig, 2003).

The method used relies on the calculation of a point utility value for each percept received. Because a single action will generally affect more than one point utility value, it is important to aggregate utility in order to capture the effects of an action on point utility values received over time. The traditional aggregation method is to form the exponential moving average of the point utility values. Let p_i be the percept sequence and t_i be the sequence of times at which the percepts arrive. Let s_i be the corresponding sequence of states. Then the corresponding sequence of point utility values is $u_i=u(s_i,p_i)$. Given the choice of exponential base λ, where $0<\lambda<1$, the exponential moving average of the sequence starting at time t is

$$\bar{q}(t) = \sum_i \lambda^{t_i - t} u_i \Theta(t_i - t) \qquad (0.1)$$

where Θ is the unit step function, which is zero when its argument is negative and one otherwise to prevent the calculation of negative utility.

Expected future aggregate utility of that action in a particular situation must be an important factor in any decision to select it. The average of the aggregate utility received when the action was selected in the past is used to estimate the expected future aggregate utility of an action. Let a_k be the action selected in situation σ_k at time t_k. Then the aggregate utility actually received after this action is given by $\overline{\mathbf{q}}(\mathbf{t}_k)$. Let $t(\sigma,a)$ be the set of all times at which action a was taken in situation σ . Then the estimator of the expected future aggregate utility of action a in situation σ will be,

$$\widehat{Q}(\sigma,a) = \sum_{t \in t(a)} \frac{q(t)}{|\, t(\sigma,a)\, |} \tag{0.2}$$

where $|\, t(\sigma,a)\, |$ is the number of elements in the set $t(\sigma,a)$. $\widehat{Q}$ is an estimator of the Q function typically defined in reinforcement learning in the special case that the set of situations is identical to the set of individual states (Papadopoulos, 2010).

C. TECHNIQUES FOR DETERMINING EXPLORATION AND EXPLOITATION RATIO

In this study, three methods for controlling the balance between exploration and exploitation are examined. The proposed techniques are demonstrated in conjunction with the Boltzmann distribution, but could easily apply to other Softmax techniques, such as the epsilon greedy method (Sutton & Barto, 1998). The Boltzmann technique assigns probabilities to all actions corresponding to their expected values, based on the value of the temperature parameter, Ʈ. A high Ʈ leads to exploratory behavior, while a low Ʈ leads to greedy (exploitative) behavior. If the probability is higher, it means its expected value is higher, and it is most likely to be taken. The probability is measured by following formula:

$$P_i = \frac{e^{\frac{U_i}{\tau}}}{\sum_j e^{\frac{U_j}{\tau}}} \tag{0.3}$$

The exploration-exploitation problem in using this function becomes one of dynamically setting the temperature parameter in a manner that allows the agent to learn something about the environment, while eventually taking advantage of this information.

1. Time Based Search then Converge

Inspired by the search-then-converge class of algorithms, this method requires that the modeler have knowledge regarding the environment in order to specify the half-life of the temperature parameter. The general form of the algorithm is shown below:

$$\tau_{new} = \frac{\tau_{Initial}}{1 + \frac{t}{t_{Exploit}}}, \tag{0.4}$$

where t is the current simulation time and $t_{Exploit}$ is the specified transition point from exploration to exploitation, equivalent to the half-life of the initial temperature. Figure 1 illustrates the shape of the decay function, with initial temperature of 1.0 and $t_{Exploit}$ values of 30 and 60.

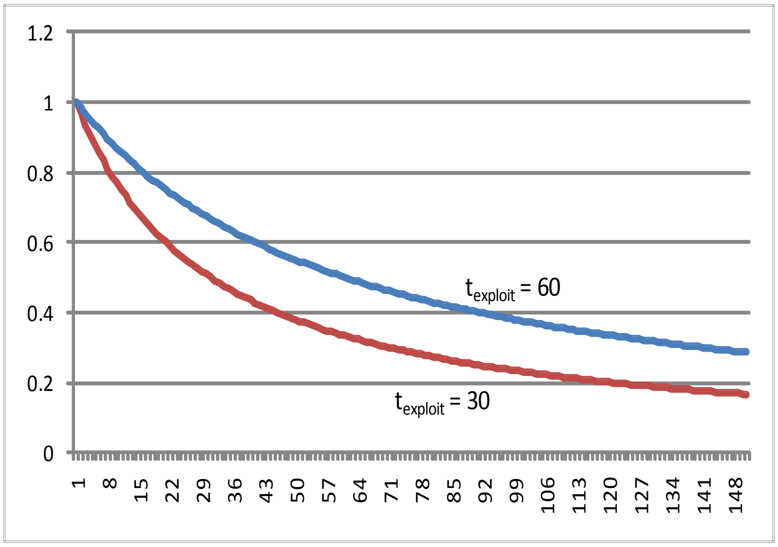


Figure 1. Decay function of temperature versus $t_{Exploit}$

The shape of the curve is controlled by the single parameter, but this still locks in the agent's behavior once the agent passes into the region that begins exploitation. A second approach based on aggregate utility is discussed below.

2. Aggregate Utility Driven Exploration

The second method also requires the user to know something about the environment in which the agent will operate. In this case rather than an arbitrary transition time from exploratory to greedy behavior, the user is required to know something about the reward structure of the environment. Keeping the same general form as the time based algorithm, this approach requires a user-specified acceptable utility. The general form of the algorithm is shown below:

$$\tau_{new} = \frac{\tau_{Initial}}{1 + \dfrac{u(t)_{aggregate}}{u_{acceptable}}} \qquad (0.5)$$

where the user divides the aggregate utility at simulation time t by the specified acceptable level of utility (Alt, J., personal communication, August 25, 2011). In dynamic environments where the aggregate utility varies over time, this algorithm results in greedy behavior when an acceptable level of behavior is reached, but provides the agent the opportunity to shift back into exploratory mode should the aggregate utility drop below the threshold, due to discounting or other effects from the environment. Sample behavior is shown at Figure 2 for notional aggregate utilities, with initial temperature of 1.0 and acceptable utility set at 3.0.

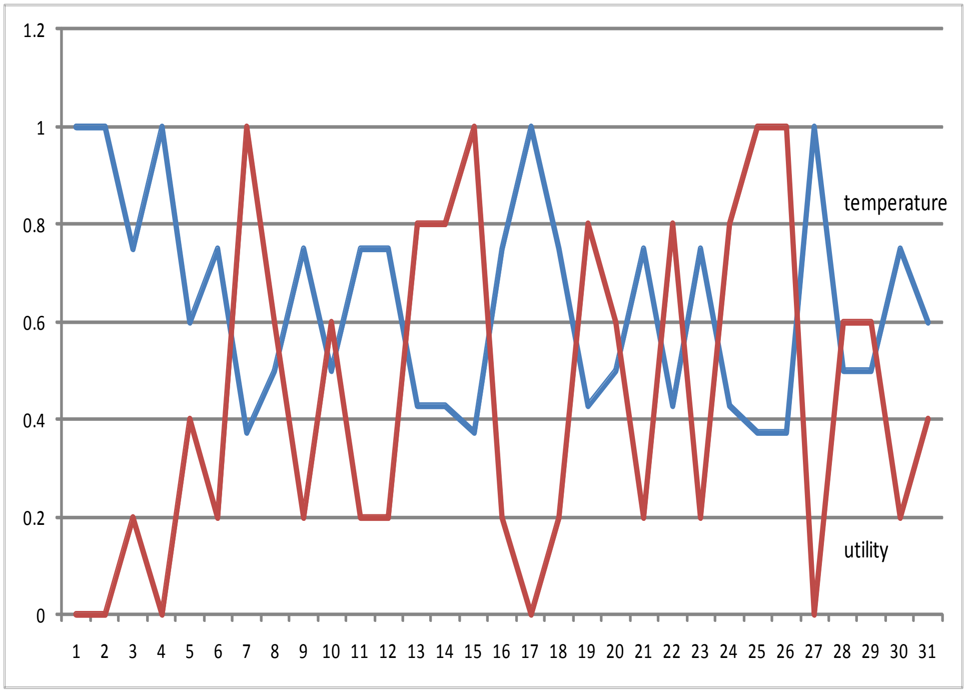


Figure 2. Sample agent behavior with utility and temperature over time

3. Combination of Time and Utility Based Technique with Happiness Function

Both techniques described above adjust the temperature based on different parameters. While the Time Based Search then Converge Technique focuses on the difference in the simulation time, the Aggregate Utility Technique prefers to focus on the utility of the agent. To obtain more sensible results, a new technique is needed. This technique is used for the first time in this thesis for environments mentioned in Chapter III. While calculating the temperature for new technique, a new function is added in order to adjust agent’s behavior based on previously taken rewards. In this situation, the new formula is defined by combination of three different formulas. Each formula is given some weighted effect on the total. After defining optimum $u_{acceptable}$ and $t_{Exploit}$, effective percentages can be examined.

$$\tau_{new} = W_1 * \tau_{TimeBased} + W_2 * \tau_{UtilityBased} + W_3 * \tau_{HappinessBased} \quad (0.6)$$

In the above formula $\sum_i^n w_i = 1$. Happiness based temperature is calculated as:

$$\tau_{HappinessBased} = \frac{1}{e^{(ShortTermHappiness - LongTermHappiness)}} \quad (0.7)$$

In the $\tau_{HappinessBased}$ formula, Short-term Happiness equals the point utility of the agent. To calculate Long-Term Memory, previous utilities of the agent are normalized and then utilities are weighted based on a new lambda discount factor. The capacity of an agent’s memory is defined by a parameter called T. Based on this parameter, an agent will check “T” number of previous utilities and based on their occurring order they will be weighted.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

This research examines three approaches for dynamically controlling the ratio between exploration and exploitation. All techniques are illustrated in conjunction with the use of a Boltzmann action selection policy (Sutton & Barto, 1998).

The first approach is based on time and the second one is based on aggregate utility. The last approach will be combination of both techniques and another function called the “happiness” function. Algorithm performance is explored in a 2 armed bandit example, a grid world example, and the CG model.

Based on each method, we design experiments, varying certain parameters that affect temperature parameter in reinforcement learning. After experiments are done, each parameter and methods that affect reinforcement learning are analyzed, all three techniques are compared with each other, and findings with respect to the research questions stated above are presented.

A. IMPLEMENTATION OF TECHNIQUES

All the techniques are implemented with Python code. For different scenarios, they are tested and the results are shown in Chapter IV, Analysis. The Python code for each technique is given Appendix A.

B. TEST BED ENVIRONMENTS

1. Two-Armed Bandit

Bandit problems have been used in agent learning to determine the balance between exploration and exploitation. Exploration of the search space is important to figure out the regions of the environment and exploitation is necessary to put the knowledge gained from exploration to use (Macready & Wolpert, 1998). In 2 armed bandit experiment, two arms have different jackpot probabilities. In this case, they are 0.1 and 0.9. However, the agent does not know which the good (0.9) arm is. An agent receives a utility of 1.0 at every jackpot. The aim of this experiment is to figure out the relationship between the best-expected utilities of agent and different scenario lengths for each technique.

2. Grid World

In the grid world environment, agents are allowed to explore an NxN grid in which the agent can occupy any one of the intersections. In every round, the agent can move up, down, left or right. At some of the intersections are rewards of unknown and varying values. The agents must find a way to maximize their rewards by devising a strategy of exploring for better rewards and exploiting the best solution currently known.

3. Cultural Geography Model

The CG model is a multi-agent simulation representing social behaviors of a population (Alt et al., 2009). Agents within the CG model select their actions according to a constant number, which is explained in Chapter II. To enhance the functionality of agents in selecting their actions and to obtain more sensible results with better utilities, this constant is changed to a dynamic parameter that depends on time and utility, applying different techniques.

C. GENERAL CONSTRAINTS OF EACH TECHNIQUE

1. Time Based Search then Converge Technique

This technique decreases temperature over simulation time. Its rate of decrease is defined by $t_{Exploit}$. If $t_{Exploit}$ is small, the temperature immediately converges and decrement steps of temperature are large. When $t_{Exploit}$ is high, decrement steps of temperature are small and temperature changes slowly. Because of that, finding a threshold $t_{Exploit}$ for each scenario length becomes the main problem for this technique. Selecting $t_{Exploit}$ that is too low makes agent explorative so that the agent cannot obtain good results based on prior successes. On the other hand, selecting $t_{Exploit}$ that is too high does not give enough time for the agent to obtain the benefits from exploration.

2. Aggregate Utility Driven Exploration

This technique decreases the temperature based on the current utility of the agent. In this case rather than an arbitrary transition time from exploratory to greedy behavior, the user is required to know something about the reward structure of the environment. Agent stays in the same temperature until it finds reward. Therefore, the agent initially does not change its temperature and goes on explorative behavior until it finds reward. There are two constrains for these technique: for short scenario lengths, the agent does not have enough time to change its behavior to exploit obtained knowledge from the environment, for long scenario lengths waiting for reward to change behavior more exploitive is costing more simulation time.

3. Combination of Time and Utility Based Technique with Happiness Function

This technique is formulated to eliminate the weakness of first two techniques and adjust the behavior of the agent with a function based on the exponential of the difference between short-term and long-term memory of the agent. At the beginning of the simulation, the agent becomes explorative and the agent adjusts its behavior in three ways, through simulation time, reward, and the effect of the reward on agent happiness, which is described in Chapter II formula 0.7.

D. EXPERIMENTS

1. Time Based Search then Converge

Parameter of interest to the modeler is the $t_{Exploit}$ parameter, functionally equivalent to the half-life of the initial temperature. In order to better understand the relationship between $t_{Exploit}$ and scenario length an experiment to systematically vary these two parameters was constructed and executed for the grid world environment with 1000 repetitions per parameter set (design point) and for the 2 armed bandit with 10000 repetitions per parameter set (design point). Twenty-six replications of three different scenario lengths are used to explore algorithm performance in CG model. For this technique, exploit start time is examined at every 10 steps starting from 1 to 750.

2. Aggregate Utility Driven Exploration

For this technique, the parameter of interest to the modeler is $u_{acceptable}$, the threshold that the agent's aggregate utility must achieve prior to the agent's behavior becoming greedy. An experiment similar to Time Based Technique described above was conducted with this algorithm in the grid world environment with 1000 repetitions and in the 2 armed bandit environment with 10000 repetitions of the experiment design examining scenario length and $u_{acceptable}$. For this technique, 26 replications of three different scenario lengths were used to explore algorithm performance in the CG model and the exploit utility value was changed at every 0.1 points starting from 0.1 to 4.0.

3. Combination of Time and Utility Based Technique with Happiness Function

For this technique, temperature is calculated based on three parameters for each action. The first parameter is the temperature obtained from the first technique for that specific action, and the second parameter is the temperature obtained from second technique. The third parameter is used to adjust temperature based on the "happiness" of the agent. This technique was tested with an experiment conducted in all three environments. For 2 armed bandit and grid world examples, 2, 10 and 50 environment sizes are used and 1000 replication is done. For all techniques, $u_{acceptable}$'s range is [0.1, 0.5, 1, 1.5, 2, and 2.5]; $t_{Exploit}$'s range is [1, 100, 200, 300, and 400]. Long-term capacity (T) is limited to 50 and discount factor (lam2) is taken 0.5. For CG model, experiment is repeated for 30 replications for each three different scenario sizes [180,360,540] with 20 agents.

IV. ANALYSIS

A. STATISTICAL ANALYSIS

In order to evaluate the performance of algorithms in each environment, three performance measures are used. These measures are regret per trial (rpt), frequency of optimum pulls (FOP), and total utility. The difference between the best possible expected reward of chosen action and expected reward playing the optimal arm is defined as total expected regret and formulated as:

$$R_T = T\mu^* - \sum_{t=1}^{T} \mu(t) \tag{0.8}$$

$\mu^* = \max_{i \in k} \mu$ is the expected return of the best action and μ is the expected reward from playing the optimal arm. In the formula, T shows trials for that playing action. Regret per trials will show how far the algorithm's performance will be from the optimal as the number of trials increase. For the grid world environment, the best aggregate reward will be achieved when the agent selects shortest path, $l = \left|x_g \text{ - } x_s\right| + \left|y_g \text{ - } y_s\right|$ between its initial position and goal position. Total regret will be calculated as:

$$R_T = \frac{T}{l} - \sum_{t=1}^{T} r_t \tag{0.9}$$

The frequency of optimal moves have been evaluated with the decrement in the distance between agent's current position and goal, $l > l_{t+1}$.This means it will count the moves that takes the agent closer the goal.

1. Time Based Search then Converge

Figure 3 shows the expected utility of the agent over t_{Exploit} for a variety of scenario lengths and various exploitation start times for the grid world example and Figure 6 shows for the 2 armed bandit example. t_{Exploit} varies by simulation length, with a point of diminishing returns and degradation in performance associated with mismatches between scenario length and $t_{Exploit}$. Figure 4 and Figure 7 show the expected utility as a function of scenario length and $t_{Exploit}$ as a contour plot for the grid world and 2 armed bandit experiments, respectively. In the grid world example, a $t_{Exploit}$ of ~100 time units suffices to provide good performance for most scenario lengths. Figure 8 shows results for CG Model with three different scenario lengths.

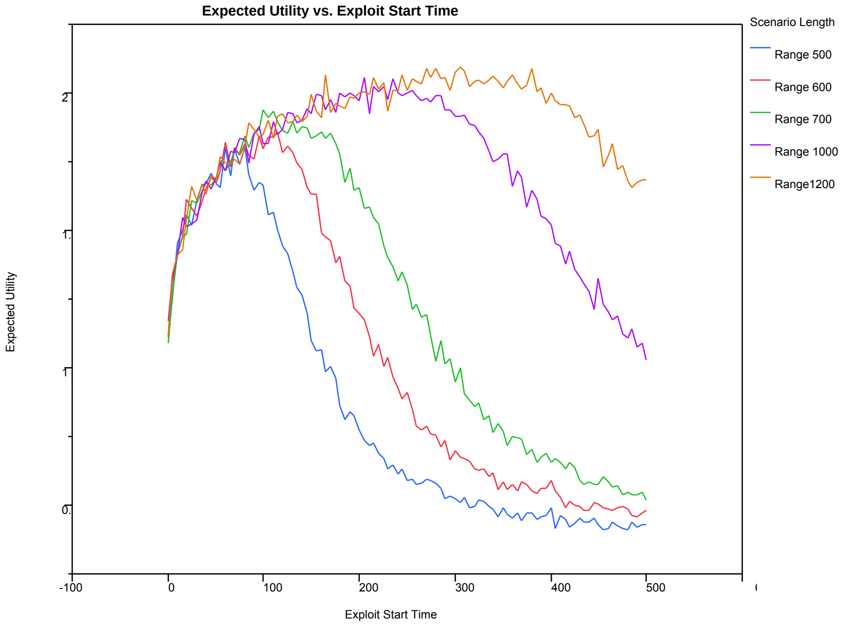


Figure 3. Expected utility as a function of $t_{Exploit}$ for 5 scenario lengths for the Grid World (GW) example.

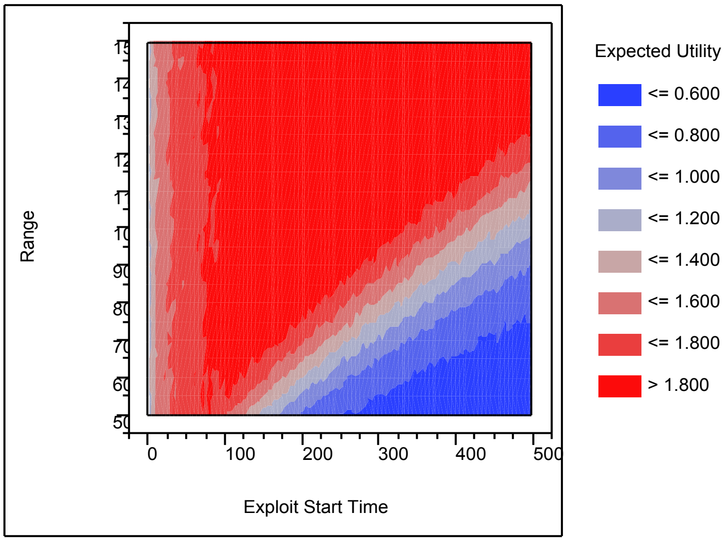


Figure 4. Contour plot of expected utility as a function of scenario length and t_{Exploit} for the GW example.

Fitting a statistical model to the results indicates a significant linear relationship between scenario length and expected utility, as would be expected. A non-linear relationship is observed between t_{Exploit} and expected utility. Efforts to normalize the residuals were unsuccessful and time constraints resulted in fitting separate models for each scenario length. As an example, the model for expected utility with a scenario length of 500 for the grid world example is shown in Figure 5 fitted with a fourth order polynomial, resulting in an Rsquare of 0.97 (Exploit Utility = 1.6646852 - 0.0043167*Exploit Time + 0.0000285*(Exploit Time-250.01)^2 + 3.5401e-8*(Exploit Time-250.01)^3 - 4.015e-10*(Exploit Time-250.01)^4).

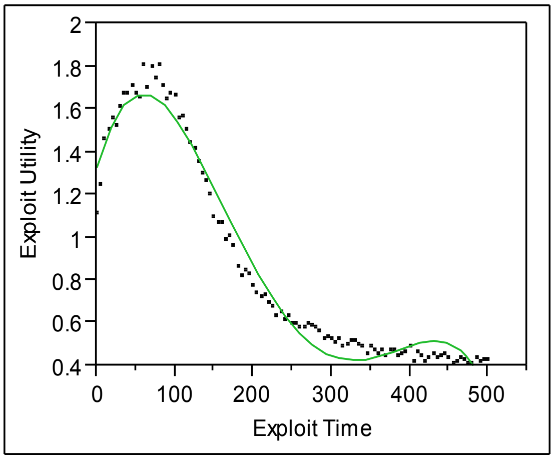


Figure 5. Bivariate fit of expected utility by $t_{Exploit}$ for the GW example.

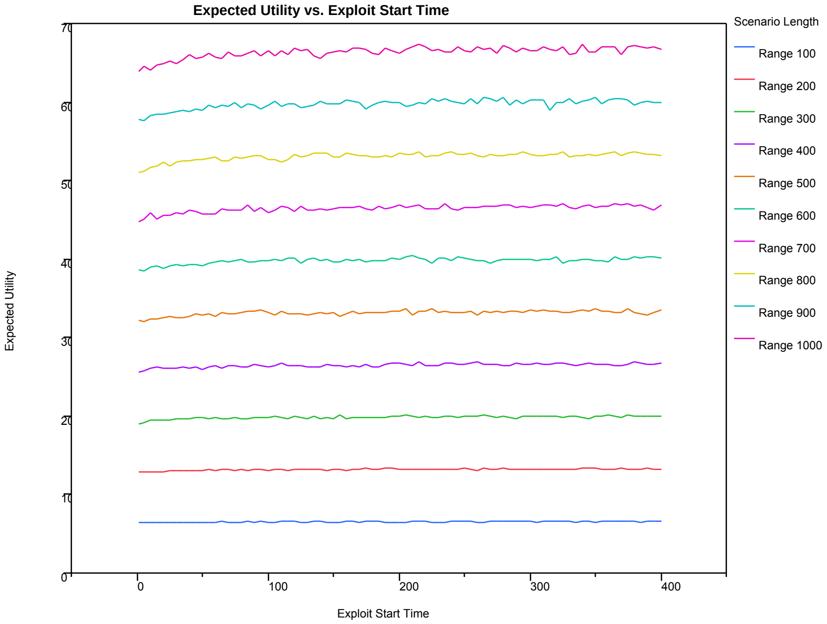


Figure 6. Expected utility as a function of $t_{Exploit}$ for 5 scenario lengths for the 2 armed bandit example.

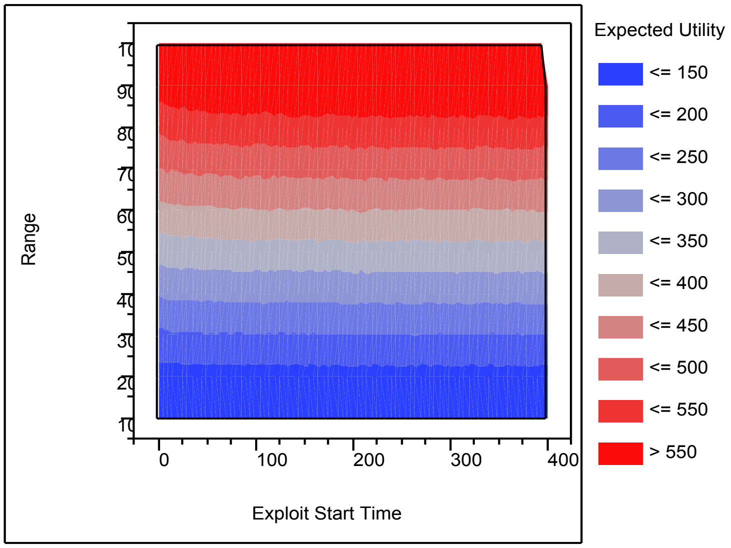


Figure 7. Contour plot of expected utility as a function of scenario length and $t_{Exploit}$ for the 2 armed bandit example.

As an example, the model for expected utility with a scenario length of 540 for CG model fitted with a fifth order polynomial, resulting in an Rsquare of 0.9278 (Expected Utility = -0.717858 + (0.0002154)*Exploit Start Time + (1.1264e-6)*(Exploit Start Time-375.013)^2 – (6.8857e-9)*(Exploit Start Time-375.013)^3 – (1.773e-11)*(Exploit Start Time-375.013)^4 +(7.068e-14)*(Exploit Start Time-375.013)^5) (Ozcan et al., 2010).

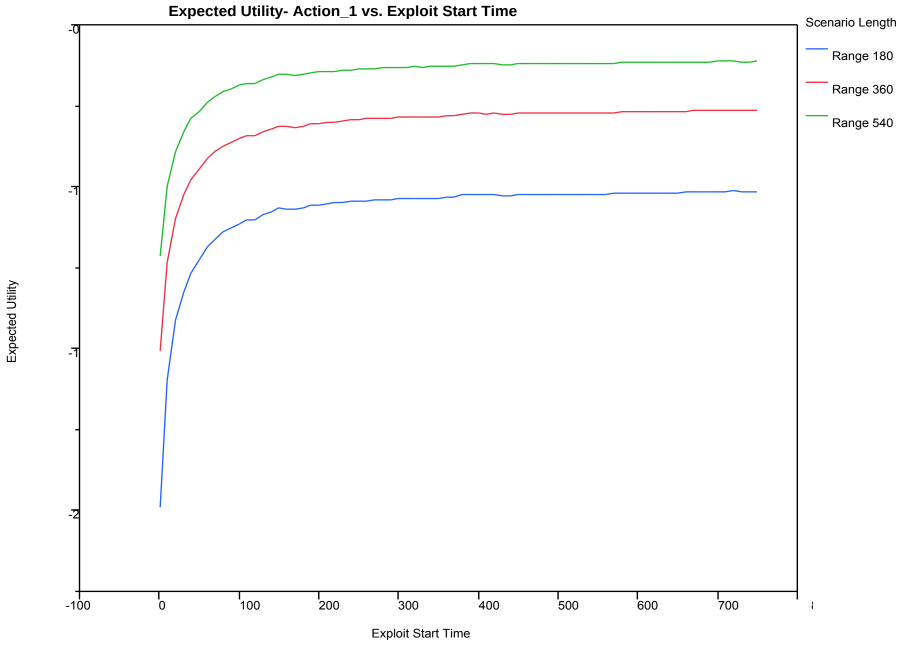


Figure 8. Expected utility as a function of $t_{Exploit}$ for 3 scenario lengths for the CG model.

2. Aggregate Utility-Driven Exploration Results

Figure 9 shows the expected utility as a function of $u_{acceptable}$ for each of the scenario lengths for the grid world example, Figure 11 shows for the 2 armed bandit example and Figure 13 shows with three scenario lengths for the CG model. Note that as scenario length goes up, the breakpoint for $u_{acceptable}$ goes up as well for the grid world example.

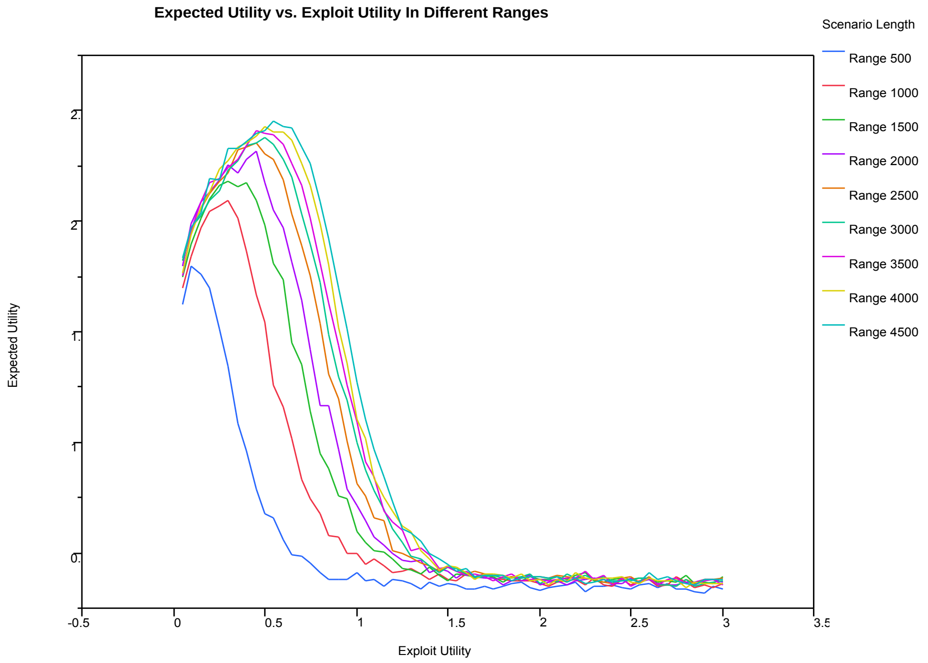


Figure 9. Expected utility as a function of $u_{acceptable}$ for 9 scenario lengths for GW example.

Figure 10 and Figure 12 illustrate the expected utility as a function of the scenario length and $u_{acceptable}$ for both environments.

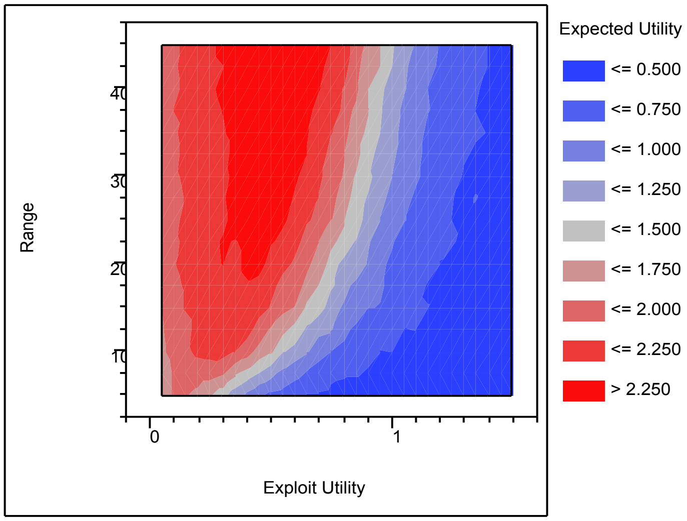


Figure 10. Contour plot of expected utility as a function of scenario length and $u_{acceptable}$ for GW example.

Similar non-linear results were observed in fitting a regression model to the grid world example results for this technique. The scenario length had very little impact on the algorithms performance in this case, Rsquare=0.26 in a bivariate fit, while $u_{acceptable}$ had a larger impact, Rsquare=0.70, but accounted for less of the variance in the response than the same term in the time based algorithm.

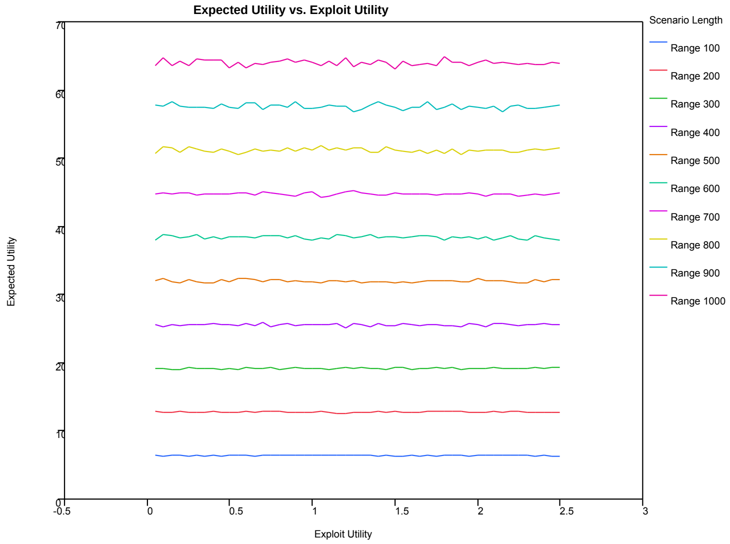


Figure 11. Expected utility as a function of $u_{acceptable}$ for 9 scenario lengths for the 2 armed bandit example.

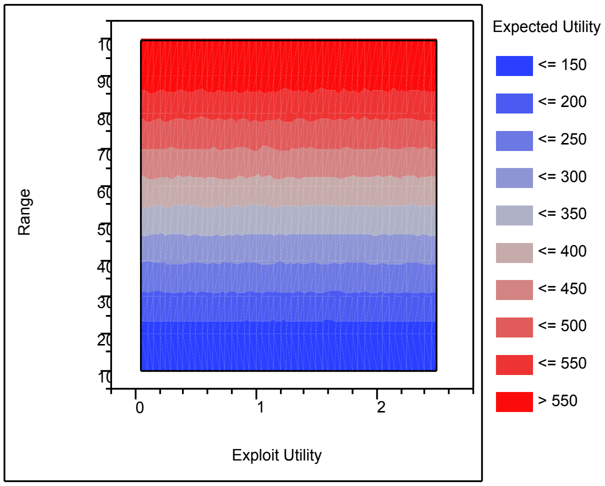


Figure 12. Contour plot of expected utility as a function of scenario length and $u_{acceptable}$ for the 2 armed bandit example.

Although $u_{acceptable}$ accounted for less of the variance in the response, with $R_{square} = 0.85$ fitted with a fifth order polynomial, than the same term in the time based algorithm, the agent obtained better utilities for each scenario lengths for CG model (Ozcan et al., 2010).

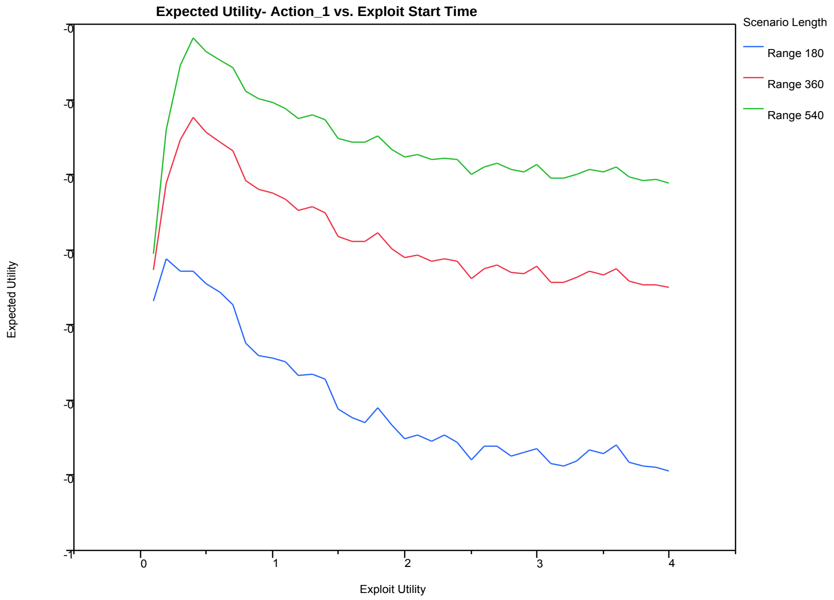


Figure 13. Expected utility as a function of $u_{acceptable}$ for 3 scenario lengths for the CG model.

3. Combination of Time and Utility Based Technique with Happiness Function

To clarify the effects of different environment sizes on the expected utility, experiments are designed with 1000 replication in three different sizes of world (2, 10, and 50) for both grid world and armed bandit examples.

Differences in the action selection sequence in both examples made agents behave differently. Changing the size of the world in GW example gives conspicuously different results than the armed bandit example in which agent has limited action selection sequence. Figure 14 and Figure 15 shows the utility change in time for both environments. It is clearly seen that both size of the world and action selection sequence have effects on expected utility of the agent. An agent prefers high $t_{Exploit}$ and $u_{acceptable}$ in order to obtain high utility. For $u_{acceptable}$

highest value (2.5) is giving highest utility and for $t_{Exploit}$ agent is selecting high value depending on the scenario length (between 200 and 400). For the CG model, an agent follows a similar action selection sequence so that the agent is obtaining similar values in time (Figure 16).

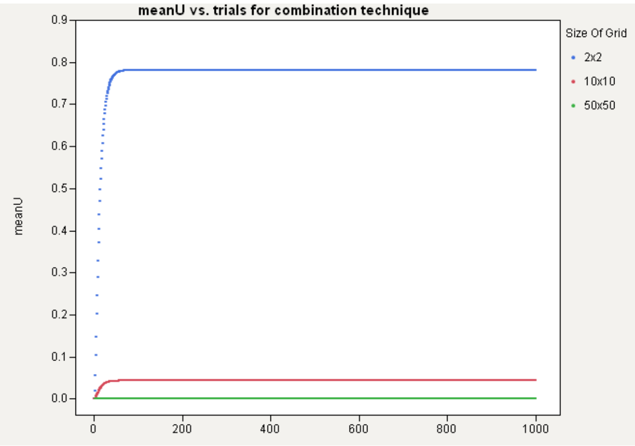


Figure 14. Mean utility as a function of Combination Technique for three different sizes of environment in GW example.

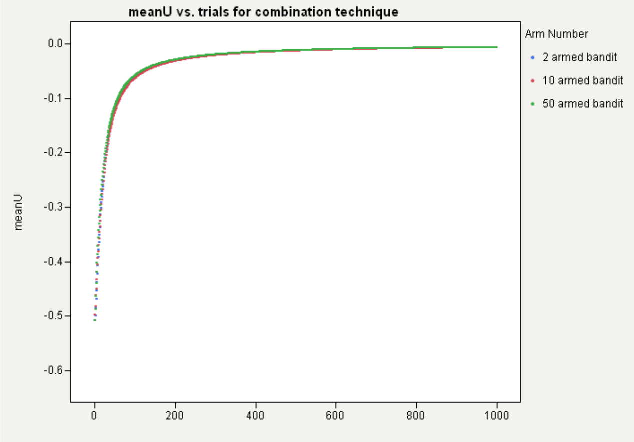


Figure 15. Mean utility as a function of Combination Technique for three different arm numbers in Armed Bandit example.

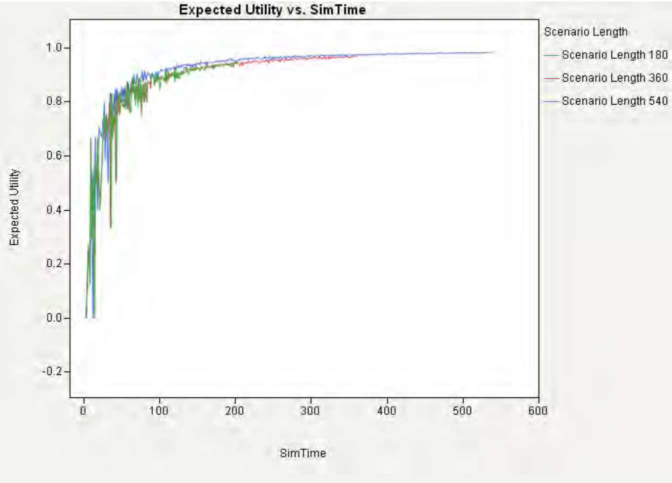


Figure 16. Mean utility as a function of Combination Technique for three different scenario lengths in the CG Model.

B. COMPARISON OF ALL TECHNIQUES WITH VARYING SIZES OF ENVIRONMENT USING TOTAL REGRET, FREQUENCY OF OPTIMAL PULLS

1. Frequency of Optimal Pulls

In the grid world example for small environments, an agent can take benefit of feedback after finding rewards by adjusting its current utility. For this reason for small environments, Aggregate Utility Driven Exploration Technique gives better results. Figure 17 and Figure 18 shows the comparison of all techniques for 2x2 and 10x10 GW examples. On the other hand, when environment gets bigger agents have difficulty in finding a reward. In this case, for longer runs the Combination of Time and Utility Based Technique with Happiness Function gives better results. On the Figure 19, it is seen that after 500 trials the FOP difference between Combination Technique and the others increases.

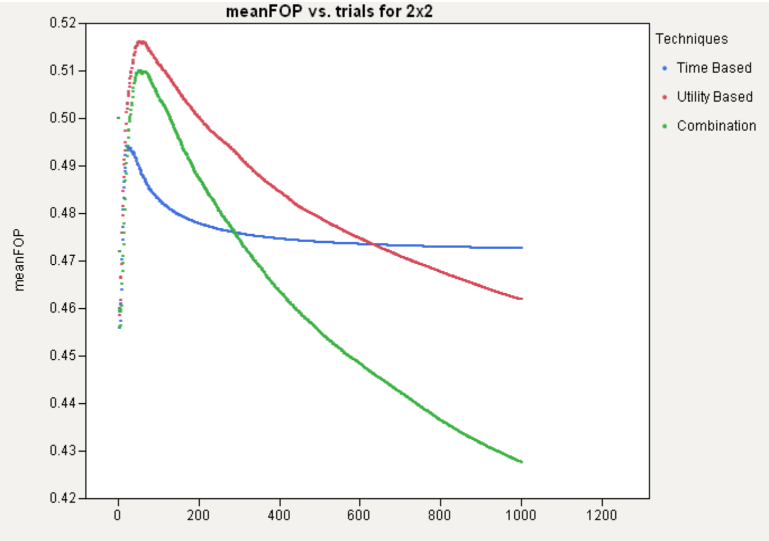


Figure 17. Comparison of mean FOP in three different techniques for 2x2 GW example.

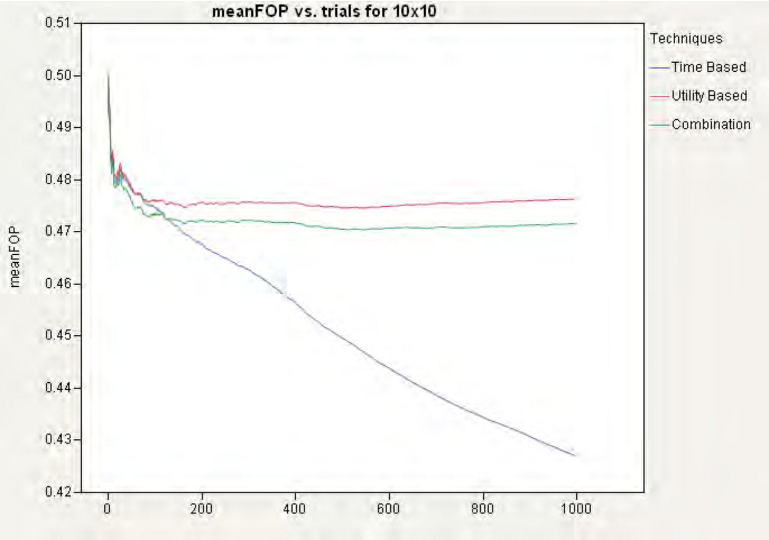


Figure 18. Comparison of mean FOP in three different techniques for 10x10 GW example.

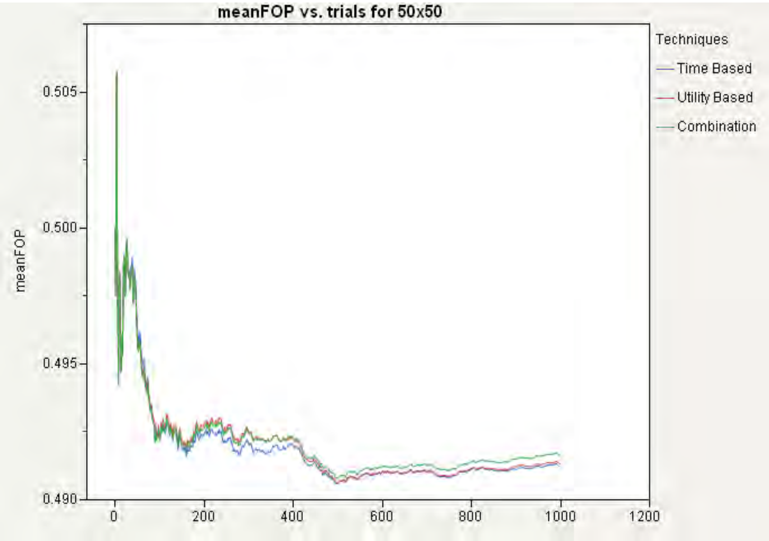


Figure 19. Comparison of mean FOP in three different techniques for 50x50 GW example.

In the armed bandit example, although for small environments all techniques give similar results when the environment gets bigger, a conspicuous difference occurs. In the Time Based Search then Converge Technique, an agent becomes greedy gradually in time so that it explores more area than the other techniques. Because it is a stochastic environment and there is uncertainty while finding the best arm, an agent needs more trials and more exploration in environment. A more explored area gives more frequency of optimal pulls. Initially the Combination Technique and Aggregate Utility Technique give better results in a short time, but for long trials Time Based Search then Converge Technique give better results (Figure 20, Figure 21 and Figure 22). For 50 armed bandit examples, this difference is higher because in Time Based Search then Converge Technique, when size of the environment increases, the agent has difficulty to find optimal path. Exploratory behavior in this technique is initially a “con,” but becomes a “pro” in longer trials.

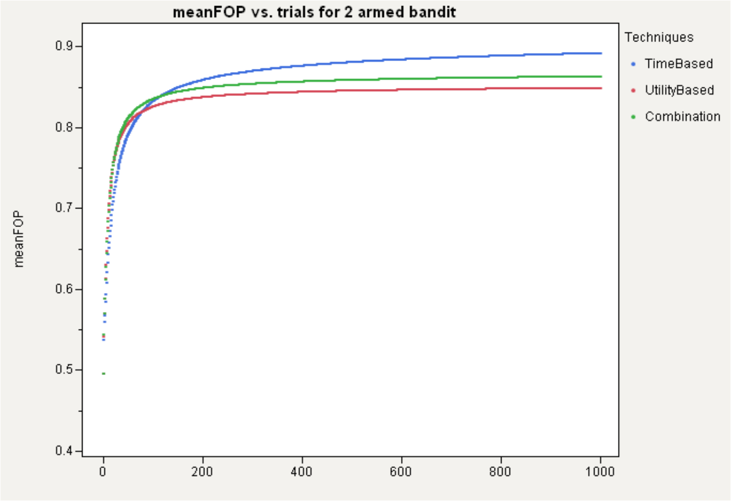


Figure 20. Comparison of mean FOP in three different techniques for two armed bandit example.

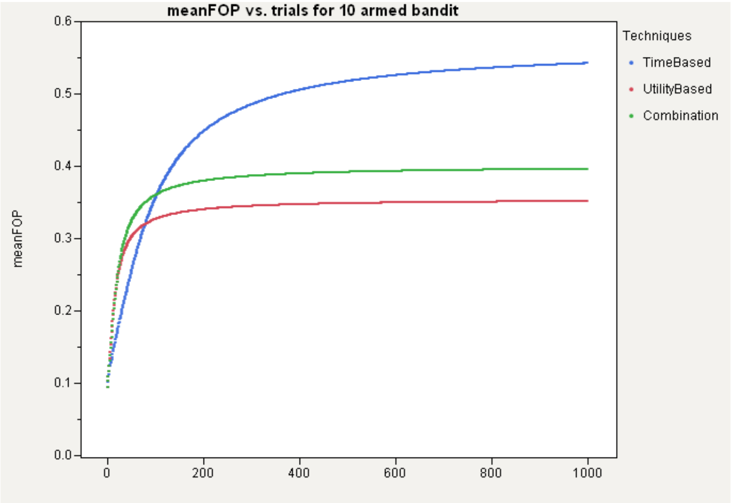


Figure 21. Comparison of mean FOP in three different techniques for 10 armed bandit example.

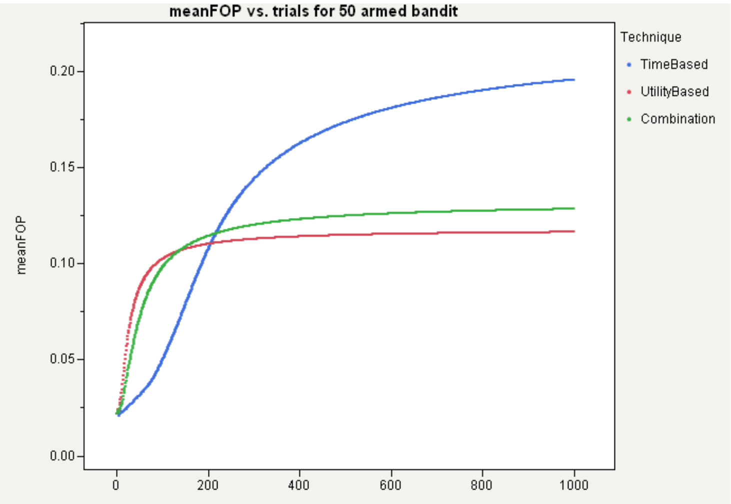


Figure 22. Comparison of mean FOP in three different techniques for 50 armed bandit example.

2. Total Regret

Both the Aggregate Utility Driven Exploration Technique and Combination of Time- and Utility Based Technique with Happiness Function become greedy based on the reward obtained by agent, so that as soon as the agent finds a reward, it becomes greedy and quits explorative behaviors for some time. Sometimes lack of enough exploration because of this greedy behavior concludes with the inability to find shorter paths. For the Time Based Search then Converge Technique greedy behavior takes some time. During this period, an agent is exploring more areas, finding shorter paths, and decreasing its regret. As a result, for both GW and Armed Bandit examples, the Time Based Search then Converge Technique results in less regret than the other two techniques. For all design points, standard error takes a maximum 0.0074 and minimum 0.0014. As seen from the armed bandit example results (Figure 25 and Figure 26), the Aggregate Utility Technique and Combination Technique reaches less regret than the Time Based Technique in a short time but when the simulation time gets longer, the Time Based Technique reaches less regret than the other techniques. For the grid world environment, although all techniques give similar results for small grid sizes, the Time Based Technique gives better results for large environments. (Figure 23 and Figure 24)

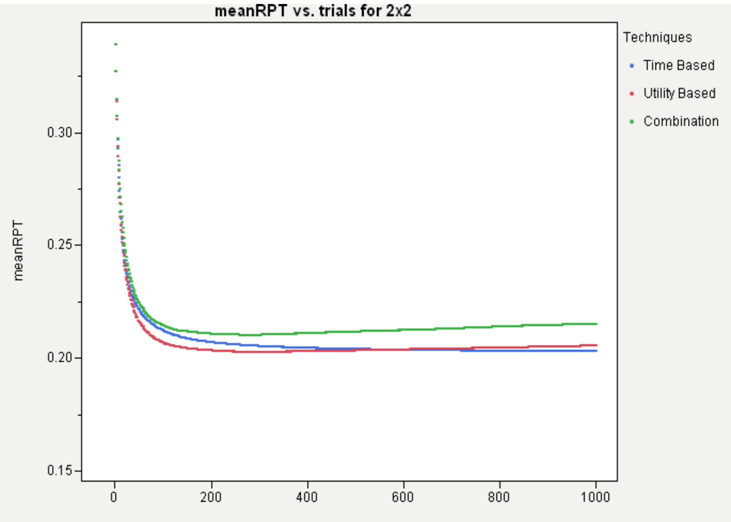


Figure 23. Comparison of mean RPT in three different techniques for 2x2 GW example.

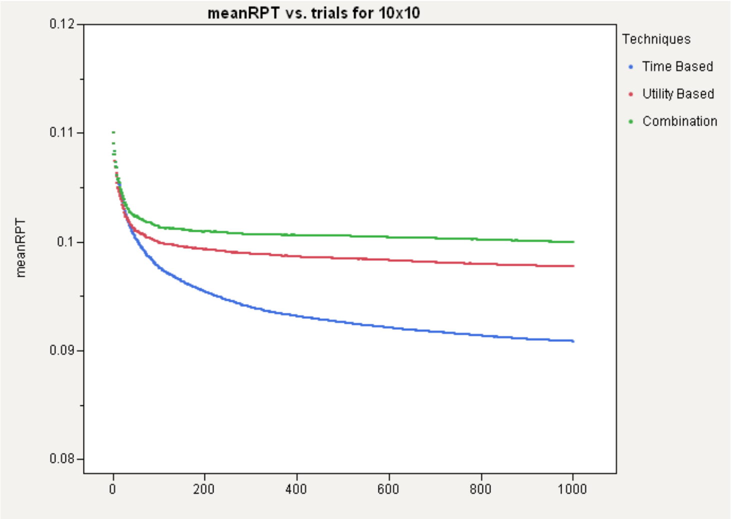


Figure 24. Comparison of mean RPT in three different techniques for 10x10 GW example.

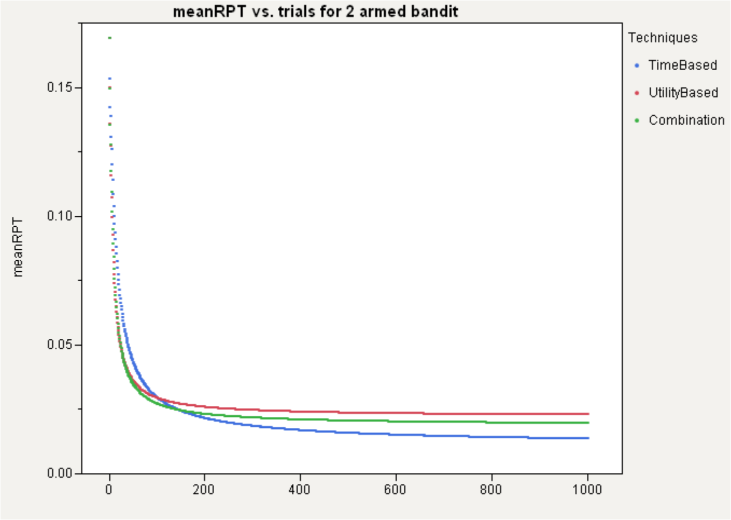


Figure 25. Comparison of mean RPT in three different techniques for 2 armed bandit example.

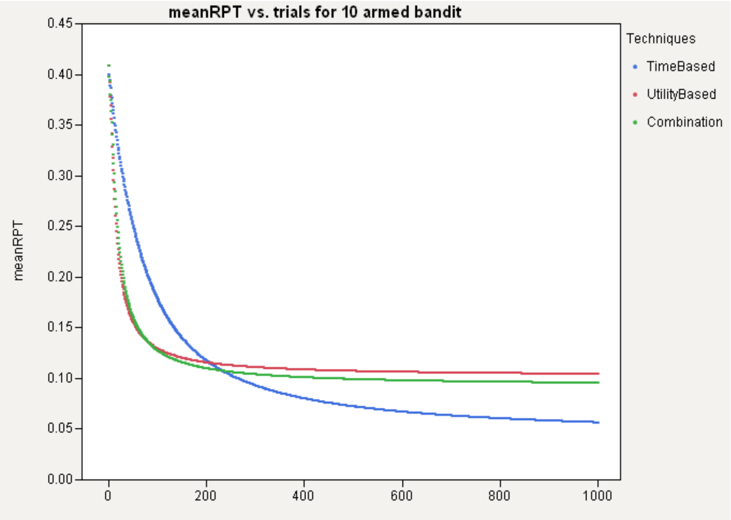


Figure 26. Comparison of mean RPT in three different techniques for 10 armed bandit example.

C. RESULTS

After designing experiments on the different scenarios, we saw some factors affect reinforcement learning directly as described below with details. Based on these factors, we focused on how to obtain the best results from each technique.

Factors that affect reinforcement learning:

Size of environment; has a negative correlation with expected utility. When you increase the size of the environment, the agent has difficulty in finding the goal and as a result, it obtains less utility. On the other hand, for the grid world example, it has negative correlation with total regret. When you increase the size of the environment, an agent finds more paths to reach to the goal so that it gets less regret. For the armed bandit example, when you increase the arm number, it needs more time to define the best arm. It needs more trials to figure it out so that increases in size of the environment causes more regret.

Scenario Length; has a positive correlation with expected utility. Increases in scenario length give the agent more time to learn the environment and obtain rewards. Scenario length also affects both exploit time and exploit utility, which defines how fast an agent will change its current behavior to an exploitive behavior.

Exploit Time ($t_{Exploit}$) and Exploit Utility ($u_{acceptable}$); have a negative correlation with expected utility and positive correlation with scenario length.

In the armed bandit example, an agent needs explorer behavior to define the best arm. For long scenario length, an agent prefers high $t_{Exploit}$ or $u_{acceptable}$, which helps in decide more accurately to find the goal (Table 1). Increases in the number of the arms causes an agent to spend more time to find the best arm and because of that, total regret gets high. For each technique, selecting high $t_{Exploit}$ or $u_{acceptable}$ gives less regret and high frequency of optimal pulls.

Table 1. Optimal parameters for each technique for Armed Bandit Example

Technique Name	Arm Number	Min Total Regret	Stderr Regret Per Trials	Exploit Utility	Exploit Time
timeBased	2	0,013537843	0,001477239	0	400
timeBased	10	0,056390152	0,002016928	0	400
timeBased	50	0,092192327	0,00166312	0	100
utilityBased	2	0,022855089	0,002242073	2,5	0
utilityBased	10	0,104292438	0,003685443	2,5	0
utilityBased	50	0,110441027	0,002937607	2,5	0
Combination	2	0,019592381	0,002059859	2,5	300
Combination	10	0,09548164	0,003538952	2,5	400
Combination	50	0,094612496	0,002490541	2,5	200

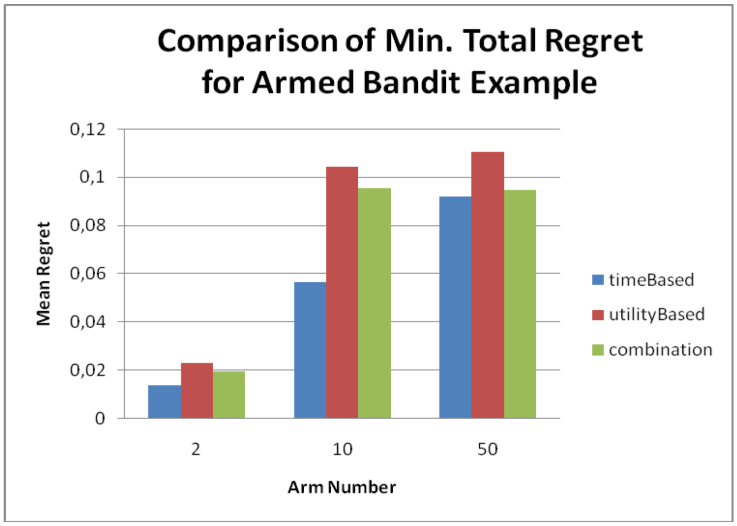


Figure 27. Comparison of each technique based on Min. Total Regret for Armed Bandit Example.

In Figure 27, the mean regret of 1000 pulls for the design points that give lowest regret for different arm numbers are shown for each technique. Table 2 shows significant differences among distributions of these design points within and between techniques. Because of that, although the mean of regret values for each technique does not look significantly different, the distributions of regret values for each technique are statistically significant.

Table 2. Armed Bandit Significance Table between and within techniques.

	Armed Bandit Significance Table		
Within Techniques	2 v 10 arms	2 v 50 arms	10 v 50 arms
Time Based	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001
Utility Based	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001
Combination	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001
Between Techniques	Time Based v Utility Based	Time Based v Combination	Utility Based v Combination
2 arms	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001
10 arms	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001
50 arms	Prob > |t|= <.0001	Prob > |t|= <.0001	Prob > |t|= <.0001

As seen in Table 2, distributions of regret values between and within techniques are statistically significant for the Armed Bandit example. Detailed statistical analysis for Table 2 is shown in Appendix B.

In the GW example, the goal is always on a fixed location, which means it is a deterministic environment, so that as soon as agent finds the goal it prefers behaving greedily by preferring small $\mathbf{t_{Exploit}}$ and $\boldsymbol{u_{acceptable}}$. As seen from Table 2, an agent gets the optimal value based on regret by selecting small $\mathbf{t_{Exploit}}$ and $\boldsymbol{u_{acceptable}}$. Figure 28 show that increasing in size of the environment gives more opportunity for the agent to select a shorter path. Because of that, more area gives less regret.

Table 3. Optimal parameters for each technique for GW Example

Technique Name	Size of Grid (Square)	Min Total Regret	Stderr Regret Per Trials	Exploit Utility	Exploit Time
timeBased	2	0,202891667	0,00543679	0	1
timeBased	10	0,090860241	0,001775314	0	1
timeBased	50	0,025073008	0,000629757	0	1
utilityBased	2	0,205398667	0,005039127	0,1	0
utilityBased	10	0,097729241	0,001642007	0,1	0
utilityBased	50	0,025284008	0,000643045	0,1	0
combination	2	0,215030667	0,004807874	0,1	1
combination	10	0,099976241	0,001705538	0,1	1
combination	50	0,025297008	0,000649327	0,1	200

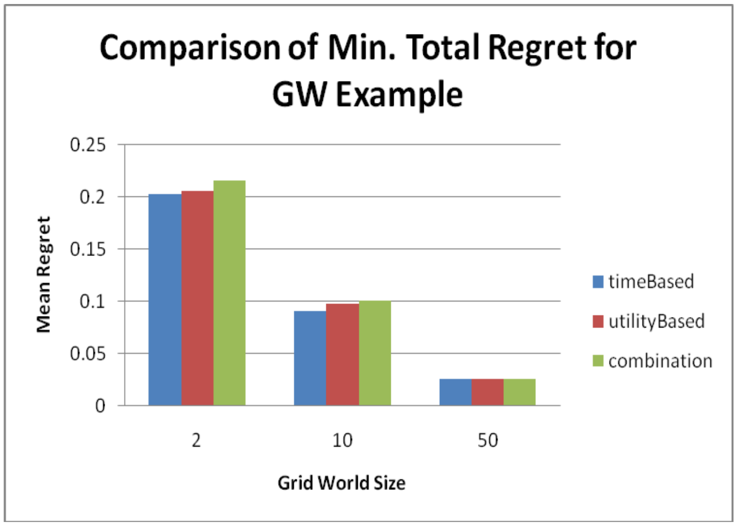


Figure 28. Comparison of each technique based on Min. Total Regret for GW Example.

In Figure 28, the mean regret of 1000 trials for the design points that give lowest regret for different sizes of environments are shown for each technique. Table 4 shows significant difference between distributions of these design points within and between techniques. As also seen from the armed bandit example, for this environment although the mean of regret values for each technique does not appear to be significantly different, the distributions of regret values for each technique are statistically significant.

Table 4. GW Significance Table between and within techniques.

<table>
<tr><th></th><th colspan="3">Grid World Significance Table</th></tr>
<tr><td>Within Techniques</td><td>2x2 v 10x10</td><td>2x2 v 50x50</td><td>10x10 v 50x50</td></tr>
<tr><td>Time Based</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
<tr><td>Utility Based</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
<tr><td>Combination</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
<tr><td>Between Techniques</td><td>Time Based v Utility Based</td><td>Time Based v Combination</td><td>Utility Based v Combination</td></tr>
<tr><td>2x2</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
<tr><td>10x10</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
<tr><td>50x50</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td><td>Prob > |t|= <.0001</td></tr>
</table>

As seen from Table 4, all differences between and within techniques are statistically significant for the GW example. Detailed statistical analysis for Table 4 is shown in Appendix C.

Experiments were repeated for the CG model with 20 agents. The Time Based Technique fit the model with R_{square}=0.88 (Appendix D) and Figure 29 shows that to obtain high expected utility, an agent should prefer high $t_{Exploit}$ and long scenario lengths.

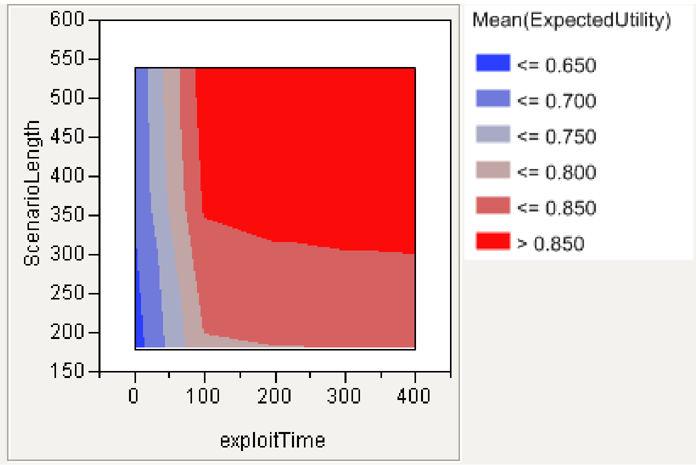


Figure 29. Contour Plot for Time Based Technique in the CG Model

Utility Based Technique fitted the model with R_{square} = 0.9483. Figure 30 shows that high exploit utility ($u_{acceptable}$) and long scenario lengths give better expected utilities.

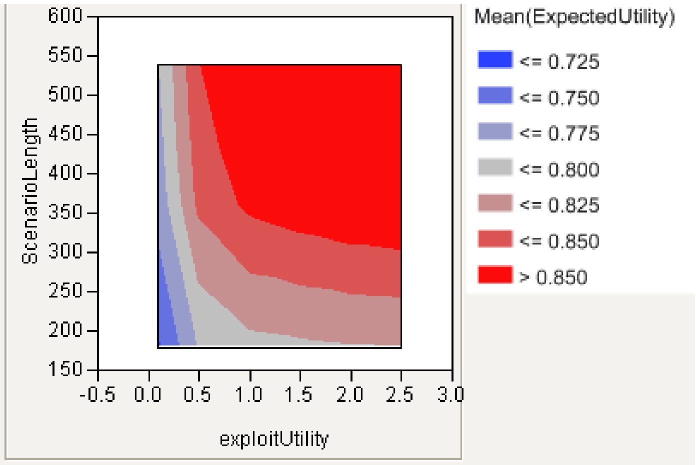


Figure 30. Contour Plot for Utility Based Technique in the CG Model

Table 5. Results of Combination Technique for CG Model.

Exploit Time	Exploit Utility	Scenario Length	Expected Utility		
1	0.1	180	0.800363955	0.878911	0.910064
100	0.5	180	0.800363955	0.878911	0.910064
200	1	180	0.800363955	0.878911	0.910064
300	1.5	180	0.800363955	0.878911	0.910064
400	2	180	0.800363955	0.878911	0.910064
1	2.5	180	0.800363955	0.878911	0.910064
100	0.1	180	0.800363955	0.878911	0.910064
200	0.5	180	0.800363955	0.878911	0.910064
300	1	180	0.800363955	0.878911	0.910064
400	1.5	180	0.800363955	0.878911	0.910064

For Combination Technique, results were not satisfactory. While changing scenario length causes different expected utilities, changing $t_{Exploit}$ and $u_{acceptable}$ does not cause any differences (Table 5). After some research has done about this problem, the cause might be determined to be the reward policy of the CG model. Some more research is needed to solve the problem.

Results showed that for small environments while Combination and Aggregate Utility Techniques give better results for small trials, in long runs, the Time Based Search then Converge Technique gives better results. In more complicated and bigger environments, the Combination Technique performed better in long runs (Table 3).

V. CONCLUSION AND FUTURE WORK

A. SUMMARY AND CONCLUSION

In this thesis, three new approaches were examined to investigate the exploration and exploitation problem in reinforcement learning from different perspectives. In each approach, it was hoped that more realistic results would be achieved by making an agent's action selection dynamic. By focusing on the correlation between the simulation's scenario lengths, sizes of environments, and expected utility of agent, a relationship that helps the user design the simulation based on his needs may be formulated. Experiments were conducted to answer the research questions.

Chapter I explained that the goal of this research was to find answers to the following questions and to recommend improvements obtainable by using new techniques:

- What techniques can be used to control the level of exploration and exploitation within cognitive agents?
- How do changes in individual agent behaviors affect the macro-level results of a test-bed simulation scenario within the Cultural Geography (CG) model?

In Chapter II, after explaining Reinforcement Learning and Boltzmann Action Selection, each technique was described in detail. Following Chapter II, experiments were designed and conducted in three different environments. Each environment has a different specialty from the others. The Armed Bandit example is the simplest non-stationary one. In this environment, an agent tries to learn with limited action selection sequence. The Grid World example is also non-stationary example with more action selection sequence. To test the techniques in a stationary environment, the Cultural Geography Model is used.

The aim of the first part of these experiments was to define which factors are important for agent learning experiment. After these experiments were conducted, factors--the size of environment, scenario length, exploit time, and exploit utility-- were found to be significantly important for each technique. Results showed that size of the environment has a negative relationship with obtained utility. A bigger environment decreases utility because agents have difficulty in finding the goal and rewards.

Exploit Time ($t_{Exploit}$), which defines how fast agent will be greedy, has a positive relationship with scenario length. For small scenario lengths, selecting small exploit time gives better results and for high scenario lengths, selecting high exploit time gives better results. It is similar with Exploit Utility ($u_{acceptable}$). Increases in scenario length need more $u_{acceptable}$ in order to obtain more utility. By constructing a contour plot of MOE over for $t_{Exploit}$, $u_{acceptable}$ and scenario length, we saw combinations for $t_{Exploit}$, $u_{acceptable}$ and a scenario length that contributes to the best and the worst expected utility.

As a main finding from this research for small environments and for short runs while Combination and Aggregate Utility Techniques give better results, for long runs the Time Based Search then Converge Technique is obtaining higher utilities from other two techniques. In more complex and bigger environments, Combination Technique performed better in long runs.

By figuring out all relationships that affect the dynamic temperature, for any simulation scenario we can select a suitable technique that gives best utility for that scenario and adjust an agent’s behavior by manipulating factors that affect that technique. This will speed up agent’s learning and adaptation to new environments.

B. FUTURE WORK

This research applied three novel approaches to controlling the level of exploration and exploitation in reinforcement learning. These early results are promising as simple approaches requiring minimal knowledge of the environment to ascertain their initial setting. Future work will apply these algorithms to more complex environments, with the intended application of representing human behavior within modeling and simulation.

Instead of using Softmax algorithm for action selection, another algorithm can be used for action selection, for example, ε-greedy algorithm. Comparison of these techniques can give better understanding in reinforcement learning.

For the Combination Technique, we used constant weights for three factors that affect the Combination Technique's temperature (0.6). By making these weights dynamic, an agent's decision strategy completely changes. By focusing on these weights, better results and better learning strategy can be obtained. We used fixed memory capacity (T) and fixed discount factor (lam2) for the agent. The effect of memory capacity on learning and discount factor can be examined for future work.

Although the algorithm for Combination Technique works for different environments, implementation of this technique in the CG model did not give satisfactory results. More research is needed on implementing this technique.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A: PYTHON CODE OF EACH TECHNIQUE

```
def tempSched( now, initialTemp, exploitTime):

    temp = initialTemp

    newTemp = temp/(1+(now/exploitTime))

    return newTemp
```

```
def tempAgUtSched( utility,initialTemp, exploitUt):

    temp = initialTemp

    newTemp = temp/(1+(utility/exploitUt))

    return  newTemp
```

```
def tempCombo(utility,temperature,now,longHappinessL):

    temp = temperature

    T=50  # Long Term Memory Capacity

    LongHappiness = 0.0

    nwt=0.0# Normalized happiness

    shortHappiness = utility

    wtHappines = 0.0# Normalized weighted happiness

    i = 0

    lam2=0.5

    nwtL =[]

    longHappinessL.append(utility)
```

```
longHappinessLCopy = longHappinessL[:]
longHappinessLCopy.reverse()
if len(longHappinessLCopy) <T :
    for u in xrange(len(longHappinessLCopy)):
        i += 1
        if max(longHappinessLCopy)== 0:
            nwt = 0.0
            nwtL.append(nwt)
        else :
            nwt=(1-((max(longHappinessLCopy)-longHappinessLCopy[u])/(max(longHappinessLCopy)-min(longHappinessLCopy)))
            nwtL.append(nwt)
        wtHappines= nwtL[u]*pow(lam2,i)
        LongHappiness+=wtHappines
        if u == T:
            break
else:
    for u in xrange(T):
        i += 1
        if max(longHappinessLCopy)== 0:
            nwt = 0.0
            nwtL.append(nwt)
        else :
            nwt=(1-((max(longHappinessLCopy)-
            longHappinessLCopy[u])/(max(longHappinessLCopy)-
            min(longHappinessLCopy)))
            nwtL.append(nwt)
        wtHappines= nwtL[u]*pow(lam2,i)
        LongHappiness+=wtHappines
        if u == T:
            break
newTemp = math.exp(shortHappiness-LongHappiness)
return newTemp
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B: ANALYSIS OF ARMED BANDIT SIGNIFICANCE WITHIN AND BETWEEN TECHNIQUES

Difference: Time Based Technique (10 Arm)-Time Based Technique (2 Arm)

Time Based Technique (10 Arm)	0.09836	t-Ratio	49.05326
Time Based Technique (2 Arm)	0.02064	DF	999
Mean Difference	0.07772	Prob > |t|	<.0001*
Std Error	0.00158	Prob > t	<.0001*
Upper 95%	0.08082	Prob < t	1.0000
Lower 95%	0.07461		
N	1000		
Correlation	0.91624		

Difference: Time Based Technique (50 Arm)-Time Based Technique (2 Arm)

Time Based Technique (50 Arm)	0.17651	t-Ratio	54.54308
Time Based Technique (2 Arm)	0.02064	DF	999
Mean Difference	0.15587	Prob > |t|	<.0001*
Std Error	0.00286	Prob > t	<.0001*
Upper 95%	0.16147	Prob < t	1.0000
Lower 95%	0.15026		
N	1000		
Correlation	0.77986		

Difference: Time Based Technique (50 Arm)-Time Based Technique (10 Arm)

Time Based Technique (50 Arm)	0.17651	t-Ratio	54.76853
Time Based Technique (10 Arm)	0.09836	DF	999
Mean Difference	0.07815	Prob > |t|	<.0001*
Std Error	0.00143	Prob > t	<.0001*
Upper 95%	0.08095	Prob < t	1.0000
Lower 95%	0.07535		
N	1000		
Correlation	0.95466		

Difference: Utility Based Technique (10 Arm)-Utility Based Technique (2 Arm)

Utility Based Technique (10 Arm)	0.11637	t-Ratio	130.4074
Utility Based Technique (2 Arm)	0.02632	DF	999
Mean Difference	0.09006	Prob > |t|	<.0001*
Std Error	0.00069	Prob > t	<.0001*
Upper 95%	0.09141	Prob < t	1.0000
Lower 95%	0.0887		
N	1000		
Correlation	0.97627		

Difference: Utility Based Technique (50 Arm)-Utility Based Technique (2 Arm)

Utility Based Technique (50 Arm)	0.13373	t-Ratio	79.87449
Utility Based Technique (2 Arm)	0.02632	DF	999
Mean Difference	0.10741	Prob > |t|	<.0001*
Std Error	0.00134	Prob > t	<.0001*
Upper 95%	0.11005	Prob < t	1.0000
Lower 95%	0.10477		
N	1000		
Correlation	0.91171		

Difference: Utility Based Technique (50 Arm)-Utility Based Technique (10 Arm)

Utility Based Technique (50 Arm)	0.13373	t-Ratio	25.1512
Utility Based Technique (10 Arm)	0.11637	DF	999
Mean Difference	0.01735	Prob > |t|	<.0001*
Std Error	0.00069	Prob > t	<.0001*
Upper 95%	0.01871	Prob < t	1.0000
Lower 95%	0.016		
N	1000		
Correlation	0.97602		

Difference: Combination Technique (10 Arm)-Combination Technique (2 Arm)

Combination Technique (10 Arm)	0.11007	t-Ratio	109.2454
Combination Technique (2 Arm)	0.02351	DF	999
Mean Difference	0.08656	Prob > |t|	<.0001*
Std Error	0.00079	Prob > t	<.0001*
Upper 95%	0.08812	Prob < t	1.0000
Lower 95%	0.08501		
N	1000		
Correlation	0.97047		

Difference: Combination Technique (50 Arm)-Combination Technique (2 Arm)

Combination Technique (50 Arm)	0.12808	t-Ratio	61.10998
Combination Technique (2 Arm)	0.02351	DF	999
Mean Difference	0.10457	Prob > |t|	<.0001*
Std Error	0.00171	Prob > t	<.0001*
Upper 95%	0.10793	Prob < t	1.0000
Lower 95%	0.10121		
N	1000		
Correlation	0.87494		

Difference: Combination Technique (50 Arm)-Combination Technique (10 Arm)

Combination Technique (50 Arm)	0.12808	t-Ratio	18.34808
Combination Technique (10 Arm)	0.11007	DF	999
Mean Difference	0.01801	Prob > |t|	<.0001*
Std Error	0.00098	Prob > t	<.0001*
Upper 95%	0.01994	Prob < t	1.0000
Lower 95%	0.01608		
N	1000		
Correlation	0.95895		

Difference: Utility Based Technique (2 Arm)-Time Based Technique (2 Arm)

Utility Based Technique (2 Arm)	0.02632	t-Ratio	31.0962
Time Based Technique (2 Arm)	0.02064	DF	999
Mean Difference	0.00568	Prob > |t|	<.0001*
Std Error	0.00018	Prob > t	<.0001*
Upper 95%	0.00604	Prob < t	1.0000
Lower 95%	0.00532		
N	1000		
Correlation	0.96498		

Difference: Combination Technique (2 Arm)-Time Based Technique (2 Arm)

Combination Technique (2 Arm)	0.02351	t-Ratio	18.29886
Time Based Technique (2 Arm)	0.02064	DF	999
Mean Difference	0.00287	Prob > |t|	<.0001*
Std Error	0.00016	Prob > t	<.0001*
Upper 95%	0.00317	Prob < t	1.0000
Lower 95%	0.00256		
N	1000		
Correlation	0.97502		

Difference: Combination Technique (2 Arm)-Utility Based Technique (2 arm)

Combination Technique (2 Arm)	0.02351	t-Ratio	-105.683
Utility Based Technique (2 arm)	0.02632	DF	999
Mean Difference	-0.0028	Prob > |t|	<.0001*
Std Error	2.66e-5	Prob > t	1.0000
Upper 95%	-0.0028	Prob < t	<.0001*
Lower 95%	-0.0029		
N	1000		
Correlation	0.9991		

Difference: Utility Based Technique (10 Arm)-Time Based Technique (10 Arm)

Utility Based Technique (10 Arm)	0.11637	t-Ratio	14.63503
Time Based Technique (10 Arm)	0.09836	DF	999
Mean Difference	0.01802	Prob > |t|	<.0001*
Std Error	0.00123	Prob > t	<.0001*
Upper 95%	0.02043	Prob < t	1.0000
Lower 95%	0.0156		
N	1000		
Correlation	0.87508		

Difference: Combination Technique (10 Arm)-Time Based Technique (10 Arm)

Combination Technique (10 Arm)	0.11007	t-Ratio	10.6337
Time Based Technique (10 Arm)	0.09836	DF	999
Mean Difference	0.01171	Prob > |t|	<.0001*
Std Error	0.0011	Prob > t	<.0001*
Upper 95%	0.01387	Prob < t	1.0000
Lower 95%	0.00955		
N	1000		
Correlation	0.90478		

Difference: Combination Technique (10 Arm)-Utility Based Technique (10 Arm)

Combination Technique (10 Arm)	0.11007	t-Ratio	-43.3295
Utility Based Technique (10 Arm)	0.11637	DF	999
Mean Difference	-0.0063	Prob > |t|	<.0001*
Std Error	0.00015	Prob > t	1.0000
Upper 95%	-0.006	Prob < t	<.0001*
Lower 95%	-0.0066		
N	1000		
Correlation	0.99712		

Difference: Utility Based Technique (50 Arm)-Time Based Technique (50 Arm)

Utility Based Technique (50 Arm)	0.13373	t-Ratio	-20.1586
Time Based Technique (50 Arm)	0.17651	DF	999
Mean Difference	-0.0428	Prob > |t|	<.0001*
Std Error	0.00212	Prob > t	1.0000
Upper 95%	-0.0386	Prob < t	<.0001*
Lower 95%	-0.0469		
N	1000		
Correlation	0.80951		

Difference: Combination Technique (50 Arm)-Time Based Technique (50 Arm)

Combination Technique (50 Arm)	0.12808	t-Ratio	-27.9878
Time Based Technique (50 Arm)	0.17651	DF	999
Mean Difference	-0.0484	Prob > |t|	<.0001*
Std Error	0.00173	Prob > t	1.0000
Upper 95%	-0.045	Prob < t	<.0001*
Lower 95%	-0.0518		
N	1000		
Correlation	0.88122		

Difference: Combination Technique (50 Arm)-Utility Based Technique (50 Arm)

Combination Technique (50 Arm)	0.12808	t-Ratio	-11.9582
Utility Based Technique (50 Arm)	0.13373	DF	999
Mean Difference	-0.0056	Prob > |t|	<.0001*
Std Error	0.00047	Prob > t	1.0000
Upper 95%	-0.0047	Prob < t	<.0001*
Lower 95%	-0.0066		
N	1000		
Correlation	0.98725		

APPENDIX C: ANALYSIS OF GRID WORLD SIGNIFICANCE WITHIN AND BETWEEN TECHNIQUES

Difference: Time Based Technique (10x10)-Time Based Technique (2x2)

Time Based Technique (10x10)	0.09367	t-Ratio	-352.082
Time Based Technique (2x2)	0.20733	DF	999
Mean Difference	-0.1137	Prob > |t|	<.0001*
Std Error	0.00032	Prob > t	1.0000
Upper 95%	-0.113	Prob < t	<.0001*
Lower 95%	-0.1143		
N	1000		
Correlation	0.81967		

Difference: Time Based Technique (50x50)-Time Based Technique (2x2)

Time Based Technique (50x50)	0.02517	t-Ratio	-458.065
Time Based Technique (2x2)	0.20733	DF	999
Mean Difference	-0.1822	Prob > |t|	<.0001*
Std Error	0.0004	Prob > t	1.0000
Upper 95%	-0.1814	Prob < t	<.0001*
Lower 95%	-0.1829		
N	1000		
Correlation	0.64556		

Difference: Time Based Technique (50x50)-Time Based Technique (10x10)

Time Based Technique (50x50)	0.02517	t-Ratio	-705.459
Time Based Technique (10x10)	0.09367	DF	999
Mean Difference	-0.0685	Prob > |t|	<.0001*
Std Error	0.0001	Prob > t	1.0000
Upper 95%	-0.0683	Prob < t	<.0001*
Lower 95%	-0.0687		
N	1000		
Correlation	0.94057		

Difference: Utiltiy Based Technique (10x10)-Utiltiy Based Technique (2x2)

Utiltiy Based Technique (10x10)	0.09883	t-Ratio	-309.302
Utiltiy Based Technique (2x2)	0.20638	DF	999
Mean Difference	-0.1075	Prob > |t|	<.0001*
Std Error	0.00035	Prob > t	1.0000
Upper 95%	-0.1069	Prob < t	<.0001*
Lower 95%	-0.1082		
N	1000		
Correlation	0.85888		

Difference: Utiltiy Based Technique (50x50)-Utiltiy Based Technique (2x2)

Utiltiy Based Technique (50x50)	0.02528	t-Ratio	-472.196
Utiltiy Based Technique (2x2)	0.20638	DF	999
Mean Difference	-0.1811	Prob > |t|	<.0001*
Std Error	0.00038	Prob > t	1.0000
Upper 95%	-0.1803	Prob < t	<.0001*
Lower 95%	-0.1818		
N	1000		
Correlation	0.81702		

Difference: Utiltiy Based Technique (50x50)-Utiltiy Based Technique (10x10)

Utiltiy Based Technique (50x50)	0.02528	t-Ratio	-1722.53
Utiltiy Based Technique (10x10)	0.09883	DF	999
Mean Difference	-0.0735	Prob > |t|	<.0001*
Std Error	4.27e-5	Prob > t	1.0000
Upper 95%	-0.0735	Prob < t	<.0001*
Lower 95%	-0.0736		
N	1000		
Correlation	0.76009		

Difference: Combination Technique (10x10)-Combination Technique (2x2)

Combination Technique (10x10)	0.10076	t-Ratio	-334.85
Combination Technique (2x2)	0.21475	DF	999
Mean Difference	-0.114	Prob > |t|	<.0001*
Std Error	0.00034	Prob > t	1.0000
Upper 95%	-0.1133	Prob < t	<.0001*
Lower 95%	-0.1147		
N	1000		
Correlation	0.8887		

Difference: Combination Technique (50x50)-Combination Technique (2x2)

Combination Technique (50x50)	0.02532	t-Ratio	-513.649
Combination Technique (2x2)	0.21475	DF	999
Mean Difference	-0.1894	Prob > |t|	<.0001*
Std Error	0.00037	Prob > t	1.0000
Upper 95%	-0.1887	Prob < t	<.0001*
Lower 95%	-0.1902		
N	1000		
Correlation	0.74878		

Difference: Combination Technique (50x50)-Combination Technique (10x10)

Combination Technique (50x50)	0.02532	t-Ratio	-2340.91
Combination Technique (10x10)	0.10076	DF	999
Mean Difference	-0.0754	Prob > |t|	<.0001*
Std Error	3.22e-5	Prob > t	1.0000
Upper 95%	-0.0754	Prob < t	<.0001*
Lower 95%	-0.0755		
N	1000		
Correlation	0.93608		

Difference: Utiltiy Based Technique (2x2)-Time Based Technique (2x2)

Utiltiy Based Technique (2x2)	0.20638	t-Ratio	-12.7132
Time Based Technique (2x2)	0.20733	DF	999
Mean Difference	-0.0009	Prob > |t|	<.0001*
Std Error	7.47e-5	Prob > t	1.0000
Upper 95%	-0.0008	Prob < t	<.0001*
Lower 95%	-0.0011		
N	1000		
Correlation	0.98258		

Difference: Combination Technique (2x2)-Time Based Technique (2x2)

```
Combination Technique (2x2)     0.21475   t-Ratio      74.95618
Time Based Technique (2x2 )     0.20733   DF                999
Mean Difference                 0.00742   Prob > |t|    <.0001*
Std Error                        0.0001   Prob > t      <.0001*
Upper 95%                       0.00761   Prob < t       1.0000
Lower 95%                       0.00722
N                                  1000
Correlation                     0.96988
```

Difference: Combination Technique (2x2)-Utiltiy Based Technique (2x2)

```
Combination Technique (2x2)     0.21475   t-Ratio      270.6862
Utiltiy Based Technique (2x2)   0.20638   DF                999
Mean Difference                 0.00837   Prob > |t|    <.0001*
Std Error                       0.00003   Prob > t      <.0001*
Upper 95%                       0.00843   Prob < t       1.0000
Lower 95%                       0.00831
N                                  1000
Correlation                     0.99742
```

Difference: Utiltiy Based Technique (10x10)-Time Based Technique (10x10)

```
Utiltiy Based Technique (10x10)     0.09883   t-Ratio      87.59321
Time Based Technique (10x10 )       0.09367   DF                999
Mean Difference                     0.00516   Prob > |t|    <.0001*
Std Error                           5.89e-5   Prob > t      <.0001*
Upper 95%                           0.00527   Prob < t       1.0000
Lower 95%                           0.00504
N                                      1000
Correlation                         0.96484
```

Difference: Combination Technique (10x10)-Time Based Technique (10x10)

```
Combination Technique (10x10)       0.10076   t-Ratio      101.0077
Time Based Technique (10x10 )       0.09367   DF                999
Mean Difference                     0.00709   Prob > |t|    <.0001*
Std Error                           0.00007   Prob > t      <.0001*
Upper 95%                           0.00722   Prob < t       1.0000
Lower 95%                           0.00695
N                                      1000
Correlation                         0.92948
```

Difference: Combination Technique (10x10)-Utiltiy Based Technique (10x10)

```
Combination Technique (10x10)       0.10076   t-Ratio      167.5739
Utiltiy Based Technique (10x10)     0.09883   DF                999
Mean Difference                     0.00193   Prob > |t|    <.0001*
Std Error                           1.15e-5   Prob > t      <.0001*
Upper 95%                           0.00195   Prob < t       1.0000
Lower 95%                            0.0019
N                                      1000
Correlation                         0.99244
```

Difference: Utiltiy Based Technique (50x50)-Time Based Technique (50x50)

```
Utiltiy Based Technique (50x50)        0.02528   t-Ratio        44.57402
Time Based Technique (50x50 )          0.02517   DF                  999
Mean Difference                        0.00011   Prob > |t|      <.0001*
Std Error                               2.47e-6  Prob > t        <.0001*
Upper 95%                              0.00012   Prob < t         1.0000
Lower 95%                              0.00011
N                                         1000
Correlation                            0.44899
```

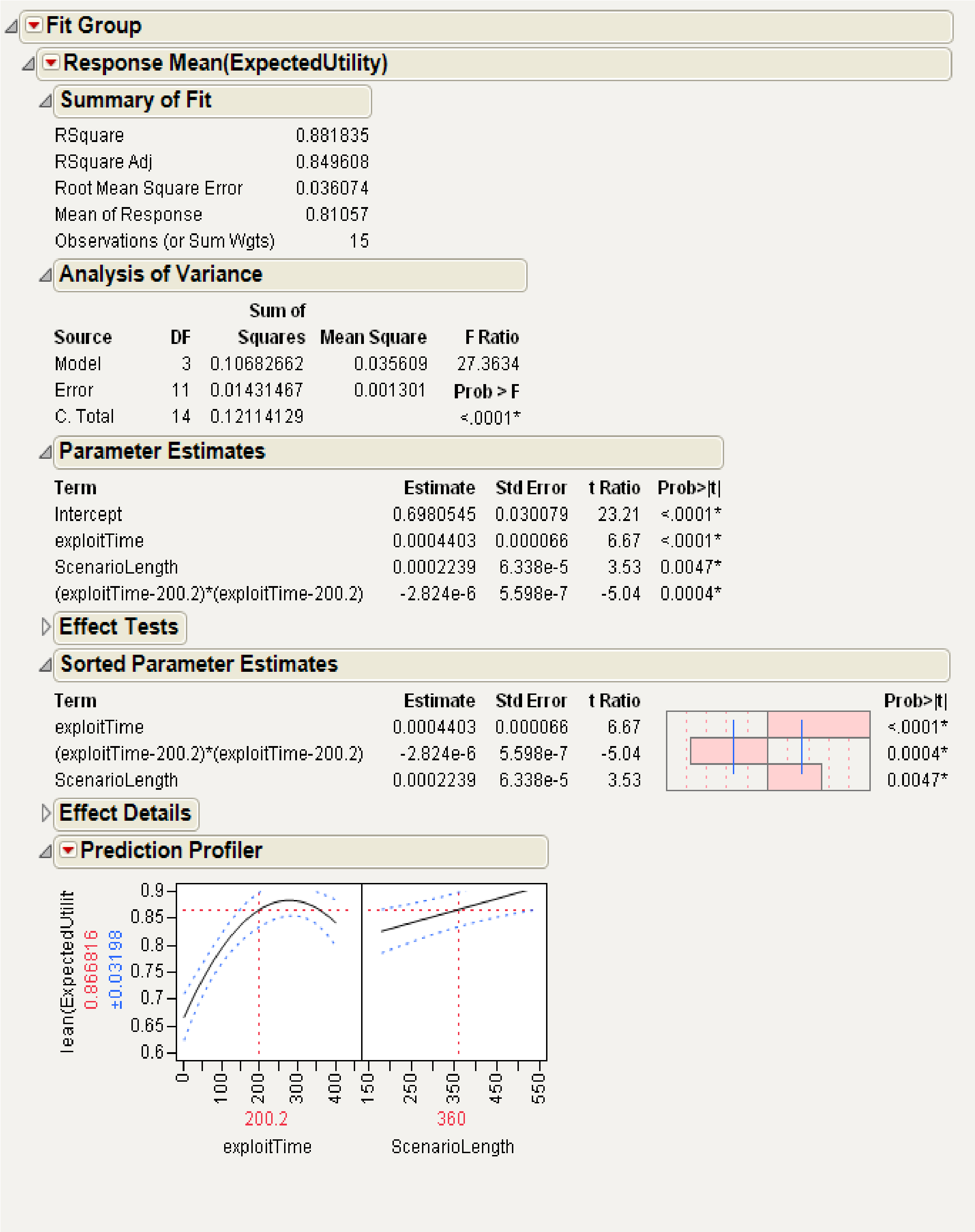
Difference: Combination Technique (50x50)-Time Based Technique (50x50)

```
Combination Technique (50x50)          0.02532   t-Ratio        66.97713
Time Based Technique (50x50 )          0.02517   DF                  999
Mean Difference                        0.00014   Prob > |t|      <.0001*
Std Error                               2.15e-6  Prob > t        <.0001*
Upper 95%                              0.00015   Prob < t         1.0000
Lower 95%                              0.00014
N                                         1000
Correlation                            0.85866
```

Difference: Combination Technique (50x50)-Utiltiy Based Technique (50x50)

```
Combination Technique (50x50)          0.02532   t-Ratio        59.49743
Utiltiy Based Technique (50x50)        0.02528   DF                  999
Mean Difference                         3.41e-5  Prob > |t|      <.0001*
Std Error                               5.73e-7  Prob > t        <.0001*
Upper 95%                               3.52e-5  Prob < t         1.0000
Lower 95%                               3.29e-5
N                                         1000
Correlation                            0.80695
```

APPENDIX D: TIME BASED TECHNIQUE ANALYSIS FOR THE CG MODEL



Fit Group

Response Mean(ExpectedUtility)

Summary of Fit

RSquare	0.881835
RSquare Adj	0.849608
Root Mean Square Error	0.036074
Mean of Response	0.81057
Observations (or Sum Wgts)	15

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	0.10682662	0.035609	27.3634
Error	11	0.01431467	0.001301	Prob > F
C. Total	14	0.12114129		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.6980545	0.030079	23.21	<.0001*
exploitTime	0.0004403	0.000066	6.67	<.0001*
ScenarioLength	0.0002239	6.338e-5	3.53	0.0047*
(exploitTime-200.2)*(exploitTime-200.2)	-2.824e-6	5.598e-7	-5.04	0.0004*

Effect Tests

Sorted Parameter Estimates

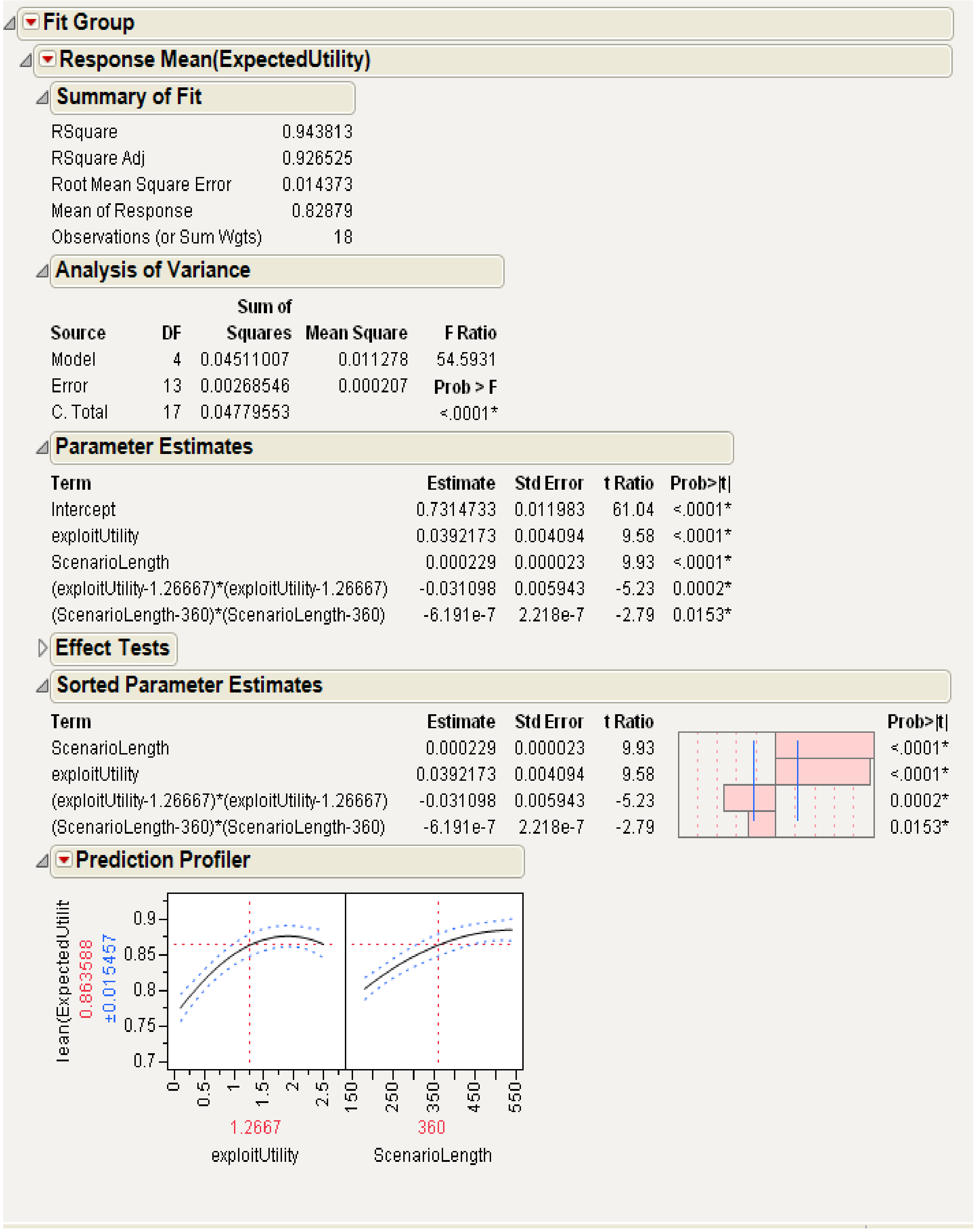
Term	Estimate	Std Error	t Ratio		Prob>|t|
exploitTime	0.0004403	0.000066	6.67		<.0001*
(exploitTime-200.2)*(exploitTime-200.2)	-2.824e-6	5.598e-7	-5.04		0.0004*
ScenarioLength	0.0002239	6.338e-5	3.53		0.0047*

Effect Details

Prediction Profiler

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E: UTILITY BASED TECHNIQUE ANALYSIS FOR THE CG MODEL



Fit Group

Response Mean(ExpectedUtility)

Summary of Fit

RSquare	0.943813
RSquare Adj	0.926525
Root Mean Square Error	0.014373
Mean of Response	0.82879
Observations (or Sum Wgts)	18

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	0.04511007	0.011278	54.5931
Error	13	0.00268546	0.000207	Prob > F
C. Total	17	0.04779553		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7314733	0.011983	61.04	<.0001*
exploitUtility	0.0392173	0.004094	9.58	<.0001*
ScenarioLength	0.000229	0.000023	9.93	<.0001*
(exploitUtility-1.26667)*(exploitUtility-1.26667)	-0.031098	0.005943	-5.23	0.0002*
(ScenarioLength-360)*(ScenarioLength-360)	-6.191e-7	2.218e-7	-2.79	0.0153*

Effect Tests

Sorted Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
ScenarioLength	0.000229	0.000023	9.93	<.0001*
exploitUtility	0.0392173	0.004094	9.58	<.0001*
(exploitUtility-1.26667)*(exploitUtility-1.26667)	-0.031098	0.005943	-5.23	0.0002*
(ScenarioLength-360)*(ScenarioLength-360)	-6.191e-7	2.218e-7	-2.79	0.0153*

Prediction Profiler

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Alt, J., Jackson, L., Hudak, D., & Lieberman, S. (2009). The cultural geography model: Evaluating the impact of tactical operational outcomes on a civilian population in an irregular warfare environment. *JDMS Methodology, Technology 6*(4), 185–199.

Macready, W. G., & Wolpert, D. H. (1998). Bandit problems and the exploration/exploitation tradeoff*: IEEE Transactions On Evolutionary Computation, 2*(1), 01–21.

Ozcan, O., Alt, J., & Darken, C. J. (2010). *Balancing exploration and exploitation in agent learning: Proceedings of the twenty-fourth international Florida artificial intelligence research society conference*. Palm Beach, FL: FLAIRS.

Papadopoulos, S. (2010, September). *Reinforcement learning: A new approach for the cultural geography model* (master's thesis). Naval Postgraduate School, Monterey, CA.

Russel, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). San Francisco: Prentice Hall.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Hava Kuvvetleri Komutanligi
   BILKARDES Sube
   Ankara, Turkey

4. Hava Harp Okulu Kutuphanesi
   Yesilyurt
   Istanbul, Turkey

5. Kara Harp Okulu Kutuphanesi
   Ankara, Turkey

6. Deniz Harp Okulu Kutuphanesi
   Tuzla
   Istanbul, Turkey

7. Dr. Chris Darken
   MOVES Institute
   Naval Postgraduate School
   Monterey, California

8. Lieutenant Colonel Jonathan Alt
   TRADOC Analysis Center
   Monterey, California