A Prospect Theory approach to the sparse reward problem

Lucas James Irwin
ljirwin@princeton.edu
Department of Computer Science
Princeton University

Rachit Dubey
rdubey@princeton.edu
Department of Computer Science
Princeton University

Tom Griffiths
tomg@princeton.edu
Department of Computer Science
Princeton University

Abstract

The problem of sparsity in reinforcement learning has serious implications for the behavior of AI as researchers embark on the road towards developing artificial general intelligence. In situations where rewards are sparse, an agent encounters a dearth of essential feedback, hindering the acquisition of human-aligned optimal policies and resulting in suboptimal behavior. This dilemma prompts an fascinating question from a cognitive science perspective: what guiding principles should designers employ when formulating an agent's intrinsic motivations? Here, we explore the benefits of using an intrinsic reward function inspired by Kahneman's and Tversky's Prospect Theory which is risk-averse and treats positive and negative rewards differently. Our results reveal that the Prospect Theory reward function is a good form of intrinsic motivation to use in dense, sparse and risky environments, suggesting that future reward architectures would benefit from incorporating the cognitive insights of Prospect Theory into their models.

Keywords: Reinforcement Learning; Reward shaping; Intrinsic motivation; Prospect Theory

Introduction

Recent breakthroughs in Artificial Intelligence (AI) have elevated the significance of the cognitive question of what motivates an organism. As we develop larger and more capable AI models, it becomes increasingly important to scrutinize the principles motivating the decisions made by AI agents exhibiting potentially superintelligent abilities. The AI safety community has raised concerns over the significant risks highly capable models will pose to society, from unintentional harmful behavior to deception to the loss of control. (Amodei, Olah, Steinhardt, Christiano, Schulman & Mané, 2016) In light of this, researchers have stressed the importance of aligning models with human values to prevent potential catastrophes resulting from their deployment (Hendrycks, Mazeika & Woodside, 2023). Growing calls have also been made for a multidisciplinary approach to AI alignment, with particular emphasis placed on the value of insights from cognitive science and psychology.

Reinforcement Learning (RL) – a subfield of AI – provides a natural framework for studying the question of optimal motivation, since RL trains an artificially intelligent agent to learn to maximize its rewards via its interaction with an environment. (Sutton & Barto, 1998) It is also among the most powerful and promising AI methods used in applications ranging from DeepMind's AlphaGo which beat Go world champion, Lee Sedol, in 2019 to Reinforcement Learning from Human Feedback (RLHF) - the method used by OpenAI to align ChatGPT. (Christiano, Leike, Brown, Martic, Legg & Amodei, 2017) Despite this, it often suffers from the "problem of sparsity"; Sparse environments (where rewards are rare or delayed) provide insufficient feedback for extrinsically motivated RL agents to learn from. (Kulkarni, Narasimhan, Saeedi & Tenenbaum, 2016) While extrinsically-motivated agents learn well in dense environments, they do not possess a large enough exploration incentive to perform well in sparse settings. This compels them to take morally dubious routes towards achieving their ultimate goals, resulting in alarming behavior from a human perspective. (Langosco, Koch, Sharkey, Pfau, Orseau & Krueger, 2022)

We propose that this problem can be addressed by designing so-called "subjective" reward functions which provide intrinsic motivations based on principles taken from human psychology. Our work takes inspiration from Kahneman's and Tversky's Prospect Theory (PT) to align an RL agent with human intuitions through a method known as "reward shaping". Reward shaping consists of designing "good" intrinsic reward functions for RL agents as a means of discovering improved learning patterns. Instead of restricting an agent's learning to the external rewards it achieves in its environment, the designer imbues the agent with an intrinsic reward function which rewards and punishes intermediate behavior throughout the agent's learning process. Given PT's emergence as a leading psychological theory of human decision-making, we postulate that an investigation into the effects of a PT intrinsic reward function could yield useful insights which will in future be applicable to the alignment of highly capable models.

By comparing the performance of an RL agent equipped with a PT intrinsic reward function to a regular RL algorithm called ε-greedy Q-learning, we discover several important benefits of using PT to train RL agents. We train both models on 3 categories of environments (dense, sparse and risky) and observe the PT model outperforming regular RL in every setting. We also observe the PT model avoiding highly undesirable actions while training on the risky environment (Cliff Walking), and our findings strongly suggest that the cognitive insights of PT could prove useful to solving the sparse reward problem.

Background

Recent work in the RL field has begun to embrace the benefits of intrinsic motivation (Chentanez, Barto & Singh, 2004) (Sorg, Lewis & Singh, 2010) (Dubey, Griffiths & Dayan, 2021) and hence inspires the focus of our work. Sorg

et al. (2004) began to test agents with altered intrinsic reward functions to learn good functions which could generalize to multiple tasks. Chentanez et al. (2010) took advantage of evolutionary algorithms to learn optimal reward functions, laying the foundation for the distinction between intrinsic and extrinsic motivation. And Dubey et al. (2021) 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.

As the popularity of deep learning and Deep-Q-Networks has increased, so too has interest in applying intrinsic reward functions to Deep Reinforcement Learning. Mohamed et. al (2015) developed a method using convolutional networks and variational inference with an intrinsic motivation metric known as "empowerment" and Kulkarni et al. (2016) presented the hierarchical 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. Deep reinforcement learning (Deep-RL) has only just begun to utilize intrinsic motivation as a means of improving performance and aligning agents with human values. (Lake, Ullman, Tenenbaum, & Gershman, 2017) The potential upside of exploring such benefits in Deep-RL provides strong inspiration for our work in developing a Prospect Theory reward function as a solution to the sparse reward problem.

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. It forms the basis of Reinforcement Learning. 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: (γ = discount factor , ε= ε-greedy exploration, α = learning rate) which are defined as follows:

- γ: the discount factor. This hyper-parameter weighs the importance of future rewards relative to the current reward at time step, t. $\gamma \in [0,1]$.
- ε: This hyper-parameter balances exploration and exploitation. With probability ε, the agent chooses an action uniformly at random at each state. With probability $(1-\varepsilon)$, the agent chooses the action with the highest estimated reward. $\varepsilon \in [0,1]$.
- α: the learning rate. This hyper-parameter governs how fast the agent learns from new information. By updating the rewards with new information based on α, an RL algorithm eventually converges to the optimal policy. $\alpha \in [0,1]$.

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: (Sutton & Barto, 1998)

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

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

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

A **policy**, π, 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. (Sutton & Barto, 1998)

An **optimal policy, π*,** 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. (Sutton & Barto, 1998)

Reward Shaping

Reward Shaping is the process of using supplemental rewards to reward or punish an agent's intermediate behavior in sparse or complex environments. This can be achieved by using a subjective reward function alongside the objective external rewards received by an agent.

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. (Sorg, Lewis & Singh, 2010) 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.

Prospect Theory

The "Prospect Theory" utilized in our work is a theory from behavioral economics and psychology which falls under the broader theory of "risky choice". The theory of risky choice assumes that human decision-making can 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 $\sum_{i=1}^{n} p_i = 1$. (Peterson, Bourgin, Agrawal, Reichman & Griffiths, 2021)

$$(A,B) \mapsto P(A)$$

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, we will consider the risky choice theory known as Prospect Theory.

Prospect Theory is a descriptive theory of risky choice developed as a response to the weaknesses of Expected Utility Theory. (Kahneman & Tversky, 1979) 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**. (Kahneman & Tversky, 1992) Loss aversion holds that losses scale faster than equivalent gains when humans make decisions in choice problems.

Prospect Theory models 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. (Kahneman & Tversky, 1992)

A Prospect Theoretic Reward Function for Reinforcement Learning

The Prospect Theory reward function we introduce 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. 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.

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. (Kahneman & Tversky, 1992) 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.

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:

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

where x is the magnitude of the outcome, α and β are positive parameters that reflect the non-linear shapes of the value function for gains and losses,and λ 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. (Kahneman & Tversky, 1992)

Prospect theory Q-learning

Q-learning is a model-free reinforcement learning algorithm which we use in our work to train RL agents. 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**:

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

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; α is the learning rate. $\alpha \in [0,1]$; γ is 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 and $\max_{a'} Q(s_{t+1},a')$ is the maximum Q-value for all possible actions a' in the next state, s_{t+1}. (Sutton & Barto, 1998)

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

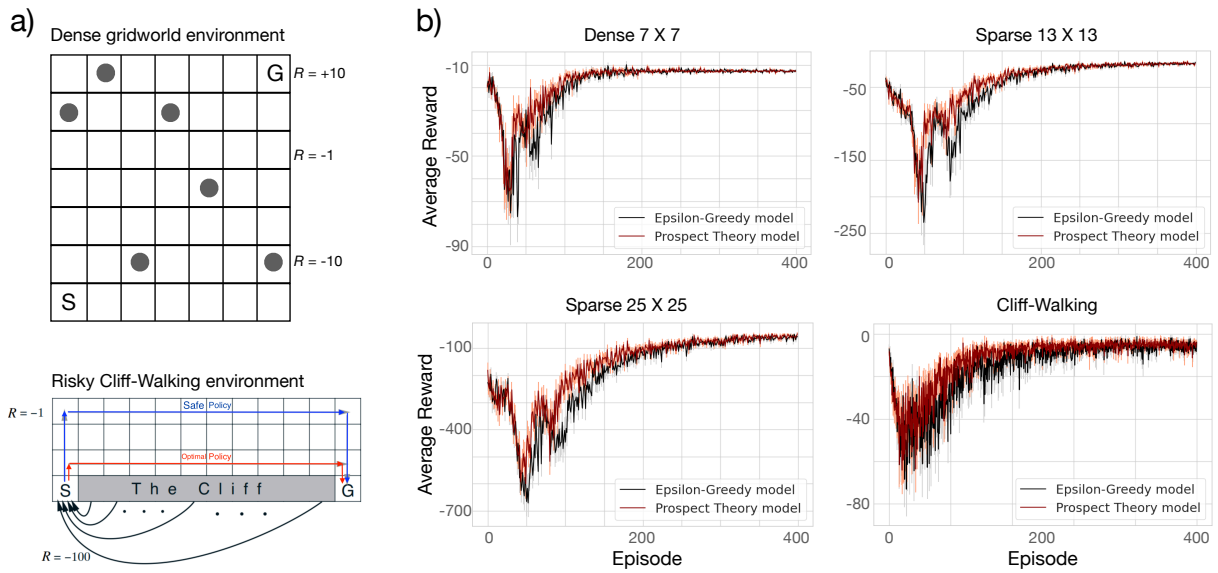


Figure 1: **(a) Environments used for our simulations.** Top: A sample gridworld environment where the goal state provides a reward of +10, poison states (shown in grey circles) provide a reward of -10, and all other states provide a reward of -1. Bottom: Cliff Walking environment adaptated from Sutton and Barto to simulate "risky" environments. **(b) Results.** Average cumulative reward per episode and 95% confidence intervals (in lighter color) plotted for the ε-greedy and the Prospect model for the different environments.

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. As the agent interacts with its environment and gains more knowledge, the Q-values are updated by the Bellman rule and are guaranteed to produce Q-values which represent the optimal policy.

In our Prospect theory model, the Q-updates are calculated by using a subjective reward, r'_t, which reflects the perceived psychological value of the reward under the assumptions of Prospect Theory. Specifically, $r'_t = r_t^{0.88}$ if $r_t > 0$ and $r'_t = -2.25 \cdot (-r_t)^{0.88}$ if $r_t < 0$. We set the learning rate, $\alpha = 0.1$, the discount factor, $\lambda = 0.99$ and epsilon, $\varepsilon = 0.1$ for all our models as these are commonly used values for these parameters. (Sutton & Barto, 1998)

Testing the Model

Environments

We test a regular ε-greedy, Q-learning model as a control and our Prospect Theory model on 3 different categories of environments: (1) dense, (2) sparse and (3) risky. For dense environments (refer to Figure 1(a)), we run the experiment on a 7-by-7 grid-world environment with the following reward states: 1 positive "food" state with a reward of +10, and 6 negative "poison" states with negative rewards of -10. All other states yield a reward of -1 and the agent reaches its intended state with 100% probability. We 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. We run each experiment 50 times and set the number of episodes to 400 and the max steps per episode to 1000. For sparse environments, we run the experiment on a 13-by-13 grid-world environment with the following reward states: 1 positive "food" state with a reward of +10, and 11 negative "poison" states with negative rewards of -10 as well as a 25-by-25 gridworld with the same reward structure (all other states provide a reward of -1). 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 risky environments, we run the experiment on the "Cliff Walking" environment introduced in Sutton and Barto. The Cliff-Walking environment is illustrated in Figure 1(a) 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 negative 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 Prospect Theory model displays performance improvements in deterministic settings.

Evaluation

To evaluate the performance of each optimized agent on each environment setting, we compute the average cumulative reward achieved by the agent over 400 episodes. This process is repeated for 50 runs and the average is then plotted along with 95% confidence intervals. To better visualize the be-

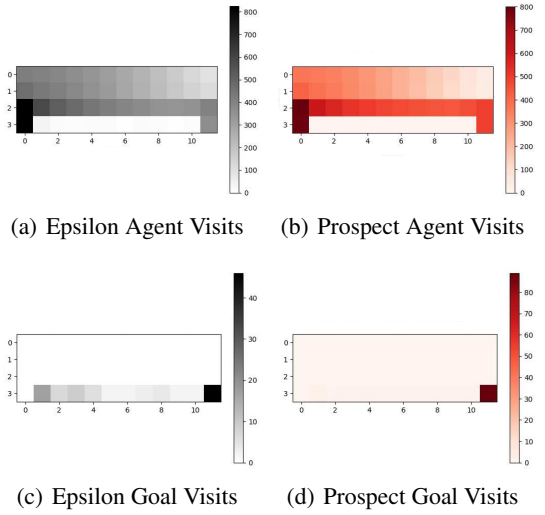
havior of each model, we plot 2 types of heat maps for the Cliff-Walking environment (refer to Figure 2). The first heat map shows the number of visits each agent makes to any state in the environment for the first 1000 episodes, and the second shows the number of visits each agent makes to each of the "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, we run two-sampled t-tests to compare the mean rewards obtained by the PT model to the mean reward achieved by the regular ε-greedy agent in each setting to confirm whether our results are statistically significant.

Results

The results achieved were promising, and provided clear signs of the benefits of using a Prospect Theory reward function. As seen in Figure 1(b), an agent's learning pattern was more efficient when it was equipped with the PT reward function. The average cumulative reward over 50 runs of the ε-greedy agent is plotted in black while the PT agent is plotted in red. We find that the average cumulative reward achieved by the agents endowed with the PT reward function was consistently higher than that obtained by the control model over 400 training episodes.

Since the plots can be hard to parse, we record some additional quantitative results illustrating the superior performance of the PT models. In the 13-by-13 grid-world environment, the PT model performs 8.36% better than the ε-greedy model for the first 1000 episodes and for the 25-by-25 grid-world the PT model performs 18.43% better for the first 1000 episodes and 24.02% for the first 100 episodes. To confirm the statistical significance of these results, we run a two-sample t-test for the mean of 50 runs of 200 episodes of the PT model against the mean of 50 runs of 200 episodes of the regular ε-greedy model. The performance improvement was significant. For the 13-by-13 sparse environment, the PT model (M = -62.7, SD = 33.7) performed better than the ε-greedy model (M = -77.8, SD = 41.1); t(-62) = -8.99, $p < 0.001$. This was also true for the dense environment (t(-27): -8.06, $p < 0.001$), the 25-by-25 sparse environment(t(-241) = -9.96, $p < 0.001$) as well as the Cliff Walking environment (t(-128) = 12.2, $p < 0.001$).

Figure 2 allows us to visualize exactly how the Prospect Theory reward function is achieving better performance. The figure plots heat maps reflecting the number of times agents of each model visit particular states in the 4-by-12 Cliff Walking environment. Figures 2 (a) - (b) clearly show that the distribution of visits of the PT model are more tightly concentrated around the optimal policy of traversing along the edge of the "cliff" than the ε-greedy model. Similarly, Figures 2 (c) - (d) show that the PT model is almost guaranteed to always reach the "food" state during the first 100 episodes, whereas the ε-greedy model occasionally falls off the "cliff". This reveals that the PT model avoids negative states far more reliably than regular RL models, indicating the model's potential suitability for use as an AI alignment tool.

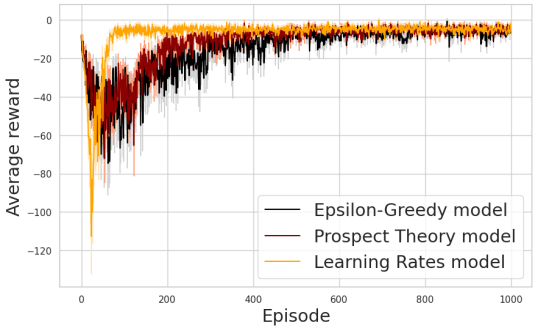


(a) Epsilon Agent Visits (b) Prospect Agent Visits

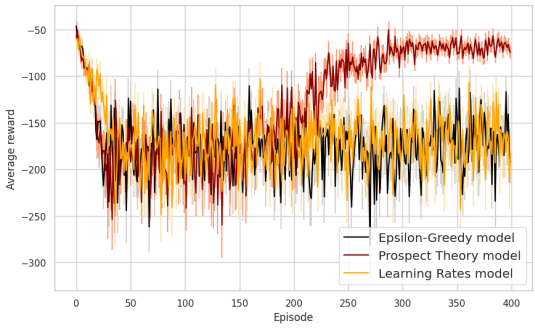
(c) Epsilon Goal Visits (d) Prospect Goal Visits

Figure 2: Heat Maps displaying the number of times agents visited all states/terminal states (a) Epsilon model state visits for the first 1000 episodes b) Prospect model state visits for the first 1000 episodes (c) Epsilon model terminal state visits for the first 100 episodes (d) Prospect model terminal state visits for the first 100 episodes

Finally, we compare our Prospect Theory reward function with an RL agent which uses separate Learning Rates for positive and negative rewards based on existing literature. (Gershman, 2015) The model hyper-parameters are identical to those used in the PT and ε-greedy models, with the only difference being that the model uses a learning rate, $\alpha = 0.5$ if $r_t < 0$ and $\alpha = 0.1$ if $r_t >= 0$. In other words, the model simulates the effect of loss aversion by assigning greater weight to negative Q-updates. In the context of the Cliff Walking environment, we ran a two sample t-test and observed a notable performance improvement of the Learning Rates model (M = -7.41, SD = 11.2) over the PT model (M = -12.0, SD = 11.9); t(-7.41) = -8.86, $p < 0.001$ when the reward states states associated with the "Cliff" were given a moderately negative reward (r = -10). However, when the "Cliff" reward states were assigned a highly negative reward (r = -100), the PT model (M = -128.1, SD = 54.6) exhibited better performance compared to the Learning rates model (M = -166.8, SD = 31.9); t(-128) = 12.2, $p < 0.001$. This observation suggests that the PT model may be particularly effective in scenarios with extremely negative outcomes, further cementing its potential suitability as an AI alignment tool in risky environments.



(a) Cliff Reward = - 10



(b) Cliff Reward = - 100

Figure 3: Average cumulative reward per episode plotted in black, orange, and red and 95% confidence intervals plotted in light gray, light orange and pink for the ε-greedy, PT and LR models in the Cliff Walking environment. (a) ε-greedy vs PT vs LR for cliff reward states = -10. (b) ε-greedy vs PT vs LR for cliff reward states = -100.

Discussion

In this work, we explore the benefits of using a Prospect Theoretic reward function to address the problem of sparsity in RL. By comparing the performance of an RL agent equipped with a PT reward function on dense, sparse and risky environments to a regular ε-greedy Q-learning agent, we discover several key benefits of using PT as an agent's intrinsic motivation. Firstly, we find that PT improves an RL agent's performance in a wide variety of settings. The performance improvement of the PT Model is greatest for sparse, stochastic environments such as the 25-by-25 grid-world, in which it outperforms ε-greedy Q-learning learning by nearly 25%.

In addition to exhibiting superior performance, we find that our PT reward function also motivates human-aligned behavior in RL agents. In contrast to regular ε-greedy Q-learning, PT incentivizes an RL agent to avoid catastrophic and morally unacceptable actions throughout its learning process. This strongly indicates the suitability of Prospect Theory reward functions as a potential approach to aligning future highly capable AI models.

We propose several possible cognitive hypotheses for our PT model's success over ε-greedy learning. 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 lethal autonomous weapons or self-driving cars. Secondly, PT 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 PT allows for more **flexibility**, enabling optimizers to discover improved forms of subjective reward functions which result in higher average cumulative rewards. In particular, by distinguishing between positive and negative rewards Prospect Theory reward functions mirror complex human intuitions on risk, and thus exhibit greater performance and alignment with human values.

While our work 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 reward function should be tested on larger, more complex environments such as Atari games. Additonally, the beneficial effects of Prospect Theory for the performance and alignment of RL models can be more efficiently achieved in some cases by using a model with separate learning rates. This motivates further study into applying the PT insights of risk sensitivity and loss aversion to reward shaping using alternative methods.

In conclusion, our work introduces the Prospect Theory reward function as a potential reward-shaping method in Reinforcement Learning. The Prospect Theory model's superior performance in dense, sparse and risky 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, we propose that 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. Though these cognitive insights may not be applicable to every use case, our work has clearly shown that they are useful in aligning an AI agent's actions with human intuitions in risky environments with highly negative outcomes. As AI researchers develop larger and more capable AI models over the next few decades, we hypothesize that the field will likely benefit from incorporating the cognitive insights described here into their models.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Chentanez, N., Barto, A., & Singh, S. (2004). Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, *17*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *International conference on machine learning* (pp. 12004–12019).

Dubey, R., Griffiths, T. L., & Dayan, P. (2022). The pursuit of happiness: A reinforcement learning perspective on habituation and comparisons. *PLoS computational biology*, *18*(8), e1010316.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic bulletin & review*, *22*, 1320–1327.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 278.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., & Tenenbaum, J. (2016). Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, *29*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Mohamed, S., & Jimenez Rezende, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, *28*.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Sorg, J., Lewis, R. L., & Singh, S. (2010). Reward design via online gradient ascent. *Advances in Neural Information Processing Systems*, *23*.

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

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.