AI-driven Prices for Sustainable Production Markets

Panayiotis Danassis  
Artificial Intelligence Laboratory  
EPFL  
Switzerland  
panayiotis.danassis@alumni.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  
EPFL  
Switzerland  
boi.faltings@epfl.ch

Abstract

The depletion of common-pool resources like fisheries is an indicative example of a market failure. Markets do not account for negative externalities; indirect costs that some participants impose on others, such as the cost of over-appropriation (which diminishes future stock, and thus harvest, for everyone). Quantifying appropriate interventions to market prices has proven to be quite challenging. 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. Our policymaker allows us to tune the prices with regard to diverse objectives such as sustainability and resource wastefulness, fairness, buyers’ and sellers’ welfare, etc. As a highlight of our findings, our policymaker is significantly more successful in maintaining resource sustainability, compared to the market equilibrium outcome, in scarce resource environments.

1 Introduction

Sustainability and the preservation of the earth’s natural resources constitute one of the most pressing issues and grand challenges in modern societies. For decades, the overarching priority of neoclassical economics has been economic growth (Barro and Sala-i Martin, 2003). Yet, the canonical assumptions of infinite and replenishable resources are practically unfounded and threaten the existence of the critical resources upon which our society ultimately depends. It is becoming increasingly clear that we need to shift our production patterns, decouple the economic growth from environmental degradation, and increase resource efficiency.

Competitive markets, founded in the works of [Walras 1874] and [Fisher 1892], constitute the fundamental mechanism of allocation; the means that products are sold and bought. Market theory ([Arrow and Debreu 1954]) suggests that free markets will reach an efficient stable outcome (market equilibrium) in which supply equals demand, and all participants are maximally satisfied by the bundles of goods that they buy or sell at the chosen prices.

Nevertheless, free markets, fail to account for negative externalities (Laffont 2008), which lead to market failure ([Bator 1958]). These externalities refer to indirect costs that are not reflected in the market equilibrium prices. A representative example of such inefficiencies is the environmental harm caused by pollution and overexploitation of natural resources, e.g., air pollution from burning fossil fuels.

1E.g., see https://www.un.org/sustainabledevelopment/
fuels, water pollution from industrial effluents, antibiotic resistance due to overuse of antibiotics in industrial farms, etc. Another prominent example of a negative externality is the depletion of the stock of fish due to overfishing. For example, according to OECD, about 25% of fish stocks globally are at risk (OECD, 2020). Fisheries constitute a common-pool resource of finite yield (i.e., it is challenging and/or costly to exclude individuals from appropriating), thus they are vulnerable to the tragedy of the commons. Fishermen do not consider the costs to others; they harvest more than is efficient (i.e., they deplete the resource faster than it can regenerate), leading to environmental degradation (ecological market failure) and may eventually threaten entire ecosystems with extinction (e.g., see the Atlantic cod fishery (Daly and Farley, 2011)).

There are many approaches to reconcile the economics of the free market with societal and environmental externalities. For example, policy-makers can correct for the inefficiencies by employing command-and-control legislation (e.g., (Wikipedia, 2022)), permit markets (Crocker, 1966; Dales, 1968) (e.g., (EU, 2022a)), or taxation (e.g., (EU, 2022b)). A classic example of the latter is Pigouvian taxes (Pigou, 1924), i.e., taxes that are equal to the external damage caused by the production decisions. While such interventions are clearly necessary, selecting and quantifying the appropriate ones has proven to be quite challenging. For example, in the case of common-fisheries, approaches aiming to determine the “optimal” level of annual harvest and subsequently control fishing to achieve that quota have often failed to prevent overfishing (Clark, 2006). Similarly, determining the marginal social cost of a negative externality and converting it into a monetary value can be quite impractical (Baumol, 1972).

We propose a practical and effective technique for calculating concrete market prices and allocations via a deep reinforcement learning policymaker agent, operating in an environment of other learning agents. These prices can serve as a clear-cut guideline for intervention, and can then be implemented by a variety of mechanisms; e.g., policy-makers can tax (or subsidize) the difference between the current market price and the computed price, or buy/sell from reserves. This new approach grants us the ability to design and test novel policies (via tuning the parameters and simulating the multi-agent environment) to tackle a plethora of real-world problems in various disciplines under a host of objectives, such as the problem of sustainable production (renewable energy, CO₂ markets, natural resource preservation, etc.).

1.1 Our Contributions

In this paper, we use deep reinforcement learning for policy making. We study the emergent behaviors as a group of deep learners interact in a complex and realistic market, where 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. In particular:

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.

2. **We introduced a novel multi-agent socio-economic environment** that combines established principles of competitive markets with the challenges of resource scarcity and the tragedy of the commons. This is paramount to understand the impact of self-interested appropriation and develop sustainable strategies. While we use a common-fishery as a test-bench, our approach is general and can be employed in any production market.

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.

---

2There are many such examples of influencing the supply/demand, e.g., (CNN, 2021; WPR, 2018; BBC, 2021).
1.2 Discussion & Related Work

**Competitive Markets** The origins of competitive market theory date back to the pioneering ideas of Walras [1874] and Fisher [1892]. Arrow and Debreu [1954], defined and studied the standard, most general model of competitive markets, and proved the existence of a market equilibrium (ME). The market that we consider in this paper is a special case of the Arrow-Debreu model, due to Fisher [1892], 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. We chose the (linear) “Fisher market” as our benchmark because, contrary to general Arrow-Debreu markets, computing a ME can be done in polynomial time.

The ME is theoretically reached via the continuous adjustment of supply and demand dictated by the market’s “invisible hand” (Smith [1791]) (“tâtonnment process”). Yet, note that the ME is reached only under a strict range of assumptions (e.g., participants are price-takers, there is no collusion, etc.), and the tâtonnment process is highly dependent on several initial parameters and can therefore be slow, and even impractical to compute (Codenotti et al., 2006; Chen et al., 2017, 2009). We also remark that while similar in spirit, the ME is a different notion from the well-known notion of the Nash equilibrium (Nash, 1950); 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 ME results assume that agents are not strategic and therefore do not attempt to influence the prices of the markets (price-takers). In the presence of rational agents, the outcome of the market can be fundamentally different (Brânzei et al., 2014).

**Learning Agents** The fundamental assumptions of price-taking and perfect rationality are challenged in recent years by the emergence of learning agents. As autonomous agents proliferate, they will be called upon to interact in ever more complex environments, and as such, will play a key part in sustainable production. In fact, learning agents have become ubiquitous in socio-economical and socio-ecological systems in recent years (e.g., (Danassis et al., 2021; Zheng et al., 2020)). This has led to the emergence of machina economica (Parkes and Wellman, 2015), 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 (Tardos, 2019)) that enterprises use learning agents as forms of bounded rationality (Rubinstein and Dalkaard, 1998). The success of multi-agent deep reinforcement learning has led to a growing interest in modeling machina economicae agents as independent deep reinforcement learning agents, learning from observational data alone without any economic modeling assumptions. Our work is one of the first to design a policymaker 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., 2019; Shen et al., 2019; Cai et al., 2018; Zheng et al., 2020).

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 (Hardin, 1968) and the challenge of sustainability (resource scarcity) – and (ii) a complex and realistic market (with a dynamic set of buyers and sellers, endowments, and utilities), where the appropriators sell their harvest.

2.1 The Common Fishery Model

We adopt the fishery model of (Danassis et al., 2021), which is based on an abstracted bio-economic model for real-world commercial fisheries (Clark, 2006; Diekert, 2012). 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 to account for multiple resources and harvesters with varying skill levels. Our model describes the dynamics of the stock of a set of common-pool resources, as a group of appropriators harvest over time. The harvest depends on (i) the effort exerted, and (ii) the ease of harvesting that particular resource at that point in time, which depends on its stock level. The stock replenishes at a rate dependent on the current stock level.
More formally, let $\mathcal{N}$ denote the set of appropriators (harvesters) and $\mathcal{R}$ the set of resources. Let $\eta_n = [\eta_{n,1}, \ldots, \eta_{n,r}, \ldots, \eta_{n,R}]$, where $\eta_{n,r} \in [0,1]$ denotes the skill (competence) of harvester $n$ for harvesting resource $r$. At each time-step $t$, every agent exerts a vector of efforts $\phi_{n,t} = [\phi_{n,1,t}, \ldots, \phi_{n,r,t}, \ldots, \phi_{n,R,t}]$, where $\phi_{n,r,t} \in [0, \Phi_{\text{max}}]$ is the effort exerted to harvest resource $r$.

Let $\epsilon_{n,t} = \phi_{n,t} \cdot \eta_n = [\epsilon_{n,1,t}, \ldots, \epsilon_{n,r,t}, \ldots, \epsilon_{n,R,t}]$ 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. [7] where $q_{r,t} \in [0, \infty)$ denotes the stock level at time-step $t$, $q_r(\cdot)$ denotes the catchability coefficient (Eq. [2]), and $S^eq$ is the equilibrium stock of the resource.

$$H_r(E_{r,t}, s_{r,t}) = \begin{cases} q_r(s_{r,t})E_{r,t}, & \text{if } q_r(s_{r,t})E_{r,t} \leq s_{r,t} \\ s_{r,t}, & \text{otherwise} \end{cases}$$

(1)

$$q_r(x) = \begin{cases} \frac{x}{25r^2}, & \text{if } x \leq 2S^eq \\ 1, & \text{otherwise} \end{cases}$$

(2)

Each environment can only sustain a finite amount of stock. If left unharvested, the stock will stabilize at $S^eq$. 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.

$$s_{r,t+1} = F(s_{r,t} - H_r(E_{r,t}, s_{r,t}))$$

(3)

$$F(x) = xe^{g_r(1 - \frac{x}{S^eq})}$$

(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 ($\$/harvesting (e.g., operational cost, taxes, etc.). Here lies the “tragedy”: the benefits from harvesting are private ($(p_{r,t}h_{n,r,t}(\cdot))$, but the loss is borne by all (in terms of a reduced stock, see Eq. [5]).

$$h_{n,r,t}(\epsilon_{n,r,t}, s_{r,t}) = \frac{\epsilon_{n,r,t}}{E_{r,t}} H_r(E_{r,t}, s_{r,t})$$

(5)

$$u_{n,r,t}(\epsilon_{n,r,t}, s_{r,t}) = p_{r,t}h_{n,r,t}(\epsilon_{n,r,t}, s_{r,t}) - c_{n,t}$$

(6)

Having a realistic environment that exhibits resource scarcity, “the tragedy of the commons”, and the challenge of sustainability is not only important for the sake of realism, but it can potentially drastically impede the learning process. The benefits of harvesting can lead to greedy agents, which in turn deplete the resources early in the episode. This will result in short episodes, limiting the learning per episode.

### 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 $\epsilon_r$ denoting the total quantity of good $r \in \mathcal{R}$ brought collectively by the sellers. Every buyer brings a monetary endowment, or simply a budget of $\beta_b$, for $b \in \mathcal{B}$. Additionally, every buyer $b$ has a valuation $v_{b,r}$ for each unit of good $r$. An allocation $x$ 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 $\sum_{b \in \mathcal{B}} x_{b,r} \leq \epsilon_r$, for any good $r$. We will consider linear Fisher markets, where the utility of a buyer given allocation $x$ is defined as:

$$u_b(x) = \sum_{r \in \mathcal{R}} x_{b,r} v_{b,r}$$

(7)

---

1. In our model $\eta_{n,r}$ does not depend on time, but one can consider agents that increase their skill level as they harvest. Moreover one can introduce castes and consider the problem of social mobility.

2. In contrast to alternative CPR games in the literature (e.g., [Zheng et al., 2020]) where resources re-spawn after being depleted.
2.2.1 Market Equilibrium

These markets are also often called Eisenberg-Gale Markets (Eisenberg and Gale, 1959)\(^5\). In such markets, a (competitive) market equilibrium is a pair \((x, p)\) 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 (market clearance). For Eisenberg-Gale markets, in particular, a market equilibrium can be found by solving the following convex optimization program:

\[
\begin{align*}
\text{max} & \quad \sum_{b \in B} \beta_b \cdot \log(u_b) \\
\text{s.t.} & \quad u_b = \sum_{r \in R} v_{b,r} \cdot x_{b,r}, \quad \forall b \in B \\
& \quad \sum_{b \in B} x_{b,r} \leq e_r, \quad \forall r \in R \\
& \quad x_{b,r} \geq 0, \quad \forall b \in B, \quad r \in R
\end{align*}
\]

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 (Jain and Vazirani, 2010; Bei et al., 2016; Chakrabarty et al., 2006).

2.3 Simulation Settings: Fishery & Market

Resources We simulated two scenarios, one with more plentiful resources, and a scarce resources scenario, using the findings of (Danassis et al., 2021) as a guide on the selection of the proper \(S^e_r\) values. See Appendix A for more details. Finally, we set \(\Phi_{\text{max}} = 1, g_r = 1,\) and \(s_0 = S^e_r\) (i.e., the stock starts from the equilibrium population).

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., \(c_{n,t} = 0, \forall n \in N, \forall t\).

Buyers Every time-step, a new set of buyers appears at the market, with budgets and valuations drawn uniformly at random on \([0, 1]\). While we will be referring to individual buyers throughout the text, our analysis extends trivially to the case where each buyer represents a class of buyers with similar budgets and valuations. The assumption that buyers in a market appear in groups with common characteristics is common in both theory and practice, and it is in fact the basis of the well-established market segmentation approach (Wedel and Kamakura, 2000).

3 Agent Architecture

3.1 Multi-Agent Deep Reinforcement Learning

We consider a decentralized multi-agent reinforcement learning scenario in a partially observable general-sum Markov game (e.g., (Littman, 1994; Shapley, 1953)). 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, using 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 (Schulman et al., 2017). For a formal definition please see Appendix B.

3.2 Harvesters’ Architecture

Each harvester takes as input (observation) a tuple \((p_{t-1}, \Phi_{n,t-1}, u_{n,t-1}(\cdot))\) consisting of the vector of prices for the resources, the vector of individual effort exerted for every resource, and reward

\(^5\)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.
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., $G(x, y) = x + U(0, y)$. 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 (Anshelevich et al., 2021) and references therein). The added noise approach corresponds to the case where agents 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 ground truth, pioneered by (Mallows, 1957) but in fact dating back to the works of Marquis de Condorcet, more than two centuries ago.

### 3.3.2 Multi-objective Optimization

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

$$ w_h \frac{1}{|N|} \sum_{n \in N} u_{n,t}(\cdot) + w_b \frac{1}{|B|} \sum_{b \in B} u_{b,t}(\cdot) + w_s \min_{r \in R} (\min(s_{r,t} - S^q_r, 0)) + w_f Fair(x) $$

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, $\sum_{n \in N} u_{n,t}(\cdot)$), the buyers’ social welfare objective (sum of utilities, $\sum_{b \in B} u_{b,t}(\cdot)$), the sustainability objective (defined in this work as the maximum negative deviation from the equilibrium stock, $\min_{r \in R} (\min(s_{r,t} - S^q_r, 0))$, and the fairness objective ($Fair(x)$). Given the broad literature on fairness, we evaluated three different well-established fairness indices: the Jain index (Jain et al., 1998), the Gini coefficient (Gini, 1912), and the Atkinson index (Atkinson, 1970). 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.

---

4For brevity, we only report results on the Jain index. Similar results were obtained for the other indices (see Appendix E).
3.4 Buyers’ Architecture

The buyers are not learning agents; rather, they maximize their utility. To allocate resources to buyers (for the cases where the price vector $p$ is computed by the policymaker, i.e., the non-market equilibrium scenario), we solve the constraint optimization program of (10) that assigns each buyer a bundle based on their budget constraints, aiming to maximize the social welfare of the buyers:

$$\max \sum_{b \in B} u_b(x)$$

subject to

$$u_b(x) = \sum_{r \in R} x_{b,r} v_{b,r}, \ \forall b \in B$$

$$\sum_{r \in R} x_{b,r} p_r \leq \beta_b, \ \forall b \in B$$

$$\sum_{b \in B} x_{b,r} \leq e_r, \ \forall r \in R$$

$$x_{b,r} \geq 0, \ \forall b \in B, \ r \in R$$

3.5 Scalability

We simulated an environment with 8 harvesters, 8 buyer classes, and 4 resources ($N = 8, R = 4, B = 8$, where we overload $B$ to denote the number of classes of buyers).

Our approach can scale gracefully to infinitely large number of harvesters and buyers. This is because the policymaker observes the cumulative harvest per resource (not individual harvest, see Section 3.3) and it is common practice to split buyers into classes (see Section 2.3) with similar budgets and valuations. Thus, assuming that the number (types) of resources and buyer classes stays the same, the size/capacity of the policymaker’s network does not need to grow. For completeness, we also simulated a larger scale scenario ($N = 12, R = 6, B = 12$) and achieved similar results (see Appendix E).

4 Simulation Results

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 8).

We evaluated the “vanilla” policymaker ($w_h = w_b = w_s = w_f = 1$), and four extreme cases where we optimize only one objective, i.e., (i) $w_h = 1$, (ii) $w_b = 1$, (iii) $w_s = 1$, and (iv) $w_f = 1$ (the rest of the weights are set to 0). The latter offers clear-cut results, but – as we will show in Section 4.2 – it can potentially lead to adverse effects. In practice, the use of simulations can enable the testing of economic policies at large-scale, and the ability to evaluate a range of different parameters, allowing the designer to ultimately select the weights that optimize the desired combination of objectives.

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 in the Tables are the average values of the last 400 episodes over those trials. (MA)DRL also lacks common practices for statistical testing (Henderson et al., 2018; Hernandez-Leal et al., 2019). In this work, we opted to use the Student’s T-test (Student, 1908) due to its robustness (Colas et al., 2019); p-values can be found in the Appendix E. All of the reported results that improve the baseline have p-values < 0.05.

Reproducibility Reproducibility is a major challenge in (MA)DRL due to different sources of stochasticity. To minimize those sources we used RLlib, an open-source MADRL library (see Appendix B).
Figure 1: Evolution of several metrics over the number of training episodes. The orange line is the baseline (market equilibrium prices). The blue line refers to the vanilla policymaker where each objective in the reward function has the same weight (see Section 4). The green line refers to the policymaker that only optimizes the specific objective of each figure (i.e., in 1a we set \( w_h = 1 \), in 1b we set \( w_b = 1 \), and in 1c, 1d, 1e, and 1f we set \( w_h = 1 \) and the rest of the weights to 0). The red and purple lines in 1c refer to a policymaker with obfuscated valuations (see Section 3.3.1). In 1f we have a scarce resource setting (see Section 2.3). Shaded areas represent one standard deviation.

4.1 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 MEP 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 policymaker achieving a fair allocation (Jain index \( \geq 0.97 \)). 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 Section 4.4.

4.2 Harvesters’ and Buyers’ 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 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 does 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 to the designer to ultimately select the weights that best serve the desired combination of objectives.

---

\(^7\)The higher the better; Jain index of 1 indicates a fair allocation.
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 4), 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 3.3.1).

| Policymaker           | Vanilla | \(w_h = 1\) | \(w_b = 1\) | \(w_s = 1\) | \(w_f = 1\) | Noisy (50) | Noisy (10) |
|-----------------------|---------|--------------|--------------|--------------|-------------|------------|------------|
| Harvesters’ SW        | -15.30  | -2.64        | -10.58       | -12.83       | -23.40      | -21.73     |            |
| Buyers’ SW            | -0.61   | -0.05        | -0.64        | -0.72        | -1.16       | -1.04      |            |
| Stock Difference       | -7.44   | -1.74        | -72.91       | -34.14       | -11.35      | -9.71      |            |
| Harvesters’ Fairness  | -0.61   | -0.05        | -0.64        | -0.72        | -1.16       | -1.04      |            |
| Buyers’ Fairness      | -0.12   | -0.18        | -0.05        | -0.09        | -0.27       | -0.31      |            |

Finally, we report results on noisy buyers’ valuations (see Section 3.3.1, split into 50 and 10 bins (last two columns of Table 1 and Fig. 1c; see Appendix E for the rest). Noisy valuations lead to only a small drop in the buyers’ and harvesters’ SW (≈ 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.

4.3 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 \(\geq 0.987\), see Appendix E 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.

4.4 Sustainability

We measure sustainability as the maximum negative deviation from the equilibrium stock (see Section 3.3.2). The introduced policymaker results in the emergence of significantly and consistently more sustainable harvesting strategies. Fig. 1c 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 \((w_s = 1)\) improves the difference to 26%.

More interesting is Fig. 1e where we simulate a scarce resource environment (see Section 2.3). 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 9.70% of the episodes, with episodes lasting as low as 48 time-steps (out of 500, which is the maximum possible). 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 on most other objectives, as seen in Table 1. 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 to the designer to select weights that best serve the desired combination of objectives.

\[\text{Note that the stock difference has negative values (negative deviation from the equilibrium stock) thus, in this metric, large negative numbers denote better performance of the policymaker.}\]
4.5 Wasted Resources and Leftover Budget

We also measured the percentage of wasted resources (harvested resources that remain unsold), see Fig. 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).

Regarding the buyers’ leftover budget (see Appendix E), the vanilla policymaker leaves 6% of the budget unused, while optimizing for the harvesters’ revenue leaves only 0.6% of the budget (by design, the MEP use the entire budget).

The wasted resources and the leftover budget represent, in a sense, the over-supply and over-demand of the policymaker’s allocation. It is clear by the low values for both metrics, and the results reported so far, that our policymaker reaches allocations that are qualitative close to the market equilibrium (at least in regard to the reported metrics), while optimizing for negative externalities (such as sustainability).

4.6 Extent of Intervention

Finally, we want to concretely quantify the level of intervention needed to carry out the computed prices. The average (across the 500 time-steps of the final episode) relative price difference \( p_{r,t}^p - p_{r,t}^{ME} / p_{r,t}^{ME} \), where \( p_{r,t}^p \) (\( p_{r,t}^{ME} \)) denotes the policymaker’s (ME) price for resource \( r \) at time-step \( t \) for the vanilla policymaker is between 170% – 437% (for the 4 resources). We can further decrease this difference, by optimizing an additional objective, i.e., adding the following term to Eq. 9: \(-w_i|p_{r,t}^p - p_{r,t}^{ME}|\), where \( w_i \) is a hyper-parameter. Adding this to the vanilla policymaker (all weights 1), reduces the relative price difference to 17% – 20%. We also discretized the price difference into three bins: low (≤ 5%), medium (5% – 20%), and high (> 20%) intervention. The vast majority of the instances (≈ 65%) are in the first two bins (the split amongst the bins is (18.99%, 45.42%, 35.59%)).

In most practical applications, the ME prices required for the aforementioned optimization can not be known in advance. What is known, though, is the current market price of a resource. Thus, we can use the latter to ensure that the prices produced by the policymaker will only require a small intervention to the state of the market.

5 Conclusion

We proposed a practical approach to computing market prices and allocations via deep reinforcement learning, allowing us to counteract negative environmental externalities. We demonstrate significant improvements, especially towards solving the challenge of sustainability of common-pool resources. Our work constitutes an important first step in studying markets composed of learning agents, which are becoming ubiquitous in recent years.

References

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.

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

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

Robert J Barro and Xavier I Sala-i Martin. *Economic growth*. MIT press, 2003.

Francis M Bator. The anatomy of market failure. *The quarterly journal of economics*, 72(3):351–379, 1958.

9Of course, one can use a more involved function of the prices.
William J Baumol. On taxation and the control of externalities. *The American Economic Review*, 62 (3):307–322, 1972.

BBC. Us to release oil reserves in attempt to lower prices, 2021. URL [https://www.bbc.com/news/business-59353194](https://www.bbc.com/news/business-59353194) [Online; accessed 7-February-2022].

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 für Informatik GmbH, Dagstuhl Publishing, 2016.

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

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.

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.

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.

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 Economies*, pages 239–250. Springer, 2006.

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.

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

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

CNN. Facing shortages, canada taps its strategic reserves of maple syrup, 2021. URL [https://edition.cnn.com/2021/11/28/business/canada-maple-syrup-reserves-shortage/index.html](https://edition.cnn.com/2021/11/28/business/canada-maple-syrup-reserves-shortage/index.html) [Online; accessed 7-February-2022].

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.

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.

Thomas D Crocker. The structuring of atmospheric pollution control systems. *The economics of air pollution*, 61:81–84, 1966.

JH Dales. Pollution, property & prices; an essay in policy-making and economics. 1968.

Herman E Daly and Joshua Farley. *Ecological economics: principles and applications*. Island press, 2011.

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.

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

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

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

EU. Eu emissions trading system, 2022a. URL https://ec.europa.eu/clima/eu-action/eu-emissions-trading-system-eu-ets_en [Online; accessed 7-February-2022].

EU. Eu green taxation, 2022b. URL https://ec.europa.eu/taxation_customs/green-taxation-0_en [Online; accessed 7-February-2022].

Irving Fisher. *Mathematical Investigations in the Theory of Value and Prices*. PhD thesis, Yale University, 1892.

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

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

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

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

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

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

Jean-Jacques Laffont. Externalities. *New Palgrave Dictionary of Economics*, 2008.

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.

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

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.

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

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

OECD. Oecd review of fisheries 2020, 2020. URL https://www.oecd-ilibrary.org/agriculture-and-food/oecd-review-of-fisheries-2020_7946bc8a-en [Online; accessed 7-February-2022].
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

Arthur Cecil Pigou. The Economics of Welfare. Macmillan, 1924.

Ariel Rubinstein and Carl-johann Dalgaard. Modeling bounded rationality. MIT press, 1998.

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

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

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.

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

Student. The probable error of a mean. Biometrika, pages 1–25, 1908.

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

Leon Walras. Éléments d’Économie Politique Pure. 1874.

Michel Wedel and Wagner A Kamakura. Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media, 2000.

Wikipedia. Clean air act (united states) — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=Clean_Air_Act_(United_States)&oldid=1069225593; [Online; accessed 7-February-2022].

WPR. Usda plans to buy $50m in milk to reduce surplus, 2018. URL https://www.wpr.org/usda-plans-buy-50m-milk-reduce-surplus; [Online; accessed 7-February-2022].

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.
Appendix

Contents

In this appendix we include several details that have been omitted from the main text due to space limitations. In particular:

- In Section A we describe how we selected the equilibrium population values for the two simulation scenarios (the one with plentiful and the one with scarce resources).
- In Section B we provide details on the agents’ architecture. Specifically, we provide a more formal definition of multi-agent reinforcement learning, details on the learning algorithm used for the policy approximation, the training hyper-parameters, the termination condition, and finally, we list the hardware resources used for training.
- In Section C we describe the employed fairness metrics.
- In Section D we briefly discuss the societal impact of this work.
- Finally, in Section E we provide additional simulation results omitted from the main text.

A Selecting the Equilibrium Population

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), $\forall r \in \mathcal{R}$. The findings of [Danassis et al., 2021] provide a guide on the selection of the $S_r^{eq}$ values. Specifically, we set $S_r^{eq} = M_s KN$, 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. We simulated two scenarios, one with $M_s = 0.8$, and a scarce resources scenario with $M_s = 0.45$.\footnote{Yet, coordination is far from trivial; see [Danassis et al., 2021].}

B Agents’ Architecture: Additional Details

B.1 Multi-Agent Deep Reinforcement Learning

We consider a decentralized multi-agent reinforcement learning scenario in a partially observable general-sum Markov game (e.g., [Littman, 1994; Shapley, 1953]). 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 $\mathcal{N} = \{1, \ldots, N\}$ 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 $O^n : \mathcal{S} \rightarrow \mathbb{R}^d$ 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 $a = \times_{n \in \mathcal{N}} a^n$, where $a^n \in \mathcal{A}^n$, the joint action. The states change according to a transition function $T : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \rightarrow \Delta(\mathcal{S})$, 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 $s_t \in \mathcal{S}$ and joint action $a_t$. The latter is given by the reward function $r^n : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \rightarrow \mathbb{R}$. 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 $\pi = \times_{n \in \mathcal{N}} \pi^n$ denote the joint policy. The goal for each agent is to maximize the long term discounted payoff, as given by $V^n_{\pi}(\sigma_0) = E\left[\sum_{t=0}^{\infty} \gamma^t r^n(\sigma_t, a_t)|\sigma_0 \sim \pi, a_t \sim T(\sigma_t, a_t)\right]$, where $\gamma$ is the discount factor and $\sigma_0$ is the initial state.

B.2 Learning Algorithm

The policies for all agents (harvesters and the policymaker) are trained using the Proximal Policy Optimization (PPO) algorithm [Schulman et al., 2017]. 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 3.4 of the main text.\footnote{In the simulations of [Danassis et al., 2021], for $N = 8$ and $M_s \leq 0.4$, the agents failed to find a sustainable strategy, and always depleted the resource.}
B.3 Implementation, Hyper-parameters, and Reproducibility

Reproducibility is a major challenge in (MA)DRL due to different sources of stochasticity, e.g., hyperparameters, model architecture, implementation details, etc. (Henderson et al. 2018; Hernandez-Leal et al. 2019; Engstrom et al. 2020). Recent work has demonstrated that code-level optimizations play an important role in performance, both in terms of achieved reward and underlying algorithmic behavior (Engstrom et al. 2020). 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 as our implementation framework. All the hyper-parameters were left to the default values specified in Ray and RLlib. For completeness, Table 2 presents a list of the most relevant of them.

| Parameter                        | Value       |
|----------------------------------|-------------|
| Learning Rate ($\alpha$)         | 0.0001      |
| Clipping Parameter               | 0.3         |
| Value Function Clipping Parameter| 10.0        |
| KL Target                        | 0.01        |
| Discount Factor ($\gamma$)       | 0.99        |
| GAE Parameter Lambda             | 1.0         |
| Value Function Loss Coefficient  | 1.0         |
| Entropy Coefficient              | 0.0         |

B.4 Termination Condition

An episode terminates when either (a) a fixed number of time-steps $T_{\text{max}} = 500$ 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.

B.5 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.

C Fairness Metrics

We employed three well established fairness metrics; the Jain index (Jain et al. 1998), the Gini coefficient (Gini 1912), and the Atkinson index (Atkinson 1970).

(a) The Jain index (Jain et al. 1998): 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 allocation game of $N$ agents, such that the $n^{\text{th}}$ agent is allotted $x_n$, the Jain index is given by Eq. (11) $J(x) \in [0, 1]$. An allocation $x = (x_1, \ldots, x_N)^\top$ is considered fair, iff $J(x) = 1$.

$$J(x) = \frac{\left(\sum_{n=1}^{N} x_n\right)^2}{N \sum_{n=1}^{N} x_n^2}$$  \hspace{0.5cm} (11)

(b) The Gini coefficient (Gini 1912): 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
Figure 2: Evolution of the mean leftover budget over the number of training episodes. The orange line is the baseline (market equilibrium prices). The blue line refers to the vanilla policymaker where each objective in the reward function has the same weight. The green line refers to the policymaker that optimizes for the harvesters’ reward (i.e., $w_h = 1$ and the rest of the weights to 0). Shaded areas represent one standard deviation.

The Atkinson index [Atkinson, 1970]: Is a measure of the amount of social utility to be gained by complete redistribution of a given income distribution, for a given $\epsilon$. In our work, we used $\epsilon = 1$. For an allocation game of $N$ agents, such that the $n^{th}$ agent is allotted $x_n$, the Atkinson index of $\epsilon = 1$ is given by Eq. 13 $A(x) \in [0, 1]$. An Atkinson index of zero expresses perfect equality, i.e., an allocation is fair iff $A(x) = 0$.

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

### D 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

### E Additional Simulation Results

#### Results in Detail

In Table 5 we provide a thorough account of the simulation results of the main text.

#### Leftover Budget

Figure 2 depicts the average (across the 8 trials) leftover budget over the number of training episodes.

#### Larger Scale Simulations

Table 3 presents the results of the larger simulation with 12 harvesters, 12 buyer classes, and 6 resources ($N = 12, R = 6, B = 12$). Note that the policymaker achieves similar results as in the smaller test-case (first two columns of Table 3).

#### Extent of Intervention

In Table 4 we show that optimizing for the extent of the intervention to the market does not significantly affect the results for the original objectives.
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 vanilla 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 present the results for an environment with 8 harvesters, 8 buyer classes, and 4 resources \((N = 8, R = 4, B = 8)\), while the last two represent a larger scale scenario with 12 harvesters, 12 buyer classes, and 6 resources \((N = 12, R = 6, B = 12)\). Note that the results of the first two columns are the same as the ones in Table 5, and have only been included to facilitate comparison between the two test-cases.

| Metric                  | Vanilla Policymaker | Interventions | p-value | p-value |
|-------------------------|---------------------|---------------|---------|---------|
| Harvesters’ Social Welfare | -7.44 1.21e-08     | 11.2 4.15e-08 |         |         |
| Buyers’ Social Welfare  | -7.01 2.26e-06     | -4.51 5.26e-05 |         |         |
| Stock Difference        | -15.30 2.72e-09    | -19.31 2.24e-16 |         |         |
| Harvesters’ Fairness Jain | -0.61 3.71e-03    | -0.45 6.37e-04 |         |         |
| Harvesters’ Fairness Gini | -2.78 2.87e-04   | -2.4 1.15e-05  |         |         |
| Harvesters’ Fairness Atkinson | -0.29 4.45e-03 | -0.21 6.26e-04 |         |         |
| Buyers’ Fairness Jain  | -0.12 5.48e-05     | -0.04 5.44e-06 |         |         |
| Buyers’ Fairness Gini  | -1.49 3.26e-07     | -0.78 3.57e-08 |         |         |
| Buyers’ Fairness Atkinson | -0.06 5.50e-05 | -0.02 5.63e-06 |         |         |

Table 4: Numerical results of the last 400 episodes of each training trial (averaged over the 8 trials). The first column refers to the vanilla policymaker, while the second to the vanilla policymaker that also optimizes for the extent of the intervention to the market (see Section 4.6 of the main text).

| Metric                  | Vanilla | Interventions |
|-------------------------|---------|---------------|
| Harvesters’ Social Welfare | 64.38   | 65.59         |
| Buyers’ Social Welfare  | 75.66   | 81.45         |
| Stock Difference        | -28.69  | -33.64        |
| Wasted Percentage       | 14.70   | 2.64          |
| Harvesters’ Fairness Jain | 0.98    | 0.99          |
| Harvesters’ Fairness Gini | 0.94    | 0.97          |
| Harvesters’ Fairness Atkinson | 0.99    | 0.99          |
| Buyers’ Fairness Jain  | 0.99    | 0.99          |
| Buyers’ Fairness Gini  | 0.98    | 0.99          |
| Buyers’ Fairness Atkinson | 0.99    | 0.99          |
Table 5: 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 \((100(\frac{X_{\text{policymaker}} - Y_{\text{market eq.}}}{Y_{\text{market eq.}}}))\), for each of the metrics presented in each row.

Each even column shows the Student’s T-test p-values.

The first two columns refer to the vanilla policymaker, where each objective in the reward function has the same weight (see Section 4 of the main text), and each of the following 8 columns refer 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 3.3.1 of the main text). 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.

| Metric                  | 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   | 1.2e-08 | -1.74       | 4.16e-07 | -72.91      | 4.87e-17 | -31.37      | 7.00e-06 | -34.14      | 1.67e-06 | -11.35     | 2.33e-05 | -9.71      | 2.95e-03 | -8.20      | 2.50e-01 | -10.07     | 9.92e-04 |
| Buyers’ Social Welfare  | -7.01   | 2.26e-06| -2.47       | 6.65e-08 | 15.42       | 9.72e-13 | 1.23        | 2.28e-04 | 2.88        | 2.14e-05 | -9.73      | 9.88e-05 | -11.51     | 2.25e-04 | -13.68     | 3.89e-06 | -11.56     | 2.00e-04 |
| Stock Difference        | -15.30  | 2.72e-09| -2.64       | 6.26e-08 | -10.58      | 2.34e-02 | -2.13       | 2.36e-03 | -12.99      | 1.53e-01 | -2.34      | 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       | 6.83e-01 | -0.14       | 2.15e-02 | -1.16      | 6.44e-03 | -1.04      | 2.70e-02 | -1.33      | 1.02e-03 | -1.76      | 1.55e-05 |
| Harvesters’ Fairness Gini | -2.78   | 2.87e-04| -0.54       | 2.09e-03 | -2.86       | 9.16e-01 | -3.08       | 6.93e-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      | 6.30e-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 |