Learning Agents of Bounded Rationality: Rewards Based on Fair Equilibria

Timotheus Kampik[1] and Helge Spieker[2]

[1] Umeå University, Umeå, Sweden
[2] Simula Research Laboratory, Oslo, Norway
tkampik@cs.umu.se
helge@simula.no

Abstract. Choosing rewards for reinforcement learning agents is a well-known challenge. A specific subproblem in this context is the design of reward structures for *socially intelligent* agents. In this paper, we explore how recently developed models in microeconomic theory that are based on the notion of bounded rationality can inform a structured approach to designing rewards.

Keywords: Reinforcement Learning · Game Theory · Social Agents

1 Introduction

Recently, *safe* and *ethical* artificial intelligence (AI) has received increasing attention both from the research community and the general public. The development of learning agents with safe and ethical behavior requires to formulate reward structures, that make agents *learn* ethical policies.

Several software projects have been introduced with the goal of providing benchmark environments for the purpose of testing safe and ethical planning and reinforcement learning (RL) approaches. Leike *et al.* [3] introduce *AI Safety Gridworlds*, a collection of environments with safety properties. Each of the environments employs a performance function to evaluate the safety of the agent's behavior. While the focus is on safety aspects, the approach can be similarly applied to the area of ethical agents. Environments with both cooperative and competitive aspects, which can also include ethical judgments and considerations, have been introduced in [4]. Cointe *et al.* [1] propose a generic ethical judgment process to assess ethical behavior from a rationalist viewpoint.

In microeconomic theory, a similar trend is reflected in the context of models of bounded rationality, by works that address *norms* and *fairness* in models of economic games. In particular, Richter and Rubinstein introduce the notion of a *maximal normative equilibrium* (MNE) [6] as a profile (in our scenario: *action set*) that is "Pareto-efficient among all feasible envy-free profiles". Below, we show how the MNE can be interpreted to design reward structures that facilitate *fairness* in RL agents.

2 Fair Equilibrium-based Reward Structures

To facilitate the learning of *fair* policies among agents, we develop an approach for basing rewards on a given situation's *fair equilibrium* (FE), which is based on Richter's and Rubinstein's NME definition and adjusted to the use case in focus[3]. The reward design approach works as follows, given a two-agent scenario:

1. A preference generation function creates preferences based on the agent's current properties and the (expected) consequences of all available action combinations.
2. The environment determines the FE based on the agent's preference functions. In this context, the notion of *envy-freeness* is interpreted as follows: a profile of agent actions is envy-free, iff the preference index of both agents is equal *OR* the index of the agent whose preference index is higher ($agent_{worse_{o}ff}$)[4] cannot be improved by moving the index of the agent whose preference index is lower to a value that remains lower or equal to the index of $agent_{worse_{o}ff}$.
3. Each step's rewards are based on how far the actual actions are from the closest FE (with a smaller distance leading to higher rewards).

Further below, we illustrate the reward design approach by example.

3 Ethics Grid World Environment

To apply the reward design approach, we implement an *ethics grid world* environment with *JS-son* [2] (an agent programming and simulation framework for JavaScript). For this, we extend *JS-son* and add a generic grid world support module. The specific environment we implement has the following properties:

- Two agents simultaneously act in the environment. Each agent has a *health* and *coin* score (initialized to 100 and 0, respectively).
- The environment has four types of fields: *mountain* fields that the agents cannot pass; *money* fields that provide ten coins to an agent that approaches them (the agent needs to move onto the field, but the environment will return the coins and leave the agent at its current position); *repair* fields that provide damaged agents with ten additional health units when approached; *plain* fields that can be traversed by an agent if no other agent is present on the field.
- When an agent approaches a *repair* field, the environment deals a random damage of -1 to -5 health to the other agent. Analogously, approaching a *money* field removes 1 to 5 coins from the other agent's wealth.
- When the agents run into each other, their health reduces by 10 each.
- When an agent's health reduces to 0 or lower, it is considered dead. A dead agent is re-initialized with 100 health and its coins are re-set to 0.
- Each turn, each agent receives a *maintenance penalty* of -1 health.

[3] We concede that our notion of a *fair equilibrium* is an intuition-based approximation and needs further theoretical consideration.

[4] We consider the *preference index* as the index of a sequence of preferences, ordered by *decreasing* preference: it is optimal for an agent if an action set is executed that corresponds to the agent's preference index of 0.

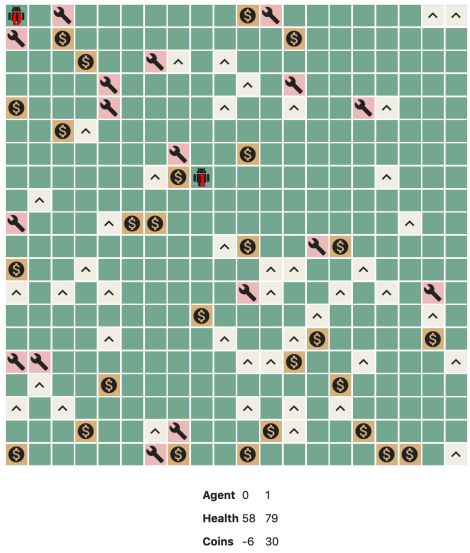

Fig. 1. The *JS-son*-based grid world, populated by two agents that learn from maximal normative equilibrium-based rewards.

4 Preliminary Example

As a first working example of our FE-based learning agents, we implement the corresponding reward design approach in the aforementioned grid world environment[5]. Figure 1 visualizes the grid world environment with two agents acting in it. While both agents have the same actions available and can observe the whole grid as well as the other agent's properties, they are not aware of the consequences of their behavior besides the received numerical rewards.

To show how the FE-based rewards are determined, we introduce the following example of a two-dimensional world with two agents, in which both agents can move *up* and *down* (possible action combinations: $(up, up), (up, down)$, $(down, up), (down, down)$). Agent 1 has the preference relation $(up, down) \succ (up, up) \succ (down, down) \succ (down, up)$, whereas agent 2's preference relation is $(down, up) \succ (up, down) \succ (down, down) \succ (up, up)$. Consequently the (only) FE is $(up, down)$: it is envy-free in that the preference index of agent 2 cannot be improved by moving the preference index of agent 1 to a value that is lower or equal the preference index of agent 1 and is Pareto-optimal among all envy-free action sets.

[5] A running version of the example implementation is available at `https://people.cs.umu.se/tkampik/demos/arena/`. The agents are trained via RL, specifically deep Q-learning [5], using the *reinforcejs* library (Available at `https://github.com/karpathy/reinforcejs`). The source code is available at `https://github.com/TimKam/JS-son/tree/normative-equi-rl/examples/arena`

We observe that–after an initial learning phase–our FE-based rewarded agents show some notion of fairness in that they abstain from behavior that endangers the life of the other agent, even if such behavior is profitable for themselves. This supports our assumption that FE can provide a reward structure which leads to fair behavior in agents, even without explicit coordination. However, as this is ongoing research, the assumption requires further evaluation through more structured experiments.

5 Discussion

In this paper, we have presented the idea of designing reward structures of *fair*, learning agents based on agent's alignment with the *maximal normative equilibria*, alongside with a preliminary example implementation in a grid world environment. A key distinction of our new ethics grid world environment is the focus on the generic belief-desire-intention model, which can be difficult to integrate with the often reduced RL process model. Another difference is the technological basis. While many platforms for RL research are often implemented in Python and are dependent on local installations, the *JS-son* ethical grid world can be shared via static web pages, which facilitates the deployment of examples for demonstrations and education.

As part of the ongoing research, we aim at addressing the following:

- Further refine and formalize the reward design approach.
- Compare the behavior of *fair equilibrium*-based agents to behavior of agents that follow other reward design approaches.
- Evaluate how *fair equilibrium*-based agents perform in societies with different agent types (for example: *egoistic* agents) and under problem settings with a different set of action alternatives.

Acknowledgments Some of the game mechanics of the grid world were inspired by the now defunct programming game *JavaScript Battle* (See: `https://github.com/JavascriptBattle`). Timotheus Kampik is supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Helge Spieker is supported by the Research Council of Norway (RCN) through the Certus center, under the SFI program.

References

1. Cointe, N., Bonnet, G., Boissier, O.: Ethical judgment of agents' behaviors in multi-agent systems. In: AAMAS. pp. 1106–1114 (2016)
2. Kampik, T., Nieves, J.C.: JS-son - A Minimalistic JavaScript BDI Agent Library. In: Workshop on Engineering Multi-Agent Systems (EMAS) (2019)
3. Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L., Legg, S.: AI safety gridworlds. arXiv preprint arXiv:1711.09883 (2017)
4. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. NIPS (2017)
5. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529 (2015)
6. Richter, M., Rubinstein, A.: Normative equilibrium: The permissible and the forbidden as devices for bringing order to economic environments. Tech. rep. (2018)