Project Proposal: Incorporating Human Biases into Reward Modeling in Reinforcement Learning

1 Introduction

Traditional reinforcement learning (RL) algorithms assume that agents operate in environments with rational reward structures. However, human decision-making often exhibits systematic biases, such as loss aversion, overconfidence, and probability weighting, as described by Prospect Theory. By integrating these human biases into the reward modeling of RL agents, we can develop systems that better reflect true human preferences and improve human-agent interaction. This project aims to modify the reward functions in RL to account for human cognitive biases, thereby enhancing the alignment of AI systems with human values.

2 Objectives

- **Incorporate Human Biases into Reward Functions:** Adjust the reward signals in RL environments to mimic human biases like loss aversion.
- **Evaluate Agent Behavior:** Analyze how the modified reward functions affect the learning efficiency and decision-making of RL agents.
- **Compare with Traditional RL Models:** Benchmark the performance of bias-adjusted agents against standard RL agents to assess improvements in human-aligned behavior.

3 Feasibility and Scope

3.1 Feasibility

The project is manageable within the course timeline. It involves modifying existing RL algorithms and environments without the need for extensive data collection or complex infrastructure.

3.2 Scope

- **Literature Review:** Study Prospect Theory and related behavioral economics concepts to understand how humans perceive gains and losses.
- **Implementation:** Integrate bias-adjusted reward functions into standard RL algorithms like Q-learning or policy gradients.
- **Testing Environments:** Use simulated environments where human-like decision-making is critical, such as financial decision-making tasks or navigation with risk factors.

4 Methodology

- **Modify Reward Functions:**
  - Implement loss aversion by weighing negative rewards more heavily than positive rewards.
  - Adjust probability weighting to reflect how humans perceive unlikely events.
- **Algorithm Integration:**
  - Apply the modified reward functions within RL algorithms.
  - Ensure the agent can still learn effectively with the new reward structure.
- **Experimentation:**
  - Run experiments comparing the behavior of agents with and without bias-adjusted rewards.
  - Use metrics such as convergence speed, cumulative reward, and policy optimality.

5 Potential Contributions and Outcomes

5.1 Contributions

- Provide empirical evidence on how human biases impact RL agent behavior.
- Offer insights into designing RL systems that are more aligned with human decision-making processes.
- Expand the understanding of how behavioral economics can inform AI development.

5.2 Outcomes

- A set of modified RL algorithms incorporating human biases.
- Analysis demonstrating the effects of these biases on agent performance.
- Recommendations for future research on integrating human cognitive factors into RL.

6 Timeline

- **Week 4:** Conduct literature review on Prospect Theory and RL.
- **Weeks 5-6:** Develop and implement bias-adjusted reward functions.
- **Weeks 7-8:** Integrate modified rewards into RL algorithms and begin experimentation.
- **Week 9:** Analyze results and compare with traditional RL models.
- **Weeks 9–10:** Compile findings and prepare the final report.

7 Resources Required

- **Software:** Access to RL libraries (e.g., OpenAI Gym, TensorFlow, PyTorch).
- **Computing Power:** Standard computational resources; no specialized hardware required.
- **Guidance:** Consultation with instructors or experts in behavioral economics and RL.

Alternative Project Options

Option 1: Modeling Human Intervention Costs in Reinforcement Learning

Introduction

Human interventions in RL can guide agents toward desired behaviors but come at a cognitive and time cost. This project aims to model the cost of human interventions within the RL framework, encouraging agents to learn efficiently with minimal reliance on human assistance.

Feasibility and Scope

- **Feasibility:** The project is achievable within the course duration, involving modifications to existing RL algorithms to include intervention costs.
- **Scope:**
  - Define a cost function representing human effort.
  - Modify RL algorithms to balance performance with intervention costs.
  - Implement active learning strategies where agents request help judiciously.

Potential Contributions and Outcomes

- Develop an RL framework that quantifies and minimizes human intervention costs.
- Demonstrate efficiency improvements in agent learning.
- Provide insights into optimizing human-agent collaboration.

Option 2: Inverse Reinforcement Learning with Behavioral Models

Introduction

Inverse Reinforcement Learning (IRL) seeks to infer the reward function an expert is optimizing based on observed behavior. By incorporating models of bounded rationality and cognitive biases, this project aims to enhance IRL algorithms to account for suboptimal human demonstrations influenced by behavioral economics principles.

Feasibility and Scope

- **Feasibility:** While challenging due to the complexity of IRL and behavioral modeling, careful scoping can make the project manageable.
- **Scope:**
  - Integrate behavioral models into existing IRL algorithms.
  - Simulate or collect data reflecting biased human behavior.
  - Evaluate the ability of the enhanced IRL algorithm to infer accurate reward functions.

Potential Contributions and Outcomes

- Improved IRL algorithms that consider human cognitive biases.
- Insights into how bounded rationality affects reward inference.
- Potential methodologies for more robust human-agent learning interactions.

Option 3: Designing Incentive-Compatible Learning Systems

Introduction

Aligning the learning objectives of RL agents with human incentives is crucial for effective collaboration. This project proposes using mechanism design principles from economics to create reward structures that motivate both agents and humans to cooperate efficiently.

Feasibility and Scope

- **Feasibility:** The project is ambitious but feasible with a focused approach on specific incentive mechanisms.
- **Scope:**
  - Study mechanism design and contract theory relevant to RL.
  - Develop reward structures that align agent behavior with human incentives.
  - Test the system in environments requiring human-agent cooperation.

Potential Contributions and Outcomes

- Creation of incentive-compatible RL frameworks.
- Enhanced understanding of aligning AI systems with human goals.
- Foundations for future research on cooperative AI and mechanism design in RL.

8 Conclusion

The primary project proposal focuses on incorporating human biases into reward modeling, offering a feasible and impactful opportunity to contribute to the field of reinforcement learning. The alternative options provide additional avenues for exploration, each with its own potential for novel contributions. By carefully selecting and executing one of these projects, significant insights can be gained into the integration of behavioral economics and RL, ultimately advancing the development of more human-aligned AI systems.