Author: Meric Ikiz
Contact information: 2zmer3@gmail.com
Github repository: https://github.com/mericikiz/ irl-cognitive
Supervisors: Luciano Cavalcante Siebert, Angelo Caregnato Neto
29.09.2023

EFFECT OF RISK RELATED COGNITIVE BIASES IN INVERSE REINFORCEMENT LEARNING

TUDelft

1. Background

Inverse Reinforcement Learning (IRL)

- In reinforcement learning an agent interacts with the environment through different actions and learns the optimal behavior (policy) observing obtained rewards. [1]
- Reward function is often hard to define precisely.
- IRL aims to learn reward function from expert demonstrations which are collected from humans. [6]

Cognitive Biases

- Humans deviate from rationality in systematic ways called cognitive biases. [2] [3]
- Many cognitive biases are effective but the main focus of this research is group of biases that affect attitudes towards risk and uncertainty.

Loss aversion

- Tendency to overweight losses : the pain of losing is much higher than pleasure of gaining something of same utility. [2] [3]
- Avoiding losses at the expense of rewards, leading to risk averse behavior. However leading to high risk decisions to avoid further losses is also possible. [8]

Research Question: To what extent can IRL learn rewards from expert demonstrations with loss and risk aversion?

Models and Theories that will be used

1. Expert Cognitive Model

System 1 and 2 Model	Prospect Theory
They have different utility functions and perceptions, System 1 is more intuitive, uses shortcuts while System 2 plans more long term and is less impulsive	Treats losses and gains asymmetrically, overestimates low probability events and underestimates high probability events

2. Maximum Entropy IRL algorithm (MEIRL)

For a set of demonstrations there are infinitely many fitting reward functions. Using the principle of maximum entropy, MEIRL finds the solution with the least amount of bias. [9]

2. Methodology

Simulating and Interpreting Expert Demonstrations

- System 1 and 2 rewards = R1, R2,
- System 2 is assumed to have a perfectly rational view of the world with known rewards received with certainty while R1 is received with a probability P1. This is a strong assumption made to observe System 1 effects in isolation.
- Additionally System 1's view of the environment is reevaluated through Prospect Theory filter before use.
- At every point, the decision is a compromise between the two systems.

Subjective Reward Assesment of R1 $r_{\text{subj}}(x)$

$$r_{\text{subj}}(x) = \begin{cases} (x-b)^{\alpha} & \text{if } x > b, \\ -\kappa(b-x)^{\beta} & \text{if } x < b. \end{cases}$$

x objective reward for System 1's preferences

α, β represents diminishing sensitivity to gains and losses

b baseline that the agent compares new rewards against

κ degree of loss aversion

Subjective Probability Assesment of P1 $w(p)$

$$w(p) = \frac{p^{\eta}}{p^{\eta} + (1-p)^{\eta}} = \text{decision weight}$$

η degree of over-weighting of small and under-weighting of large probabilities

p objective probability of receiving this reward

[2][3][8]

Final Expert Decision making [7]

- Pass R1 and P1 through Prospect Theory filter and multiply to get $RP1_{\text{subj}}$
- Perform value iteration on $RP1_{\text{subj}}$ to get the most optimal actions for System 1 preferences at each decision point. Call this V1*
- Initialize V1 and V2 arbitrarily (value iterations for System 1 and 2)
- At each decision point (state) choose compromise action and find:

$$V_{\text{combined}} = V_2 - \psi(V_1^* - V_1)$$

ψ cognitive control cost representing mental effort needed to deviate from System 1's optimal course of action

Application of MEIRL [9]

- Calculate state visitation frequencies of the expert SVF_{expert} according to expert trajectory samples generated from policy adhering to V_{combined}
- Initialize guessed rewards per state arbitrarily
- Compare SVF_{expert} with the SVF generated by inferred reward, gradient
- Update according to gradient and repeat until convergence

3. Experiments and Results

Experiment 0: Small Grid World with Two Terminal States

System 1 Rewards and Probabilities | System 2 Rewards | Joint Value Iteration | Expert Trajectories (200 Samples)

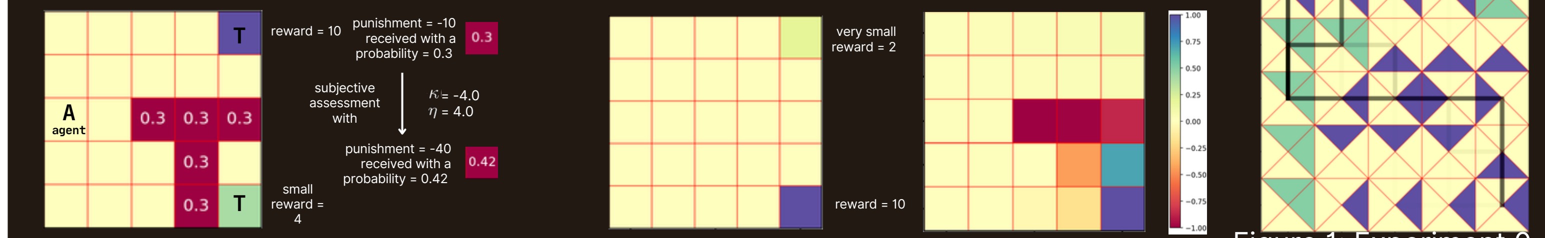
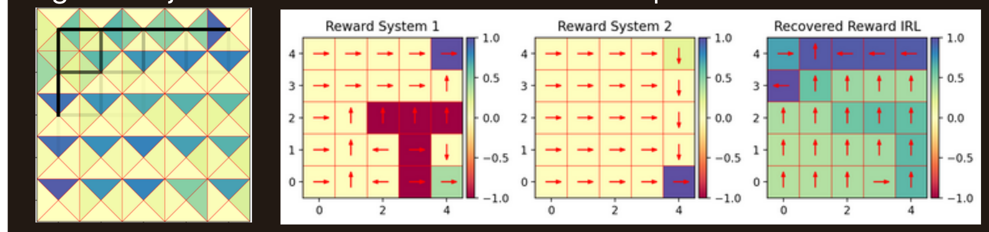


Figure 1: Experiment 0

Agent Trajectories and Inferred Reward Comparison with actual



- The most significant gap between expert and IRL is caused by cognitive control, because this is a factor that plays a dynamic role.
- Because of the expert's hesitancy it ends up collecting even further negative punishment than it needs for getting to the reward.

Experiment 1: Limiting Choices

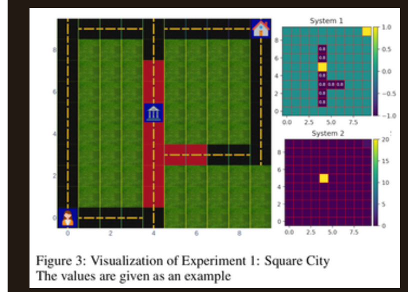


Figure 3: Visualization of Experiment 1: Square City The values are given as an example

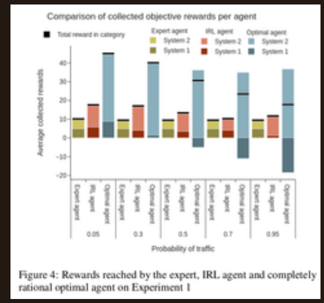


Figure 4: Rewards reached by the expert, IRL agent and completely rational optimal agent on Experiment 1

- Loss aversion and risk aversion significantly impact the subjective reward and decision weights, leading to a gap between expert demonstrations and the optimal objective rewards.
- The expert agent avoids negative outcomes losing on rewards, resulting in a lower overall sum of total rewards.
- The IRL agent's System 1 rewards closely follow the expert, although System 2 rewards change. But at very high risk System 1 rewards are also different.
- While the expert agent keeps a balanced reward profile, the IRL agent does not differentiate, especially when the punishment is more likely. This is expected as the IRL agent does not know there are two different reward functions and can only see the end behavior.
- The most significant gap between expert and IRL is caused by cognitive control, because this is a factor that plays a dynamic role

4. Conclusions

Limitations

- Limited nature of the MEIRL algorithm: MEIRL only considers trajectories and assumes decision making is fairly static.
- No actual human data was collected. Therefore we only simulate the cognitive bias we believe exist and test for them.
- More complex environments, tasks and agent models would be more realistic but introduce more complexity and require different search techniques.

Conclusions

- Although the IRL agent can make similar trajectories, it cannot infer any underlying motivations or relations between them. This is expected from MEIRL, thus more sophisticated models needed.
- The agent is not very consistent, especially when faced with a lot of options. More careful environment planning is needed.
- It is important to study IRL with cognitive biases to not consider expert as optimal as it can cause loss of nuance or even the main goal of the agent.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. Proceedings of the twenty-first international conference on Machine learning, page 1, 2004.
[2] Daniel Kahneman. Thinking, Fast and Slow. Farrar, Straus and Giroux, 2011.
[3] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. Econometrica, 47(2):263–291, 1979.
[4] Qiang Liu. irl-maxent. GitHub repository, 2023. [Online; accessed May 12, 2023].
[5] E. Mazumdar, L.J. Ratliff, T. Fiez, and S. Shankar Sastry. Gradient-based inverse risk-sensitive reinforcement learning. In 2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017, pages 5796–5801, 2018.
[6] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 663–670. Morgan Kaufmann Publishers Inc., 2000.
[7] Alexander Peysakhovich. Reinforcement learning and inverse reinforcement learning with system 1 and system 2. arXiv preprint arXiv:1805.00909, 2018.
[8] Peter P. Wakker. Prospect Theory for Risk and Ambiguity. Cambridge University Press, 2010.
[9] BrianD.Ziebart,AndrewMaas, J.AndrewBagnell,and Anind K. Dey. Maximum entropy inverse reinforcement learning. In Proc. AAAI, pages 1433–1438, 2008.