# Agents with aspirations: A Prospect Theory approach to intrinsic motivation in Reinforcement Learning

***Lucas James Irwin***

Senior Thesis

Department of Computer Science

Princeton University

April 19, 2023


Advised by Professor Tom Griffiths

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Arts in Computer Science

Princeton University

I hereby declare that this Thesis represents my own work in accordance with University regulations.

Lucas James Irwin

To my family.

**Agents with aspirations**

**Lucas James Irwin**

## ABSTRACT

For centuries, philosophers have debated the ways in which human minds should optimally manage the trade-off between achieving long-term external goals and maintaining immediate internal happiness. Various different approaches have been taken, from perspectives which prioritize being content with one's lot to ambitious calls for setting high aspirations. When investigated in the context of intrinsically motivated reinforcement learning (RL), the question of how we should optimally manage our motivations yields interesting results which shed light on the optimal behavior of both humans and machines. This thesis explores the effect of shaping the reward function of an RL agent in order to investigate what forms of intrinsic motivation improve an agent's ability to learn in different environments. By testing two reward functions based on Dubey et al.'s aspiration model and Kahneman and Tversky's Prospect Theory, I discover a new reward function which I call the Prospect Theory reward function that performs optimally in dense, sparse and complex environments such as the "Cliff Walking" environment from Sutton and Barto. The parameters of the reward functions are tuned with Bayesian optimization and the performance of the resulting agents is evaluated by comparing the average cumulative reward achieved by each agent. Results reveal that the Prospect Theory value function is a good form of intrinsic motivation to use in multiple categories of environments, suggesting that future RL architectures would benefit from incorporating Prospect Theory into their reward functions.

# Acknowledgments

I would like to thank my advisor Professor Tom Griffiths as well as Rachit Dubey and Sreejan Kumar for their guidance.

# Contents

# 1 Introduction

For centuries, philosophers have debated the ways in which human minds should optimally manage the trade-off between achieving long-term external goals and maintaining immediate internal happiness. [**10**] [**11**] On the one hand, ancient Greek Stoic philosophers such as Epictetus have claimed that we possess absolute control over our level of motivation, to the extent that we can "hack" our aspirations to be happy regardless of the external rewards we receive. [**11**] On the other hand, 17th century English philosopher, Thomas Hobbes, argued that we are driven by a restless desire to achieve power after power, and thus posited that we will always benefit from setting our aspiration level very high. [**10**]

Until recently, the study of optimal motivation has been largely confined to purely theoretical approaches but a sub-field of artificial intelligence known as intrinsically motivated reinforcement learning now enables us to study this trade-off in a more rigorous manner than was once possible. Reinforcement Learning (RL) provides a natural framework for studying the question of optimal motivation, since RL trains an agent to learn an optimal strategy via the rewards it receives from its interaction with an environment. Recent work by Dubey et al. [**1**] has shown that adding an aspiration component to the reward function used by a reinforcement learning agent can improve its performance by providing the agent with an increased exploration incentive. [**1**] Interestingly, this work also showed that an agent performs optimally when the aspiration coefficient is both nonzero and set to be neither too high nor too low, suggesting that an optimal RL model should possess a modified reward function which models the effect of aspiration on the agent's "perceived" rewards.

This thesis builds on the work of Dubey et al. by exploring the effect of reward

shaping in RL to investigate what forms of intrinsic motivation are most beneficial to an agent's performance in different environments. By testing two reward functions based on Dubey et al.'s aspiration model and a new reward function based on Kahneman and Tversky's Prospect Theory, [**21**] [**24**] this work illustrates that a Prospect Theory value function is a good form of intrinsic motivation to use, especially in sparse environments. The training process consists of three main steps. First, the parameters of the reward functions are tuned with Bayesian optimization applied to a Q-learning agent which learns a policy in simulated, grid-world environments similar to the ones used in Dubey et al. [**1**]. Next, the optimized values are used to test the aspiration and Prospect Theory reward functions on Q-learning agents in the same environments. Finally, the performance of each model is evaluated by comparing the average cumulative reward obtained over 50 runs of 1000 episodes to the reward obtained by a regular $\epsilon$-greedy RL model with the same learning rate and discount factor.

Results are significant, confirming both Dubey et al.'s work while also producing new findings which show that a Prospect Theory-inspired reward function performs better than both $\epsilon$-greedy and aspiration-based models. This effect is even more pronounced in sparse environments with limited or delayed feedback, which are commonly challenging for vanilla RL agents to deal with, [**15**] as well as complex environments such as Sutton and Barto's "Cliff Walking". [**6**] Overall, this thesis suggests that the Prospect Theory value function is a good form of intrinsic motivation to use in multiple categories of environments, strongly implying that future projects would benefit from incorporating the insights of Prospect Theory into their optimal reward design strategies.

This work is especially exciting when considered in the context of modern views

on artificial general intelligence (AGI). In a recent paper, Silver et al. propose the bold yet enticing claim that intelligence can be fully understood as an agent maximizing rewards via its interaction with an environment. [**26**] Their hypothesis implies that an effective reward-maximizing agent running on "as-yet-undiscovered algorithms" [**26**] could develop abilities associated with natural intelligence. In particular, Silver et al. claim that reward itself is sufficient to instruct a variety of behaviours including lanugage, perception, social intelligence and generalization, suggesting that a future, powerful RL agent could be the solution to artificial general intelligence. If this hypothesis is true, the identification of Prospect Theory as an effective form of intrinsic motivation provided in this thesis could one day be relevant to the design of such an AGI.

# 2 Problem Background and Related Work

Reinforcement Learning (RL) is a broad field which encompasses a variety of different algorithms and learning architectures. This thesis will narrow its focus to the sub-field known as "optimal reward design". Optimal reward design consists of designing "good" intrinsic reward functions for RL agents as a means of maximizing the average cumulative reward achieved by an agent over multiple epochs. This is especially useful in environments where rewards are sparse or delayed since these environments provide insufficient feedback for extrinsically motivated RL agents. [**15**] While extrinsically-motivated agents learn well in dense environments, they do not possess a large enough exploration incentive to perform well in sparse and other complicated settings and hence take a long time to train (sometimes to no avail). [**15**] This motivates the design of alternative reward functions since a genuinely flexible artificial intelligence should possess the ability to learn independently regardless of the underlying structure of its environment. [**16**]

Recent work in the RL field has begun to embrace the benefits of intrinsic motivation [**1**] [**3**] [**5**] [**15**] and hence inspires the focus of this thesis. In 2004, Singh et al. began to test agents with altered intrinsic reward functions to learn good functions which could generalize to multiple tasks. [**13**] In a similar 2010 paper, Singh et al. took advantage of evolutionary algorithms to learn optimal reward functions, laying the foundation for the distinction between intrinsic and extrinsic motivation. [**14**] In 2021, Dubey et al. tested three different formulations of optimal reward functions and found that a reward function which accounted for the intrinsic "aspiration" of the agent outperformed all other models tested. [**1**]

As the popularity of deep learning and Deep-Q-Networks has increased, so too has interest in applying the optimal reward design problem to Deep Reinforcement Learning. In 2015, Mohamed et al. developed a method using convolutional networks and variational inference with an intrinsic motivation metric known as "empowerment". [**16**] In 2016, Kulkarni et al. presented the hierachical DQN (h-DQN) – a framework which allowed for the efficient exploration of complicated environments by dividing the learning process into two q-value functions which learnt intrinsic goals and extrinsic goals independently. [**15**] Deep reinforcement learning (Deep-RL) has only just begun to utilize intrinsic motivation as a means of improving performance [**8**] and the potential upside of exploring such benefits in Deep-RL also inspires the work of this thesis.

In the next few sections, I will introduce some fundamental mathematical concepts and frameworks which underpin the work of this thesis. A key part of this section will focus on the aspiration reward function presented in the Dubey et al. paper [**1**] as well as the theories of risky choice presented by Kahneman and Tversky. [**21**] [**24**] This will provide context for the two reward functions used in this thesis. I will

also devote time towards providing a formal overview of reinforcement learning.

# 3 RL Formalism

## 3.1 Markov Decision Processes

A Markov Decision Process (**MDP**) is a mathematical framework for modeling stochastic decision-making processes where the outcomes are either random or deterministic (controlled by the agent). Formally, an MDP consists of a 5-tuple $(S, A, P, r, s_0)$ where each component is defined in the following way:

- $S$: the set of possible states, $s$.
- $A$: the set of possible actions, $a$.
- $P(s_{t+1}|s_t, a_t)$: the probability of transitioning from state, $s_t$, to state, $s_{t+1}$ after taking action $a_t$. ("Transition probability")
- $r$: the immediate reward, $r_t$, achieved after taking action, $a_t$, at time $t$ and transitioning from state, $s_t$, to $s_{t+1}$. This is governed by the reward function.
- $s_0$: the start state.

In the context of reinforcement learning, an MDP is extended to include the following hyper-parameters: ($\gamma$ = discount factor , $\epsilon$= $\epsilon$-greedy exploration, $\alpha$ = learning rate) which are defined as follows:

- **gamma**, $\gamma$: the discount factor. This hyper-parameter weighs the importance of future rewards relative to the current reward at time step, $t$. $\gamma \in [0, 1]$.
- **epsilon**, $\epsilon$: This hyper-parameter balances exploration and exploitation. With probability $\epsilon$, the agent chooses an action uniformly at random at each state. With probability $(1 - \epsilon)$, the agent chooses the action with the highest estimated reward. $\epsilon \in [0, 1]$.

- **alpha**, $\alpha$: the learning rate. This hyper-parameter governs how fast the agent learns from new information. By updating the rewards with new information based on $\alpha$, an RL algorithm eventually converges to the optimal policy. $\alpha \in [0, 1]$.

## 3.2 Optimal Policy:

In general, the goal of an RL agent is to maximize its expected cumulative long-term reward. Since agent-environment interactions potentially continue indefinitely, this can be formulated as the expected discounted sum of rewards over an infinite time horizon:

$$R_t = r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \cdots + \gamma^\infty \cdot r_\infty \tag{1}$$

where $R_t$ is the cumulative sum of the rewards, $r_t$ is the reward at time step, $t$, and $\gamma$ is the discount factor. Or more concisely as:

$$R_t = \sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i} \tag{2}$$

A **policy**, $\pi$, is simply a mapping from states, $s \in S$, and actions, $a \in S_a$, to a probability distribution, $\pi(a|s)$, which represents the probability of taking action, $a$ in state $s$. [**6**]

An **optimal policy, $\pi$*,** is a deterministic policy which maximizes the cumulative discounted reward, $R_t$, achieved at all states, $s \in S$, and thus produces the optimal value function (see below). It can be determined with algorithms such as value iteration, policy iteration and Q-learning. [**6**]

## 3.3 Value Functions:

Value functions provide an estimate of how beneficial it is for an RL agent to be in a given state, $s$. In general, the value of state, $s$, at time-step, $t$, while following policy, $\pi$, is the reward the agent expects to obtain by beginning in state $s$ and following policy $\pi$ thereafter. This is known as the **state-value function** of a policy and is illustrated by the following equation:

$$V^{\pi}(s) = \mathbb{E}\pi\left[R_t \mid S_t = s\right] = \mathbb{E}\pi\left[\sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i} \mid S_t = s\right] \tag{3}$$

where $\mathbb{E}\pi$ represents the expected value of a random variable assuming that an agent follows policy $\pi$, $R_t$ is the cumulative sum of rewards and $t$ is the time-step. The value for any terminal states is always zero. [**6**]

In a similar fashion, we can define the **action-value function** of a policy, $Q_{\pi}(s, a)$ as the reward the agent expects to achieve by beginning at state, $s$, taking action, $a$, and following policy $\pi$ thereafter.

$$Q^{\pi}(s, a) = \mathbb{E}\pi\left[R_t \mid S_t = s, A_t = a\right] = \mathbb{E}\pi\left[\sum_{i=0}^{\infty} \gamma^i \cdot r_{t+i} \mid S_t = s, A_t = a\right] \tag{4}$$

where the initial action is fixed, and $\mathbb{E}\pi$ represents the expected value of a random variable assuming that an agent follows policy $\pi$. The state-value function and action-value function can be approximated from experience. In an MDP, there is guaranteed to be at least one deterministic policy which maximizes the cumulative expected discounted reward at every state. The optimal action-value function can achieved by following the optimal policy, $\pi$*, where $\pi$* is associated with an optimal Q-value function in the following way:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \text{ where } \pi^*(s) = \arg\max_{a} Q^*(s,a) \tag{5}$$

$Q^*(s,a)$ can be reliably discovered using $Q$-learning which is defined below.

## 3.4 Model-free vs model-based RL:

In Reinforcement Learning, there are two main approaches to solving problems:

**Model-free:** In model-free learning, agents learn an optimal policy without ever constructing a model of their environment (a state transition function). Agents learn optimal policies which maximize the cumulative discounted reward by trial-and-error with algorithms that approximate the state transition function over time. Model-free methods tend to be more computationally efficient and more straightforward to implement but can require longer training times to perform well. Q-learning is a form of model-free reinforcement learning. [**6**]

**Model-based:** In model-based learning, agents construct and learn an explicit model of their environment and use it to learn the best actions to take via planning. [**9**] Once the model has been learned, the agent utilizes it to predict future states and their associated rewards. Model-based methods tend to be more sample efficient than model-free methods but demand greater computational resources to construct an explicit model of their environment. Policy iteration is a form of model-based learning. [**6**]

Both methods are effective at discovering optimal policies and the choice of method depends on the specific problem being solved. In the context of this thesis, model-free Q-learning was selected as the preferred method due to its simplicity and its use in previous work. [**1**]

## 3.5 Q-learning

Q-learning is a model-free reinforcement learning algorithm that is commonly used as an approach to calculating the optimal action-value function. [**1**] [**6**] Q-learning learns the optimal policy by computing Q-values for every state-action pair, $Q(s, a)$, in an $n$-by-$m$ Q-table where $n =$ the number of states and $m =$ the number of actions in that state. The Q-values are initialized either randomly or as zero (with all terminal states initialized to zero), and are updated using the **Bellman equation**:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left( r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right) \tag{6}$$

where each component is defined in the following way:

- $Q(s_t, a_t)$ is the Q-value associated with taking action $a$ in state $s$ at time $t$.
- $\alpha$: the learning rate. $\alpha \in [0, 1]$.
- $\gamma$: the discount factor. $\gamma \in [0, 1]$.
- $r_t$ is the immediate the reward received for taking action $a$ in state $s$ at time $t$.
- $\max_{a'} Q(s_{t+1}, a')$ is the maximum Q-value for all possible actions a' in the next state, $s_{t+1}$

In other words, the Q-values of each state-action pair $Q(s_t, a_t)$ are the expected cumulative reward the agent assumes it will receive if it takes action $a$ in state $s$ at time $t$ and then pursues the optimal policy thereafter. The agent begins without any knowledge of the environment and takes actions to explore it according to a specified exploration policy such as $\epsilon$-greedy exploration. As the agent interacts with its environment and gains more knowledge, the Bellman rule is guaranteed to produce Q-values which converge to the optimal action-value function and hence

represent the optimal policy. [**6**] The hyper-parameters ($\epsilon$ $\alpha$ and $\gamma$) are often tuned to their optimal values for the best performance, and it can often be beneficial to decay $\epsilon$ as the agent gains experience. [**6**]

An illustration of the $Q$-learning algorithm is provided below for even greater clarity (Adapted from [**6**]):

**Algorithm: Q-Learning**

Initialize $Q(s, a)$, arbitrarily $\forall s \in S$, $a \in A$, and $Q(\text{terminal states}, \cdot) = 0$

**Repeat (for each episode):**

Initialize $s$.

**Repeat (for each step of episode):**

With probability $\epsilon$: Choose $a \in A$ from $s$ uniformly at random.

With probability $1 - \epsilon$ Choose $a \in A$ from $s$ using policy derived from $Q$ ( $\epsilon$-greedy).

Take action $a$, observe $r$, $s'$.

$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)$.

$s \leftarrow s'$.

**until** $s$ is terminal.

In this demonstration, $s$ represents the agent's current state, $a$ represents the action taken by the agent in that state, $s'$ represents the state the agent transitions to and $r$ represents the reward received in that transition. $\alpha$ is the learning rate and $\gamma$ is the discount factor. The Q-learning algorithm updates the $Q$-values for each state-action pair using the Bellman rule and the exploration policy for choosing actions is an $\epsilon$-greedy strategy. Q-learning continues this process for every episode until a terminal state is reached.

### 3.6 Deep Q-Learning:

**Deep Q-learning** is a model-free RL algorithm which combines ideas from deep-learning and reinforcement learning to approximate an optimal policy. [**8**] A deep neural network is trained to learn the optimal Q-function $Q(s, a)$ where the inputs are states and actions and the output is the Q-value of the current state-action pair. Gradient descent is used to minimize the difference between the network's predicted Q-values and the actual Q-values computed with the Bellman equation. Upon training, an RL agent can then use the predicted Q-table to follow the optimal policy.

## 4 Reward Shaping

### 4.1 Optimal Reward Function

A **subjective reward function** is a transformation of the external reward actually achieved by an agent in its interaction with an environment to a subjective reward based on the mathematical form of some concept often taken from psychology. [**1**] [**13**] [**14**] This is analogous to the way in which humans "perceive the world" in a skewed way which may be detached from the reality of their environment. Just as human inductive biases can often result in beneficial behaviors, a subjective reward function can benefit an agent by steering it towards improved learning patterns.

Given this, an **optimal reward function** can be defined in the following way. Let $A$ be an RL agent and $R$ be the set of possible reward functions consisting of the original reward function and all subjective reward functions. A reward function $x \in R$ is defined as a mapping of states, $S$, to scalar values corresponding to the rewards at those states, $r_s$. A reward function results in a history, $H$, when it is

utilized to adapt agent $A$ to a sampled environment, $e$, within the distribution over Markov Decision Processes environments in which we desire our agent to perform optimally. Let $F(H)$ also be a fitness function which evaluates each history, $H$, by mapping it to a scalar value. Then, an optimal reward function is defined as the specific reward function, $x \in R$ which maximizes $F(H)$ over all Markov Decision Process environments in which we want our agent to perform optimally. [**2**] Generally, $F(H)$ can be defined as the average cumulative reward achieved by an agent over a certain number of episodes.

## 4.2 Aspiration Reward Function

As a large portion of my work will be extending the paper by Dubey et al. [**1**], it is worthwhile to devote some attention towards summarizing the model proposed in that work. In the paper, the authors develop a reward function equivalent to the following:

$$R = w_1 \cdot \text{objective} + w_2 \cdot \text{compare} \tag{7}$$

where objective $= r_t$ or the reward at time $t$, compare $= r_t - \rho$ where $\rho$ is the aspiration level which is hard-coded by the designer, and $[w_1, w_2]$ are the weights associated with the objective reward and the aspiration level. [**1**] Therefore the aspiration reward function can be written as follows:

$$R = w_1 \cdot r_t + w_2 \cdot (r_t - \rho) \tag{8}$$

The authors perform a exhaustive search over $w_1$,$w_2$ and $\rho$ and compare the fitness function of each of those models to determine the optimal value of the hyper-parameters. Other hyper-parameters such as the value of $\epsilon$ in $\epsilon$-greedy learning and the learning rate, $\alpha$, are optimized using grid search. [**1**]

In theory, the aspiration reward function should provide an agent with an additional, intrinsic "exploration incentive" since the agent evaluates the objective reward received at each time step, $t$, with its aspiration, $\rho$, and hence assigns negative values to all states which are not positive terminal states above the aspiration level. [**1**] This accelerates the agent's learning by encouraging the agent to explore new states and hence advantages the agent in non-stationary and sparse environments. [**1**]

## 4.3 Prospect Theory

The second reward function explored in this paper takes advantage of Kahneman and Tversky's work on human decision-making. To explain its effectiveness, I will give an overview of the risky choice theory which it is based on.

The theory of risky choice assumes that human decision-making can often be formalized as risky choice problems. A **choice problem** is just a mapping from a gamble pair $(A, B)$ to a probability, $P(A)$, where each gamble or **prospect** consists of a list of outcomes, $x_i$ and their associated probabilities, $p_i$ such that $\sum_{i=1}^{n} p_i = 1$. [**20**]

$$(A, B) \mapsto P(A) \tag{9}$$

A theory of risky choice seeks to describe a mapping which can predict the choices of human beings with greatest accuracy. In the context of reward shaping, it is useful to consider three theories of risky choice: expected value (EV), expected utility (EU) and prospect theory (PT)), each of which reduces to the previous theory via identity functions. [**20**] The following section will give a brief overview of the hierarchy of risky choice theories to provide context for the prospect theory reward function utilized in this thesis.

### 4.3.1 Expected Value

Expected value theory (EV) lies at the first level of the hiearchy. Expected value assigns a higher value to the gambles with the greatest expectation. It can be calculated as follows:

$$E[u] = \sum_i x_i \cdot p_i \tag{10}$$

where $x_i$ is the outcome and $p_i$ is its associated probability.

### 4.3.2 Expected Utility Theory

Expected Utility Theory is a classical theory of risky choice which was formerly the most widely accepted theory in psychology and economics. [**21**] [**22**] In contrast to expected value, expected utility theory defines a value function which subjectively weighs the values of different outcomes. The equation for expected utility is:

$$E[u] = \sum_{i=1}^{n} v(x_i) \cdot p_i \tag{11}$$

where $x_i$ is the outcome, $p_i$ is the associated probability and $v(.)$ is the value (or utility) function. The theory also assumes that people are risk averse according to the principle of diminishing sensitivity.[**21**] [**23**] In other words, they prefer a certain outcome with lower expected utility to an uncertain outcome with higher expected utility. This also implies that the utility function is concave. [**23**]

Expected Utility Theory reduces to expected value if we define the value function as the identity function $v(x) = x$. [**20**]

### 4.3.3 Prospect Theory

Prospect Theory is a descriptive theory of risky choice developed as a response to the weaknesses of expected utility theory. [**21**] [**24**] By modeling preferences using an S-shaped utility function which is steeper for losses than gains, it aims to explain the differences in the risk-averse and risk-seeking behavior of human agents by considering the effect of **loss aversion**. [**21**] [**23**] Loss aversion holds that losses scale faster than equivalent gains when humans make decisions in choice problems. [**23**]

In contrast to expected utility theory, prospect theory holds that people have a subjective view of probability as well as outcomes, and therefore defines a probability weighting function which allows for the under-weighting of low probabilities and the over-weighting of high ones. [**21**] [**24**] Prospect Theory models hence take the following form:

$$E[u] = \sum_{i=1}^{n} v(x_i) \cdot \pi(p_i)$$

Where $v(.)$ is the value function, $p_i$ is the probability of each outcome,$x_i$, and $\pi(.)$ is the probability weighing function. $\pi(.)$ is strictly increasing such that $\pi(0) = 0$ and $\pi(1) = 1$ and it can be different for gains and losses. [**25**] Since this thesis only focuses on model-free, Q-learning, the probability weighing function will not be relevant for our purposes since we do not have an explicit model of the environment's transition probabilities.

Prospect Theory reduces to Expected Utility Theory if we define the probability weighting function to be the identity function $\pi(x) = x$. [**20**]

### 4.3.4 Prospect Theory Reward Function:

As we have seen, prospect theory shapes the value function, $v(x_i)$, and the probability weighting function, $\pi(p_i)$, to reflect human psychology. [**19**] In the context of this thesis, we shall focus solely on the value function since we do not construct a probability transition function in model-free Q-learning.

The Prospect Theory reward function used in this thesis has two main features. First, it is S-shaped to reflect human responses to gains and losses. The function is concave for gains since humans tend to be risk averse when it comes to positive rewards. For instance, consider two gambles. (1) 100% chance of winning \$100 (2) 50% chance of winning \$200 and 50% chance of winning \$0. Most rational individuals would choose gamble (1) despite the fact that the expected value of the two gambles is equal. This implies a concave utility curve for gains since the marginal increase of an additional unit of gain decreases as the value of the gain increases. [**19**] On the other hand, the value function curve is convex for losses since humans tend to be risk-seeking when it comes to negative rewards. For instance, consider the following gamble: (1) 100% chance of losing \$100 and (2) 50% chance of losing \$200. Most rational individuals prefer gamble (2) and hence are risk-seeking when it comes to losses, confirming the notion that losses should by convex. [**19**]

In addition to an S-shaped utility curve, the Prospect Theory reward function is also steeper for losses than it is for gains to reflect the influence of loss aversion. In their work, Kahneman and Tversky theorized that people are more sensitive to losses than they are to gains. Consider a third gamble: (1) 50% chance of gaining \$100 and 50% chance of losing \$100 (2) no bet. Most people prefer (2) despite the fact that the expected value of the two options is equal. This implies that the subjective value of (1) is less than zero, proving that the subjective value of losses scales faster than that of gains.

Table 1: Hyper-parameter bounds

| | | | |
|---|---|---|---|
| Epsilon-greedy | $\epsilon \in [0, 1]$ | | |
| Aspiration | $\rho \in [0, 100]$ | $w1 \in [0, 1]$ | $w2 \in [0, 1]$ |
| Prospect | $\alpha \in [0, 1]$ | $\beta \in [0, 3]$ | $\lambda \in [1, 5]$ |

Given these features, a good form of the prospect theory reward function can be parameterized as a power function with separate parameters for gains and losses: [**19**]

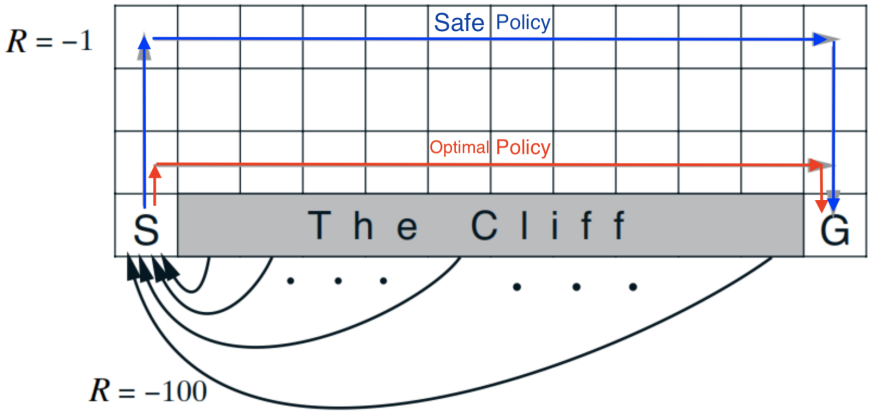
$$V(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-x)^{\beta} & \text{if } x < 0 \end{cases} \tag{12}$$

where $x$ is the magnitude of the outcome, $\alpha$ and $\beta$ are positive parameters that reflect the non-linear shapes of the value function for gains and losses,and $\lambda$ is the scaling coefficient of loss aversion. In their work, Kahneman and Tversky provided the values of $\alpha = .88$, $\beta = .88$, and $\lambda = 2.25$ based on experiments they ran on a sample of college students, [**19**] but we will use the optimized values obtained via Bayesian optimization.

# 5 Approach:

## 5.1 Training

To test the effect of the aspiration and Prospect Theory reward functions, I take a standard hyper-parameter optimization approach. Various hyper-parameter optimization methods were tested as the "meta-learners". Gradient-based methods such as stochastic gradient descent (SGD) promised the best performance but were unfortunately unusable due to the discrete nature of the Q-learning algorithm. Evo-




[6]

Figure 1: Cliff Walking environment from Sutton and Barton

lutionary algorithms were also a strong option but were ultimately dropped due to their slow performance. Bayesian optimization was therefore chosen to tune the aspiration and prospect theory hyper-parameters in all experiments due to its superior efficiency and straightforward implementation compared to other methods. For the aspiration model I tune the aspiration, $\rho$, and the weighting coefficients of the reward function $w_1, w_2$. For Prospect Theory I tune the non-linear exponents for gains and losses,[$\alpha$, $\beta$], and the linear scaling coefficient of the utility function for losses, $\lambda$. I run Bayesian optimization for 100-200 iterations on each tested environment with hyper-parameter bounds equal to those listed in Table 1. As a baseline I used $\epsilon$-greedy Q-Learning with an optimized value of $\epsilon$ and $\gamma = 0.99$. Additionally I set, $\gamma = 0.99$ and $\alpha = 0.1$ for all models.

## 5.2 Environments

I test the aspiration and Prospect Theory reward functions on 3 different categories of environments: (1) dense, (2) sparse and (3) complex. The layouts of the environments used to represent "dense" and "sparse" settings is inspired by the layouts proposed in Dubey et al. [**1**] For dense environments, I run the experiment on a

7-by-7 grid-world environment with 8 reward states: 1 positive "food" state with a reward of +10, and 6 negative "poison" states with negative rewards of -10. Since the effect of aspiration is starkest in stochastic environments, [**1**] I simulate stochasticity by assigning the agent a random start state anywhere in the first column and a random "food" state in one of the four corners. Multiple reward states were necessary since other environments were too simple for aspiration to have an effect. I run each experiment 50 times and set the number of episodes to 1000 and the max steps per episode to 1000. For sparse environments, I run the experiment on a 13-by-13 grid-world environment with 12 reward states: 1 positive "food" state with a reward of +10, and 11 negative "poison" states with negative rewards of -10 ad well as a 25-by-25 gridworld with the same reward structure. Stochasticity is incorporated in the same way by allowing for a random start position anywhere in the first column and a random "food" position in one of the four corners.

For complex environments, I run the experiment on the "Cliff Walking" environment provided in Sutton and Barto. The Cliff-Walking environment is illustrated in Figure 1 and consists of a 4-by-12 grid-world with a start state in position $(3, 0)$ and single goal state in position $(3, 11)$. $(3, j)$ for $1 \leq j \leq 10$ are all "cliff" states with associated rewards of -100. The optimal policy consists of traversing along the edge of the "cliff" to reach the goal state, with an associated reward of -12.0, while "safe policies" consist of other paths which avoid the cliff and reach the goal without necessarily traversing along the edge of the cliff. In this case, the environment is kept deterministic to see if the aspiration and Prospect Theory models display performance improvements in deterministic settings.

## 5.3 Evaluation

To evaluate the performance of each optimized agent on each environment setting, I compute the average cumulative reward achieved by the agent over 1000 episodes. This process is repeated for 50 runs and the average is then plotted along with 95% confidence intervals. The performance of the aspiration and Prospect Theory models is compared to the average cumulative reward achieved by a regular $\epsilon$-greedy Q-learning agent by calculating the performance improvement of each model in two ways. Firstly, the average cumulative reward obtained over the first 1000, first 200, first 100 and last 100 episodes is reported for each model in each environment setting. Secondly, the % improvement in the average cumulative reward achieved by each model over the $\epsilon$-greedy model is reported for the same episodes as above.

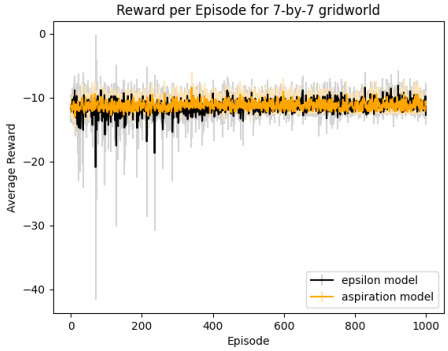
Additionally, I plot 2 types of heat maps for the sparse 25-by-25 gridworld and the Cliff-Walking environment to visualize the different behavior of the three agents. The first type of heat map shows the number of visits each agent makes to any state in the environment for the first 1000 episodes, and the second type shows the number of visits each agent makes to each of the "terminal" states in the environment for the first 100 episodes. These maps allow us to visualize how often each agent takes the optimal policy as well as how often each agent ends in the positive reward state. Finally, I run a t-test to compare the mean reward obtained by the Prospect Theory and aspiration models to the mean reward achieved by the regular $\epsilon$-greedy agent. This is computed for 50 runs of 200 episodes. I report the t-statistic and p-value for each model (aspiration vs $\epsilon$-greedy and Prospect Theory vs $\epsilon$-greedy) in order to determine whether the results are statistically significant.

# 6 Results:

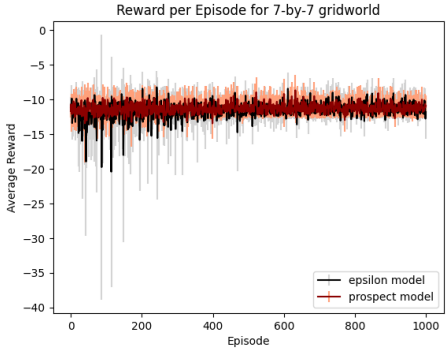
The two reward models proposed in this thesis exhibited promising performance relative to regular $\epsilon$-greedy Q-learning. The aspiration model and the Prospect Theory model beat $\epsilon$-greedy learning in both dense and sparse environments, with the most significant results achieved for the sparse, 25-by-25 gridworld. Significant results were also achieved in the complex environment setting, with the Prospect Theory and aspiration models outperforming baseline results reported for the Cliff-Walking environment in Sutton and Barto. [**6**] Furthermore, the experiments reveal the Prospect Theory model as superior to the aspiration model in every environment, suggesting that the challenge of defining "good' reward functions can be reduced to the question of finding good forms of the Prospect Theory reward function. Overall, the results suggest that optimal reward design could benefit from incorporating the assumptions of Prospect Theory into the reward functions of model-free RL agents, especially in sparse and complex environments where rewards are delayed and feedback is limited.

## 6.1 Dense environments

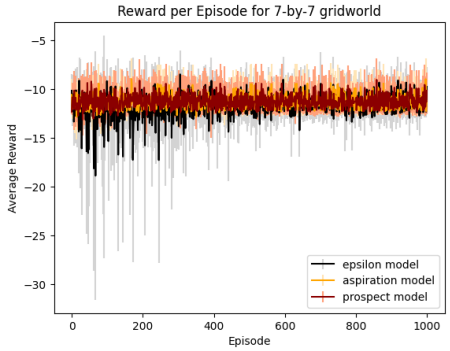
The results achieved in the dense environment setting were promising, and provided the first sign of the benefits of a Prospect Theory reward function. As seen in Figure 2, an agent's learning pattern was more efficient when it was equipped with both the aspiration and Prospect Theory reward functions. Figure 2(a) and 2(b) display the average cumulative reward achieved by the aspiration model and the Prospect model against the average cumulative reward achieved by a regular $\epsilon$-greedy Q-learning model. Clearly, the average cumulative reward achieved by the agents with optimal aspiration weights and optimal Prospect Theory hyperparameters was consistently higher than that obtained by the control model over the average of 50 runs of 1000 training episodes. Figure 2(c) plots the performance




(a) Epsilon vs Aspiration (7x7)




(b) Epsilon vs Prospect (7x7)




(c) Epsilon vs Aspiration vs Prospect (7x7)

Figure 2: Average cumulative reward per episode plotted with 95% confidence intervals. (a) Epsilon model vs Aspiration model in a 7-by-7 gridworld (b) Epsilon model vs Prospect model in a 7-by-7 gridworld (c) Epsilon model vs Aspiration model vs Prospect model in a 7-by-7 gridworld

of all three models to illustrate the superior performance of Prospect Theory over the aspiration model.

Since the plots can be hard to parse, I report some additional quantitative results illustrating the superior performance of the Prospect Theory and aspiration models. Table 2 displays the average cumulative reward achieved by each model over the first 100, 200 and 1000 episodes as well as the last 100 episodes, with the results of the Prospect Theory model highlighted in bold to illustrate the model's superior performance over both the aspiration and $\epsilon$-greedy models. This performance improvement is further illustrated in Table 3 which presents the percentage im-

provement in the performance of the aspiration and Prospect Theory models over the epsilon model.

To confirm the statistical significance of these results, Table 4 reports the t-statistic and p-value for the mean of 50 runs of 200 episodes of the aspiration model and the Prospect model against the mean of 50 runs of 200 episodes of the regular $\epsilon$-greedy model. The t-statistic measures the difference between the mean of the cumulative reward obtained by the two models. The values of 6.13 and 7.29 indicate that the mean reward for the aspiration and Prospect Theory models is significantly higher than the mean reward achieved by the $\epsilon$-greedy model. Furthermore, the p-values provide further evidence of the magnitude of the models' performance improvements by reporting the probability of reporting the observed difference in means if the null hypothesis was true. In our case, the null hypothesis holds that the performance of the two models and the $\epsilon$-greedy model are equal. The extremely low p-values reported ($p < 0.001$ for both models) indicate that there is strong evidence to reject the null hypothesis and hence we are able to conclude that the aspiration and Prospect models exhibit a statistically significant improvement in the average cumulative reward achieved in dense environments.

Overall, the results for dense environments were promising, confirming Dubey et al.'s results by showing that the aspiration model outperforms the $\epsilon$-greedy model, while also producing new results which show that the Prospect Theory model performs even better than the aspiration model. This provided the first sign of the Prospect model's superior performance as a form of intrinsic motivation to use in model-free RL.

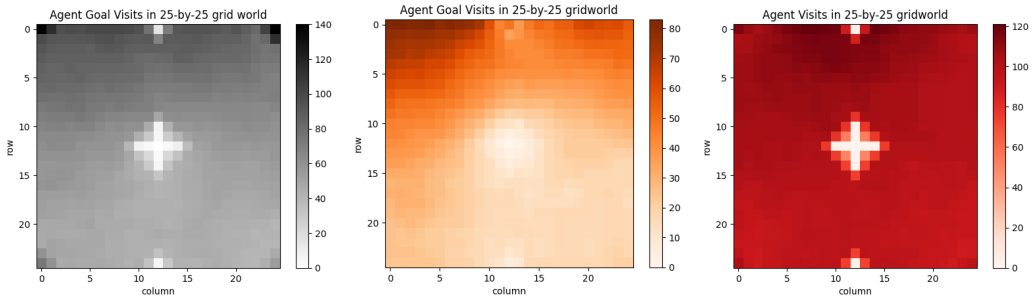
| Environment | | Epsilon-Greedy | Aspiration | Prospect |
|---|---|---|---|---|
| **7-by-7** | First 1000 | −11.6898 | −11.2826 | **-11.2388** |
| | First 200 | −12.2778 | −11.3848 | **-11.2816** |
| | First 100 | −12.4112 | −11.4016 | **-11.3090** |
| | Last 100 | −11.3302 | −11.2802 | **-11.1824** |
| **13-by-13** | First 1000 | −17.3886 | −15.9861 | **-15.9350** |
| | First 200 | −18.392 | 16.2546 | **-16.1914** |
| | First 100 | −18.6222 | −16.3256 | **-16.3146** |
| | Last 100 | −16.6356 | −15.9184 | **-15.7212** |
| **25-by-25** | First 1000 | −32.3982 | −27.2454 | **-26.4277** |
| | First 200 | −36.3790 | −28.9437 | **-27.6663** |
| | First 100 | −37.1716 | −29.5280 | **-28.2418** |
| | Last 100 | −29.0760 | −26.2372 | **-25.7854** |

Table 2: Results for sparse environments. Average cumulative reward displayed for the epsilon-greedy, aspiration and prospect models for the first 1000, first 200, first 100 and last 100 training episodes.

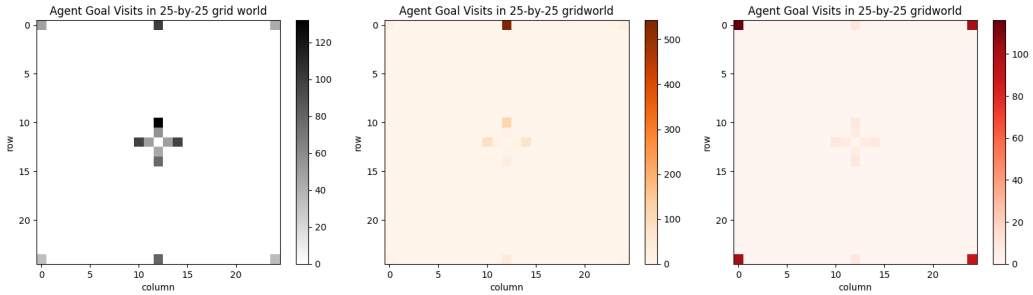
## 6.2 Sparse environments

The positive effects of the aspiration and Prospect Theory reward functions on the performance of model-free RL become even clearer in sparse environments. From the results obtained, we continue to observe the two broad trends reported above. On the one hand, the experiments confirm the results of Dubey et. al [**1**] as the aspiration model outperforms the epsilon model on every metric. The performance improvement is significant in many cases as can be seen in table 3. In the 13-by-13 gridworld, the aspiration models achieves an average cumulative reward which is 12.33% higher than the $\epsilon$-greedy model for the first 100 episodes and 8.07% higher for the first 1000 episodes. For the 25-by-25 gridworld the results are even more significant. The aspiration model performs 20.56% better for the first 100 episodes and 15.90% better for the first 1000 episodes.

Secondly, we contine to observe results which show that the Prospect Theory model outperforms both the $\epsilon$-greedy model and the aspiration model on all metrics in




(a) Epsilon Agent Visits (b) Aspiration Agent Visits (c) Prospect Agent Visits




(d) Epsilon Goal Visits (e) Aspiration Goal Visits (f) Prospect Goal Visits

Figure 3: Heat Maps displaying the number of times agents visited all states/terminal states (a) Epsilon model state visits for the first 1000 episodes (b) Aspiration model state visits for the first 1000 episodes c) Prospect model state visits for the first 1000 episodes (d) Epsilon model terminal state visits for the first 100 episodes (e) Aspiration model terminal state visits for the first 100 episodes (f) Prospect model terminal state visits for the first 100 episodes

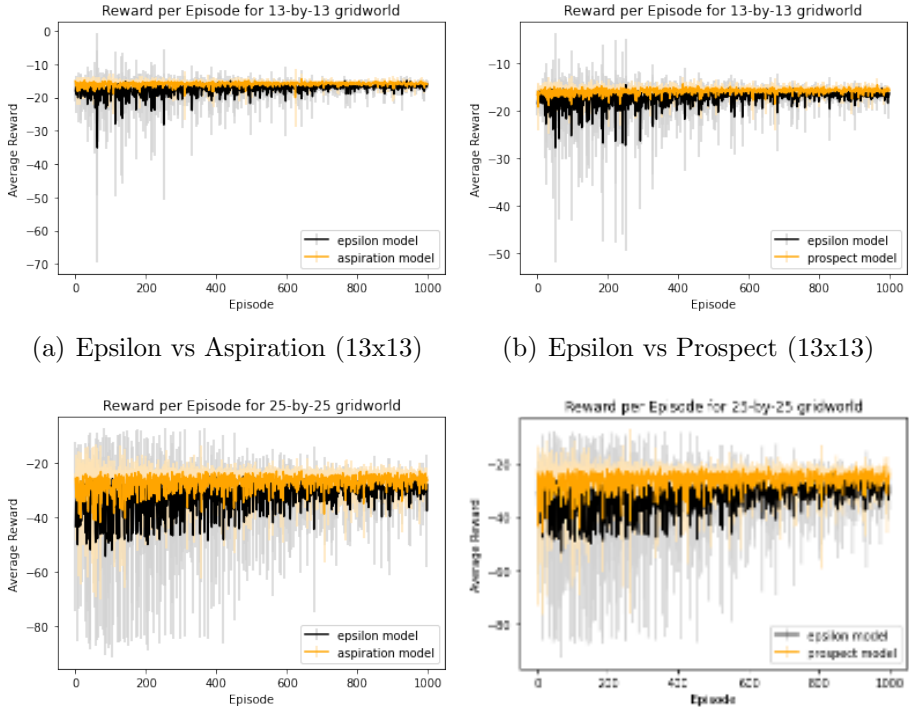
both experiments. The performance improvement is even more significant than that of the aspiration model. In the 13-by-13 grid-world environment, the Prospect model performs 8.36% (>8.07%) better than the $\epsilon$-greedy model for the first 1000 episodes and for the 25-by-25 grid-world the Prospect model performs 18.43% (> 15.90%) better for the first 1000 episodes and 24.02% (> 20.56%) for the first 100 episodes. Figure 3 allows us to clearly visualize exactly how the Prospect model is achieving superior performance. The figure plots heat maps reflecting the number of times agents of each model visit particular states in the 25-by-25 grid world. Figures 3 (a)-(c) display the distribution of visits over all possible states in the first 1000 episodes, while Figures 3 (d)-(f) display the distribution of visits to negative and positive "reward" states in the first 100 episodes. As can be seen, the

| **Environment** | | % Aspiration | % Prospect |
|---|---|---|---|
| **7-by-7** | First 1000 | 3.48% | 3.86% |
| | First 200 | 7.28% | 8.11% |
| | First 100 | 8.13% | 8.88% |
| | Last 100 | 0.44% | 1.30% |
| **13-by-13** | First 1000 | 8.07% | 8.36% |
| | First 200 | 11.62% | 11.96% |
| | First 100 | 12.33% | 12.39% |
| | Last 100 | 4.31% | 5.50% |
| **25-by-25** | First 1000 | 15.90% | 18.43% |
| | First 200 | 20.44% | 23.95% |
| | First 100 | 20.56% | 24.02% |
| | Last 100 | 9.76% | 11.32% |

Table 3: Results showing % improvement in the average cumulative reward achieved in sparse environments for the first 1000, first 200, first 100 and last 100 training episodes. The % Aspiration column displays the % increase in average cumulative reward achieved by the aspiration model and the % Prospect column display the % increase achieved by the prospect model.

Prospect Model is clearly avoiding the negative reward states more often than the other models in figure (c) (the negative states are located in the central white cross and the two white squares at the top and bottom of the grid) while also visiting the positive "food" states more frequently than the other models in figure (f) (the "food" states are randomly located in the four corners). A t-test is also performed and the p-values for the Prospect Theory and aspiration model are $p < 0.001$, once again allowing us to reject the null hypothesis and to conclude that the performance improvement exhibited by the two models is significant.

This is a very strong result which suggests that Kahneman and Tversky's value function is a good form of intrinsic motivation to use in sparse environments. It also suggests that a Prospect Theory reward function is preferable to an aspiration reward function in sparse environments, reflecting the strengths of applying Prospect Theory to RL architectures.




(a) Epsilon vs Aspiration (13x13)

(b) Epsilon vs Prospect (13x13)

(c) Epsilon vs Aspiration (25x25)

(d) Epsilon vs Prospect (25x25)

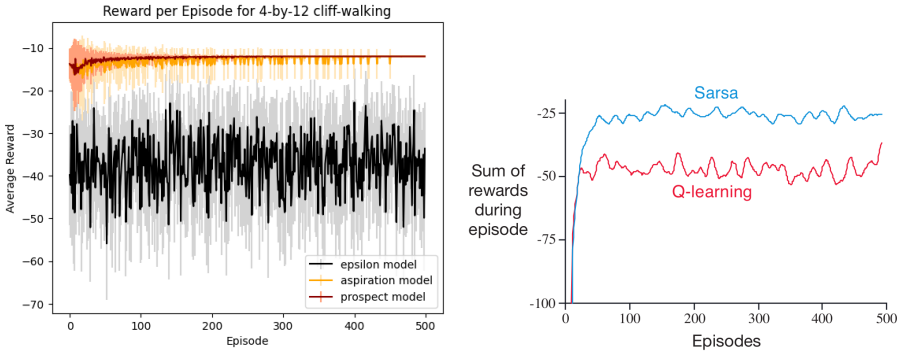
Figure 4: Average cumulative reward per episode plotted in black and orange and 95% confidence intervals plotted in light gray and light orange. (a) Epsilon model vs Aspiration model in a 13-by-13 gridworld (b) Epsilon model vs Prospect model in a 13-by-13 gridworld (c) Epsilon model vs Aspiration model in a 25-by-25 gridworld (d) Epsilon model vs Prospect model in a 25-by-25 gridworld.

## 6.3 Cliff-Walking

The results achieved in the complex, "Cliff-Walking" environment were also significant, suggesting that a Prospect Theory reward function advantages an agent in

| **Environment** | Model | t-statistic | p-value |
|---|---|---|---|
| **7-by-7** | aspiration | 6.14 | p <0.001 |
| | prospect | 7.29 | p <0.001 |
| **13-by-13** | aspiration | 52.24 | p <0.001 |
| | prospect | 59.34 | p <0.001 |
| **25-by-25** | aspiration | 95.01 | p <0.001 |
| | prospect | 103.40 | p <0.001 |

Table 4: Results of t-tests. t-statistics and p-values shown for each model in each environment setting

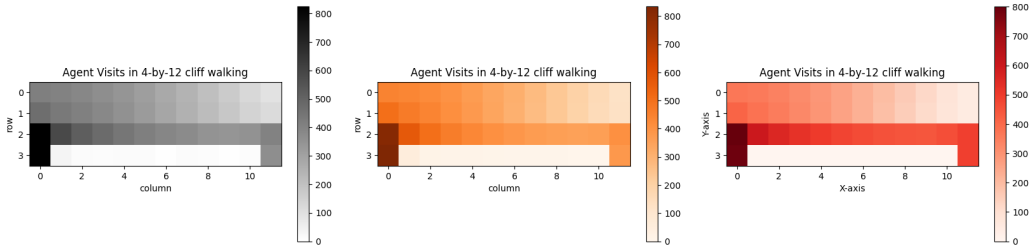



(a) Prospect Theory Experiment (b) Sutton and Barton Experiment

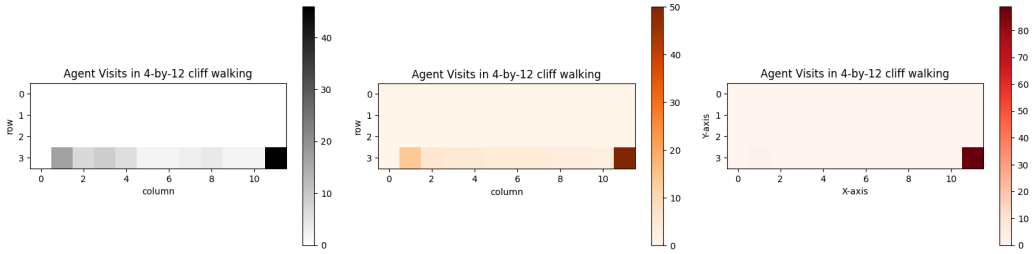
Figure 5: (a) Average cumulative reward per episode achieved by the $\epsilon$-greedy, aspiration and Prospect models plotted with 95% confidence intervals (b) Cumulative reward per episode achieved for $\epsilon$-greedy Q-learning and SARSA in Sutton and Barto

a wide variety of settings. The performance improvement of the Prospect Theory and aspiration models is illustrated in Figure 5. On the left, Figure 5 (a) displays the results obtained for RL agents with optimized Prospect Theory and aspiration model reward functions in red and orange respectively, as well as a regular, $\epsilon$-greedy RL agent with $\epsilon = 0.1$ in black. Meanwhile, Figure 5 (b) displays the results from Sutton and Barto for a Q-learning and SARSA agent both with $\epsilon = 0.1$. The results obtained for the Q-learning agent used in this thesis are similar to those achieved in Sutton and Barto, while the results for both the Prospect Theory and aspiration models are clearly superior to both the Q-learning and SARSA models.

The Prospect Theory model converges to the optimal policy after only 300 episodes, achieving an average cumulative reward of -12.0 which is the reward associated with traversing along the edge of the cliff to reach the goal. This beats both the Q-learning model which takes the optimal path more often than the SARSA model but achieves a lower average cumulative reward due to the $\epsilon$-greedy exploration policy which causes it to fall off the cliff randomly. [**6**] Sutton and Barto note that decaying the value of $\epsilon$ can cause the Q-learning agent to converge more rapidly, [**6**]




(a) Epsilon Agent Visits (b) Aspiration Agent Visits (c) Prospect Agent Visits




(d) Epsilon Goal Visits (e) Aspiration Goal Visits (f) Prospect Goal Visits

Figure 6: Heat Maps displaying the number of times agents visited all states/terminal states (a) Epsilon model state visits for the first 1000 episodes (b) Aspiration model state visits for the first 1000 episodes c) Prospect model state visits for the first 1000 episodes (d) Epsilon model terminal state visits for the first 100 episodes (e) Aspiration model terminal state visits for the first 100 episodes (f) Prospect model terminal state visits for the first 100 episodes

but the results obtained in this experiment reveal that no such decay is necessary in a Prospect Theory model which is able to learn the optimal policy independently.

Figure 6 plots a heat map identical to the one used for the 25-by-25 grid-world which allows us to visualize the Prospect Theory model's performance against the other two models. Figures 6 (a)-(c) clearly show that the distribution of visits of the Prospect Theory model are more tightly concentrated around the optimal policy of traversing along the edge of the "cliff" than both the aspiration model and the $\epsilon$-greedy model. Similarly, Figures 6 (d) - (f) show that the Prospect Theory model is almost guaranteed to always reach the "food" state during the first 100 episodes, whereas the aspiration model occassionally falls of the "cliff" and the $\epsilon$-greedy model is equally as likely to fall of the cliff as it is to reach the goal state.

We also confirm the statistical significance of these results as the p-values obtained were $p < 0.001$ for both the Prospect Theory model and the aspiration model.

Overall, the results obtained for the Cliff-Walking environment confirm results achieved in the dense and sparse settings. The $\epsilon$-greedy model is outperformed by the aspiration model which in turn is outperformed by the Prospect model. The results also show that the Prospect model is an effective model to use in deterministic and complex environments, proving that the model is a good form of intrinsic motivation to use in a wide variety of settings.

# 7 Discussion:

In this thesis, I set out to investigate two forms of intrinsic motivation (aspiration and Prospect Theory) which promised to provide good solutions to the optimal reward problem. By tuning the parameters of each reward function using Bayesian optimization on three different categories of environments (dense, sparse and complex), I achieve significant results which accomplish two broad goals.

Firstly, the results confirm the work of Dubey et al. since the aspiration model performs better than the $\epsilon$-greedy model in every environment setting. Such findings were anticipated as aspiration is hypothesized to provide the agent with an "exploration incentive" [**1**] which acts as the agent's intrinsic motivation to continue exploring an environment despite the fact that rewards are sparse and delayed. Intrinsic aspiration is shown to advantage agents which would otherwise encounter difficulty learning in an environment with limited external feedback and this thesis cements aspiration's place as an effective form of intrinsic motivation to use in multiple categories of environments.

Secondly, this thesis proposes a new optimal reward function based on the concept of Kahneman and Tversky's Prospect Theory which is shown to result in more efficient learning patterns than both $\epsilon$-greedy learning and Dubey et al.'s aspiration model in all environments tested. The performance improvement of the Prospect Model over aspiration and $\epsilon$-greedy Q-learning is greatest for sparse, stochastic environments such as the 25-by-25 grid-world used in this thesis, in which it outperforms $\epsilon$-greedy learning by nearly 25%. Moreover, the Prospect Theory reward function performs optimally in complex environments such as the "Cliff Walking" environment proposed in Sutton and Barto. This is demonstrated by the model's earlier convergence to the optimal policy after only 300 episodes as well as the higher frequency with which it takes the optimal policy which is visualized via heat maps displaying the distribution of agent visits.

I propose several possible hypotheses for the Prospect model's success over the $\epsilon$-greedy and aspiration models. Firstly, Prospect Theory reward functions account for **risk preferences** which are crucially relevant to many environments. For instance, in an environment where the risks of taking gambles are high, the reward function incentivizes the agent to avoid taking such actions. This could be highly relevant to agents operating in morally sensitive situations such as self-driving cars facing complex variants of the "trolley problem" in which they have to decide which humans should be killed or injured. Secondly, Prospect Theory reward functions account for **loss aversion**, incentivizing an agent to avoid losses more than gains. This likely has a significant effect on the Q-values of states adjacent to negative rewards, and thus dis-incentivizes the agent from repeating past mistakes.

Finally, the non-linear scaling of rewards which occurs in Prospect Theory allows for more **flexibility** than the linear aspiration component added to the reward function in aspiration models. This allows optimizers to discover improved forms

of subjective reward functions which result in higher average cumulative rewards. Additionally, unlike aspiration models, the Prospect Theory reward function distinguishes between positive and negative rewards and considers the magnitude of rewards, thus addressing two key limitations of aspiration.

While this thesis has yielded significant results, it still has several limitations which remain to be improved upon in future work. Firstly, the experiments are limited to Q-learning in relatively small environments, with the largest environment consisting of only 625 possible states. To generalize these results, the Prospect Theory model should be tested on larger, more complex environments such as Atari games. Additionally, an investigation of the effect of a Prospect Theory reward function in Deep reinforcement learning environments would also be worthwhile. The success of a Prospect Theory reward function in sparse environments suggests that Prospect Theory is likely to exhibit even greater performance improvements in large environments trained on deep Q-networks and the lack of current papers on the topic suggests that the potential upside of such an investigation could be considerable. Finally, more sophisticated optimization methods including gradient-based methods could be used in a Deep-RL setting to meta-learn the optimal parameters of the Prospect Theory reward function. In the meta-RL domain, the Prospect Theory parameters could also be dynamically updated to adapt to different environment settings and produce a more flexible model closer to an AGI.

In conclusion, this thesis introduces the Prospect Theory model as a solution to the optimal reward problem in model-free RL. The Prospect model's superior performance in dense, sparse and complex environments observed in these experiments suggests that the challenge of defining "good' reward functions can be reduced to the question of finding good forms of the Prospect Theory reward function. As such,

I propose that model-free RL agents' subjective reward functions should abide by the three assumptions of Prospect Theory: They should be (1) risk-averse to gains (2) risk-seeking to losses and (3) abide by the concept of loss aversion.

Centuries of philosophical debate on how to best manage our motivations ranging from Stoic appeals for moderation to Hobbesian calls for a restless pursuit of power have preceeded the work of this thesis. The progress made here in discovering a new optimal reward function is especially exciting when considered in the context of Silver et al.'s views on Artificial General Intelligence which suggest that a future, powerful RL agent could exhibit intelligence similar to that found in humans. If this hypothesis is true, the identification of Prospect Theory as an effective form of intrinsic motivation provided in this thesis could one day influence the design of such an AGI and thus bring us closer to solving intelligence.

## Code

The code used to complete this project is located at: github.com/lucasjirwin/Intrinsically-motivated-Prospect-Theory-RL-agent

# 8 References


[1] Dubey, Rachit, et al. "The Pursuit of Happiness: A Reinforcement Learning Perspective on Habituation and Comparisons.", 31 Dec. 2021

[2] Singh, Satinder, Richard L. Lewis, and Andrew G. Barto. Where do rewards come from. In Proceedings of the annual conference of the cognitive science society, pp. 2601-2606. Cognitive Science Society, 2009

[3] Sorg, J., Lewis, R. L. Singh, S. Reward design via online gradient ascent. Adv. Neural Inf. Process. Syst. 23, 2190–2198 (2010).

[4] Vilalta, Ricardo, and Youssef Drissi. A perspective view and survey of meta-learning. Artificial intelligence review 18, no. 2 (2002): 77-95.

[5] Zou, H., Ren, T., Yan, D., Su, H. Zhu, J. Reward shaping via meta-learning. arXiv preprint arXiv:1901.09330 (2019)

[6] Sutton, R. S., Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglous, I., King, H., Kumaran, D., Wierstra, D. Hassabis, D. (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–33. [arBML, MB, DGe]

[8] Lake, B., Ullman, T., Tenenbaum, J., Gershman, S. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, E253. doi:10.1017/S0140525X16001837

[9] T. Hospedales, A. Antoniou, P. Micaelli and A. Storkey, Meta-Learning in Neural Networks: A Survey, in IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 44, no. 09, pp. 5149-5169, 2022. doi: 10.1109/TPAMI.2021.3079209

[10] Hobbes, T. (2008). Leviathan (J. C. A. Gaskin, Ed.). Oxford University Press, pt. 1, ch. 11

[11] Epictetus. (2014). Discourses, fragments, handbook, Oxford University Press

[12] Trepel C, Fox CR, Poldrack RA. Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. Brain research. Cognitive Brain

Research. 2005 Apr;23(1)

[13] S. P. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In Advances in neural information processing systems, pages 1281–1288, 2004. .

[14] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. Autonomous Mental Development, IEEE Transactions on, 2(2):70–82, 2010.

[15] Kulkarni, T.D., Narasimhan, K., Saeedi, A., Tenenbaum, J.B. (2016). Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation. NIPS.

[16] Kulkarni, T.D., Saeedi, A., Gautam, S., Gershman, S.J. (2016). Deep Successor Reinforcement Learning. ArXiv, abs/1606.02396.

[17] Mohamed, S., Jimenez Rezende, D. (2015). Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. NIPS.

[18] Colas, C., Fournier, P., Chetouani, M., Sigaud, O. amp; Oudeyer, P.. (2019). CURIOUS: Intrinsically Motivated Modular Multi-Goal Reinforcement Learning, Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research

[19] Trepel, C., Fox, C. R., Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. Brain research. Cognitive brain research, 23(1), 34–50.

[20] Peterson, J., et al., (2021), Using large-scale experiments and machine learning to discover theories of human decision-making. Science 372, 1209-1214

[21] Kahneman, D., Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–291.

[22] Thaler, R. H., Tversky, A., Kahneman, D., Schwartz, A. (1997). The Effect of Myopia and Loss Aversion on Risk Taking: An Experimental Test. The Quarterly Journal of Economics, 112(2), 647–661.

[23] Shafir, E., Tversky, A. (1995). Decision making. In E. E. Smith D. N. Osherson (Eds.), Thinking: An invitation to cognitive science (pp. 77–100). The MIT Press.

[24] Tversky, A. Kahneman, D., (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and Uncertainty, 5(4), 297–323.

[25] Kőszegi, B., Rabin, M. (2006). A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics, 121(4), 1133–1165.

[26] Silver, D., Singh, S., Precup, D., Sutton, R. S. (2021). Reward is enough. Artificial Intelligence, 299, 103535.