Multi-Principal Assistance Games: Definition and Collegial Mechanisms

Arnaud Fickinger
Department of EECS
University of California, Berkeley
arnaud.fickinger@berkeley.edu

Simon Zhuang
Department of EECS
University of California, Berkeley
simonzhuang@berkeley.edu

Andrew Critch
Department of EECS
University of California, Berkeley
critch@berkeley.edu

Dylan Hadfield-Menell
Department of EECS
University of California, Berkeley
dhm@berkeley.edu

Stuart Russell
Department of EECS
University of California, Berkeley
russell@berkeley.edu

Abstract

We introduce the concept of a *multi-principal assistance game* (MPAG), and circumvent an obstacle in social choice theory — Gibbard's theorem — by using a sufficiently "collegial" preference inference mechanism. In an MPAG, a single agent assists N human principals who may have widely different preferences. MPAGs generalize *assistance games*, also known as cooperative inverse reinforcement learning games. We analyze in particular a generalization of apprenticeship learning in which the humans first perform some work to obtain utility and demonstrate their preferences, and then the robot acts to further maximize the sum of human payoffs. We show in this setting that if the game is sufficiently *collegial* — i.e., if the humans are responsible for obtaining a sufficient fraction of the rewards through their own actions — then their preferences are straightforwardly revealed through their work. This revelation mechanism is non-dictatorial, does not limit the possible outcomes to two alternatives, and is dominant-strategy incentive-compatible.

1 Introduction

The growing presence of AI systems that collaborate and coexist with humans in society highlights the emerging need to ensure that the actions of AI systems benefit society as a whole. This question is formalized as the *value alignment* problem in the AI safety literature [1], which emphasizes the need to align the increasingly powerful and autonomous systems with those of their human principal(s). However, humans are prone to misspecify their objectives which can lead to unexpected behaviors [1]; hence research in value alignment has focused on deriving preferences from human actions. In the body of research in value alignment and human robot interaction, the majority of the work involves scenarios with one human and one AI system. It is an appealing setting because the robot and the human share

Accepted at the Cooperative AI Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

the same goal. Therefore, methods in this setting such as inverse reinforcement learning [2, 3, 4], inverse reward design [5], and LILA [6] revolve around how an AI system can optimally learn the preferences of the human and apply these results to novel environments. Similarly, the human's incentive is to optimally teach the robot its own preferences. The combination of a learning AI system and a teaching human yields the *assistance game* (also known as the cooperative inverse reinforcement learning game) [7].

However, AI systems in the real world do not fit this *one human, one AI* paradigm. Recommendation systems, autonomous vehicles, and parole algorithms do not exist in a vacuum—they often influence and are influenced by multiple humans. Hence, we consider a variation on assistance games where a robot acts on behalf of multiple humans, which we call the *multi-principal assistance game* (MPAG). The key difference between this and the scenario with only one human is that, in general, different humans have different preferences, so it is impossible to align the AI to perfectly match the preferences of everyone. The problem of aggregating individual preferences for making collective decisions has been studied by economists and philosophers for more than two hundred years and constitutes the heart of social choice theory [8].

Even with a given aggregation method, however, the inference process itself is challenged by the presence of selfish agents. While the robot acts to optimize the aggregate of utilities, each human acts to optimize their own utility. Therefore, unlike the single-principal assistance game, the multi-principal assistance game is no longer fully cooperative. This creates a problem for existing value alignment algorithms. These algorithms work under the assumption that the demonstrations and information provided are truly representative of the human's preferences. However, the misalignment between the AI system and each human's preferences yields a perverse incentive for the humans: can they misrepresent their preferences to gain a more desirable outcome?

In this work we introduce the alignment problem of an AI system with multiple principals and establish a strong connection with results in computational social choice theory and mechanism design. We consider a subclass of MPAGs that generalizes apprenticeship learning [3]. In *multi-principal apprenticeship learning*, the robot observes trajectories from multiple humans and then produces a trajectory that maximizes a social aggregate of the inferred rewards.

Our contributions are as follows:

- We introduce the problem of learning from multiple strategic demonstrators with possibly wildly divergent reward and formulate an impossibility result in this context.
- We introduce an algorithm that manipulates feature-matching *Learning From Demonstration* (LfD) algorithm in polynomial time, thus challenging the fact that computational hardness can be a barrier to manipulation in this context.
- We relate this problem to a real-world example, emphasizing the need to push research towards this direction.
- We propose a *collegial* mechanism to circumvent the impossibility result in this context. Specifically, a collegial mechanism exploits the fact that when collected in the field, demonstrations can have consequences for their demonstrators that are independent of the behavior of the AI system.

1.1 Related Work

Value Alignment. The need for AI systems to align with the preferences of humans is well documented in AI safety literature [1]. A first line of work formulates goal inference as an inverse planning problem [9]. For example, Inverse Reinforcement Learning (IRL) computes a reward such that the observed trajectory is optimal in the underlying Markov Decision Process (MDP) [2] [10]. A common assumption of inverse planning methods is that the robot does not influence the decision-making of the human. However, previous work has shown that the presence of a robot has a significant influence on humans [11] [12]. Furthermore, it has been shown that the robot can benefit from interacting with the human

to infer the goal. For example, Hadfield-Menell et al. have shown that if we formulate goal inference as a game between the human and the robot, observing the optimal trajectory of the human is not a Nash equilibrium of the game in general [7]. Our work extends this idea to the multi-agent setting: we show that when a robot acts on behalf of multiple humans using LfD tools for single-agent alignment, the best strategy of the humans depends on the strategy of the other humans. This motivates the need for developing LfD tools specifically for multi-agent alignment.

Computational Social Choice. Social Choice Theory (SCT) is a branch of Economics that studies the aggregation of individual preference towards a collective choice and encompasses many real-world scenarios like voting, fair allocation and auctions. A famous result of SCT is Gibbard's impossibility theorem which loosely states that any non-trivial[1] process of collective decision is subject to manipulation [13]. Much effort in the Computational Social Choice and Mechanism Design communities has been focused on identifying situation where Gibbard can be circumvented and develop computational tool against manipulation [14]. A first line of work exploits the fact that there are some restrictions on the domain of preferences such that Gibbard doesn't hold anymore. Two widely studied domain restrictions are the single-peaked preferences in the voting literature [15], requiring the utility functions to be uni-modal, and the quasi-linear utilities in the auction literature [16], requiring money transfer between the users and the system to be applicable to the real world. Yet real-world demonstrations usually come from multi-modal reward function and we don't consider money transfer between the AI system and the users here, thus this restrictions are not applicable in our context. A second line of work exploits the fact that it might be computationally hard to manipulate a system. Yet we show that it is not hard to manipulate the value alignment methods we are considering in this paper by proposing an algorithm that computes a best-response in polynomial time. In this work we propose a natural way to circumvent Gibbard when learning from demonstration: collecting demonstrations that are meaningful for the demonstrator.

Cost of Lying. As we will see, collegial mechanisms incur a natural cost of lying to the demonstrator. Cost of lying has been introduced in many different scenarios, eg. guilt aversion [17], altruism [18] and reciprocity [19]. A widely studied model in economics is the model of partial verification, where the system can detect a lie when it is too far from the truth [20], in that case inflicting an infinite cost to the liar. Our model can be seen as a soft version of partial verification [21]. A line of work in the voting literature is interested in costly voting, where voters can pay more or less to express the degree of their preferences [22]. Yet it is hard to link the utility to the willingness to pay, especially when voters have unequal wealth.

Learning from multiple demonstrators. Few works have been interested in learning a single-agent task from multiple demonstrators. Castro et al. introduce a maximum margin algorithm that exploit the fact that the multiple demonstrators have different known levels of expertise [23]. Noothigattu et al. show that feature matching algorithms recover a good policy when the demonstrators are optimal with respect the a random perturbation of the same underlying reward [24]. In contrast, we show that feature matching algorithms are easily manipulable by a strategic demonstrator. None of these works consider strategic demonstrators.

Human-Robot Team Robot evolving in a multi-human environment has already been studied by the Human-Robot Interaction community. Much work has focused on trust building and resource allocation [25]. A common assumption is that the robot and the humans have a common payoff known to the robot. Our work generalize this setting to general-sum payoffs possibly unknown to the robot.

[1] A process is non-trivial when it is neither dictatorial nor limiting the possible outcome to two options only.

2 Impossibility Result for Multi-Agent LfD Methods

2.1 Application of Gibbard's Theorem to LfD

We consider a finite Markov Decision Process without reward (S, A, P, μ_0, T) where:

- S is a finite set of states.
- A is a finite set of actions.
- $P : S \times A \times S \to [0, 1]$ is a stochastic transition function.
- μ_0 is a initial state distribution
- T is a finite horizon

In IRL, an AI system observes an expert trajectory $\tau \in (S \times A)^T$ and computes a reward $R : S \times A \to \mathbb{R}$ that makes this trajectory optimal [2]. Apprenticeship Learning (AL) methods uses this reward as a proxy to compute a stochastic policy $\pi : S \times A \to [0, 1]$ that best imitates the expert [3].

We propose *multi-principal apprenticeship learning* as a generalization of AL. We suppose that there are N demonstrators, each with a private reward function $R_i : S \times A \to \mathbb{R}$ and providing one trajectory τ_i to the AI system. The AI system observe all trajectories, compute a stochastic policy using an AL method for example and follows the policy to produce a trajectory $\tau_R \in (S \times A)^T$. The goal of the AI system is to maximize a social welfare function W of the true rewards:

$$\tau_R^* \in \arg\max_{\tau} W(R_1(\tau), ..., R_N(\tau)) \tag{1}$$

Social welfare functions are a heavily studied field, examples include the *utilitarian* criterion $W_U(R_1, ..., R_N) = \sum_h R_h$ [26] and the *egalitarian* criterion $W_E(R_1, ..., R_N) = \min_h R_h$ [27, 28]. In the remainder of the paper we consider the utilitarian criterion.

The process leading to τ_R can be represented by a stochastic function $g : ((S \times A)^T)^N \to \triangle((S \times A)^T)$. The objective of human i is to lead the AI system towards a trajectory that maximize their own utility:

$$\tau_i^* \in \arg\max_{\tau_i} \mathbb{E} R_i(g(\tau_i, \tau_{-i})) \tag{2}$$

Since $(S \times A)^T$ is a finite non-empty set, g is a *game form* as defined by Gibbard and we can apply his impossibility result for non-deterministic process [13]:

Theorem 1 (Gibbard 1978)**.** *On the domain of versatile*[2] *trajectories, any straightforward*[3] *mechanism must be a probability mixture of mechanisms of two kind:*

- *Duple mechanisms, where the set of possible trajectories are restricted to two.*
- *Unilateral games, where one human gets to choose among a certain set of possible lotteries over trajectories.*

Thus the only straightforward LfD mechanisms are not acceptable mechanisms. A first solution widely explored in the mechanism design literature would be to constraint the domain of preferences. Yet reward functions for real-world tasks can have various structure and be highly multi-modal, and we don't consider here money transfer between the demonstrators and the AI system. If we can't constraint the domain of preferences, we can hope that manipulating an AI system observing trajectories is computationally hard. Yet we challenge this hope in the following section by introducing an algorithm that manipulates feature matching algorithms, a widely used family of LfD algorithms, in polynomial time.

[2] A trajectory is versatile if the set of utility profile for which it is dominant has interior points.

[3] A straightforward mechanism induces a game where every player has a weakly dominant strategy. It is equivalent to say that it is not manipulable.

2.2 A polynomial-time algorithm to manipulate feature-matching algorithms

Current methods in learning from demonstrations are not adapted for dealing with multiple humans because they are easily manipulable. For example, if an AI system uses a feature matching algorithm on an aggregation of demonstrations coming from different humans, we can find a demonstration such that the robot policy is biased towards a single demonstrator in polynomial time. More specifically, we have the following result:

Proposition 1. *Suppose that the AI system models the humans as noisily-optimal planners with linear features and computes the rewards that maximize the likelihood of the aggregated demonstrations [10]. The best-response trajectory of a human can be computed in polynomial time.*

To show that, we transform the manipulation problem into a least square problem. The objective of the AI system is the following:

$$\begin{aligned} \omega^* &= \max_{\omega} P(\tilde{\tau}|\omega, \rho_0) \\ P(\tilde{\tau}|\omega, \rho_0) &= \prod_{i=1}^{N} \frac{e^{\phi(\tau^i)^T\omega}}{Z(\omega, \rho_0)} \\ Z(\omega, \rho_0) &= \sum_{\tau, s_0 \sim \rho_0} e^{\phi(\tau)^T\omega} \end{aligned} \tag{3}$$

where N is the number of demonstrators, $\tilde{\tau} = (\tau_1, ..., \tau_n)$ is the aggregate of trajectories and $\phi(\tau^i)^T\omega = \sum_{t=1}^{T} \phi(s_t^i)^T\omega$ is the cumulative return of τ^i under reward ω.

By taking the gradient to zero, we see that this concave objective is maximized when the expected feature count of the computed reward's optimal policy is equal to the empirical feature count of the aggregated demonstrations:

$$\mathbb{E}_{\tau \sim \pi^*(w^*), s_0 \sim \rho_0}(\phi(\tau)) = \frac{\sum_{i=1}^{N} \phi(\tau^i)}{N} \tag{4}$$

This gives the strategic demonstrator a simple procedure to bias the system towards their own interest: they give a demonstration such that the empirical feature count of the aggregated demonstrations is the closest possible to their own policy's expected feature count. Formally, this translates into the following least squares objective:

$$\tau_i^*(\tau_{-i}) = \min_{\tau_i} \left\| \phi(\tau_i) - (N\mathbb{E}_{\tau \sim \pi^*(w_i), s_0 \sim \rho_0}(\phi(\tau)) - \sum_{j \neq i} \phi(\tau_j)) \right\|_2^2 \tag{5}$$

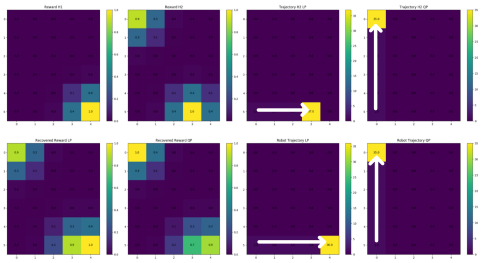
Concretely, a strategic demonstrator will act like they like (dislike) something more than they actually do to push the AI system towards (away from) a particular outcome.

An efficient way to solve this problem is to find the occupancy measure that minimizes the following constrained least squares problem:

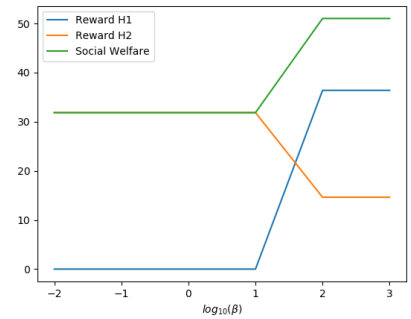
$$\begin{aligned} \min_{\rho_{s,a}^t} \quad & || \sum_{s,a,t} \rho_{s,a}^t \phi(s) - (N\mathbb{E}[\phi|w] - \sum_{j \neq i} \phi(\tau_j))||^2 \\ \text{subject to} \quad & \sum_a \rho_{s',a}^{t+1} = \sum_{s,a} P(s, a, s') \rho_{s,a}^t \ \forall s', \forall t \in [0, T-1] \\ & \sum_a \rho_{s,a}^0 = \mu_0[s] \ \forall s \end{aligned} \tag{6}$$

This is a linearly constrained least squares problem solvable in polynomial time. We can compute the best-response trajectory directly from the occupancy measure in linear time. We test our solver in a gridworld environment (see Fig. 1a) and observe that:

- The best-response trajectory can be computed in very short time, thus computation time is not a barrier to manipulation.
- The best-response trajectory is not what the agent would have picked in isolation and depend on other humans' trajectories. In other words, this LfD method is not straightforward in general, in accordance with Gibbard's theorem.



(a) First row: R_1; R_2; H_2's honest trajectory; H_2's strategic trajectory. Second row: Recovered reward when H_2 is honest; Recovered reward when H_2 is strategic; Robot's trajectory when H_2 is honest; Robot's trajectory when H_2 is strategic (H_1 is always honest).



(b) Return of H_1 and H_2 and social welfare from the robot's trajectory against β (H_1 is always honest).

2.3 An Example of Real-World Misalignment: Microsoft Tay

In 2016, Microsoft released a Twitter chatbot, Tay, designed to learn to converse via tweets. It took less than 24 hours for a group of prankster users to train Tay to mix racist comments into its discourse. Tay had at least three conceptual problems:

1. Manipulable inputs. Tay was not trained on chat logs 'in the wild'; it was trained by humans who knew there was a system that could be manipulated to achieve goals outside its intended purpose.
2. User/creator misalignment. TAY's creators primarily wanted TAY to imitate a normal person, not entertain people (although entertainment was useful to gain more engagement and data). By contrast, its users primarily wanted to be entertained. This means there was a misalignment between the creators as principals and the users as principals.
3. User/user misalignment. Prankster users wanted offensive entertainment, presenting a misalignment between different groups of the users. The pranksters were able to increase TAY's level of racism to an unusual degree using:
   (a) extreme features (highly racist inputs)
   (b) extreme numbers of inputs

In this work, we are concerned primarily with problems analogous to problem 1 and problem 3(a). Problem 1 raises the question of how to elicit honest responses, i.e., incentive compatibility. Problem 3(a) raises the question of users exaggerating their preferences to move the value of a 'compromise' policy closer to their desired policy, which is a kind of manipulability for many learning procedures that we have recovered in the previous subsection by computing the best-response to a feature-matching LfD algorithm (see Equ. 5). As we will see in the next section, these problems can both be mitigated by requiring demonstrations to have natural consequences outside of the AI's policy, thereby eliciting more 'normal' behavior from demonstrators.

3 Circumventing Gibbard's Theorem with Collegial Mechanisms

3.1 Exploiting the Consequential Nature of Real-World Demonstrations

Even if it is not hard to compute what trajectory would bias a system towards an individual interest, there are scenarios in real life where a human would prefer to stay honest. This is due to the fact that when collected in the field, demonstrations might have consequences for their demonstrators that are independent of the behavior of the AI system. In our setting, this can be formulated using the same objective with an additional term, the direct reward the human will get by choosing an action with a coefficient β that quantifies the importance

that the human puts on the direct consequences of its demonstration relatively to the robot's action:

$$\tau_i^* \in \arg\max_{\tau_i} \beta R_i(\tau_i) + \mathbb{E}R_i(g(\tau_i, \tau_{-i})) \tag{7}$$

To obtain a clear bound on β we suppose that the reward function are integer-valued: $R_i : S \times A \to \mathbb{N}$ and that there is $M \in \mathbb{N}$ such that: $\forall i, \forall (s,a) \in S \times A, R_i(s,a) \leq M$. Notice that the Gibbard's impossibility result stated in the previous section still holds when we take integer-valued reward functions with a fixed upper-bound. We have the following result:

Proposition 2. *If $\beta > M$, then every mechanism is straightforward.*

This result shows that we can circumvent Gibbard's theorem by looking for situation where demonstrations are the most meaningful, incurring a natural cost of lying for the demonstrator.

A similar bound can be obtained for real-valued reward functions. For every human i, we define $R_i^* = \max_{s,a} R_i(s,a)$, we assume that there is (s,a) such that $R(s,a) < R_i^*$ and we define $\gamma_i = \min_{(s,a)}\{R_i^* - R_i(s,a) : R(s,a) < R_i^*\}$. We also define $\gamma = \min_i \gamma_i$. Since we consider a finite MDP with a finite number of humans we have $\gamma > 0$. We have the following result:

Proposition 3. *If $\beta > \frac{M}{\gamma}$, then every mechanism is straightforward.*

Even when $\frac{M}{\gamma} \to \infty$ and $\beta < \frac{M}{\gamma}$, collecting meaningful demonstrations can significantly reduce the manipulability of a mechanism. To see that, we consider a plurality voting system with random tiebreak with 3 voters and 3 alternatives and compare the proportion of manipulable profile when $\beta = 0$ and $\beta = 1$ for the utilities domain $\{R : \{1,2,3\} \to [0,1], R(1) + R(2) + R(3) = 1\}$ under which Gibbard's theorem still holds and such that $\gamma \to 0$. Using a geometric argument on the 2-simplex we show that:

Proposition 4. *In a system using plurality voting with random tiebreak with 3 voters and 3 alternatives, $\frac{1}{3}$ of the simplex is manipulable[4] when $\beta = 0$ while only $\frac{1}{9}$ of the simplex is manipulable when $\beta = 1$.*

We can efficiently compute the best-response trajectory when $\beta > 0$ by adding a term to the objective of the previous optimization problem (see Equ. 6):

$$\begin{aligned} \max_{\rho^t_{s,a}} \quad & \beta \sum_{s,a,t} \gamma^t \rho^t_{s,a} \phi(s)^T w - || \sum_{s,a,t} \rho^t_{s,a} \phi(s) - (N\mathbb{E}[\phi|w] - \sum_{j \neq i} \phi(\tau_j))||^2 \\ \text{subject to} \quad & \sum_a \rho^{t+1}_{s',a} = \sum_{s,a} P(s,a,s') \rho^t_{s,a} \ \forall s', \forall t \in [0, T-1] \\ & \sum_a \rho^0_{s,a} = \mu_0[s] \ \forall s \end{aligned} \tag{8}$$

We recover a regularized dual of the linear program formulation for finite-horizon discounted Markov Decision Process [29].

We plot the social welfare obtained by the robot's policy in our gridworld setting against the importance that the strategic demonstrator put on their demonstration (see Fig 1b). We observe that when β is higher than 100, H_2 is incentivized to be honest and the social welfare increases significantly.

Thinking back about our real-world example, if the Tay bot had been reading from people's work account instead of anonymous Twitter feed, the problem would not have occurred, since there's a greater negative utility to the human for providing profane examples in the former case.

[4] We say that a utility function is manipulable when the associated best response strategy depends on the strategy of the other humans.

3.2 Towards Efficient Mechanisms with Collaboration beyond Demonstrations

So far we have considered a subclass of MPAGs where the AI system is passively observing the humans. Although it enables a clear comparison with the social choice theory, it is arguably not the best way to learn human values. A challenge of learning from multiple demonstrators is that demonstrations give only one mode of the reward function. Yet to maximize the social welfare we certainly need more information: there is cardinal utility profile such that the social maximizing action is sub-optimal with respect to each of the individual utilities[5].

In this section we widen the considered class of MPAGs to yield an approximately efficient[6] mechanism. Specifically, we assume that the AI system learns human values through a human-robot collaborative task. Consequently, the AI system has an influence on the utility the human get when demonstrating their preferences.

We obtain a non-trivial asymptotic worst case bound on the social welfare in a simplified model of human-robot collaboration. We consider a stateless sequential setting where at each time step, the robot can choose one human (among N humans) to collaborate with. During the collaboration, the human chooses one action (among M actions) and at the end of the step, the robot chooses weather the human get the associated reward. The robot chooses at which time step to stop and then chooses an action. We wish to maximize the social welfare of this action.

A robot mechanism $\mathcal{M}$ is given by a human selection criterion, a reward allocation criterion, a stopping time criterion and an action selection criterion. We define the distortion[7] of the robot's mechanism as:

$$\triangle(\mathcal{M}) = \max_{R} \frac{\max_a \sum_h R_h(a)}{\mathbb{E} \sum_h R_h(a_{\mathcal{M}}(R))} \tag{9}$$

We propose a mechanism that achieves a non-trivial asymptotic distortion (see Algorithm 1). In broad outline, the robot chooses a human and allocates reward only if the human did not choose the action before. Periodically, the robot chooses a random action with probability $1 - \frac{1}{2^{\frac{1}{M}}}$.

Exploiting the fact that humans plan with a discount factor strictly less than 1 and using recent tools from the ordinal voting literature[30], we obtain the following bound:

Proposition 5. $\triangle(\mathcal{M}) = O(\sqrt{M \log M})$

4 Conclusion

In this paper, we explore an area of concern in the study of AI alignment—ensuring that AI systems are designed so that humans agents are incentivized to interact with AI systems in a "honest" way. We view our main contributions as follows:

- Propositions 2, 3 and 4 show that collegial preference inference can yield numerous desirable properties including incentive-compatibility.
- Proposition 5 reveals the asymptotic performance of a mechanism coupling collegial preference inference with human-robot collaboration.

These results appear to be reasons for optimism in the domain of mechanism design for multi-principal assistance games. Meanwhile, the overall problem of preventing manipulative behavior in multi-human AI systems is open and presents many opportunities for further work. Our methods are applied to fairly simple problems; there exists a need to generalize these results to more general theoretical settings and more complicated situations in the real world.

[5] Consider the utility profile $\{(0.6, 0.4, 0), (0, 0.4, 0.6)\}$ in a stateless MDP with 3 actions and 2 humans.

[6] A mechanism is efficient if it maximizes the social welfare.

[7] The distortion is a notion introduced in the ordinal voting literature.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pages 663–670, 2000.

[3] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[4] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[5] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in neural information processing systems*, pages 6765–6774, 2017.

[6] Mark Woodward, Chelsea Finn, and Karol Hausman. Learning to interactively learn and assist. *CoRR*, abs/1906.10187, 2019.

[7] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[8] Amartya Sen. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181, 1986.

[9] Chris L Baker, Joshua B Tenenbaum, and Rebecca R Saxe. Goal inference as inverse planning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

[10] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[11] Ben Robins, Kerstin Dautenhahn, Rene Te Boekhorst, and Aude Billard. Effects of repeated exposure to a humanoid robot on children with autism. In *Designing a more inclusive world*, pages 225–236. Springer, 2004.

[12] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, 19(1-2):61–84, 2004.

[13] Allan Gibbard. Straightforwardness of game forms with lotteries as outcomes. *Econometrica: Journal of the Econometric Society*, pages 595–614, 1978.

[14] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[15] Duncan Black et al. The theory of committees and elections. 1958.

[16] Theodore Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

[17] Pierpaolo Battigalli, Gary Charness, and Martin Dufwenberg. Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232, 2013.

[18] Rudolf Kerschbamer, Daniel Neururer, and Alexander Gruber. Do altruists lie less? *Journal of Economic Behavior & Organization*, 157:560–579, 2019.

[19] Ernst Fehr and Simon Gächter. Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives*, 14(3):159–181, 2000.

[20] Ioannis Caragiannis, Edith Elkind, Mario Szegedy, and Lan Yu. Mechanism design: from partial to probabilistic verification. In *Proceedings of the 13th acm conference on electronic commerce*, pages 266–283, 2012.

[21] Andrew Kephart and Vincent Conitzer. The revelation principle for mechanism design with reporting costs. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, page 85–102, New York, NY, USA, 2016. Association for Computing Machinery.

[22] Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pages 33–37, 2018.

[23] Pablo Samuel Castro, Shijian Li, and Daqing Zhang. Inverse reinforcement learning with multiple ranked experts. *arXiv preprint arXiv:1907.13411*, 2019.

[24] Ritesh Noothigattu, Tom Yan, and Ariel D Procaccia. Inverse reinforcement learning from like-minded teachers.

[25] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Reinforcement learning with fairness constraints for resource distribution in human-robot teams. *arXiv preprint arXiv:1907.00313*, 2019.

[26] C. Liu, X. Xu, and D. Hu. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398, 2015.

[27] Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2014.

[28] Dritan Nace and Michal Pióro. Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial. *IEEE Communications Surveys & Tutorials*, 10(4):5–17, 2008.

[29] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[30] Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190 – 213, 2015.

Algorithm 1: Approximately Efficient Mechanism

Input : Number of humans N; Number of arms M
Output : Robot action a_R

```
for h ∈ [1, N] do
|  Actions[h] ⟵ [1, M]
end
for a ∈ [1, M] do
|  Score[a] ⟵ 0
end
for t ∈ [1, M] do
|  for h ∈ [1, N] do
|  |  Let human h choose an action a
|  |  if a ∈ Actions[h] then
|  |  |  Execute a
|  |  |  Scores[a] += 1/t
|  |  |  Remove a from Actions[h]
|  |  end
|  end
|  With probability 1 − 1/(2·1/M), return an arm sampled uniformly on [1, M].
end
```

Return arm a with probability $\frac{Score[a]}{\sum_{a'} Score[a']}$