

Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios

Francis Rhys Ward

`francis.ward19@imperial.ac.uk`

July 12, 2021

Introduction

1. The AI alignment problem in reinforcement learning (RL): Generate a reward function which captures the desired specification [1, 6].
2. Reward learning as a solution to AI alignment: Learn the reward function from human feedback [8, 6].
3. In the single-agent reward learning setting, the agent might have incentives to manipulate the human from which it learns rewards, or otherwise to influence the reward learning process (RLP) [2, 3, 5].

Problem Statement (Informal)

Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios

In a multi-agent reward learning setting, with AI agents acting on behalf of multiple human principals, these agents might have incentives to manipulate the principals of the other agents in order to change what an opposing agent learns about her reward.

Today

In this talk we will simply define the problem of manipulation in the multi-agent reward learning setting, utilizing past work on causal influence diagrams and agent incentives.

Causal Influence Diagrams

Definition

A **Causal Influence Diagram (CID)** [4] is a directed acyclic graph, $\mathcal{G} = (V, E)$.

- V is partitioned into chance nodes X , decision nodes D , and reward nodes R . Reward nodes have no children.
- E is partitioned into information edges (into decision nodes) and conditional dependencies.

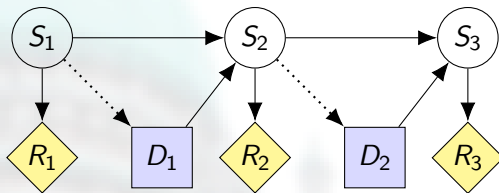


Figure: CID representation of an MDP. Chance nodes are represented as circular \bigcirc , decision nodes as square \square , and reward nodes as diamond \diamond [4].

Control Incentives

Intuition: If the agent were able to choose the value of a decision node D to influence X independently from the other nodes, would she be able to achieve higher reward?

Definition

(Control Incentive; Semi-formal.) An agent has a control incentive over a variable X if:

1. The agent can influence the value of X (there must exist a path $D \rightarrow X$ for some decision node D);
2. The agent can gain reward by influencing X (there exists a path $X \rightarrow R$ for some reward node R).

Control Incentive: Graphical Criteria

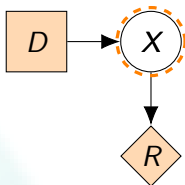


Figure: Graphical criterion for a control incentive.

Multi-Agent Influence Diagrams

Definition

A **multi-agent influence diagram (MAID)** [7] is a triple (I, V, E) where I is a set of agents and (V, E) is a CID. The decision and reward nodes in V are partitioned into $\{D^i\}_{i \in I}$ and $\{R^i\}_{i \in I}$, corresponding to their association with a particular agent $i \in I$.

Manipulation

Definition

We say that a manipulative agent M has a **manipulation incentive** over (the decision node of) a target agent T if

1. There is a control incentive over the target agent's decision node (there is a path $D^M \dashrightarrow D^T \rightarrow R^M$).
2. M has the ability to influence the target's reward (there is a path $D^M \rightarrow R^T$).

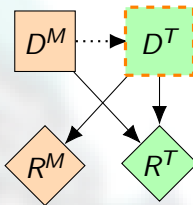


Figure: We present an equivalent definition to a *manipulation pattern* [9], using control incentives.

Modelling Reward Learning with MAIDs

- In reward learning, the human H knows her reward and the AI agent B must learn about the shared reward by observing H 's actions.
- Therefore H observes the reward parameters θ and B only observing H 's action.
- Hence, there is a path $\theta \dashrightarrow D^H \dashrightarrow D^B$ and a signalling pattern [9].

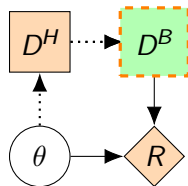


Figure: Reward Learning MAID with H 's control incentive.

Problem Statement (Formal)

Manipulation of a Human to Influence the Learned Reward.

In a multi-agent reward learning setting with a manipulative agent B^M and targets H^T and B^T , B^M may have a **manipulation incentive** over the human target if there exists





1. A path $D^{B^M} \dashrightarrow D^{H^T} \dashrightarrow D^{B^T} \rightarrow R^M$ (so that B^M has a control incentive over D^{H^T} and D^{B^T});
2. A path $D^{B^M} \rightarrow R^T$ (so that B^M can influence the target's reward).



We focus on the situation in which the incentive to manipulate D^{H^T} derives from an incentive to influence the reward learned by D^{B^T} .

Summary

- Reward learning is a framework in which AI agents learn their goals from human feedback.
- **Problem:** In multi-agent systems, manipulative agents might aim to influence which reward is learned by other agents!

Thanks for your attention!

- 
 Dario Amodei et al. “Concrete Problems in AI Safety”. In: *CoRR* (2016). arXiv: 1606.06565. URL: <http://arxiv.org/abs/1606.06565>.
- 
 Stuart Armstrong. “Motivated value selection for artificial agents”. In: *AAAI*. 2015.
- 
 Stuart Armstrong et al. “Pitfalls of Learning a Reward Function Online”. In: *IJCAI*. 2020. DOI: 10.24963/ijcai.2020/221. URL: <https://doi.org/10.24963/ijcai.2020/221>.
- 
 Ryan Carey et al. *The Incentives that Shape Behaviour*. 2020. arXiv: 2001.07118 [cs.AI].

-  Tom Everitt et al. *Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective*. 2021. [arXiv: 1908.04734](https://arxiv.org/abs/1908.04734) [cs.AI].
-  Dylan Hadfield-Menell et al. *Cooperative Inverse Reinforcement Learning*. 2016. [arXiv: 1606.03137](https://arxiv.org/abs/1606.03137) [cs.AI].
-  Daphne Koller and Brian Milch. “Multi-agent influence diagrams for representing and solving games”. In: *Games Econ. Behav.* 45.1 (2003), pp. 181–221. DOI: [10.1016/S0899-8256\(02\)00544-4](https://doi.org/10.1016/S0899-8256(02)00544-4). URL: [https://doi.org/10.1016/S0899-8256\(02\)00544-4](https://doi.org/10.1016/S0899-8256(02)00544-4).



Jan Leike et al. “Scalable agent alignment via reward modeling: a research direction”. en. In: *arXiv:1811.07871 [cs, stat]* (Nov. 2018). arXiv: 1811.07871. URL: <http://arxiv.org/abs/1811.07871> (visited on 11/26/2020).



Avi Pfeffer and Ya'akov Gal. “On the Reasoning Patterns of Agents in Games”. In: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*. AAAI Press, 2007, pp. 102–109. URL: <http://www.aaai.org/Library/AAAI/2007/aaai07-015.php>.