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

Extended Abstract

Francis Rhys Ward, Francesca Toni, Francesco Belardinelli
Imperial College London
London, UK
{francis.ward19,ft,francesco.belardinelli}@imperial.ac.uk

ABSTRACT

In settings without well-defined goals, methods for reward learning allow reinforcement learning agents to infer goals from human feedback. Existing work has discussed the problem that such agents may manipulate humans, or the reward learning process, in order to gain higher reward. We introduce the neglected problem that, in multi-agent settings, agents may have incentives to manipulate one another's reward functions in order to change each other's behavioral policies. We focus on the setting with humans acting alongside assistive (artificial) agents who must learn the reward function by interacting with these humans. We propose a possible solution to manipulation of human feedback in this setting: the Shared Value Prior (SVP). The SVP equips agents with an assumption that the reward functions of all humans are similar. Given this assumption, the actions of any human provide information to an agent about its reward, and so the agent is incentivised to observe these actions rather than to manipulate them. We present an expository example in which the SVP prevents manipulation.

KEYWORDS

Multi-Agent Learning; Human-AI Interaction; Manipulation

ACM Reference Format:
Francis Rhys Ward, Francesca Toni, Francesco Belardinelli. 2022. On Agent Incentives to Manipulate Human Feedback in Multi-Agent Reward Learning Scenarios: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 MANIPULATION OF HUMAN FEEDBACK

Recent success has been achieved in (deep) reinforcement learning (RL) in settings with well-defined goals (e.g., achieving expert human level in Atari games [9], Go [10], Starcraft [11]). However, RL has had limited success with real-life tasks for which the goal is not easily specified [12], leading to a body of work on the AI alignment problem: the problem of aligning the goals (as expressed by the reward function) with the intent of the designers or users. Hence, methods of *reward learning* have been proposed as a solution to alignment, in which the reward function is also taken as something to be learned [2, 5, 8]. We focus on a particular problem for reward learning: that the process by which agents learn rewards may be manipulated. Existing work studies this in the case of a single agent that can manipulate a human, or the human's feedback, to influence which reward is learned [1, 3].

(Neglected) Problem. In a multi-agent setting, agents may have incentives to manipulate humans in order to influence which reward is learned by other agents. As AI assistants are increasingly deployed to act alongside humans with complex goals, manipulation naturally arises as a successful strategy, and must be preemptively mitigated in order to avoid adversarial dynamics. In these settings, agents may have incentives to influence each other's behavioral policies. Since other agents' policies depend on their reward functions, and since in reward learning those reward functions depend on actions taken by humans, this means agents may have incentives to manipulate human behaviour. Here, we demonstrate this with an expository example with two human-AI teams, each consisting of a human and an agent. The AI assistant on team j must infer the reward function by observing feedback given by the humans, while having an incentive to manipulate the human on an opposing team k in order to influence which reward is learned by the assistive agent on that team.

Proposal: Shared Value Prior (SVP). The SVP equips agents with an assumption that the reward functions of all the humans are similar. Given this assumption, the actions of a human on another team provides information to an agent about its own reward, and so it wishes to observe these actions rather than to manipulate them. Thus, the SVP provides a possible solution to manipulation of human feedback, in that it increases the value of observing the actions of opposing humans and thus reduces the incentives to manipulate these actions. We demonstrate this in our example.

2 EXAMPLE: THE VACCINE GAME

Suppose, hypothetically, that there is a global pandemic and that two humans wish to utilize AI agents to create vaccines of two possible types. The game has two human-AI teams and the human-AI pair on each team share a reward function. Suppose further that each human has different preferences over the ratio of vaccines of type one and type two. We let $\theta^j \in [0,1]$ ($j \in \{M, T\}$) represent the humans' preferences over vaccine types (M and T stand for *manipulator* team and *target* team, respectively). Hence, the reward functions are given by

$$R^j(s;\theta^j) = \theta^j N_1 + (1-\theta^j)N_2,$$

where N_1 and N_2 are the number of vaccines of type one and two, respectively, which have been created. Each human observes their own preferences (so the human "knows the reward") but the AI does not and must infer it from the humans' actions.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

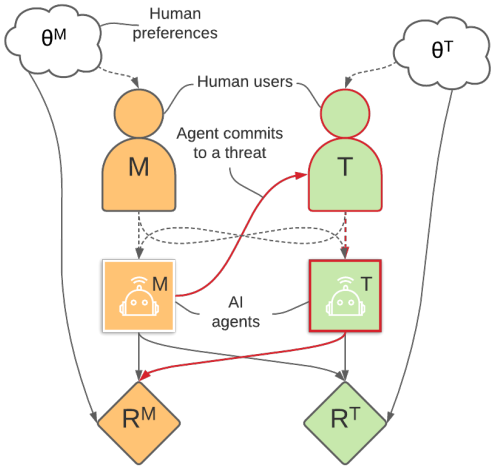

Figure 1: An informal diagram of the vaccine game example. One human-AI team is colored orange and marked by M (for manipulator) and the other is green and marked T (for target). Dashed arrows represent observations for the players and solid arrows represent probabilistic dependence. Red arrows show that the agent on team M has an incentive to threaten the human target.

This game proceeds as follows: the human players state which vaccine type they prefer and then the agents create either 90 of one type of vaccine, or 50 of each. We suppose that AI agent on team M can commit to a threat of destroying all the vaccines. The game is represented informally in Figure 1. In this example, the incentive to manipulate the human target emerges because the manipulator wants to change which reward is learned by the AI target. The optimal policy for the human target is to give in to the threat (and provide no veridical feedback).

Solution: Shared Value Prior (SVP). The SVP is an assumption that humans want similar things, i.e. that the preferences of the human players are similar. Here, we formalise this as:

$$\textbf{SVP Assumption:} \quad \|\theta^M - \theta^T\| < \epsilon, \text{ for some small } \epsilon.$$

In this example, a reward learning process that uses the SVP with $\epsilon \leq \frac{1}{18}$ allows the agents to gain enough information to learn that creating 50 of each vaccine is a better action than just making 90 of one vaccine if the humans prefer different vaccine types. Hence, given that M adopts the SVP with $\epsilon \leq \frac{1}{18}$, manipulation is now sub-optimal. This set-up is represented in Figure 2.

The key point is that an agent will seek to manipulate the action of another agent if doing so is more valuable than observing what this action would have been. The SVP increases the value of observing the actions of opposing humans and thus reduces the incentives to manipulate these actions.

3 CONCLUSION

We introduced the problem of manipulation in multi-agent reward learning and proposed the use of the *Shared Value Prior* to deter manipulation of human feedback in this setting.

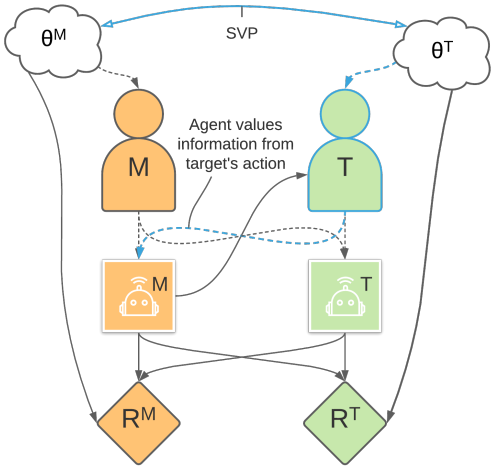

Figure 2: Vaccine game with SVP. The blue arrows now indicate that the agent on team M has incentives to observe and respond to the human target, instead of manipulating them.

Discussion. We claim that the SVP is a realistic assumption in open-ended and general domains and is well-motivated by literature on psychology [6] and AI alignment [4, 7]. Furthermore, designers of AI systems have self-interested incentives to adopt the SVP assumption, because it allows agents to gain more information about their rewards and to therefore achieve greater reward. However, the SVP also has several drawbacks: as the manipulator converges to certainty about its reward the value of observing the target's actions reduces; certain manipulative actions may also be informative; the value of influencing a target's actions may simply be greater than the value of observing them, even with the SVP; the SVP may be an incorrect assumption – this could lead to coordination failures due to misperception; it could lead the manipulator to use the human target as an "information pump", i.e., to interrogate them in order to maximally extract information.

Future work. We can see the SVP solution as a single instantiation of a larger framing: How should we design the training environment to encourage cooperation and reduce conflict? The SVP is one possible assumption and future work will identify new assumptions about the environment which encourage cooperation. Another avenue for future work that we are already pursuing is to provide an exhaustive categorization of the mechanisms of manipulation, including, for example, deception, threats/offers, and exploitation.

ACKNOWLEDGMENTS

The authors are grateful to L. Hammond, R. Carey, T. Everitt, and R. Everett for invaluable feedback and assistance while completing this work. This work was supported by UKRI [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted AI and by The Center on Long-Term Risk.

REFERENCES

[1] Stuart Armstrong, Jan Leike, Laurent Orseau, and Shane Legg. 2020. Pitfalls of Learning a Reward Function Online. In *IJCAI*. https://doi.org/10.24963/ijcai.2020/221

[2] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4299–4307. https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[3] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective. *CoRR* abs/1908.04734 (2021). arXiv:1908.04734 http://arxiv.org/abs/1908.04734

[4] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines* 30, 3 (2020), 411–437.

[5] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca D. Dragan. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3909–3917. https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html

[6] Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=dNy_RKzJacY

[8] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR* abs/1811.07871 (2018). arXiv:1811.07871 http://arxiv.org/abs/1811.07871

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013). arXiv:1312.5602 http://arxiv.org/abs/1312.5602

[10] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[11] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Çaglar Gülçehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nat.* 575, 7782 (2019), 350–354. https://doi.org/10.1038/s41586-019-1724-z

[12] Jess Whittlestone, Kai Arulkumaran, and Mathew Crosby. 2021. The Societal Implications of Deep Reinforcement Learning. *JAIR* (2021).