National Research University Higher School of Economics

as a manuscript

Dmitry Ivanov

Application of machine learning to game theory problems: auctions and Markov games

PhD Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Moscow – 2024

The PhD dissertation was prepared at the National Research University Higher School of Economics

Academic Supervisor: **Alexander Nesterov**, PhD; associate professor of economics, head of the international laboratory of game theory and decision making, National Research University Higher School of Economics

Game theory provides a mathematical framework to model strategic interactions between multiple parties, revealing deep insights into competitive and cooperative scenarios alike. Core to game theory (and economics in general) is the premise of rationality of all agents involved. In relation to humans, an ideal agent that optimally processes information, incurs no computational costs, avoids errors, exhibits no biases, and overall acts perfectly with respect to their goals, is often referred to as *homo economicus*. Parkes and Wellman (2015) astutely observed that Artificial Intelligence (AI) agents could be a better fit to these ideals, and coined a term *machina economicus* as a synthetic antipode to the perfectly rational human agent.

Of course, neither species exists.[1] While humans deviate from the rationality premise in an uncountable number of ways, modern generative AI models are known to fail on trivial problems (like counting R's in 'Strawberry'), hallucinate (Zhang et al., 2023; Huang et al., 2023), and even exhibit the same cognitive biases as humans (Schramowski et al., 2022; Acerbi and Stubbersfield, 2023). Still, game theory has long provided valuable models of human behavior, making its extension to AI a natural progression. The potential for synergies is therefore immense.

As AI becomes increasingly integrated in all facets of society, it is imperative to develop methods tailored for analyzing, understanding, and guiding interactions of AI agents, especially in the presence of distinct and potentially conflicting incentives. Game theory and economics offer a rich array of tools that can be adapted for this purpose (Conitzer, 2019; Hadfield-Menell and Hadfield, 2019), as has already been demonstrated in such diverse areas as classification (Ghalme et al., 2021), recommender systems (Bahar et al., 2020), multi-agent reinforcement learning (Leibo et al., 2017), and even large language models (Duetting et al., 2024).

The reverse direction is no less exciting: machine learning opens up new avenues for tackling game-theoretic problems that were previously infeasible. One such advancement is the emerging field of differentiable economics (Dütting et al., 2024), which employs deep learning techniques in areas like auction design (Dütting et al., 2019; Curry et al., 2023) and contract design (Wang et al., 2024).

This dissertation showcases examples from both directions, demonstrating the reciprocal enrichment of machine learning and game theory.

[1] As of the time of writing, singularity has yet to occur.

Relevance and Significance

My first study advances the field of automated design of revenue-maximizing auctions through deep learning. The classic approach vastly employed in the literature is to derive analytic solutions by applying pen-and-paper theoretic analysis to subsets of problems or even particular problem instances (Myerson, 1981; Manelli and Vincent, 2006; Pavlov, 2011; Giannakopoulos and Koutsoupias, 2014; Daskalakis et al., 2015; Yao, 2017; Haghpanah and Hartline, 2021). This involves narrowing down the problem space through specifying auction parameters, such as the number of items being sold, the number of participants, and/or the distributions of valuations of each participant over each item bundle. Besides the scrutiny required to analyze each particular setting, as well as the unrealistic requirement of access to private information, this approach becomes infeasible even in seemingly innocent settings involving only two participants and two items.

As an alternative, automated auction design (Conitzer and Sandholm, 2002, 2003, 2004) takes a computational perspective and employs data-driven methods in order to approximate optimal solutions in any setting. A breakthrough in this field is the celebrated RegretNet framework (Dütting et al., 2019), which parameterizes the auction mechanism as a neural network. RegretNet takes the agents' bids for all items as input, which it processes through a multi-layered perceptron to output probabilistic item allocations between participants, as well as payments for each participant. It is trained using a nuanced loss function that reflects a mixture of two objectives: revenue (maximize the total of payments) and bidder truthfulness (minimize regret, a quantitative measure of participants' incentives to misreport their bids).

I build upon RegretNet by introducing two independent improvements. Firstly, I present RegretFormer, a neural architecture leveraging attention layers, which offers better performance and generalization capabilities than the prior alternatives. Secondly, I propose a novel loss function optimized through dual gradient descent, simplifying hyperparameter tuning and providing a clear, interpretable mechanism to balance the trade-off between the two objectives. Both improvements are validated through an extensive and intricate empirical study that goes beyond the standard comparison of performance metrics. Overall, this work presents a new state-of-the-art approach to automated auction design.

In **my second study**, I critically examine the prevalent assumption in Multi-Agent Reinforcement Learning (MARL) that equates the cooperation of self-interested agents with social welfare maximization. The dominant view on the problem of cooperation is purely computational, allowing unbounded intervention into the agents' objectives, e.g. by shaping rewards (Peysakhovich and Lerer, 2018a,b; Hughes et al., 2018; Jaques et al., 2019; Wang et al., 2019; Eccles et al., 2019; Jiang and Lu, 2019; Durugkar et al., 2020; Yang et al., 2020; Zimmer et al., 2021; Phan et al., 2022), or private information, e.g. by sharing parameters (Gupta et al., 2017). Given the complexity of temporally and spatially extended mixed-motive environments typically studied through MARL (and formalized as Markov games, Leibo et al. (2017)), this conventional approach is convenient in simplifying both training and validation. However, it overlooks the importance of respecting agents' individuality and susceptibility to exploitation by selfish actors. Challenging this norm, I argue that cooperation should emerge from the strategic decision-making of rational agents as a socially beneficial equilibrium, robust against deviations for personal gains.

Inspired by advances in game theory (Monderer and Tennenholtz, 2009), I propose using mediators as an implementation of this refined concept of cooperation. Mediators are benevolent entities that may act on behalf of the agents who consent to the mediation. Crucially, if an agent does not find mediation acceptable, it may choose to act in the shared environment itself. However, in this case, the mediator will not consider this agent's welfare when acting for other agents (who did agree to the mediation). This complex interplay requires the mediator to carefully balance all agents' incentives and guide them towards mutually beneficial equilibria implemented through unanimous mediation. To train the mediator and the agents, I parameterize both parties as neural networks, formulate their interaction as an optimization problem constrained by agents' incentives, and solve it using the policy gradient.

I demonstrate the effectiveness of this strategy in achieving cooperative equilibria without compromising individual agency in classic social dilemmas and public good games, as well as their sequential modifications with analytically intractable state spaces. This novel methodology opens new avenues for creating more resilient and equitable agent interactions in complex mixed-motive environments.

Finally, **my third study** contributes to the field of personalized ML, which concerns tailoring a model's decisions to individuals' unique characteristics and preferences (den Hengst et al., 2020). Specifically, I focus on personalization opportunities in high-stakes domains like healthcare and autonomous driving. In these domains, the deployment of any automated solution necessitates a rigorous regulatory approval process (Breton et al., 2020), making personalization to each user infeasible. To address this, I propose a framework coined represented Markov Decision Processes (r-MDPs), which is designed to strike a delicate balance between the need for personalization and the regulatory constraints. This framework models a scenario where a population of users, each with distinct preferences, may choose from a limited set of representative policies to act in a single-agent MDP on their behalf. The task of the designer then comprises two interdependent aspects: train the representative policies (the computational aspect) and match each user to a policy such that the overall social welfare is maximized (the game-theoretic aspect). Once the policies are manufactured in a simulator, they can be submitted for approval by regulatory entities, and finally deployed in the real world.

Delving deeper into the problem, I recognize the intractability of directly solving r-MDPs due to the exponential complexity introduced by the need to select the most appropriate policies for each user from a constrained set. To address this, I draw inspiration from classical clustering algorithms, such as K-means and Expectation-Maximization (MacQueen, 1967; Dempster et al., 1977; Lloyd, 1982), formulating two deep reinforcement learning algorithms that iteratively refine policy assignments and optimize the policies. These algorithms are supported by robust theoretical underpinnings: each iteration, they monotonically improve, and thus eventually converge to local maxima of social welfare.

The empirical investigations span across diverse simulated environments, from toy but demonstrative Resource Gathering (Barrett and Narayanan, 2008) to complex control tasks in MuJoCo (Todorov et al., 2012), demonstrating the versatility and effectiveness of the algorithms in delivering personalized policies under stringent budget constraints. These results not only validate the practicality of my approach in offering meaningful personalization within regulated domains but also illuminate the path for future explorations into extending these methodologies to real-world applications, further bridging the gap between the theoretical ideals of machine learning and the pragmatic demands of regulatory compliance.

Research Objectives

1. Advance the frontiers of automated auction design through deep learning via the use of self-attention layers.
2. Showcase a game-theoretic perspective on cooperation in mixed-motive Markovian environments through the use of mediators.
3. Propose a compromise approach to personalized RL tailored for domains where deployment of distinct policies is costly.

Key Results

Based on the studies described above, I formulate the following **key results to be defended**:

1. The proposed RegretFormer architecture based on self-attention layers is the new state-of-the-art in the automated auction design. Furthermore, the proposed loss function modification based on dual gradient descent is less sensitive to hyperparameters and unambiguously controls the revenue-regret trade-off.
2. Mediators can be applied in mixed-motive MARL to create new socially beneficial equilibria. These equilibria can be identified with my algorithm by applying policy gradient to a constrained optimization problem that I specified.
3. Meaningful personalization of ML models to a population of users can be achieved with only a handful of solutions. In the context of RL, the policies representing these solutions can be trained with my algorithms that combine the high-level structure of K-means and EM clustering with policy optimization through policy gradient.

Personal contribution

These results were achieved in collaboration with experts in the field and bright students. However, in all studies, I was a core contributor, as evidenced by my first authorship in all three publications that constitute this dissertation.

The first study I did with a team of peers. I led the project and actively contributed to formulating research directions and hypotheses, as well as to implementing algorithmic developments and experiments. The core contributions of the study – the state-of-the-art architecture and the improved loss function – are based on my ideas. I actively contributed to writing the paper.

I worked on the second study with students. I led this project, formulating the research direction of applying mediators in MARL, deriving a constrained optimization problem, proposing to solve it using policy gradient, and designing experiments. The students handled the codebase, implementing the algorithm and most of the experiments based on my directions and ideas. The paper was written entirely by me.

The third study was done in collaboration with an academic expert in the fields of ML and game theory, who formulated the practical problem of personalization in high-stakes domains and proposed a clustering-inspired RL solution. I took the research from there, proposing a modified version of the algorithm (both of which made it into the publication), designing experiments, and implementing the codebase. The paper was mostly written by me, barring part of the introduction.

Publications and Approbation of Research

I have a total of seven publications in proceedings of international peer-reviewed conferences. Three of these publications constitute this dissertation.

First-tier publications

1. **Ivanov, D.**, Safiulin, I., Filippov, I., & Balabaeva, K. (2022). Optimal-er auctions through attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 34734-34747.

2. **Ivanov, D.**, Zisman, I., & Chernyshev, K. (2023). Mediated Multi-Agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Vol. 22, pp. 49-57.

3. **Ivanov, D.**, & Ben-Porat, O. (2024). Personalized Reinforcement Learning with a Budget of Policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 12735-12743.

Reports at conferences and seminars

1. Poster Presentation at *the 36th Conference on Neural Information Processing Systems (NeurIPS)*, December 2022, New Orleans, USA (virtual). Optimal-er auctions through attention.

2. Poster Presentation at *the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2023, London, UK. Mediated Multi-Agent Reinforcement Learning.

3. Presentation at *an internal research seminar in DeepMind*, June 2023, London, UK. Mediated Multi-Agent Reinforcement Learning.

4. Pre-recorded presentation at *the 38th AAAI Conference on Artificial Intelligence*, February 2024, Vancouver, Canada. Personalized Reinforcement Learning with a Budget of Policies.

Content of Works

Optimal-er Auctions through Attention

Auction Design

In this part of my thesis, I examine auction mechanisms where a group of bidders, each denoted by $N = 1, ..., n$, express their interests in a collection of items, labeled $M = 1, ..., m$, through valuation functions v_i. These functions represent how each bidder evaluates the items, with a key assumption being the additivity of valuations: the total value a bidder assigns to a set of items is the sum of the values they assign to each individual item.

The core of this study revolves around understanding how bidders, each with their valuation functions drawn from specific distributions, interact within the auction framework. An auctioneer, equipped with a sample of past valuation profiles, attempts to design an optimal auction, yet is challenged by the lack of direct knowledge about the bidders' true valuations or their distributions.

The formulation of the auction includes rules for item allocation between participants and payments of each participant, striving for mechanisms where bidders are incentivized to bid their true valuations – a concept known as dominant strategy incentive compatibility (DSIC). Additionally, the auction aims to be ex-post individually rational (IR), ensuring that bidders do not regret their participation regardless of the outcome.

The task of designing optimal auctions, particularly in settings with multiple items, is treated as an optimization problem. The objective is to maximize expected revenue under DSIC and IR constraints. While the problem is well-understood for single-item auctions, extending the principles to multi-item auctions introduces significant complexity without straightforward solutions.

RegretNet

Building on the innovative work of RegretNet (Dütting et al., 2019), my thesis explores advancements in the realm of optimal auction design through deep learning. The core of RegretNet is its dual-network architecture, comprising an allocation network and a payment network. These networks function by processing a structured input – a bid matrix representing the bids of all participants for all items – through multiple fully connected layers to produce meaningful outputs. Specifically, the allocation network determines the probabilities of item allocation among bidders, translating bid matrices into allocation probabilities. Meanwhile, the payment network calculates the payments each bidder must make, again based on the bid matrix, determining the portion of a bidder's expected utility to be transferred to the auctioneer.

A novel aspect of RegretNet is its optimization objective. The architecture is designed to maximize revenue subject to the DSIC constraint. This constraint is operationalized through the concept of ex-post regret, which measures a surplus of utility a bidder would gain by optimally deviating from their true valuation in their bid. RegretNet aims to minimize this regret to zero, ensuring that bidders have no incentive to misreport their valuations. The optimization process is facilitated by the augmented Lagrangian method, which balances the two conflicting objectives of revenue maximization and regret minimization.

My thesis builds upon this foundation, aiming to refine and extend the capabilities of deep learning models in the complex landscape of auction theory.

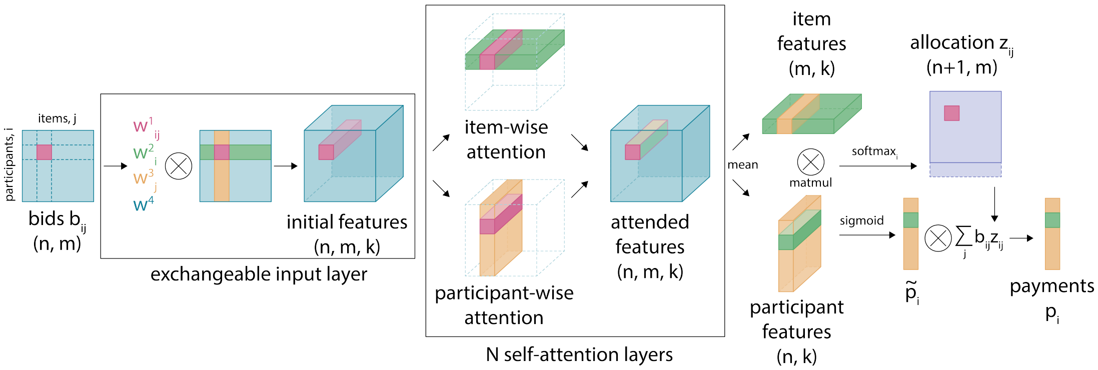


Figure 1: RegretFormer

My Modifications of RegretNet

I introduce two significant advancements to the RegretNet framework for optimal auction design: the RegretFormer architecture based on self-attention layers and an alternative constrained objective for a more tractable loss function.

On the one hand, the RegretNet's architecture, while pioneering, faces challenges related to the sensitivity of auction outcomes to the order of items and participants in the bid matrix, its limitation to a constant number of participants and items, and the expressiveness of its fully connected layers. These issues hinder its practical applicability and ability to generalize across different auction settings.

To overcome these limitations, I propose RegretFormer, a novel architecture that incorporates attention layers. The architecture is illustrated in Figure 1. Specifically, the self-attention layers are applied to a feature map produced from the bid matrix both item-wise and participant-wise. The outputs from these self-attention layers are combined via a fully connected layer, and this process can be repeated multiple times. The final step involves processing these outputs to produce the allocation matrix and payment vector. This design ensures that the architecture remains agnostic to the order of bids and enables its applicability to auctions with varying numbers of items and participants. Furthermore, the expressivity of attention layers improves performance on large problems.

On the other hand, RegretNet's original training procedure relies heavily on the precise tuning of hyperparameters to manage the trade-off between its dual objectives. This process is not only cumbersome but also fraught with the potential for performance degradation if the parameters are not optimally set (Rahme et al., 2021b).

Table 1: Architecture comparison

R_{max}	setting	RegretNet		EquivariantNet		RegretFormer	
		revenue	regret	revenue	regret	revenue	regret
10^{-3}	1x2	0.572	0.0007	**0.586**	0.00065	0.571	0.00075
	2x2	0.889	0.00055	0.878	0.0008	**0.908**	0.00054
	2x3	1.317	0.00102	1.365	0.00084	**1.416**	0.00089
	2x5	2.339	0.00142	2.437	0.00146	**2.453**	0.00102
	3x10	5.59	0.00204	5.744	0.00167	**6.121**	0.00179
10^{-4}	1x2	0.551	0.00007	0.548	0.00013	**0.556**	0.00014
	2x2	0.825	0.00005	0.75	0.00005	**0.861**	0.00006
	2x3	1.249	0.00007	1.226	0.0001	**1.327**	0.00011
	2x5	2.121	0.00013	2.168	0.00017	**2.339**	0.00015
	3x10	5.02	0.00062	5.12	0.00025	**5.745**	0.00022

To overcome these challenges, I propose a simplified and more intuitive framework that prioritizes revenue maximization within a predefined regret budget. This new objective is formalized as a relaxed version of the constrained optimization problem, aiming to minimize the average regret across all bids without exceeding a specified maximum threshold. Furthermore, this alternative formulation introduces an automatic adjustment mechanism for the Lagrange multiplier through dual gradient descent.

This revised approach presents two significant advantages. Firstly, it eliminates the need to balance conflicting objectives through several hyperparameters, simplifying hyperparameter tuning. Secondly, the explicit setting of a regret budget makes the design process more straightforward and less sensitive to hyperparameter variations. Empirically, this method has proven to be robust across various settings, requiring minimal adjustments to the regret budget parameter.

Through these modifications, I address some of the limitations of the original RegretNet, offering a clearer, more efficient path to optimal auction design.

Table 2: Ratio of the estimated regret to the regret budget; it should be close to 1

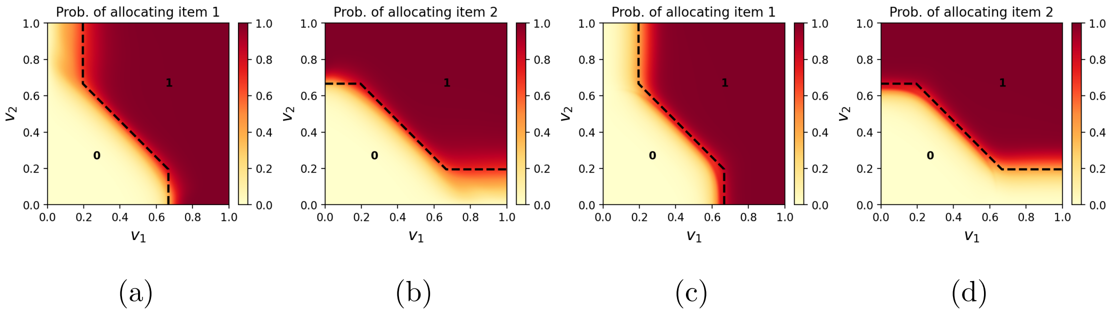
R_{max}	setting	RegretNet		EquivariantNet		RegretFormer	
		train	valid	train	valid	train	valid
10^{-3}	1x2	1.12	1.22	1.04	1.11	1.01	1.31
	2x2	0.97	1.24	1.41	1.82	0.89	1.19
	2x3	1.07	1.55	1.11	1.23	1.02	1.26
	2x5	0.94	1.21	1.11	1.2	0.8	0.83
	3x10	0.89	1.09	0.9	0.87	1.03	0.88
10^{-4}	1x2	0.94	1.27	0.92	2.37	1.31	2.52
	2x2	0.95	1.94	1.73	1.33	0.93	1.39
	2x3	1.52	1.12	1.57	1.63	1.6	1.66
	2x5	1.04	1.23	1.02	1.57	0.95	1.28
	3x10	0.9	3.71	1.05	1.46	0.88	1.15

Experiments

I conduct a series of experiments to evaluate the effectiveness of RegretFormer in comparison to RegretNet and EquivariantNet (Rahme et al., 2021a) across various auction settings, as well as the effectiveness of the loss function modification at controlling the revenue-regret trade-off. An in-depth comparison is provided in the main body of the thesis.

In Table 1, I report experiments differing only in the number of participants (n) and items (m), represented as $n \times m$. Valuations of all participants for all items are independently drawn from the uniform distribution $U[0, 1]$ The experiments span five distinct setups: 1×2, 2×2, 2×3, 2×5, and 3×10, with the 1×2 setting referencing the well-analyzed Manelli-Vincent auction, which has a known optimal revenue of 0.55. For the remaining configurations, optimal revenues remain unidentified.

The empirical results underscore RegretFormer's superior performance in generating higher revenue across all setups except the simplest 1×2, where the performance gap between the models is negligible. This gap widens significantly in more complex settings, suggesting that while the permutation-equivariance inherent to RegretFormer contributes to its success, the enhanced expressivity afforded by attention layers is likely the primary driver of its outperformance. Furthermore, both Regret-



(a) (b) (c) (d)

Figure 2: Allocation probabilities in 1x2: (a, b) RegretNet; (c, d) RegretFormer

Net and RegretFormer are shown to approximate the optimal allocation probabilities in the 1×2 setting effectively (Fig. 2).

Table 2 examines the precision of my approach in maintaining the pre-specified regret budget, which is central to the modified objective introduced earlier. Through this analysis, I demonstrate that the ratio of estimated normalized total regret to the specified regret budget approaches the ideal value of 1 during training. Still, a validation phase that involves more precise regret estimation reveals some deviations, suggesting that increasing the number of optimization steps during training could further refine the approach, despite the trade-off in training duration.

Mediated Multi-Agent Reinforcement Learning

Markov Games and Sequential Social Dilemmas

The next part of my thesis delves into the framework of Markov Games as a pivotal structure for understanding interactions within Multi-Agent Reinforcement Learning (MARL) environments. Markov Games, essentially an extension of the single-agent Markov Decision Process (MDP) to the multi-agent context, comprise a set of agents, each equipped with its own reward function. These games encapsulate scenarios where agents, based on the current state of the MDP, make simultaneous decisions according to their policies. The collective actions of all agents then guide the transition of the underlying MDP to a new state, reflecting the interconnected impact of each agent's decisions.

A critical aspect of learning dynamics in Markov Games is their convergence towards joint policies that represent some equilibria (typically, subgame perfect equilibria, also known as Markov perfect equilibria Maskin and Tirole (2001)). In an equilibrium, no agent can unilaterally change its strategy to benefit itself, thereby ensuring stability within the game's strategic landscape.

Furthermore, my research focuses on a particular subset of Markov Games known as Sequential Social Dilemmas (SSDs). SSDs are characterized by inherent conflicts of interest among agents, which, if not navigated carefully, lead to socially suboptimal outcomes. These dilemmas highlight the tension between individual incentives and the collective good, often resulting in scenarios where the pursuit of personal rewards undermines the potential for achieving the best possible outcome for the group as a whole and each individual within.

Mediators

A mediator acts as an additional entity within the game that represents a subset of agents called *coalition* by acting on their behalf. This representation is contingent upon the agents' voluntary *commitment*, allowing for a dynamic that respects agents' autonomy—agents can choose to engage with the mediator or act independently.

Mediator's strategy is defined for any possible coalition. Contrary to the fixed-strategy approach traditionally associated with mediators Monderer and Tennenholtz (2009), my adaptation involves treating the mediator with RL in parallel with the other agents in the environment. This dynamic iteration gives rise to what I term *Markov mediators.* In my implementation, the Markov mediator observes the same information as the agents in the coalition and acts for them to maximize their aggregate reward. This definition encapsulates the mediator's ability to observe the state of the game from the perspective of its coalition and act in a manner that enhances collective outcomes, essentially guiding the group towards socially beneficial equilibria. The agents periodically decide whether to commit to entering the coalition or act independently, thus preserving the agents' autonomy.

A mediator's policy that incentivizes all agents to commit and prevents unilateral beneficial deviations is called a *mediated equilibrium.* It serves as an alternative solution concept to subgame-perfect or Nash equilibria.

Deep Mediated MARL

Both the agents and the mediator are trained via Actor-Critic frameworks. The actor represents the policy, whereas the critic represents an approximation of the

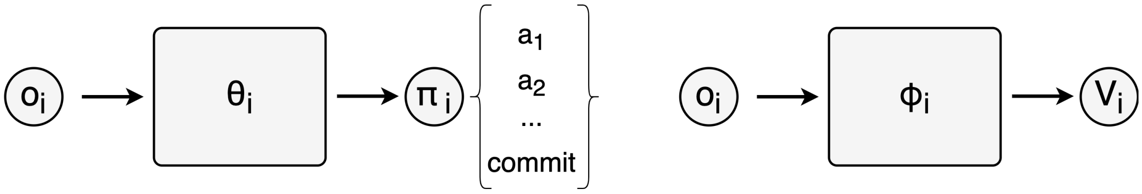


Figure 3: Schematic illustration of the architectures of the actor (left) and the critic (right) of the agents

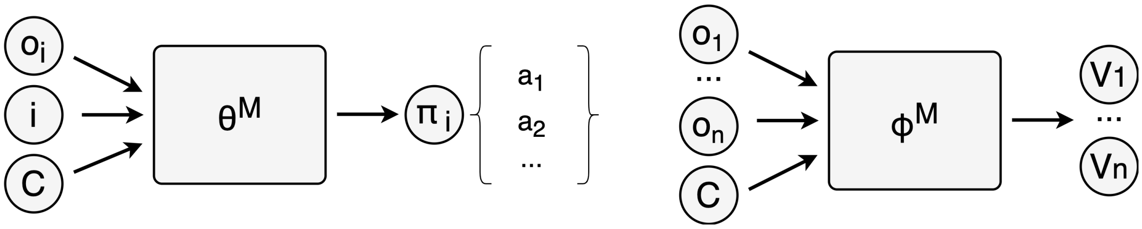


Figure 4: Schematic illustration of the architectures of the actor (left) and the critic (right) of the mediator

value function. Both actor and critic can be parameterized with neural networks. The architectures are illustrated in Figures 3 and 4. The notation is as follows: o_i denotes an observation of agent i (which is a function of the state of the MDP), π_i denotes a policy for agent i, C denotes the coalition, V_i denotes an approximation of the value function of agent i, and θ and ϕ denote neural network parameters.

A naive approach to training the mediator is to maximize social welfare for its coalition, disregarding individual agents' incentives. This approach, while focused on collective good, does not inherently motivate agents to commit, as it may not align with their self-interests. Recognizing this, I introduce two constraints to align the mediator's objectives with individual agents' motivations. Firstly, the Incentive-Compatibility (IC) Constraint ensures agents benefit from joining the coalition, receiving at least as much payoff as they would independently. Secondly, the Encouragement (E) Constraint prevents non-committing agents from exploiting the mediator, ensuring their payoff does not exceed what they would receive if they had committed. To integrate these constraints and effectively train the mediator, I employ the method of Lagrange multipliers within the policy gradient framework.

Experiments

In one of the pivotal experiments of my study, I explore the phenomenon of free-riding in multi-agent games, a situation where some agents exploit the cooperative

Table 3: Results in one-step Public Good Game; c and m denote *contribute* and *commit* actions, $|C|$ denotes coalition size, $\tilde{\pi}$ and $\widetilde{\pi^M}$ denote the averaged policies of the agents and the mediator.

$^{(*)}$ no mediator $^{(\dagger)}$ naive mediator $^{(\ddagger)}$ constrained mediator

PGG	$N = 3$	$N = 10$	$N = 25$
reward$^{(*)}$	0.012	0.0	0.0
reward$^{(\dagger)}$	0.652	0.005	0.014
$\tilde{\pi}(m)^{(\dagger)}$	0.658	0.159	0.121
$\tilde{\pi}^M(c)^{(\dagger)}$	0.985	0.001	0.02
$\pi^M(c \mid \|C\| = 2)^{(\dagger)}$	0.993	-	-
$\pi^M(c \mid \|C\| = 3)^{(\dagger)}$	0.999	-	-
reward$^{(\ddagger)}$	0.891	0.827	0.817
$\tilde{\pi}(m)^{(\ddagger)}$	0.916	0.961	0.933
$\tilde{\pi}^M(c)^{(\ddagger)}$	0.959	0.858	0.817
$\pi^M(c \mid \|C\| = 2)^{(\ddagger)}$	0.774	-	-
$\pi^M(c \mid \|C\| = 3)^{(\ddagger)}$	0.996	-	-

efforts of others for personal gain. This challenge becomes particularly evident in environments with more than two agents, where the introduction of a mediator could inadvertently enable free-riding, thereby undermining collective welfare. To illustrate this, I utilize the Public Goods Game (PGG) as a testbed, contrasting the outcomes with a Naive mediator against those with a mediator that enforces the Encouragement (E) constraint – referred to as the Constrained mediator.

In the PGG, each of N agents possesses a unit of utility they can choose to contribute to a public good or withhold (defect). The total contribution is amplified by a factor n greater than one but less than N, then evenly redistributed among all agents, creating incentives to defect.

The experiment's results clearly demonstrate the dynamics at play (Table 3). For a game with $N = 3$ agents and an amplification factor of $n = 2$, different mediator strategies lead to distinct outcomes. Without any mediator, agents default to defecting. A Naive mediator manages to encourage cooperation between two agents, yet this opens the door for the third agent to free-ride on their contribution. In contrast, the Constrained mediator successfully guides the game towards a socially optimal equilibrium by encouraging a balanced mix of cooperation and defection, effectively addressing the free-riding issue through a policy that adapts to agent behavior to prevent exploitation.

The results are consistent even as the number of agents increases, with the Constrained mediator consistently fostering cooperation across all agents by learning a policy that reciprocally adjusts to punish or deter free-riding behavior. This experiment highlights the Constrained mediator's capability to navigate the complexities of multi-agent cooperation, promoting an equilibrium that balances individual incentives with collective welfare.

The main text also contains experiments with a sequential modification of PGG, where agents' endowments are preserved throughout rounds, and the public good may grow exponentially if agents cooperate. This allows to validate scalability of the proposed approach.

Personalized RL with a Budget of Policies

Represented Markov Decision Processes

Represented Markov Decision Processes (r-MDPs) are introduced in my study as an extension of standard MDPs to facilitate personalized machine learning solutions, particularly in contexts where regulatory constraints and the necessity for personalized decision-making intersect. An r-MDP is defined as a tuple $\mathcal{M}_r = (S, A, \mathcal{T}, \mathcal{T}_0, \gamma, N, K, (r^i)_{i \in N})$, incorporating elements from standard signle-agent MDPs such as states S, actions A, transition dynamics $\mathcal{T}$, initial state distribution $\mathcal{T}_0$, and discount factor γ. Additionally, it introduces N as the set of agents with n members, K as the set of representative policies limited by a budget k, and r^i, the individual reward function for each agent i.

In the r-MDP framework, agents do not engage directly with the environment. Instead, each agent is associated with a representative from the set K, who acts on their behalf. This model employs a dual objective: to optimally assign agents to representatives (α^i) and to train these representatives' policies (π^j) to maximize the overall utilitarian social welfare. The challenge lies in maximizing the expected cumulative rewards for all agents through their representative policies, taking into account the probability of each agent being represented by a given representative.

The novelty of r-MDPs stems from their focus on abstracting the direct interaction between agents and the environment, distinguishing between the "actors" (representatives) in the environment and the agents themselves. Representatives serve the purpose of maximizing the welfare of the agents they represent, without having intrinsic reward functions. This abstraction allows for a focused approach to maximizing social welfare under the constraints of policy budgets and regulatory considerations.

Note that the MDP remains single-agent, and each representative effectively acts in its own copy of the MDP with identical dynamics but distinct reward functions.

My approach to solving r-MDPs

My methodology for solving r-MDPs addresses the curse of dimensionality posed by the exponential growth of potential assignments as the number of agents increases. Direct optimization of the joint assignment requires searching over a set with cardinality proportional to K^n, which quickly becomes infeasible with large n, due to the sheer scale of possible combinations.

To manage this, my approach involves breaking down the problem into more manageable components. Initially, I consider a scenario where the policies of representatives, π^j, are fixed. Under this condition, the task of maximizing social welfare simplifies to assigning each agent i to the representative j^* that yields the highest expected welfare.

Another angle of simplification assumes that the assignments α^i are set and focuses on optimizing the representatives' policies. In this scenario, enhancing social welfare equates to solving a series of MDPs, each tailored to a specific representative. This approach transforms the overarching optimization challenge into a collection of individual problems, each concentrated on optimizing the policy for one representative at a time.

Combining these simplifications leads to a factorized approach, where the optimization of agent assignments and representative policies are conducted independently yet iteratively – each assignment phase is optimized based on the current set of policies, and then policies are refined given the latest assignments. This methodology aims to iteratively approximate the optimal solution to the joint objective of maximizing social welfare, navigating the complexities of r-MDPs with a strategic division of the optimization process into more tractable sub-tasks.

As a specific implementation of the factorized approach, I introduce an algorithm that draws inspiration from the principles of the K-means and Expectation-Maximization (EM) clustering algorithms. This EM-like algorithm is specifically tailored to navigate the optimization landscape of r-MDPs. The algorithm operates through a cyclical process consisting of two main phases: the Expectation (E-step) and Maximization (M-step).

During the E-step, akin to assigning points to clusters in clustering algorithms, agents are allocated to representatives based on current policy performance. This assignment process is facilitated by an $n \times k$ table, $\tilde{Q}$, which stores approximations of the Q-values for each agent-representative pair. Agents are then greedily reassigned to their optimal representatives, determined by the highest Q-value approximation in $\tilde{Q}$, thus optimizing their expected utility from the current policy landscape.

Following the reassignment of agents, the M-step focuses on refining the representatives' policies to improve performance based on the new assignments. This policy update is conducted using Proximal Policy Optimization (PPO, Schulman et al. (2017)), ensuring that representatives' strategies evolve to better serve the collective welfare of the agents they represent.

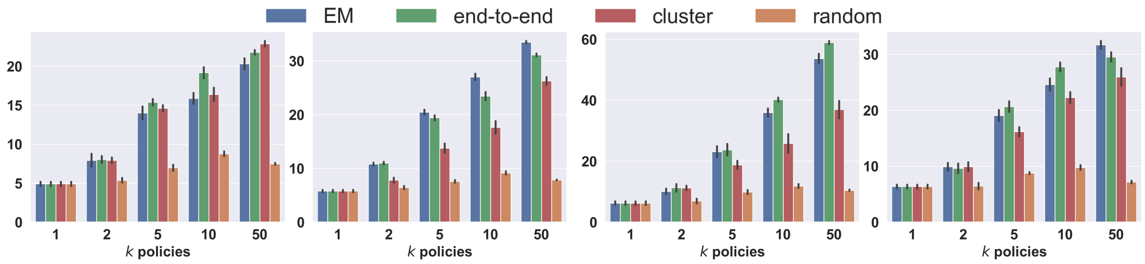
Mirroring the iterative refinement process of K-means, this EM-like algorithm is designed to converge to a local optimum of utilitarian social welfare within the r-MDP framework. This convergence is not just an empirical observation but is also formally established as a theorem.

Additionally, I propose a modification of the EM-like algorithm that relaxes the greedy agent reassignment during the E-step.

Experiments

To validate the effectiveness of the proposed algorithms under complex conditions, I conducted experiments within the MuJoCo simulation environments – specifically, HalfCheetah, Ant, Hopper, and Walker2d. These environments involve the control of robots using continuous actions within high-dimensional state spaces.

To frame these environments within the r-MDP context, I introduced $n = 100$ agents, each being assigned a uniformly sampled target velocity. The agents' rewards were then determined based on how closely the robot's velocity matched their assigned



(a) Ant (b) HalfCheetah (c) Hopper (d) Walker2d

Figure 5: Social welfare achieved by ours (EM, end-to-end) and baseline algorithms in MuJoCo environments

target velocity at each time step, introducing a personalized aspect to the challenge that each agent faces within the common single-agent environment.

The experiments explored various policy budgets, k, including 1, 2, 5, 10, and 50, to understand the impact of policy constraints on our ability to effectively personalize the agents' experiences and outcomes within these complex simulations.

The results, presented in Figure 5, highlight the superior performance of my algorithms – both the EM-like algorithm presented above and its end-to-end variation – over existing clustering approaches from the personalization literature (Hassouni et al., 2018) across different policy budgets. Notably, both algorithms not only significantly surpassed the random assignments but also consistently outperformed the clustering baseline in nearly all environments and settings tested. The exception was within the Ant environment, where the clustering baseline showed competitive performance, indicating a particular interaction between the environment's complexity and the baseline's method of personalization.

These findings underscore the robustness and adaptability of the proposed algorithms, demonstrating their potential to achieve meaningful personalization in machine learning applications, even under the stringent conditions posed by complex, high-dimensional tasks and limited policy budgets.

Conclusion

This thesis bridges the realms of game theory and AI, demonstrating through three distinct studies how deep learning and reinforcement learning can be harnessed to address and solve complex problems within the intersection of these fields. Each paper contributes to our understanding and capabilities in designing AI systems that can effectively navigate and optimize within multi-agent game-theoretic frameworks.

The first study introduces RegretFormer, a novel deep learning architecture for optimal auction design that surpasses existing methods. By rethinking the objective formulation of RegretNet, this work not only advances the state-of-the-art in automated auction design but also simplifies the optimization process, reducing the burden of hyperparameter tuning and suggesting validation procedures that could benefit future research in regret-based optimization.

The second study challenges conventional perspectives on cooperation in multi-agent reinforcement learning environments, advocating for the integration of mediators to achieve equilibrium-driven cooperation. By adapting the game-theoretic concept of mediators to the context of Markov games, this study explores conditional cooperation beyond simple cooperative dynamics, introducing a constrained optimization approach that enhances both social and individual welfare. The potential applications of mediators in MARL are vast, and this research opens multiple avenues for future exploration, from applying mediators in more complex environments to combining them with cryptographic technologies for decentralized execution.

In the third study, the focus shifts to the challenge of personalizing AI solutions within regulatory constraints through the concept of represented Markov Decision Processes. The development and validation of two deep reinforcement learning algorithms demonstrate the feasibility of achieving personalization under policy budget constraints, highlighting the potential for these approaches to be extended to real-world applications. Moreover, the game-theoretic view of the problem as social welfare optimization lays the groundwork for follow-up studies. For example, extensions could incorporate fairness and outside options into personalized reinforcement learning, aiming to ensure that personalization enhances, rather than compromises, both equity and social welfare.

Collectively, these studies underscore the synergistic potential of combining game theory with artificial intelligence to create multi-agent systems that are not only intelligent and adaptive but also robust to manipulations, equitable, and prosocial. From fostering cooperation in MARL to personalizing solutions in high-stakes domains, this thesis exemplifies how game-theoretic principles can guide and enhance AI research, offering a roadmap for future investigations at the intersection of these two pivotal fields. Moreover, the study on automated auction design exemplifies the application of AI to enrich a fundamentally game-theoretic problem, demonstrating the transformative impact of deep learning on nontrivial practical applications.

References

Acerbi, A. and Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Bahar, G., Ben-Porat, O., Leyton-Brown, K., and Tennenholtz, M. (2020). Fiduciary bandits. In *International Conference on Machine Learning*, pages 518–527. PMLR.

Barrett, L. and Narayanan, S. (2008). Learning all optimal policies with multiple criteria. In *Proceedings of the 25th international conference on Machine learning*, pages 41–47.

Breton, M. D., Kanapka, L. G., Beck, R. W., Ekhlaspour, L., Forlenza, G. P., Cengiz, E., Schoelwer, M., Ruedy, K. J., Jost, E., Carria, L., et al. (2020). A randomized trial of closed-loop control in children with type 1 diabetes. *New England Journal of Medicine*, 383(9):836–845.

Conitzer, V. (2019). Designing preferences, beliefs, and identities for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9755–9759.

Conitzer, V. and Sandholm, T. (2002). Complexity of mechanism design. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 103–110.

Conitzer, V. and Sandholm, T. (2003). Automated mechanism design: Complexity results stemming from the single-agent setting. In *Proceedings of the 5th international conference on Electronic commerce*, pages 17–24.

Conitzer, V. and Sandholm, T. (2004). Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 132–141.

Curry, M., Sandholm, T., and Dickerson, J. (2023). Differentiable economics for randomized affine maximizer auctions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2633–2641.

Daskalakis, C., Deckelbaum, A., and Tzamos, C. (2015). Strong duality for a multiple-good monopolist. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 449–450.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

den Hengst, F., Grua, E. M., el Hassouni, A., and Hoogendoorn, M. (2020). Reinforcement learning for personalization: A systematic literature review. *Data Science*, 3(2):107–147.

Duetting, P., Mirrokni, V., Paes Leme, R., Xu, H., and Zuo, S. (2024). Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 144–155.

Durugkar, I., Liebman, E., and Stone, P. (2020). Balancing individual preferences and shared objectives in multiagent reinforcement learning. *Good Systems-Published Research*.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D., and Ravindranath, S. S. (2019). Optimal auctions through deep learning. In *International Conference on Machine Learning*, pages 1706–1715. PMLR.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C., and Ravindranath, S. S. (2024). Optimal auctions through deep learning: Advances in differentiable economics. *Journal of the ACM*, 71(1):1–53.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., and Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. (2021). Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR.

Giannakopoulos, Y. and Koutsoupias, E. (2014). Duality and optimality of auctions for uniform distributions. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 259–276.

Gupta, J. K., Egorov, M., and Kochenderfer, M. (2017). Cooperative multi-agent control using deep reinforcement learning. In *International conference on autonomous agents and multiagent systems*, pages 66–83. Springer.

Hadfield-Menell, D. and Hadfield, G. K. (2019). Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422.

Haghpanah, N. and Hartline, J. (2021). When is pure bundling optimal? *The Review of Economic Studies*, 88(3):1127–1156.

Hassouni, A. e., Hoogendoorn, M., van Otterlo, M., and Barbaro, E. (2018). Personalization of health interventions using cluster-based reinforcement learning. In *PRIMA 2018: Principles and Practice of Multi-Agent Systems: 21st International Conference, Tokyo, Japan, October 29-November 2, 2018, Proceedings 21*, pages 467–475. Springer.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR.

Jiang, J. and Lu, Z. (2019). Learning fairness in multi-agent systems. *Advances in Neural Information Processing Systems*, 32.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

Manelli, A. M. and Vincent, D. R. (2006). Bundling as an optimal selling mechanism for a multiple-good monopolist. *Journal of Economic Theory*, 127(1):1–35.

Maskin, E. and Tirole, J. (2001). Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219.

Monderer, D. and Tennenholtz, M. (2009). Strong mediated equilibrium. *Artificial Intelligence*, 173(1):180–195.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Parkes, D. C. and Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272.

Pavlov, G. (2011). Optimal mechanism for selling two goods. *The BE Journal of Theoretical Economics*, 11(1):0000102202193517041664.

Peysakhovich, A. and Lerer, A. (2018a). Consequentialist conditional cooperation in social dilemmas with imperfect information. In *International Conference on Learning Representations*.

Peysakhovich, A. and Lerer, A. (2018b). Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2043–2044. International Foundation for Autonomous Agents and Multiagent Systems.

Phan, T., Sommer, F., Altmann, P., Ritz, F., Belzner, L., and Linnhoff-Popien, C. (2022). Emergent cooperation from mutual acknowledgment exchange. In

Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pages 1047–1055.

Rahme, J., Jelassi, S., Bruna, J., and Weinberg, S. M. (2021a). A permutation-equivariant neural network architecture for auction design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5664–5672.

Rahme, J., Jelassi, S., and Weinberg, S. M. (2021b). Auction learning as a two-player game. In *International Conference on Learning Representations*.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., and Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.

Wang, J. X., Hughes, E., Fernando, C., Czarnecki, W. M., Duéñez-Guzmán, E. A., and Leibo, J. Z. (2019). Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 683–692. International Foundation for Autonomous Agents and Multiagent Systems.

Wang, T., Duetting, P., Ivanov, D., Talgam-Cohen, I., and Parkes, D. C. (2024). Deep contract design via discontinuous networks. *Advances in Neural Information Processing Systems*, 36.

Yang, J., Li, A., Farajtabar, M., Sunehag, P., Hughes, E., and Zha, H. (2020). Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems*, 33:15208–15219.

Yao, A. C.-C. (2017). Dominant-strategy versus bayesian multi-item auctions: Maximum revenue determination and comparison. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 3–20.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zimmer, M., Glanois, C., Siddique, U., and Weng, P. (2021). Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 12967–12978. PMLR.