Achieving Diverse Objectives with AI-driven Prices in Deep Reinforcement Learning Multi-agent Markets

Panayiotis Danassis
Artificial Intelligence Laboratory
École Polytechnique Fédérale de Lausanne
Switzerland
panayiotis.danassis@epfl.ch

Aris Filos-Ratsikas
Department of Computer Science
University of Liverpool
United Kingdom
aris.filos-ratsikas@liverpool.ac.uk

Boi Faltings
Artificial Intelligence Laboratory
École Polytechnique Fédérale de Lausanne
Switzerland
boi.faltings@epfl.ch

Abstract

We propose a practical approach to computing market prices and allocations via a *deep reinforcement learning policymaker* agent, operating in an environment of other learning agents. Compared to the idealized market equilibrium outcome – which we use as a benchmark – our policymaker is much more flexible, allowing us to *tune* the prices with regard to diverse objectives such as sustainability and resource wastefulness, fairness, buyers' and sellers' welfare, etc. To evaluate our approach, we design a realistic market with multiple and diverse buyers and sellers. Additionally, the sellers, which are deep learning agents themselves, compete for resources in a common-pool appropriation environment based on bio-economic models of commercial fisheries.

We demonstrate that: (a) The introduced policymaker is able to achieve comparable performance to the market equilibrium, showcasing the potential of such approaches in markets where the equilibrium prices can not be efficiently computed. (b) Our policymaker can notably outperform the equilibrium solution on certain metrics, while at the same time maintaining comparable performance for the remaining ones. (c) As a highlight of our findings, our policymaker is significantly more successful in maintaining resource *sustainability*, compared to the market outcome, in scarce resource environments.

1 Introduction

The theory of competitive markets, founded in the works of Walras [59], Fisher [9] and Arrow and Debreu [4] is one of the most prominent economic models of the 20th century. The *market equilibrium* – a stable outcome in which supply equals demand, and all participants are maximally satisfied by the bundles of goods that they buy or sell – is meant to capture the outcome of a free market, dictated by the market's "invisible hand" [56], which adjusts the prices until market clearance is achieved via the so-called *tâtonnment process* [59] (i.e., continuous adjustment of supply and demand). However, the fundamental principles of this theory are based on often idealized assumptions, namely that the participants are price-takers – and thus they do not influence the prices of the market – and that this continuous adjustment of prices will indeed lead to the desired outcome. In reality, there have been several examples, especially in markets with a limited number of sellers, where instances of collusion

Preprint. Under review.

and price manipulation have been observed;[1] practices which are known to deter the market from its intended equilibrium outcome. Furthermore, it has been shown (e.g., see [8, 18]) that in fundamental market models, the convergence of the tâtonnment process is highly dependent on several initial parameters and can therefore be slow, and even the centralized computation of market equilibria for certain markets can be computationally hard [20, 17, 16], thus in many cases *impractical* to compute.

Additionally, even under these idealized assumptions, the market equilibrium is geared towards very specific goals, namely fairness for the participants and economic efficiency given the set of chosen prices. However, from a more global perspective, there are several other important objectives which are not immediately captured by the market ecosystem, since it is based heavily on the economic principles of supply and demand. For example, one of the most pressing issues in modern societies is *sustainability* and the preservation of the Earth's natural resources.[2] Clearly, the extent to which natural resources are harvested is correlated with the potential monetary gains that the sellers of those resources can achieve in the market. With these "exogenous" objectives being of paramount importance, it is only natural to assume that some form of intervention to the reign of free markets is not only inevitable, but also fully justified.

A clear-cut way of imposing some institutional or governmental control on the outcomes of the markets is via taxation on products and sales. Taxes can indeed be an effective measure, but they also sometimes have adverse effects, such as driving businesses away from the local markets, and often come in contrast with the principles of free markets. In many scenarios, the ideal means of intervention would be less intrusive, simultaneously catering to the exogenous objectives and salvaging some of the attractive properties of the market equilibrium. Additionally, we would like these means to be fully effective in a world where the market participants are *learning agents*.

Learning agents have become ubiquitous in socio-economical and socio-ecological systems in recent years (e.g., [22, 64, 62]). This has lead to the emergence of machina economicus [48], an approximate counterpart of homo economicus – the perfectly rational agent of neoclassical economics – given computational barriers and the lack of common knowledge. For example, with the emergence of machine learning, it has been observed (e.g., see [58]) that enterprises use learning agents as forms of bounded rationality [52] (see also [1]). The success of multi-agent deep reinforcement learning has lead to a growing interest in modeling machina economicus agents as *independent* deep reinforcement learning agents that need to interact, learn, cooperate, coordinate, and compete with other learning agents in ever more complex, non-stationary environments (e.g., [22, 62, 49, 38, 50, 32, 51, 35, 60, 42, 37]). Reinforcement learning is also considered a candidate theory of animal habit learning [38, 47]. Moreover, it does not make use of any economic modeling assumptions, rather it learns from observational data alone, making it ideal for multi-objective optimization in complex domains.

In this new regime, it is unclear whether the market's "laissez-faire" will lead to desirable outcomes, and how robust the notion of the market equilibrium is. Instead, we take an alternative approach, using deep reinforcement learning for policy making. In particular, we study the *emergent behaviors* as a group deep learners interact in a complex and realistic market. Our market model consists of all the established market parameters (i.e., a dynamic set of buyers and sellers, endowments, and utilities), as well as a realistic exogenous common-pool resource appropriation component, which exhibits properties related to the tragedy of the commons [29]. In our model, both the pricing policy and the harvesting behaviors are learned *simultaneously*. Neither the policy maker nor the harvesters have prior knowledge / assumptions of domain dynamics or economic theory, and every agent only makes use of information that it can individually observe. Our policymaker can be used to optimize any desired social outcome allowing us to *tune* the prices with regard to diverse objectives such as sustainability and resource wastefulness, fairness, buyers' and sellers' welfare, etc.

1.1 Our Contributions

(1) We propose a practical approach to computing market prices and allocations via a *deep reinforcement learning policymaker* agent, that allows us to *tune* the prices with regard to diverse objectives such as sustainability and resource wastefulness, fairness and buyers' and sellers' welfare.

[1] E.g., see `http://news.bbc.co.uk/1/hi/business/7132108.stm` or `https://fortune.com/2015/06/30/apple-conspired-with-book-publishers-appeals-court-confirms/`.

[2] E.g., see OECD's 25 Climate Actions `https://www.oecd.org/environment/25-climate-actions.pdf`, or UN's sustainable development goals `https://www.un.org/sustainabledevelopment/`.

(2) We introduced a novel multi-agent socio-economic environment combining established principles of competitive markets with the challenges of resource scarcity and the tragedy of the commons.

(3) We provide a thorough (quantitative & qualitative) analysis on the learned policies and demonstrate that they can achieve significant improvements over the market equilibrium benchmark for several objectives, while maintaining comparable performance for the rest. As a highlight of our results, we show that our policymaker fares notably better in terms of sustainability of resources, essentially without compromising any of the remaining objectives.

Given that it is often quite hard to experiment with real-world pricing policies, traditional work in economics often results to simplifying assumptions which are hard to validate. Our approach provides an alternative route, enabling experimentation (via tuning of the parameters and simulating the multi-agent environment) to find the best possible policies.

1.2 Discussion & Related Work

The origins of competitive market theory date back to the late 1800s and the pioneering ideas of Walras [59] and Fisher [9]. Most instrumental in assembling these ideas into a cohesive theory was the work in economics in the mid-1900s, most notably by Arrow and Debreu [4], who defined and studied what we today know as the standard, most general model of competitive markets, and proved the existence of a market equilibrium under mild assumptions. The market that we consider in this paper is a special case of the Arrow-Debreu model, due to Fisher [9], where the market participants are divided into buyers and sellers, and buyers do not have intrinsic value for money, but rather use money as a means of facilitating the trade. This so-called "Fisher market" model has been extensively studied in economics but also in computer science, in terms of computation and convergence to equilibria [33, 7, 14, 63, 25, 18]. We chose the (linear) Fisher market as our benchmark because contrary to the case of general Arrow-Debreu markets, computing a market equilibrium for this market can be done in polynomial time via convex programming (see Section 2.2 for the details). We also remark that while similar in spirit, the market equilibrium is a different notion from the well-known notion of the Nash equilibrium [45]; the former is a stable point of the market supply and demand adjustment, whereas the latter is a stable point of the participants' strategic play. In particular, the classic market equilibrium results assume that agents are *not strategic* and therefore do not attempt to influence the prices of the markets (i.e., they are price-takers). It has been shown that in the presence of (perfectly rational) strategic agents, the outcome of the market can be fundamentally different from the intended outcome of the market equilibrium [12, 2, 15].

As discussed in the introduction, the last few years have witnessed a shift towards bounded rationality models; most prominently viewing rational decisions via the lens of machine learning [46, 13] and in particular reinforcement learning. Our work is one of the first to design a policy maker via deep reinforcement learning in economic environments. There is some recent work that has adopted a similar agenda, but on markedly different domains and using different approaches. Duetting et al. [24] and Shen et al. [55] consider the problem of finding optimal prices in revenue-maximizing auctions via deep reinforcement learning, and Cai et al. [13] design a policy maker for impression allocation in e-commerce platforms against learning agents. Very recently, Zheng et al. [64] consider an abstract socio-economic domain via the lens of deep reinforcement learning for policy making, focusing on taxation as a means of institutional intervention, rather than the adjustment of the prices.

Moving on to the common-pool resource appropriation component of our work, there has been great interest recently in Common-Pool Resource (CPR) problems (and more generally, social dilemmas [36]) as an application domain for Multi-agent Deep Reinforcement Learning (MADRL) [22, 49, 38, 50, 32, 51, 35, 60, 42, 37]. CPR problems offer complex environment dynamics and relate to real-world socio-ecological systems. Using deep reinforcement learning to address sustainability problems was done recently by Danassis et al. [22], who were also the first to introduce the realistic common fishery appropriation environment – based on bio-economic models of commercial fisheries – that we also employ in this work. We have extended this model to deal with multiple resources, and harvesters with diverse skill levels, as we explain in Section 2.1.

In terms of the methodology, our work falls broadly into the following three categories: Reward shaping [32, 35, 51], which refers to adding a term to the extrinsic reward an agent receives from the environment, opponent shaping [37, 42, 49], which refers to manipulating the opponent (by e.g., sharing rewards, punishments, or adapting your own actions), and automated mechanism design

[6, 13, 24], where an an external agent distributes additional rewards and punishments to promote desirable objectives on a population of artificial learners.

2 Environment and Agent Models

In this section we provide a detailed description of our complex economic model. It consists of (i) a common-pool resource appropriation game – where a group of appropriators compete over the harvesting of a set of common resources and which exhibits properties related to the tragedy of the commons [29] and the challenge of *sustainability* – and (ii) a complex and realistic market (with a dynamic set of buyers & sellers, endowments, and utilities), where the appropriators sell their harvest.

2.1 The Common Fishery Model

In this work, we adopt the common fishery model introduced by Danassis et al. [22], which is based on an abstracted bio-economic model for *real-world* commercial fisheries [19, 23]. It is important to note that we chose this environment due to its *complex dynamics*, but the proposed approach can be employed in any market and for any resources, not just fisheries. We have extended the model of [22] to account for multiple resources, and harvesters with varying skill levels. The model describes the dynamics of the stock of a set of common-pool renewable resources, as a group of appropriators harvest over time. The harvest depends on (i) the effort exerted by the agents and (ii) the ease of harvesting that particular resource at that point of time, which depends on its stock level. The stock replenishes over time with a rate dependent on the current stock level.

More formally, let $\mathcal{N}$ denote the set of appropriators and $\mathcal{R}$ the set of resources. Let $\boldsymbol{\eta}_n = [\eta_{n,1}, \dots, \eta_{n,r}, \dots, \eta_{n,R}]$, where $\eta_{n,r} \in [0, 1]$ denotes the skill[3] (competence) of harvester n for harvesting resource r. At each time-step t, every agent exerts a vector of efforts $\boldsymbol{\phi}_{n,t} = [\phi_{n,1,t}, \dots, \phi_{n,r,t}, \dots, \phi_{n,R,t}]$, where $\phi_{n,r,t} \in [0, \Phi_{max}]$ is the effort exerted to harvest resource r. Let $\boldsymbol{\varepsilon}_{n,t} = \boldsymbol{\phi}_{n,t} \cdot \boldsymbol{\eta}_n = [\varepsilon_{n,1,t}, \dots, \varepsilon_{n,r,t}, \dots, \varepsilon_{n,R,t}]$ denote the 'effective effort', and $E_{r,t} = \sum_{n \in \mathcal{N}} \epsilon_{n,r,t}$ the total effort exerted by all the harvesters at resource r at time-step t. Then, the total harvest of resource r is given by Eq. 1, where $s_{r,t} \in [0, \infty)$ denotes the stock level at time-step t, $q_r(\cdot)$ denotes the catchability coefficient (Eq. 2), and S_r^{eq} is the equilibrium stock of the resource.

$$H_r(E_{r,t}, s_{r,t}) = \begin{cases} q_r(s_{r,t})E_{r,t} & \text{, if } q_r(s_{r,t})E_{r,t} \leq s_{r,t} \\ s_{r,t} & \text{, otherwise} \end{cases} \quad (1) \qquad q_r(x) = \begin{cases} \frac{x}{2S_r^{eq}} & \text{, if } x \leq 2S_r^{eq} \\ 1 & \text{, otherwise} \end{cases} \quad (2)$$

Each environment can only sustain a finite amount of stock. If left unharvested, the stock will stabilize at S_r^{eq}. Note also that $q_r(\cdot)$, and therefore $H_r(\cdot)$, are proportional to the current stock, i.e., the higher the stock, the larger the harvest for the same total effort. The stock dynamics of each resource are governed by Eq. 3, where $F(\cdot)$ is the spawner-recruit function (Eq. 4) which governs the natural growth of the resource, and g_r is the growth rate.[4]

$$s_{r,t+1} = F(s_{r,t} - H_r(E_{r,t}, s_{r,t})) \quad (3) \qquad F(x) = xe^{g_r(1 - \frac{x}{S_r^{eq}})} \quad (4)$$

We assume that the individual harvest is proportional to the exerted effective effort (Eq. 5), and the revenue of each appropriator is given by Eq. 6, where $p_{r,t}$ is the price ($ per unit of resource), and $c_{n,t}$ is the cost ($) of harvesting (e.g., operational cost, taxes, etc.). Here lies the "tragedy": the benefits from harvesting are private ($p_{r,t}h_{n,r,t}(\cdot)$), but the loss is borne by all (in terms of a reduced stock, see Eq. 3).

$$h_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t}) = \frac{\varepsilon_{n,r,t}}{E_{r,t}} H_r(E_{r,t}, s_{r,t}) \quad (5) \qquad u_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t}) = p_{r,t}h_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t}) - c_{n,t} \quad (6)$$

[3] In our model $\eta_{n,r}$ does not depend on time, but one can consider agents that increase their skill level and become more efficient as they harvest a particular resource. Moreover one can introduce social castes and consider the problem of social mobility.

[4] According to [22], to avoid highly skewed growth models and unstable environments, $g_r \in [-W(-1/(2e)), -W_{-1}(-1/(2e))] \approx [0.232, 2.678]$, where $W_k(\cdot)$ is the Lambert W function.

2.2 The Fisher Market Model

In a Fisher market there is a set of buyers $\mathcal{B}$ and a set of divisible goods (resources) $\mathcal{R}$, sold by one or multiple sellers. Every seller brings to the market a quantity of each good, with e_r denoting the total quantity of good $r \in \mathcal{R}$ brought collectively by the sellers. Every buyers brings a monetary endowment, or simply a *budget* of β_b, for $b \in \mathcal{B}$. Additionally, every buyer b has a *valuation* $v_{b,r}$ for each unit of good r. An allocation $\mathbf{x}$ is an $|\mathcal{B}| \times |\mathcal{R}|$ matrix, where $x_{b,r}$ denotes the amount of good r that is allocated to buyer b. In a feasible allocation, it holds that $\sum_{b\in\mathcal{B}} x_{b,r} \leq e_r$, for any good r.

We will consider linear Fisher markets, where the *utility* of a buyer given allocation $\mathbf{x}$ is defined as $v_b(\mathbf{x}) = \sum_{r\in\mathcal{R}} x_{b,r}u_{b,r}$. These markets are also often called *Eisenberg-Gale Markets*[5] [26]. In such markets a *(competitive) market equilibrium* is a pair $(\mathbf{x}, \boldsymbol{p})$ of an allocation and a vector of prices, one for each good, such that at these prices each buyer is allocated a utility-maximizing bundle of goods, the budgets are entirely spent, and the goods are entirely sold; the latter condition is typically referred as *market clearance*. For Eisenberg-Gale markets in particular, a market equilibrium can be found as the solution to the following convex optimization program:

$$\begin{aligned} \max \quad & \sum_{b\in\mathcal{B}} \beta_b \cdot \log(v_b) && (7)\\ s.t. \quad & v_b = \sum_{r\in\mathcal{R}} v_{b,r} \cdot x_{b,r}, \ \ \forall\, b \in \mathcal{B} \\ & \sum_{b\in\mathcal{B}} x_{b,r} \leq e_r, \ \ \forall\, r \in \mathcal{R} \\ & x_{b,r} \geq 0, \ \ \forall\, b \in \mathcal{B}, \ \ r \in \mathcal{R} \end{aligned}$$

While the prices do not strictly appear in this formulation, they can be recovered as the Lagrangian multipliers for the second set of constraints (the feasibility constraints of the good supply). Given the above formulation, a market equilibrium in a Fisher market always exists and it can also be computed in polynomial time. In fact, there are also combinatorial algorithms for equilibrium computation in Fisher markets, e.g., see [33, 7, 14].

3 Simulation Results

3.1 Environment Settings

Common Fishery. We simulated an environment with 8 harvesters, 8 buyers, and 4 resources ($N = 8, R = 4, B = 8$). We set the maximum effort at $\Phi_{max} = 1$, the growth rate at $g_r = 1$, and the initial population at $s_0 = S_r^{eq}$ (i.e., the stock starts from the equilibrium population), for every resource $r \in \mathcal{R}$. The findings of [22] provide a guide on the selection of the S_r^{eq} values. Specifically, we set $S_r^{eq} = M_s K N$, where $K = (e^{g_r}\Phi_{max})/(2(e^{g_r} - 1)) \approx 0.79$ is a constant, and $M_s \in \mathbb{R}^+$ is a multiplier that adjusts the scarcity of the resource (difficulty of the problem). For $M_s = 1$ the resource will not get depleted, even if all agents harvest at maximum effort.[6] We simulated two scenarios, one with $M_s = 0.8$, and a *scarce resources* scenario with $M_s = 0.45$.[7]

Harvesters. We set the skill level $\eta_{n,r} = 0.5$ for all agents and resources, except for one resource for each agent, specifically $\eta_{n,r} = 1$ if $n = r$. Finally, we assume no cost in harvesting, i.e., $c_{n,t} = 0$.

Buyers. Every time-step, a new set of buyers appears at the market, with budgets and valuations drawn uniformly at random on $[0, 1]$.

[5] Strictly speaking, the term "Eisenberg-Gale Market" is often used to refer to Fisher markets with CES utility functions, which are a superclass of the linear utility functions that we consider here.

[6] Yet, the problem of coordination remains far from trivial; see [22].

[7] In the simulations of [22], for $N = 8$ and $M_s \leq 0.4$, the agents failed to find a sustainable strategy, and always depleted the resource.

3.2 Multi-Agent Deep Reinforcement Learning

We consider a *decentralized* multi-agent reinforcement learning scenario in a partially observable general-sum Markov game (e.g., [41, 54]). At each time-step, agents take actions based on a partial observation of the state space, and receive an individual reward. Each agent learns a policy independently. More formally, let $\mathcal{N} = \{1, \dots, N\}$ denote the set of agents, and $\mathcal{M}$ be an N-player, partially observable Markov game defined on a set of states $\mathcal{S}$. An observation function $\mathcal{O}^n : \mathcal{S} \rightarrow \mathbb{R}^d$ specifies agent n's d-dimensional view of the state space. Let $\mathcal{A}^n$ denote the set of actions for agent $n \in \mathcal{N}$, and $\boldsymbol{a} = \times_{\forall n \in \mathcal{N}} a^n$, where $a^n \in \mathcal{A}^n$, the joint action. The states change according to a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^N \rightarrow \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the set of discrete probability distributions over $\mathcal{S}$. Every agent n receives an individual reward based on the current state $\sigma_t \in \mathcal{S}$ and joint action $\boldsymbol{a}_t$. The latter is given by the reward function $r^n : \mathcal{S} \times \mathcal{A}^1 \times \dots \times \mathcal{A}^N \rightarrow \mathbb{R}$. Finally, each agent learns a policy $\pi^n : \mathcal{O}^n \rightarrow \Delta(\mathcal{A}^n)$ independently through their own experience of the environment (observations and rewards). Let $\boldsymbol{\pi} = \times_{\forall n \in \mathcal{N}} \pi^n$ denote the joint policy. The goal for each agent is to maximize the long term discounted payoff, as given by $V^n_{\boldsymbol{\pi}}(\sigma_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^n(\sigma_t, \boldsymbol{a}_t) | \boldsymbol{a}_t \sim \boldsymbol{\pi}_t, \sigma_{t+1} \sim \mathcal{T}(\sigma_t, \boldsymbol{a}_t)\right]$, where γ is the discount factor and σ_0 is the initial state.

Harvester Architecture. Each agent uses a two-layer (64 neurons each) neural network for the policy approximation. The input (observation $o^n = \mathcal{O}^n(S)$) is a tuple $\langle \boldsymbol{p}_{t-1}, \boldsymbol{\phi}_{n,t-1}, \boldsymbol{u}_{n,t-1}(\cdot) \rangle$ consisting of the vector of prices for the resources, the vector of individual effort exerted for every resource, and reward (cumulative out of all the resources) obtained in the previous time-step. The output is a vector of continuous action values $a_t = \boldsymbol{\phi}_{n,t} \in [0, \Phi_{max}]$ specifying the current effort level to exert for harvesting each resource. The reward received from the environment corresponds to the revenue, i.e., $r^n(\sigma_t, \boldsymbol{a}_t) = \sum_{r \in \mathcal{R}} u_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t})$.

Policymaker Architecture. The policymaker also uses a two-layer (64 neurons each) neural network for the policy approximation. The input is a tuple $\langle \boldsymbol{\varepsilon}_t, \boldsymbol{s}_t, \boldsymbol{\beta_t}, G(\boldsymbol{v}_t) \rangle$, where $\boldsymbol{\varepsilon}_t$ is the efforts exerted by all the harvesters for all the resources, $\boldsymbol{s}_t$ is the current stock level of each resource, $\boldsymbol{\beta_t}$ is the budgets of the current set of buyers (recall that a random set of buyers appears at the market at each time-step), and finally, $G(\boldsymbol{v}_t)$ are the valuations of the buyers, obfuscated by a function $G(\cdot)$. The output is a vector of continuous action values $a_t = \boldsymbol{p_t} \in [0, \infty]$ that corresponds to the prices.

To test the robustness of our policymaker in more realistic scenarios, we considered the case where the buyer's valuations are *obfuscated*. To put this into context, note that one of the idealized assumptions that allows the market equilibrium to be computed centrally is that all the information of the market is *completely and accurately* known. For good supplies and budgets, this assumption is reasonable, as these are typically observable or inferable, and qualify as "hard" information [40] (see also [11]). In contrast, the valuations of the agents are "soft" information; they are hard to elicit, since they are expressed on a cardinal scale, and are possibly even accurately unknown to the agents themselves. The literature on computational social choice theory [10] has been concerned with the effect of limited or noisy valuation information on the desired outcomes of a system.

We considered three different obfuscation functions for the buyers' valuations: (i) the identity function $G(x) = x$ (no obfuscation) – which we used in the majority of the simulations – (ii) a function that splits $[0, 1]$ into k bins, and each valuation value is replaced by the midpoint of the bin interval (average value of the endpoints), and (iii) a function that adds uniform noise on $(0, y)$, i.e., $G(x, y) = x + \mathcal{U}(0, y)$.

The bins approach corresponds to the case where the agents are not asked to provide accurate cardinal values, but instead they provide scores that somehow encode their actual values. As the literature of the distortion in computational social choice suggests, such an elicitation device is cognitively much more conceivable (see [3] and references therein). The added noise approach corresponds to the case where agents are uncertain about their own values, so they end up reporting noisy estimates of their true value. This approach is clearly reminiscent of the literature on noisy estimates of a ground truth, pioneered by Mallows [43] but in fact dating back to the works of Marquis de Condorcet, more than two centuries ago.

Finally, the policymaker's reward is the weighted average of the desired objectives, specifically:

$$w_h \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \boldsymbol{u}_{n,t}(\cdot) + w_b \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \boldsymbol{v}_{b,t}(\cdot) + w_s \min_{r \in \mathcal{R}} \left(\min(s_{r,t} - S_r^{eq}, 0) \right) + w_f Fair(\mathbf{x}) \quad (8)$$

where w_h, w_b, w_s, and $w_f \in [0, 1]$ correspond to the weights for the harvesters' social welfare objective (sum of utilities, $\sum_{n \in \mathcal{N}} \boldsymbol{u}_{n,t}(\cdot)$), the buyers' social welfare objective (sum of valuations, $\sum_{b \in \mathcal{B}} \boldsymbol{v}_{b,t}(\cdot)$), the sustainability objective (defined in this work as the maximum negative deviation from the equilibrium stock, $\min_{r \in \mathcal{R}} (\min(s_{r,t} - S_r^{eq}, 0))$), and the fairness objective ($Fair(\mathbf{x})$). Given the broad literature on fairness, we employed three different well-established fairness indices: the the Jain index [34], the Gini coefficient [28], and the Atkinson index [5]. It is important to note that the proposed technique is not limited to our choice of objectives; rather it can be used for *any combination of objectives*.

Learning Algorithm. The policies for all agents (harvesters and the policymaker) are trained using the Proximal Policy Optimization (PPO) algorithm [53]. PPO was chosen because it avoids large policy updates, ensuring a smoother training, and avoiding catastrophic failures.

As a reminder, the buyers are not learning agents; see Section 3.1.

3.2.1 Reproducibility and Reporting of Results

Reproducibility. Reproducibility is a major challenge in (MA)DRL due to different sources of stochasticity, e.g., hyper-parameters, model architecture, implementation details, etc. [30, 31, 27]. To minimize those sources, the implementation was done using RLlib[8], an open-source library for MADRL [39]. We refer the reader to Appendix A.1 for the description of the architecture, and the list of hyper-parameters.

Termination Condition. An episode terminates when either (a) a fixed number of time-steps $T_{max} = 500$ is reached, or (b) any of the resources gets depleted, i.e., the stock falls below a threshold $\delta = 10^{-4}$. We trained our agents for 2400 episodes,

Statistical Significance. All simulations were *repeated* 8 *times*. The graphs depict the average values over those 8 trials, and the shaded area represents one standard deviation of uncertainty. The reported numerical results (Table 1) are the average values of the last 400 episodes over those trials. (MA)DRL also lacks common practices for statistical testing [30, 31]. In this work, we opted to use the Student's T-test [57] due to it's robustness [21]; p-values can be found in Table 3. All of the reported results that improve the baseline have p-values < 0.05.

3.3 Results

In what follows we study the effect – with regard to diverse objectives such as sustainability and resource wastefulness, fairness, buyers' and sellers' welfare, etc. – of introducing the proposed policymaker to our complex economic system, compared to having the market equilibrium prices (as given by solving the convex optimization program of (7)).

We evaluated the "vanilla" policymaker, where $w_h = w_b = w_s = w_f = 1$, and four extreme cases where we optimize only one of the objectives, i.e., (i) $w_h = 1$ and $w_b = w_s = w_f = 0$, (ii) $w_b = 1$ and $w_h = w_s = w_f = 0$, (iii) $w_s = 1$ and $w_h = w_b = w_f = 0$, and (iv) $w_f = 1$ and $w_h = w_b = w_s = 0$. The latter offers clear-cut results, but – as we will show in Section 3.3.2 – it can potentially lead to adverse effects. In practice, use of simulations can enable the testing of economic policies at *large-scale*, and the the ability to evaluate a range of different parameters, allowing the *designer to ultimately select the weights that optimize the desired combination of objectives*.

3.3.1 Comparing the "Vanilla" Policymaker to the Market Equilibrium Prices (MEP)

Figures 1a and 1b depict the per-harvester mean reward and per-buyer mean utility, respectively, while rows 1 and 2 of Table 1 show the relative difference of the achieved social welfare (sum of utilities), as compared to the market equilibrium prices (MEP). The vanilla policymaker (blue line in the figures and first column of the table) achieves results comparable to the equilibrium prices in both cases, with a loss of only $\approx 7\%$ of social welfare. Similar results are achieved in the case of fairness – both for the sellers and buyers (last two rows of Table 1) – with both the MEP and the

[8] `https://docs.ray.io/en/latest/rllib.html`

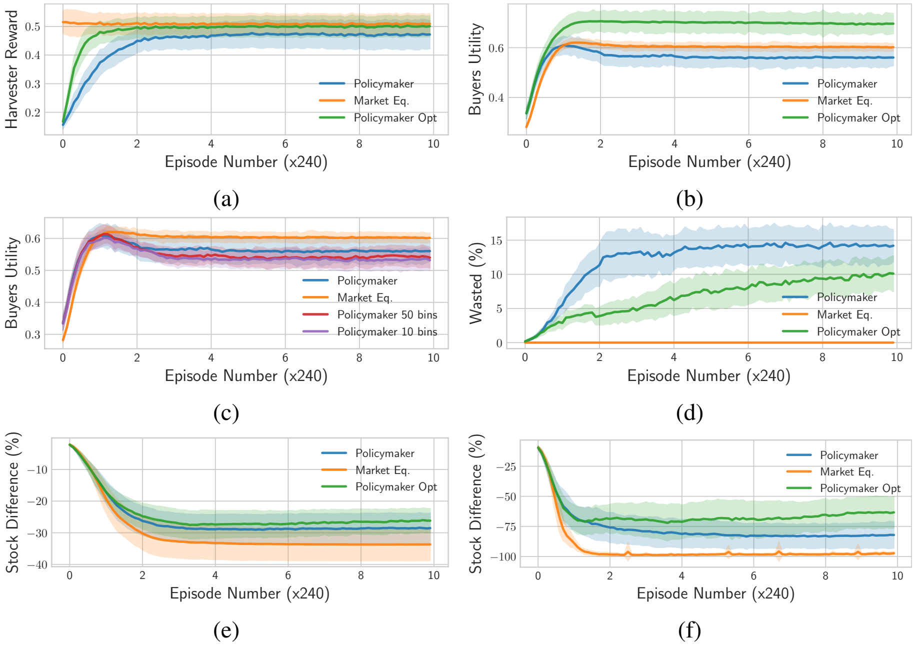

Figure 1: Evolution of several metrics over the number of training episodes. The orange line is the baseline, where the prices are the market equilibrium prices. The blue line refers to the vanilla policymaker where each objective in the reward function has the same weight (see Section 3.3). The green line refers to the policymaker that only optimizes the specific objective of each figure (i.e., in 1a we set $w_h = 1$ and $w_b = w_s = w_f = 0$, in 1b we set $w_b = 1$ and $w_h = w_s = w_f = 0$, and in 1d, 1e, and 1f we set $w_s = 1$ and $w_h = w_b = w_f = 0$). The red and purple lines in 1c refer to a policymaker with obfuscated valuations (see Section 3.2). Finally, in 1f we have a scarce resource setting (i.e., $M_s = 0.45$, see Section 3.1). Shaded areas represents one sd.

policymaker achieving a fair allocation (Jain index ≥ 0.98[9]). This is particularly important, as the market equilibrium is geared by design to optimize the aforementioned metrics, i.e., fairness for the participants and economic efficiency. Notably, the vanilla policymaker significantly outperforms the MEP when it comes to sustainability, as we describe in more detail in Section 3.3.4.

3.3.2 Harvesters' Revenue, Buyers' Utility, and Social Welfare (SW)

Optimizing specifically for the harvesters' revenue or the buyers' utility (setting $w_h = 1$ or $w_b = 1$, respectively, and the remaining weights to 0), results in the policymaker closing the gap, or even significantly outperforming the MEP (green line in Figures 1a and 1b and second and third column of Table 1). The harvesters' Social Welfare (SW) improves from -7.44% to -1.74%, while the buyers' SW exhibits a dramatic improvement from -7% to $+15.42\%$.

It is important to note, though, that contrary to the case of optimizing the sustainability or the fairness, exclusively optimizing the harvesters' SW has detrimental effects to the the buyers' SW and vice versa (see Table 1). This is because these two objectives are somewhat orthogonal; low prices lead to high buyers SW but low harvesters SW, and vice versa (although money do not have an intrinsic value in Fisher markets). In this work we showcase the potential of a vanilla policymaker, and the extreme cases of optimizing just one objective; it is up the designer to ultimately select the weights that best serve the desired combination of objectives.

Finally, we report results on noisy buyers' valuations (see Section 3.2), split into 50 and 10 bins (last two columns of Table 1, and Figure 1c; see Table 3 for the rest). Noisy valuations lead to only a small drop in the buyers' and harvesters' SW ($\approx 2-4\%$), the fairness remains the same, while sustainability improves significantly (up to 8% compared to the vanilla policymaker). This comes to show that the policymaker is robust to noisy valuations, which are much easier and practical to elicit.

[9] The higher the better; an allocation is considered totally fair iff the Jain index is 1. See Appendix A.2.

Table 1: Numerical results of the last 400 episodes of each training trial (averaged over the 8 trials). Each column represents the relative difference (%) of the particular configuration of the policymaker, as compared to the market equilibrium prices ($100(X_{\text{policymaker}} - Y_{\text{market eq.}})/Y_{\text{market eq.}}$), for each of the metrics presented in each row. The first column refers to the vanilla policymaker, where each objective in the reward function has the same weight (see Section 3.3), and each of the following 4 columns refers to a policymaker that only optimizes the specific objective in the title (having weight 0 for the rest). Finally, the last two columns refer to a vanilla policymaker with obfuscated valuations (valuations split into 50 and 10 bins respectively, see Section 3.2).

	Policymaker						
	vanilla	$w_h = 1$	$w_b = 1$	$w_s = 1$	$w_f = 1$	Noisy (50)	Noisy (10)
Harvesters' Social Welfare	-7.44	-1.74	-72.91	-31.37	-34.14	-11.35	-9.71
Buyers' Social Welfare	-7.01	-24.71	15.42	1.23	2.88	-9.73	-11.51
Stock Difference[10]	-15.30	-2.64	-10.58	-21.83	-12.99	-23.40	-21.73
Harvesters' Fairness	-0.61	-0.05	-0.64	-0.72	-0.14	-1.16	-1.04
Buyers' Fairness	-0.12	-0.18	-0.05	-0.09	-0.07	-0.27	-0.31

3.3.3 Fairness

The MEP are geared towards optimizing fairness; it is important to ensure that the introduction of the policymaker does not result in an exploiter-exploitee situation. All of the evaluated versions of the policymaker achieve a fair allocation (Jain index ≥ 0.98[9], see Table 3 for the other metrics). The relative values (Table 1) show a consistent improvement when specifically optimizing for fairness ($w_f = 1$) but, in absolute terms, all versions actually result in fair allocations.

3.3.4 Sustainability

The question of sustainability in the use of common-pool resources constitutes a critical open problem. Individuals face strong incentives to appropriate, which results in *overuse* and even *depletion* of the resource. The partial observability and the inherent pitfalls of MADRL (e.g., non-stationarity, credit assignment, global exploration, etc. [31, 44, 61]) further increase this challenge.

We measure sustainability as the maximum negative deviation from the equilibrium stock (see Section 3.2). The introduced policymaker results in the *emergence of significantly and consistently more sustainable harvesting strategies*. Figure 1e shows that the MEP maintain a population stock that is 34% below the equilibrium population (on average), while the policymaker is only 28.5%. Optimizing for sustainability ($w_s = 1$, green line) improves the difference to 26%.

More interesting is Figure 1f, where we simulate a *scarce resource* environment (see Section 3.1). In this setting, the introduction of the policymaker results in a *dramatic improvement in sustainability*. The MEP maintain a population stock that is 97.3% below the equilibrium population (on average), while the policymaker is 82.1% and optimizing for sustainability improves the difference to 63.3%; almost 35% improvement compared to MEP. In this setting, the MEP fail to result in a sustainable strategy and permanently *deplete* the resources in 9.79% of the episodes, with episodes lasting as low as 48 time-steps (out of 500). In contrast, the vanilla policymaker fails in 4.59% of the episodes (min episode length of 180 time-steps), and the version that optimizes sustainability fails in only 2.24% of the episodes (min episode length of 258 time-steps).

Importantly, optimizing for sustainability does not have detrimental effects to most other objectives, as seen in Table 1.[10] The harvesters' and buyers' fairness improve as well, and so does the buyers' welfare; only the harvesters' welfare degrades; but, as mentioned, it is up the designer to ultimately select the weights that best serve the desired combination of objectives.

Finally, we also measured the percentage of wasted resources (harvested resources that remain unsold), see Figure 1d. Of course, by design, the MEP sell the entire harvest. Optimizing for sustainability results in a decrease of the wasted resources from 14% to 10% (blue vs. green line).

[10] Note that the stock difference has negative values (negative deviation from the equilibrium stock) thus, in this metric, large negative numbers are *in favor* of the policymaker.

4 Conclusion

We proposed a practical approach to computing market prices and allocations via *deep reinforcement learning*, allowing us to optimize a host of diverse objectives such as sustainability and resource wastefulness, fairness, and buyers' and sellers' welfare. We evaluate our approach in a realistic environment that combines an established market model with a common-pool appropriation component, and demonstrate significant improvements, especially towards solving the challenge of sustainability of limited resources. Our work constitutes an important first step in studying markets composed of *learning* agents, which are becoming ubiquitous in recent years.

Societal Impact. Our approach can actively facilitate social mobility, sustainability, and fairness. As a potential negative social impact, the introduction of learning agents in socio-economic systems might bring forth an "arms-race" for the best means of production, which now shift from traditional, to computational resources and technological know-how. This can increase social inequality.

References

[1] David Abel. Concepts in bounded rationality: Perspectives from reinforcement learning. Master's thesis, Brown University, 2019.

[2] Bharat Adsul, Ch Sobhan Babu, Jugal Garg, Ruta Mehta, and Milind Sohoni. Nash equilibria in fisher market. In *International Symposium on Algorithmic Game Theory*. Springer, 2010.

[3] Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A Voudouris. Distortion in social choice problems: The first 15 years and beyond. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

[4] Kenneth J Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

[5] Anthony B Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3): 244–263, 1970. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(70)90039-6. URL `https://www.sciencedirect.com/science/article/pii/0022053170900396`.

[6] Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[7] Xiaohui Bei, Jugal Garg, Martin Hoefer, and Kurt Mehlhorn. Computing equilibria in markets with budget-additive utilities. In *24th Annual European Symposium on Algorithms, ESA 2016*, page 8. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2016.

[8] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[9] William C Brainard, Herbert E Scarf, et al. *How to compute equilibrium prices in 1891*. Citeseer, 2000.

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

[11] Simina Brânzei and Aris Filos-Ratsikas. Walrasian dynamics in multi-unit markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1812–1819, 2019.

[12] Simina Brânzei, Yiling Chen, Xiaotie Deng, Aris Filos-Ratsikas, Søren Frederiksen, and Jie Zhang. The fisher market game: Equilibrium and welfare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[13] Qingpeng Cai, Aris Filos-Ratsikas, Pingzhong Tang, and Yiwei Zhang. Reinforcement mechanism design for e-commerce. In *Proceedings of the 2018 World Wide Web Conference*, pages 1339–1348, 2018.

[14] Deeparnab Chakrabarty, Nikhil Devanur, and Vijay V Vazirani. New results on rationality and strongly polynomial time solvability in eisenberg-gale markets. In *International Workshop on Internet and Network Economics*, pages 239–250. Springer, 2006.

[15] Ning Chen, Xiaotie Deng, Bo Tang, and Hongyang Zhang. Incentives for strategic behavior in fisher market games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[16] Xi Chen, Decheng Dai, Ye Du, and Shang-Hua Teng. Settling the complexity of arrow-debreu equilibria in markets with additively separable utilities. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 273–282. IEEE, 2009.

[17] Xi Chen, Dimitris Paparas, and Mihalis Yannakakis. The complexity of non-monotone markets. *Journal of the ACM (JACM)*, 64(3):1–56, 2017.

[18] Yun Kuen Cheung, Richard Cole, and Yixin Tao. Dynamics of distributed updating in fisher markets. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 351–368, 2018.

[19] Colin W Clark. *The worldwide crisis in fisheries: economic models and human behavior*. Cambridge University Press, 2006.

[20] Bruno Codenotti, Amin Saberi, Kasturi Varadarajan, and Yinyu Ye. Leontief economies encode nonzero sum two-player games. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 659–667, 2006.

[21] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. *arXiv preprint arXiv:1904.06979*, 2019.

[22] Panayiotis Danassis, Zeki Doruk Erden, and Boi Faltings. Improved cooperation by exploiting a common signal. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 395–403, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.

[23] Florian K Diekert. The tragedy of the commons from a game-theoretic perspective. *Sustainability*, 4(8):1776–1786, 2012.

[24] Paul Duetting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1706–1715. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/duetting19a.html`.

[25] Krishnamurthy Dvijotham, Yuval Rabani, and Leonard J Schulman. Convergence of incentive-driven dynamics in fisher markets. *Games and Economic Behavior*, 2020.

[26] Edmund Eisenberg and David Gale. Consensus of subjective probabilities: The pari-mutuel method. *The Annals of Mathematical Statistics*, 30(1):165–168, 1959.

[27] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=r1etN1rtPB`.

[28] Corrado Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[29] Garrett Hardin. The tragedy of the commons. *science*, 162(3859):1243–1248, 1968.

[30] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669/16677`.

[31] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 2019.

[32] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3330–3340, Red Hook, NY, USA, 2018. Curran Associates Inc.

[33] Kamal Jain and Vijay V Vazirani. Eisenberg–gale markets: algorithms and game-theoretic properties. *Games and Economic Behavior*, 70(1):84–106, 2010.

[34] Raj Jain, Dah-Ming Chiu, and W. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *CoRR*, cs.NI/9809099, 1998. URL `http://arxiv.org/abs/cs.NI/9809099`.

[35] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. volume 97 of *Proceedings of Machine Learning Research*, pages 3040–3049, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[36] Peter Kollock. Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 24 (1):183–214, 1998.

[37] Raphael Koster, Dylan Hadfield-Menell, Gillian K. Hadfield, and Joel Z. Leibo. Silly rules improve the capacity of agents to learn stable enforcement and compliance behaviors. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1887–1888, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

[38] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. Int. Foundation for Autonomous Agents and Multiagent Systems, 2017.

[39] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Joseph Gonzalez, Ken Goldberg, and Ion Stoica. Ray RLlib: A composable and scalable reinforcement learning library. In *Deep Reinforcement Learning symposium (DeepRL @ NeurIPS)*, 2017.

[40] José María Liberti and Mitchell A Petersen. Information: Hard and soft. *Review of Corporate Finance Studies*, 8(1):1–41, 2019.

[41] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, page 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

[42] Andrei Lupu and Doina Precup. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 789–797, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

[43] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[44] Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012. doi: 10.1017/S0269888912000057.

[45] John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[46] Denis Nekipelov, Vasilis Syrgkanis, and Eva Tardos. Econometrics for learning agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 2015.

[47] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3): 139–154, 2009.

[48] David C. Parkes and Michael P. Wellman. Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272, 2015. ISSN 0036-8075. doi: 10.1126/science.aaa8403. URL `https://science.sciencemag.org/content/349/6245/267`.

[49] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, pages 3643–3652, 2017.

[50] Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BkabRiQpb`.

[51] Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 2043–2044, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[52] Ariel Rubinstein and Carl-johann Dalgaard. *Modeling bounded rationality*. MIT press, 1998.

[53] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL `http://arxiv.org/abs/1707.06347`.

[54] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953. ISSN 0027-8424. doi: 10.1073/pnas.39.10.1095. URL `http://www.pnas.org/content/39/10/1095`.

[55] Weiran Shen, Pingzhong Tang, and Song Zuo. Automated mechanism design via neural networks. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pages 215–223, 2019.

[56] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*, volume 1. Librito Mondi, 1791.

[57] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[58] Éva Tardos. Learning and efficiency of outcomes in games, seminar slides. 2019.

[59] Leon Walras. *Elements of pure economics*. Routledge, 2013.

[60] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 683–692, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

[61] Rudolf Paul Wiegand and Kenneth A. Jong. *An Analysis of Cooperative Coevolutionary Algorithms*. PhD thesis, USA, 2004. AAI3108645.

[62] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. Learning to incentivize other learning agents. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15208–15219. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/ad7ed5d47b9baceb12045a929e7e2f66-Paper.pdf`.

[63] Li Zhang. Proportional response dynamics in the fisher market. *Theoretical Computer Science*, 412(24):2691–2698, 2011.

[64] Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

A Appendix

Contents

In this appendix we include several details that have been omitted from the main text for brevity. In particular:

- In Section A.1 we provide details on the agent architecture, hyper-parameters, and computation resources.
- In Section A.2, we describe the employed fairness metrics.
- Finally, in Table 3, we provide a thorough account of the simulation results.

A.1 Agent Architecture and Hyper-parameters

Recent work has demonstrated that code-level optimizations play an important role in performance, both in terms of achieved reward and underlying algorithmic behavior [27]. To minimize those sources of stochasticity – and given that the focus of this work is in the performance of the introduced policymaker and not of the training algorithm – we opted to use RLlib[11] as our implementation framework. Each agent uses a two-layer (64 neurons each) feed-forward neural network for the policy approximation. The policies are trained using the Proximal Policy Optimization (PPO) algorithm [53]. All the hyper-parameters were left to the default values specified in Ray and RLlib[12]. For completeness, Table 2 presents a list of the most relevant of them.

Table 2: List of hyper-parameters.

Parameter	Value
Learning Rate (α)	0.0001
Clipping Parameter	0.3
Value Function Clipping Parameter	10.0
KL Target	0.01
Discount Factor (γ)	0.99
GAE Parameter Lambda	1.0
Value Function Loss Coefficient	1.0
Entropy Coefficient	0.0

A.1.1 Computational Resources

All the simulations were run on an Intel Xeon E5-2680 (Haswell) – 12 cores, 24 threads, 2.5 GHz – with 256 GB of RAM.

A.2 Fairness Metrics

We employed three of most established fairness metrics: the Jain index [34], the Gini coefficient [28], and the Atkinson index [5]:

(a) The Jain index [34]: Widely used in network engineering to determine whether users or applications receive a fair share of system resources. It exhibits a lot of desirable properties such as: population size independence, continuity, scale and metric independence, and boundedness. For an

[11] RLlib (`https://docs.ray.io/en/latest/rllib.html`) is an open-source library on top of Ray (`https://docs.ray.io/en/latest/index.html`) for Multi-Agent Deep Reinforcement Learning [39].

[12] See `https://docs.ray.io/en/latest/rllib-algorithms.html#ppo`.

allocation game of N agents, such that the n^{th} agent is alloted x_n, the Jain index is given by Eq. 9. $\mathbb{J}(\mathbf{x}) \in [0,1]$. An allocation $\mathbf{x} = (x_1, \ldots, x_N)^\top$ is considered fair, iff $\mathbb{J}(\mathbf{x}) = 1$.

$$\mathbb{J}(\mathbf{x}) = \frac{\left(\sum\limits_{n=1}^{N} x_n\right)^2}{N \sum\limits_{n=1}^{N} x_n^2} \tag{9}$$

(b) The Gini coefficient [28]: One of the most commonly used measures of inequality by economists intended to represent the wealth distribution of a population of a nation. For an allocation game of N agents, such that the n^{th} agent is alloted x_n, the Gini coefficient is given by Eq. 10. $\mathbb{G}(\mathbf{x}) \geq 0$. A Gini coefficient of zero expresses perfect equality, i.e., an allocation is fair iff $\mathbb{G}(\mathbf{x}) = 0$.

$$\mathbb{G}(\mathbf{x}) = \frac{\sum\limits_{n=1}^{N} \sum\limits_{n'=1}^{N} |x_n - x_{n'}|}{2N \sum\limits_{n=1}^{N} x_n} \tag{10}$$

(b) The Atkinson index [5]: Is a measure of the amount of social utility to be gained by complete redistribution of a given income distribution, for a give ϵ. In our work, we used $\epsilon = 1$. For an allocation game of N agents, such that the n^{th} agent is alloted x_n, the Atkinson index of $\epsilon = 1$ is given by Eq. 11. $\mathbb{A}(\mathbf{x}) \in [0,1]$. An Atkinson index of zero expresses perfect equality, i.e., an allocation is fair iff $\mathbb{A}(\mathbf{x}) = 0$.

$$\mathbb{A}(\mathbf{x}) = 1 - \frac{1}{\frac{\sum_{n=1}^{N} x_n}{N}} \left(\prod_{n=1}^{N} x_n\right)^{1/N} \tag{11}$$

Table 3: Numerical results of the last 400 episodes of each training trial (averaged over the 8 trials).
Each odd column represents the relative difference (%) of the particular configuration of the policymaker, as compared to the market equilibrium prices ($100(X_{\text{policymaker}} - Y_{\text{market eq.}})/Y_{\text{market eq.}}$), for each of the metrics presented in each row.
Each even column shows the Student's T-test p-values.
The first two columns refers to the vanilla policymaker, where each objective in the reward function has the same weight (see Section 3.3), and each of the following 8 columns refers to a policymaker that only optimizes the specific objective in the title (having weight 0 for the rest).
Finally, the last 8 columns refer to a vanilla policymaker with obfuscated valuations (see Section 3.2). The first 4 of them split the valuations into 50 and 10 bins, respectively, while the last 4 add uniform noise (5% and 10%, respectively).
The p-values are computed as follows: The p-value for the vanilla policymaker is calculated using the results from the market equilibrium prices (i.e., we measure the significance of the difference of the policymaker results compared to the MEP). The p-value for any of the following policymakers is calculated using the results from the vanilla policymaker (i.e., we measure if there is a statistically significant change between the vanilla and the optimized policymaker).
Finally, note that the stock difference has negative values (negative deviation from the equilibrium stock) thus, in this metric, large negative numbers are *in favor* of the policymaker.

										Policymaker								
	vanilla	p-value	$w_h = 1$	p-value	$w_b = 1$	p-value	$w_s = 1$	p-value	$w_f = 1$	p-value	Noisy (50)	p-value	Noisy (10)	p-value	Uni (0.05)	p-value	Uni (0.1)	p-value
Harvesters' Social Welfare	-7.44	1.21e-08	-1.74	4.16e-07	-72.91	4.87e-17	-31.37	7.00e-06	-34.14	1.67e-06	-11.35	2.38e-05	-9.71	2.95e-03	-8.20	2.50e-01	-10.07	9.92e-04
Buyers' Social Welfare	-7.01	2.26e-06	-24.71	6.65e-08	15.42	9.72e-13	1.23	2.28e-04	2.88	2.14e-05	-9.73	9.88e-03	-11.51	2.25e-04	-13.68	3.89e-06	-11.56	2.00e-04
Stock Difference	-15.30	2.72e-09	-2.64	2.66e-08	-10.58	2.34e-02	-21.83	2.36e-03	-12.99	1.53e-01	-23.40	2.72e-06	-21.73	3.42e-05	-24.68	4.90e-07	-22.28	1.41e-05
Harvesters' Fairness Jain	-0.61	3.71e-03	-0.05	6.94e-03	-0.64	8.96e-01	-0.72	6.81e-01	-0.14	2.13e-02	-1.16	6.44e-03	-1.04	2.70e-02	-1.33	1.02e-03	-1.76	1.15e-05
Harvesters' Fairness Gini	-2.78	2.87e-04	-0.54	2.09e-03	-2.86	9.16e-01	-3.08	6.99e-01	-1.29	2.48e-02	-4.57	7.44e-03	-4.09	3.85e-02	-4.75	4.01e-03	-5.52	2.87e-04
Harvesters' Fairness Atkinson	-0.29	4.45e-03	-0.03	8.55e-03	-0.29	9.53e-01	-0.33	7.54e-01	-0.07	2.32e-02	-0.59	3.16e-03	-0.48	3.67e-02	-0.66	6.50e-04	-0.93	2.68e-06
Buyers' Fairness Jain	-0.12	5.48e-05	-0.18	1.32e-01	-0.05	1.30e-02	-0.09	4.07e-01	-0.07	6.31e-02	-0.27	7.60e-06	-0.31	9.04e-07	-0.31	3.82e-07	-0.31	1.14e-06
Buyers' Fairness Gini	-1.49	3.26e-07	-1.96	1.16e-01	-0.84	7.07e-03	-1.29	4.04e-01	-1.06	4.31e-02	-2.57	1.83e-05	-2.74	6.28e-06	-2.81	1.80e-06	-2.78	2.97e-06
Buyers' Fairness Atkinson	-0.06	5.50e-05	-0.09	1.38e-01	-0.02	1.19e-02	-0.05	4.16e-01	-0.03	5.61e-02	-0.13	8.68e-06	-0.15	1.19e-06	-0.15	4.24e-07	-0.15	1.29e-06