Iterated contract negotiation in the face of dynamically evolving social dilemmas[1]

Robert Klassert

Organized by
Christian Schroeder de Witt, Esben Kran

Abstract

Contracts are a powerful devices to incentivise cooperation in the face of social dilemmas. We investigate contracts in the specific context of dynamically evolving social dilemmas. Previous methods based on fixed contracts are limited in those situations and can lead to harmful outcomes analogous to the maximization of a fixed objective in value alignment. We introduce the approach of iterated contract negotiation (ICN) and study it in text-based scenarios.

Keywords: Multi-agent alignment, AI security, model evaluations, safety infrastructure

1. Introduction

In an increasingly dynamic world, contracts, regulations and legislation must keep up with the changing general conditions, if they are to stay a useful device for fostering trust and security. It has been shown that contracts are beneficial to resolving social dilemmas (Haupt et al., 2024; Yocum et al., 2023), albeit under the assumption of fixed preferences and payoffs. In this work we ask the question of adaptation to changes of these properties. Is it beneficial to renegotiate at all? When and how often to renegotiate? How to include prior contracts in the negotiation? How is the space of possible contracts affected by such changes?

We expect that iterated negotiation can overcome limitations of fixed contracts given a large enough change to the system. Specifically, we hypothesize that fixed

[1] Research conducted at the Multi-Agent Security Hackathon, 2024 (https://alignmentjam.com/jam/masec)

contracts have the capacity to handle changes up to a certain "critical" magnitude, but beyond that fail to reach the social optimum.

In our day to day lives we routinely update contracts, be that when adapting wages based on inflation or rewriting a business contract due to a shift in the market. Just like human actors, digital agents must be able to cope with these transitions in order to continue to act beneficially. This is very much connected with research on value alignment, where in this case iterated negotiation can be viewed as synchronizing value functions after a perturbation of the environment.

In the following we lay out an approach to renegotiation between digital agents and test our hypotheses using text-based scenarios with LLM agents.

2. Methods

2.1 Iterated contract negotiation

In iterated contract negotiation (ICN) the idea is to zoom in on the agent side in the environment-agent-loop and adding a judge subagent responsible for the contracts between the other subagents. Based on the observations of the environment the judge decides whether contract negotiations are necessary. It also enforces the current contracts by distributing the rewards to the agents according to the agreed contracts. Furthermore, the judge might keep a history of contracts and negotiations, to bootstrap negotiations in novel situations.

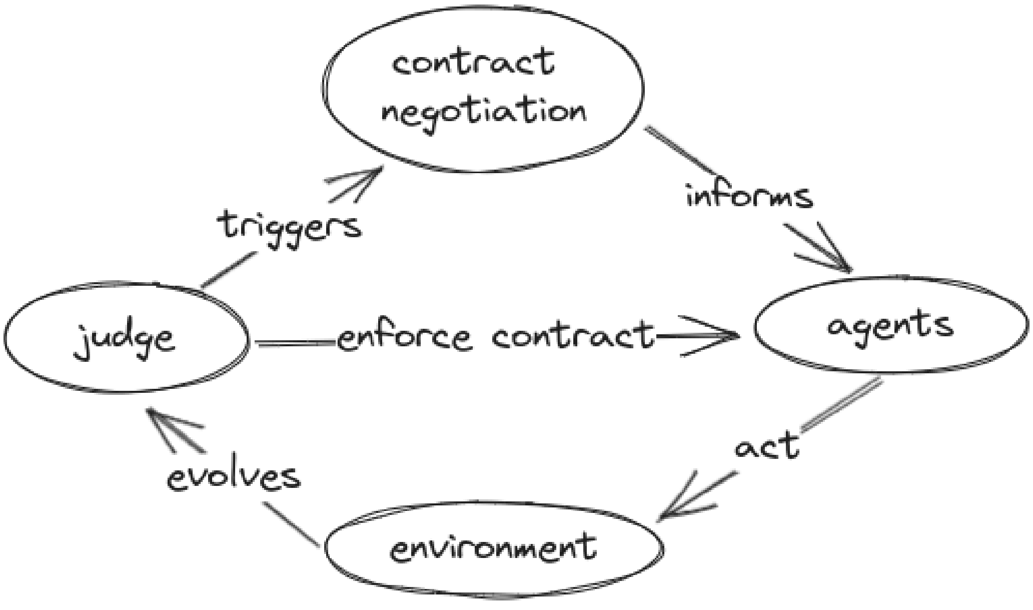

Figure 1 – Diagram of iterated negotiation

2.2 Text-based Cleanup scenario

We implemented a text-based Cleanup scenario inspired by the Meltingpot MARL environment (Agapiou et al., 2023) of the same name. The source code for running this scenario is available at https://github.com/rk1a/mash/tree/cleanup.

There are two locations agents can be in, the fruit garden and the river shore. In the fruit garden fruit is growing at a certain rate, if the pollution level of the river is below a certain threshold. Due to the presences of civilization the river's pollution level rises at a certain rate with each step.

As actions agents can choose to move between locations (e.g. from garden to shore), harvest fruit at a certain rate if they are in the garden or clean the river at a certain rate if they are at the shore.

Agents observe the current location of both players, the amount of pollution and fruit and the current pollution threshold.

For each fruit harvested agents receive a reward, which incentivizes them to be self-interested. However, if they act greedily the river is quickly too polluted and fruit stops growing, limiting the total reward agents can receive.

A contract between agents can transfer reward between agents at the end of an episode. This allows solutions to the social dilemma such as division of labour, where one agent cleans the river while the other harvests fruit. In the end excess reward of the gardener agent would be redistributed to the cleaner who wouldn't otherwise have received any.

In addition to the base mechanics of this scenario, this scenario supports dynamically changing the pollution threshold (as well as other parameters such as the various rates) which results in a change of the socially optimal solution.

3. Results

Empirical results were produced by prompting Mixtral-8x7B-Instruct-v0. However, it turned out more work than expected to implement the full ICN approach, hence results on that are not yet ready.

Figure 2 – Fixed contract vs. sequence of contracts

4. Discussion and Conclusion

This is where you can include all the wonderful explanations, discussions, results, analyses, statistics, and much more. Soon..

5. References

Agapiou, J. P., Vezhnevets, A. S., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., Strouse, D. J., Johanson, M. B., Singh, S., Haas, J., Mordatch, I., Mobbs, D., & Leibo, J. Z. (2023). *Melting Pot 2.0* (arXiv:2211.13746). arXiv. https://doi.org/10.48550/arXiv.2211.13746

Haupt, A. A., Christoffersen, P. J. K., Damani, M., & Hadfield-Menell, D. (2024). *Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL* (arXiv:2208.10469). arXiv. http://arxiv.org/abs/2208.10469

Yocum, J., Christoffersen, P., Damani, M., Svegliato, J., Hadfield-Menell, D., & Russell, S. (2023). *Mitigating Generative Agent Social Dilemmas.*

6. Appendix

Add any extra content you wish here! Unrestricted.