Using Prospect Theory to Improve Learning from Sparse Rewards

Anonymous CogSci submission

Abstract

Sparse rewards present a significant challenge for agents trained via reinforcement learning. When rewards are sparse, an agent is missing essential feedback, hindering the acquisition of effective policies and resulting in sub-optimal behavior. This problem prompts a fascinating question from a cognitive science perspective: what guiding principles should designers employ when formulating an agent's intrinsic motivations? 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 could benefit from incorporating insights from human cognition 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 classic question of what motivates an organism. As we develop more capable AI agents, it becomes increasingly important to scrutinize the principles motivating the decisions made by those agents. The AI safety community has raised concerns over the significant risks highly capable artificial agents will pose to society, from unintentional harmful behavior, to deception, to the loss of control (Amodei et al., 2016; Saunders, Sastry, Stuhlmueller, & Evans, 2017). 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 (Amodei et al., 2016).

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 maximize its rewards via its interaction with an environment (Sutton & Barto, 2018). It is also among the most powerful AI methods used in applications ranging from DeepMind's AlphaGo (Silver et al., 2017) to the Reinforcement Learning from Human Feedback method used to align chatbots (Christiano et al., 2017). Despite this, RL 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 effectively (Kulkarni, Narasimhan, Saeedi, & Tenenbaum, 2016; Dubey, Agrawal, Pathak, Griffiths, & Efros, 2018; Botvinick et al., 2019). 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 alternative routes towards achieving their ultimate goals, sometimes resulting in alarming behavior from a human perspective (Di Langosco, Koch, Sharkey, Pfau, & Krueger, 2022).

One way to address the problem of sparsity is by designing so-called "subjective" reward functions (Ng, Harada, & Russell, 1999; Singh, Lewis, & Barto, 2009; Hadfield-Menell, Milli, Abbeel, Russell, & Dragan, 2017). Instead of restricting an agent's learning to the external rewards it achieves in its environment, the designer imbues the agent with a system of intrinsic rewards that are used by the reinforcement learning algorithm. These intrinsic rewards can motivate adaptive behavior that ultimately leads agents to learn effective policies despite sparse external rewards.

Designing good subjective reward functions can be challenging. We follow recent work that bases subjective reward functions on principles taken from human psychology (Singh et al., 2009; Pathak, Agrawal, Efros, & Darrell, 2017; Pathak, Gandhi, & Gupta, 2019), taking inspiration from Prospect Theory (PT) (Tverksy & Kahneman, 1979). Given Prospect Theory's status as a leading psychological theory of human decision-making, we postulate that basing a subjective reward function on its principles could yield useful insights applicable to developing better-aligned AI agents.

In particular, Prospect Theory asserts that human decision-makers treat gains and losses differently. People tend to weigh losses heavier than gains, a factor that affects their attitudes towards risk in these different settings. We explore the consequences of implementing such a difference in RL agents, defining a subjective reward function that exactly duplicates the form that Tverksy and Kahneman (1979) found effectively captured human behavior.

By comparing the performance of an RL agent equipped with a PT intrinsic reward function to a regular RL algorithm based on ε-greedy Q-learning (Watkins & Dayan, 1992), we discover several benefits of using PT to train RL agents. We train both models on three 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. Our findings suggest that the cognitive insights of PT could prove useful in making progress on the sparse reward problem and preventing risky behaviors in AI systems, with obvious consequences for improving alignment and safety.

Background

Recent work within RL has begun to embrace the benefits of intrinsic motivation (Chentanez, Barto, & Singh, 2004; Singh et al., 2009; Dubey, Griffiths, & Dayan, 2022). Exploration of subjective reward functions has been shown to

be valuable not just for simple RL settings, but also for modern deep reinforcement learning algorithms (Mohamed & Jimenez Rezende, 2015; Kulkarni et al., 2016; Pathak et al., 2017). For instance, 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 Deep Q Network (h-DQN) – a framework which allowed for the efficient exploration of complicated environments by dividing the learning process into two value functions which learnt intrinsic goals and extrinsic goals independently. The potential upside of exploring such benefits in RL provides inspiration for our work exploring the use of a Prospect-Theoretic reward function as a potential pathway towards the sparse reward problem. In what follows, we will briefly summarize the key technical ideas behind RL and Prospect Theory.

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 (for a more detailed treatment, we refer the readers to Sutton and Barto, 2018). Formally, an MDP consists of a 5-tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where each component is defined in the following way:

- $\mathcal{S}$: the set of possible states, s.
- $\mathcal{A}$: 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, also known as the 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.
- γ: the discount factor, weighing the importance of future rewards relative to the current reward at time step, t, with $\gamma \in [0,1]$.

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:

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

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:

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

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

An **optimal policy, π*,** is a deterministic policy which maximizes the cumulative discounted reward, R_t, achieved at all states, $s \in \mathcal{S}$, and thus produces the optimal value function (see below). It can be determined with algorithms such as value iteration, policy iteration and Q-learning (see Sutton & Barto, 2018). In this paper we focus on Q-learning, which is extremely widely used both as a model of human learning and in the computer science literature.

Q-learning

Q-learning is a model-free reinforcement learning algorithm commonly used to train RL agents (Watkins & Dayan, 1992). 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 at 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; r_t is the immediate 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}. 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. A commonly used exploration policy is the **ε-greedy** policy, where $\varepsilon \in [0,1]$ is a parameter which balances exploration and exploitation. Under this policy, with probability ε the agent chooses an action uniformly at random at each state and with probability $(1-\varepsilon)$ the agent chooses the action with the highest estimated reward.

Provided sufficient opportunities for exploration, as provided by the ε-greedy strategy, it is possible to prove guarantees for the outcome of Q-learning. Specifically, 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 (Sutton & Barto, 2018). However, it may take a very long time to find this optimal policy, particularly in environments where rewards are sparse.

Subjective Reward Functions

When rewards are sparse reinforcement learning algorithms can be very slow to find an optimal policy. One way to address this is to use a **subjective reward function** either in place of or as a supplement to the objective external rewards received by an agent. This subjective reward function is typically a transformation of the external reward actually achieved by an agent in its interaction with an environment

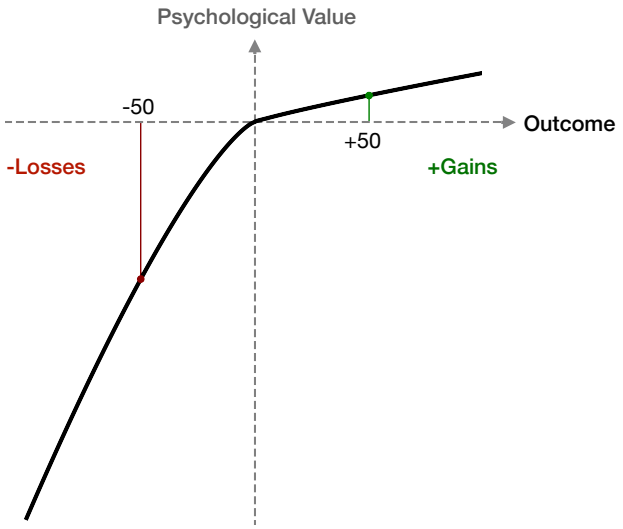

Figure 1: **Prospect Theory value function.** Illustration of the Prospect Theory value function which captures how people treat losses and gains differently. According to this theory, a $50 loss hurts more than the joy of a $50 gain.

(Ng et al., 1999; Singh et al., 2009; Sorg, Lewis, & Singh, 2010). This is analogous to the way in which humans experience rewards that may be detached from the objective rewards in their environment. Such subjective reward functions can benefit an agent by steering it towards improved policies.

Prospect Theory

Prospect Theory is an account of how people make decisions in a "risky choice" setting where they have to evaluate gambles (Tverksy & Kahneman, 1979). 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 $\sum_{i=1}^{n} p_i = 1$.

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

A theory of risky choice seeks to describe a mapping which can predict the choices of human beings with greatest accuracy.

Prospect Theory was developed as a response to the weaknesses of Expected Utility Theory (Von Neumann & Morgenstern, 1944). It expresses the value of a gamble as

$$V = \sum_{i=1}^{n} v(x_i) \cdot \pi(p_i) \tag{3}$$

where $v(.)$ is the subjective value function, x_i and outcome, p_i is the probability of that outcome, and $\pi(.)$ is the probability weighting function. $\pi(.)$ is strictly increasing such that $\pi(0) = 0$ and $\pi(1) = 1$. Our focus here will be on the subjective value function, but we anticipate that it would also be worthwhile to explore different probability weighting functions in reinforcement learning in future work.

Tverksy and Kahneman (1979) found that human choices were well-modeled by using an S-shaped subjective value function (refer to Figure 1). 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 people 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 subjective value function is convex for losses since humans tend to be **risk-seeking** when it comes to losses. For instance, consider the following gamble: (1) 100% chance of losing $100 and (2) 50% chance of losing $200. Most people prefer gamble (2) and hence are risk-seeking when it comes to losses, confirming the notion that losses should by convex.

In addition, the subjective value function is steeper for losses than gains. This makes it possible to explain people's different attitude towards the value of losses and gains. For example, consider the gamble (1) 50% chance of gaining $101 and (2) 50% chance of losing $100. Most people would prefer not to take this gamble, even though the expected yield is greater than zero. This is a form of **loss aversion**. Loss aversion can be captured by having the subjective value function for losses scale faster than for equivalent gains.

Given these features, a common form of the subjective value function assumed in Prospect Theory is a parameterized power function with separate parameters for gains and losses:

$$v(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0 \\ -\lambda(-x)^{\beta} & \text{if } x < 0 \end{cases} \tag{4}$$

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. Tversky and Kahneman (1979) provided the values of $\alpha = .88$, $\beta = .88$, and $\lambda = 2.25$ based on experiments they ran on a sample of college students.

Prospect Theory and Reinforcement Learning

Prospect Theory offers an empirically successful framework for explaining human decision-making. In doing so, it specifies how people seem to transform external rewards – in many experiments, money – into subjective rewards. This suggests the possibility that we might be able to design an effective subjective reward function for reinforcement learning based on principles from Prospect Theory. In the remainder of the paper we explore how using ideas from Prospect Theory to define a subjective reward function affects the performance of RL agents.

Prospect-Theoretic Q-learning

We use Q-learning with ε-greedy exploration as a basic model and investigate the consequences of simply modifying the reward function. In our Prospect Theory model, the Q-updates

are calculated using a subjective reward function, r'_t, which reflects the perceived psychological value of the reward under the function shown in Equation 4. 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 $\alpha = 0.1$, $\lambda = 0.99$, and $\varepsilon = 0.1$ for all our models as these are commonly used values for these parameters (Sutton & Barto, 2018).

Environments

We test our Prospect Theory model against a regular ε-greedy, Q-learning model as a control on three different categories of environments: (1) dense, (2) sparse and (3) risky. For dense environments (refer to Figure 2), 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 experiment with the "Cliff Walking" environment studied in Sutton and Barto (2018). The Cliff-Walking environment is illustrated in Figure 2 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 behavior of each model, we plot two types of heat maps for the Cliff-Walking environment (refer to Figure 4). 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

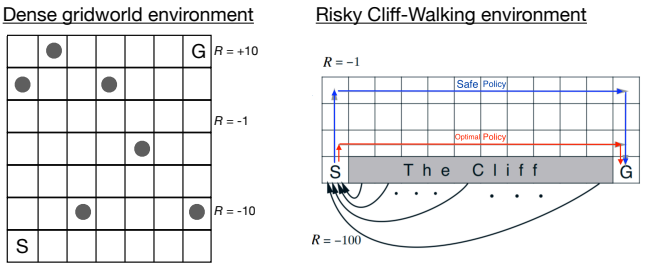

Figure 2: **Testing Prospect Theory in different environments.** Left: A sample gridworld environment where the goal state provides a reward of +10, poison states (shown in grey circles) provide a reward of -10, and all other states provide a reward of -1. Right: The Cliff Walking environment adaptated from Sutton and Barto (2018) to simulate "risky" environments.

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-sample 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 test whether our results are statistically significant.

Results

The results showed the benefits of using Prospect Theory to define a subjective reward function. As seen in Figure 3, 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 control 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 agents over 400 training episodes.

Since the plots can be hard to parse, we record some additional quantitative results illustrating the better performance of the PT agents. In the 13-by-13 grid-world environment, the PT agents performs 8.36% better than the control agents 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 ran 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 control model. The performance improvement was statistically 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 4 allows us to visualize exactly how the Prospect Theory reward function is achieving better performance.

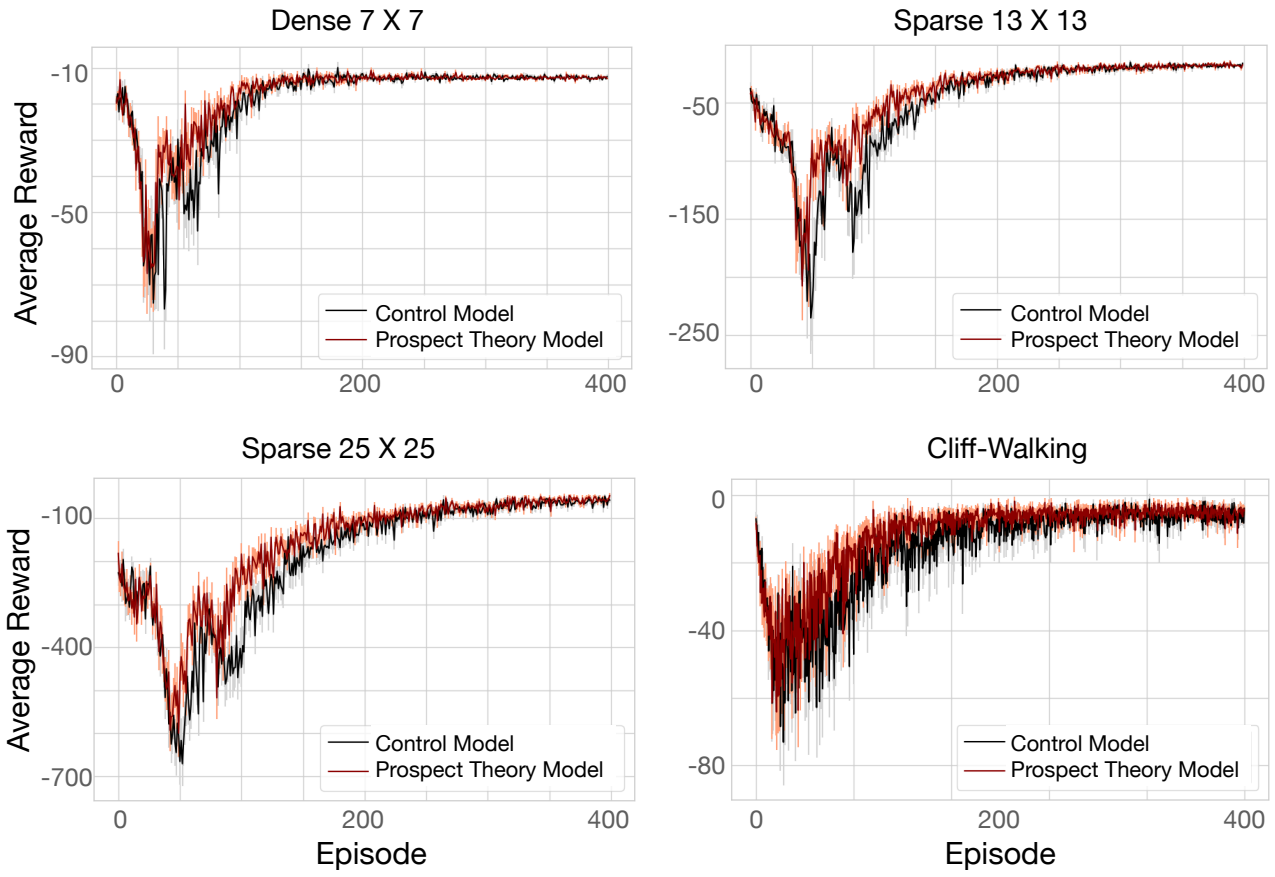

Figure 3: **Results.** Average cumulative reward per episode and 95% confidence intervals (in lighter color) plotted for the control and the Prospect Theory model for the different environments. The Prospect Theory reward function helps the Q-learning agent to converge faster and hence achieve better performance in all the environments.

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 4 (a) and (b) 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 control model. Similarly, Figures 4 (c) and (d) show that the PT model is almost guaranteed to always reach the "food" state during the first 100 episodes, whereas the control model occasionally falls off the "cliff". This reveals that the PT model avoids negative states more reliably than regular RL models.

Discussion

We explored the benefits of using a Prospect-Theoretic subjective 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 better performance, we find that our PT reward function also potentially motivates more human-aligned behavior in RL agents. In contrast to regular ε-greedy Q-learning, PT incentivizes an RL agent to avoid catastrophic and risky actions throughout its learning process. This suggests that Prospect Theory could be used to design reward functions that could better align future AI models.

Cognitive Interpretations

One advantage of using ideas from psychological research to inform the design of subjective reward functions is that it provides us with a natural way to interpret changes in the behavior of the resulting RL agents in terms of cognitive psychology. To this end, we can identify several possible cognitive hypotheses for our PT model's success over ε-greedy learning. First, the Prospect Theory subjective reward functions accounts 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 in-

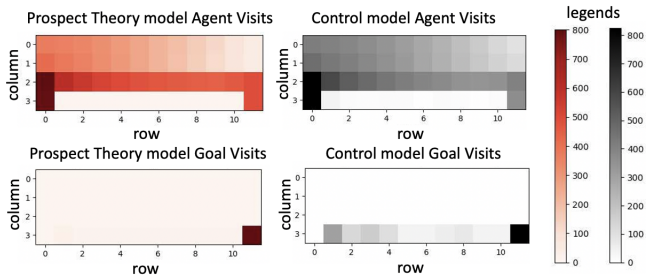

Figure 4: Heat Maps displaying the number of times agents visited all states/terminal states. (Top left) Prospect Theory model state visits for the first 1000 episodes. (Top right) Control model state visits for the first 1000 episodes. (Bottom left) Prospect Theory model terminal state visits for the first 100 episodes. (Bottom right) Control model terminal state visits for the first 100 episodes.

centivizes 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. Second, the PT subjective reward function accounts 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 disincentivizes the agent from repeating past mistakes. Finally, the non-linear scaling of rewards which occurs in PT allows for more **flexibility**, potentially enabling optimizers to discover improved forms of subjective reward functions which result in higher average cumulative rewards.

Rationality and Reward

Because it was originally proposed as an alternative to Expected Utility Theory, Prospect Theory is associated with explaining irrational decision-making on the part of humans. However, our results suggest that adopting a reward function of this form can be beneficial in supporting efficient reinforcement learning. These results are consistent with recent arguments that there may be a rational justification for the form of the subjective value function assumed in Prospect Theory (Stewart, Chater, & Brown, 2006; Bhui & Gershman, 2018). In particular, they provide a complementary approach to assessing the rationality of this form of value function, focused on what leads to highest learning performance rather than what best captures the statistics of the choice environment.

Connections to Other Models

While our approach achieves risk aversion by modifying the reward function, an alternative approach is to use separate learning rates for positive and negative rewards (Gershman, 2015; Palminteri & Lebreton, 2022). Prior work has shown the benefits of such an approach, with model-free reinforcement learning faster when the learning rate is higher for negative rewards (Shen, Tobia, Sommer, & Obermayer, 2014; Gershman, 2015). Our work complements this body of literature, showing how a reward function based on Prospect Theory can help learn faster from negative rewards and achieve risk aversion.

While modifying the scale of the reward function and modifying the learning rate can have analogous effects, there are reasons why it may be beneficial to focus on the reward function. First, it provides continuity with the literature on decision-making, where decisions are usually evaluated individually and the only explanation for differences in behavior is in terms of the shape of the reward function rather than the learning rate. Second, reward functions potentially admit a variety of transformations other than scale. For example, the concave and convex forms of the subjective value function for gains and losses respectively in Prospect Theory provides another dimension of variation that can be separated from scale. We hope to pursue this in future work.

Limitations and Future Directions

While our work offers insights into the potential of basing subjective reward functions on models of human decision-making, it still has several limitations which remain to be improved upon in future work. Most notably, 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 subjective reward function should be tested on larger, more complex environments such as Atari games (Mnih et al., 2013).

In addition, here we have focused on just one aspect of Prospect Theory – the subjective value function. Prospect Theory also makes clear predictions about how people represent probabilities, and in particular that they overweight small probabilities. Exploring the impact of introducing a probability weighting function based on Prospect Theory into the transition probabilities used in RL is an exciting direction for future work.

Conclusion

Our work uses Prospect Theory to define a novel subjective reward function that can be used in reinforcement learning. Agents using a subjective reward function based on Prospect Theory learned more quickly in various experiments, suggesting that the challenge of defining "good" reward functions can potentially be addressed by drawing on ideas from psychology. Based on our results, it seems that RL agents can benefit from three properties that characterize human decision-making: 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 setting, our work suggests that the resulting reward function is useful in a range of environments. 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*.

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, *125*(6), 985.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422.

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., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018). Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*.

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.

Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., & Dragan, A. (2017). Inverse reward design. *Advances in Neural Information Processing Systems*, *30*.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. *arXiv preprint arXiv:2306.12001*.

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.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mohamed, S., & Jimenez Rezende, D. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*, *28*.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning* (Vol. 99, pp. 278–287).

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning* (pp. 2778–2787).

Pathak, D., Gandhi, D., & Gupta, A. (2019). Self-supervised exploration via disagreement. In *International Conference on Machine Learning* (pp. 5062–5071).

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.

Saunders, W., Sastry, G., Stuhlmueller, A., & Evans, O. (2017). Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*.

Shen, Y., Tobia, M. J., Sommer, T., & Obermayer, K. (2014). Risk-sensitive reinforcement learning. *Neural Computation*, *26*(7), 1298–1328.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where do rewards come from. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 2601–2606).

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

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

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

Tverksy, A., & Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 263-292.

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

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton university press.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279–292.