AI-driven Prices for Externalities and Sustainability in Production Markets

Panayiotis Danassis
Harvard University
Cambridge, MA, USA
pdanassis@g.harvard.edu

Aris Filos-Ratsikas
University of Edinburgh
Edinburgh, United Kingdom
aris.filos-ratsikas@ed.ac.uk

Haipeng Chen
William & Mary
Williamsburg, VA, USA
hchen23@wm.edu

Milind Tambe
Harvard University
Cambridge, MA, USA
milind_tambe@harvard.edu

Boi Faltings
EPFL
Lausanne, Switzerland
boi.faltings@epfl.ch

ABSTRACT

Traditional competitive markets do not account for negative externalities; indirect costs that some participants impose on others, such as the cost of over-appropriating a common-pool resource (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

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 [3] suggests that these markets will reach an efficient stable outcome, the *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, these established market models fail to account for negative *externalities* [45], which lead to market failure [5]. 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, water pollution from industrial effluents, antibiotic resistance due to overuse of antibiotics in industrial farms, etc. Another prominent example of a negative externality – which we will use as a real-world, indicative test-case throughout the paper – is the depletion of the stock of fish due to overfishing.[1] With these "exogenous" objectives being of paramount importance, it is only natural to assume some form of intervention to the reign of these markets.

There are many approaches to reconcile the economics of the competitive market with societal and environmental externalities. For example, policy-makers can correct for the inefficiencies by employing command-and-control legislation (e.g., [73]), permit markets [20, 21] (e.g., [31]), or taxation (e.g., [32]). A classic example of the latter is Pigouvian taxes [56], 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 instance, 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 [16]. Similarly, determining the marginal social cost of a negative externality and converting it into a monetary value can be quite impractical [6]. Another approach, founded in the seminal works of Samuelson [59, 60], attempts to capture and address externalities via the introduction of *public goods* in the market, and the computation of the equilibrium points for these "extended" markets. The extent to which these public goods can capture all appropriate government functions has been largely debated (e.g., see [55, 60]) and besides, the impracticality of computing the equilibrium points of these markets analytically is even more pronounced than in "traditional" competitive markets.

An added complication when it comes to devising effective policies for sustainability and combating externalities, or any societal objective for that matter, is the fact that the interactions between the different entities in the market ecosystem are rather complex and of a repeated nature. Indeed, the appropriate mathematical modeling of these systems is that of a *stochastic* (or *Markov*) game (see Section 2), in which the actors (i.e., the policy-maker and the harvesters of natural resources) are both aiming to optimize their individual utilities over a fixed horizon. To do that, they need to optimize over their future rewards, taking into the account the effect of the actions of the other actors on their own. Solving these games analytically is both conceptually and computationally hard, even for relative simple variants of those games and well-behaved equilibrium notions (e.g., see [12, 26, 27]). For this reason, most classic works in economics and mathematics (e.g., [40, 65, 66]) have only gone as far as identifying conditions that merely establish the existence of some equilibrium, without providing any guarantees about its properties. Besides the computational burden, another significant hurdle to the analytical approach is that it typically requires full observability of the environment and the actions of the other participants, which is most often not the case in practice.

[1] For example, according to OECD, about 25% of fish stocks globally are at risk [53].

Reinforcement learning (see [44]) has been proposed and extensively used as an alternative approach for computing optimal strategies in stochastic games [48]. The idea is that the actors, as learners, interact with their environment exclusively via signals of limited information: they typically observe their rewards based on their past actions, and update their current actions accordingly, via the employment of some carefully devised learning algorithm. Note that this approach does not require observable information about the parameters of the environment or the other actors. It also does not require the derivation of analytical solutions to complex optimization problems, as "off the shelf" reinforcement learning algorithms are readily available. For these reasons, an established line of work has considered reinforcement learning as a form of bounded rationality [58] which is much more conceivable for complex environments in practice, compared to the standard "perfect" rationality of traditional economic agents (see Section 1.2.2 for more details). Finally, reinforcement learning has been shown to be generally robust to changes in a range of input parameters, making it very suitable for complex and volatile environments.

Motivated by the arguments above, we propose a practical and effective technique for calculating concrete market prices and allocations via a *deep reinforcement learning policymaker*, 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.[2] This new approach grants us the ability to abstract real-world situations into a form that makes them amenable to research, and allows for advances in the state-of-the-art. In particular, it enables us 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, CO_2 markets, natural resource preservation, etc.). Our work falls into the very recent research agenda of building agent-based models to inform policy in socio-economic environments (see [43, 76]).

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. Our goal is to investigate the feasibility of using traditional deep reinforcement learning agents as (i) a practical alternative to classical notions of rationality and market equilibria, and (ii) a means to reach stable outcomes that are comparable with the idealized market equilibrium outcome from economics, while at the same time optimizing exogenous objectives.

(2) We introduce a novel multi-agent socio-economic environment and *prove* a necessary condition for market failure. Our environment 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 an indicative, real-world test-bench, our approach is general and can be employed in *any* production market. To demonstrate the need for intervention via our policymaker, we provide an analytical example (see Section 2.3) where leaving the market entirely "free" to act according to the market equilibrium will result in the depletion of the resource in a short period of time. Our analysis also highlights the inherent challenges of theoretically computing optimal strategies for either the policymaker or the harvesters, and justifies modeling both as learning agents instead.

(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

1.2.1 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 [33], 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 (e.g., see [15]), computing a ME can be done in polynomial time. We remark that in our setting the harvesters are the recurring entities in the system, as they return to the market with (potentially) new harvest in each round. On the other hand, we assume that buyers are short-lived (i.e., they only participate in the market once) or myopic (i.e., they are trying to maximize their utility at the current market stage). Goktas et al. [36] recently defined the notion of a *stochastic* Fisher market (extending the ideas of [34] for Arrow-Debreu-type markets), where the same buyers participate in the same Fisher market over a sequence of steps, aiming to maximize their long-term utility and possibly saving part of their budget for future steps. As our focus in this work

[2] There are many such examples of influencing the supply/demand [7, 17, 74].

is on the harvesters (who are also our learning agents) rather than the buyers, we elected to study the standard, static Fisher market model instead.

The ME is theoretically reached via the continuous adjustment of supply and demand dictated by the market's "invisible hand" [64] ("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 [14, 15, 18]. We also remark that while similar in spirit, the ME is a different notion from the well-known notion of the Nash equilibrium [51]; 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 from the ME [1, 9, 13].

1.2.2 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., [24, 76]). This has led to the emergence of machina economica [54], 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 [69]) that enterprises use learning agents as forms of bounded rationality [58]. These can range from simple no-regret bandit algorithms like in [11, 52], to more complex approaches. The success of multi-agent deep reinforcement learning (DRL) has led to a growing interest in modeling machinae economicae agents as *independent* deep reinforcement learning agents, learning from observational data alone without any economic modeling assumptions. As an example, in a recent work, Zheng et al. [76] use DRL to study taxation policies, but in a simpler domain (they use a synthetic appropriation game – compared to our real-world fishery model – and do not consider a concrete market model like our Fisher market). Moreover, the resources can not be permanently depleted, thus they do not address the challenge of sustainability and the tragedy of the commons, and they allow for weight sharing during training, while we have completely independent learning agents. There is other recent work that has adopted a similar agenda, but in markedly different domains and using different approaches [11, 29, 43, 63, 70]. Our work is one of the first to design a *practical* policymaker via deep reinforcement learning in realistic economic environments.

Finally, there has been great interest lately in Common-Pool Resource (CPR) appropriation problems as an application domain for Multi-agent Deep Reinforcement Learning (MADRL), e.g., [25, 75, 76]. CPR problems offer complex environment dynamics and relate to real-world socio-ecological systems. Danassis et al. (2021) were the first to introduce the realistic fishery model that we employ in this work (see also [23, 25]). We extended the model to deal with multiple resources, harvesters with diverse skill levels, and a realistic Fisher market.

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 [37] 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

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* [37]. 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 [22]).

We adopt the fishery model of [24], which is based on an abstracted bio-economic model for *real-world* commercial fisheries [16, 28]. 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 $\boldsymbol{\eta}_n = [\eta_{n,1}, \dots, \eta_{n,r}, \dots, \eta_{n,R}]$, 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 Equation 1, where $s_{r,t} \in [0, \infty)$ denotes the stock level at time-step t, $q_r(\cdot)$ denotes the catchability coefficient (Equation 2), and S_r^{eq} is the equilibrium stock of the resource.

$$H_r(E_{r,t}, s_{r,t}) = \begin{cases} q_r(s_{r,t})E_{r,t} & \text{, if } q_r(s_{r,t})E_{r,t} \le s_{r,t} \\ s_{r,t} & \text{, otherwise} \end{cases} \tag{1}$$

$$q_r(x) = \begin{cases} \frac{x}{2S_r^{eq}} & \text{, if } x \le 2S_r^{eq} \\ 1 & \text{, otherwise} \end{cases} \tag{2}$$

Each environment can only sustain a finite amount of stock. If left unharvested, the stock will stabilize at S_r^{eq}. Note also that $q_r(\cdot)$,

[3] 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.

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 Equation 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 (Equation 5), and the revenue of each appropriator is given by Equation 6, where $p_{r,t}$ is the price ($ per unit of resource), and $c_{n,r,t}(\cdot)$ 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 Equation 3).

$$h_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t}) = \frac{\varepsilon_{n,r,t}}{E_{r,t}} H_r(E_{r,t}, s_{r,t}) \quad (5)$$

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

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.[5]

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

2.2.1 Market Equilibrium.

These markets are also often called *Eisenberg-Gale Markets* [30].[6] 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 (market clearance). For Eisenberg-Gale markets, in particular, a market equilibrium can be found by solving a convex optimization program (see supplement).

2.3 Market Failure Under Optimal Harvesting

In this section we provide concrete examples of conditions that lead to market failure – specifically, the depletion of the resource – under market equilibrium prices and *optimal harvesting*. We remark that while market failures are a well-documented phenomenon [6, 45, 57, 68], our example shows that they can actually happen even when the harvesters are coordinating and harvesting using *optimal strategies*, which in fact incorporate the need for future harvest, to ensure not just immediate but also future rewards.

To simplify our analysis, we consider a setting where a single resource is harvested and sold and which the means of harvesting are controlled by a single entity; effectively, the latter corresponds to having a single harvester in our model. Thus, we only have a single control variable, ε_t, corresponding to the cumulative harvest.[7] This makes our example stronger, as we show that even if the harvesters could somehow conceivably coordinate to harvest optimally as a group, the resource would still be depleted.

Finding the optimal harvesting strategy consists of finding a piecewise continuous control ε_t, so as to maximize the total discounted revenue for a given time horizon T, i.e., $\max_{\varepsilon_t} \sum_{t=0}^{T} \gamma^t u_t(\varepsilon_t, s_t)$, subject to (3), where γ is the discount factor. The optimal control is given by Theorem 2.1 (please see the supplement for the proof).

THEOREM 2.1. *The optimal control $\varepsilon_t^* = \arg\max_{\varepsilon_t} \sum_{t=0}^{T} \gamma^t u_t(\varepsilon_t, s_t)$, subject to (3) and assuming linear to the effort cost function $c_t(\varepsilon_t) = c\varepsilon_t + c'$, is given by the following equation, where λ_t are the adjoint variables of the Hamiltonians:*

$$\varepsilon_t^* = \begin{cases} E_{max}, & \text{if } (\gamma^t p_t - \lambda_t) q(F(s_{t-1} - H(\varepsilon_{t-1}, s_{t-1}))) - \gamma^t c \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

From Theorem 2.1, it follows that as long as the price p_t at which the good is sold in each round is large enough, the harvesters will exert maximum effort in each round, which can lead to the depletion of the resource (we quantify the time needed for that to happen later in the section).

Now let's assume that in the absence of any policymaker, the market equilibrium (i.e., the free market outcome) is calculated in each time-step. By the clearing condition of the Fisher market [41, 50], it holds that at the chosen price $p_t = p$, the budget of each buyer is entirely spent, i.e., $x_b \cdot p = \beta_b$, where x_b is the amount of the resource allocated to buyer b. By summing over all budgets, we obtain that $\beta = \sum_b \beta_b = p$ (assuming for simplicity that the total supply of the resource is 1), i.e., the total price should be equal to the cumulative budget of all buyers. Then, according to Theorem 2.1, if the cumulative budget is sufficiently large, the harvesters will harvest with maximum effort at every time-step, eventually depleting the resource. In what follows, we quantify what "large enough" means; specifically, we provide a sufficient condition on the relation between the cumulative budgets in different time-steps that ensures maximum-effort harvest in each time-step. This is captured in the following theorem, whose proof can be found in the supplement. For simplicity, we assume fixed operational cost $c_t(\varepsilon_t) = c'$, as in [24].

THEOREM 2.2. *For $|\mathcal{R}| = |\mathcal{H}| = 1$, $c_t(\varepsilon_t) = c'$, and the model dynamics described in Section 2, the harvester will harvest with maximum effort E_{max} iff $\beta_T \geq 0$ and $\forall t < T$:*

$$\frac{\beta_t}{\max\limits_{i \in [t+1 \dots T]} \beta_i} \geq \frac{1}{2S^{eq}} \sum_{j=1}^{T-t} (\gamma e^g)^j \quad (7)$$

where $\beta_t = \sum_b \beta_{b,t}$.

[4] Assuming $s_0 \leq 2S_{eq}$, and following the derivation of [24], to avoid highly skewed growth models and unstable environments we need $g_r \leq -W_{-1}(-1/(2e)) \approx 2.678$, where $W_k(\cdot)$ is the Lambert W function.

[5] In contrast to alternative CPR games in the literature (e.g., [76]) where resources re-spawn after being depleted.

[6] 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.

[7] To improve readability, we have omitted the subscripts n, and r, in this section.

We remark that this is only a *sufficient* condition, as we have applied several simplifications to the derivation to end up with a formula from which useful information can be extracted. Computing a necessary condition appears to require extremely tedious derivations, as the bound involves computing partial derivatives of a recursive expression which results in a function composed over T time-steps. Still, as we demonstrate in the following, this sufficient condition is enough to support our claim that depletion can realistically happen in environments of interest.

According to (7), as we reach the time horizon T, a smaller budget is needed to prompt the harvester into harvesting with maximum effort. As a grounding example, consider the following concrete parameters of the model: $t = 0$ (which corresponds to the worst case), $\gamma = 0.9$, $g = -\ln(\gamma) \approx 0.105$, a time horizon of 5 years ($T = 60$, assuming a time-step is one month), and $S^{eq} = M_s K N = 0.8 \times 0.79 \times 1$ (see Section 2.4.1). In this case, (7) results in:

$$\frac{\beta_0}{\max\limits_{i \in [1 \ldots T]} \beta_i} \geq \frac{1}{2S^{eq}} \sum_{j=1}^{T} (\gamma e^g)^j = \frac{T}{2S^{eq}} \approx 47.5 \tag{8}$$

(8) indicates that there exist environments where the *cumulative* budget required to harvest with maximum effort in every time-step is not much smaller than its initial value. Given that β is indeed a cumulative measure, such environments are reasonable since, even if the budget for each buyer is (almost) fixed, the number of buyers that actually appear in the market might decrease over time.

Now suppose that the harvesters do harvest with maximum effort. The next step is to calculate the time of depletion. Let $D(s) = F(s - H_r(E_{max}, s))$, where $F(\cdot)$ and $H(\cdot)$ are given by (4) and (1), respectively. The resource will get depleted in $t_d \in \mathbb{Z}^+$ time-steps, such that $D^{t_d}(s_0) < \delta S^{eq}$, i.e., when the stock drops below a percentage of the equilibrium population.[8] Assuming $\delta = 10^{-3}$ and the same parameters specified in the example above, then $t_d = 5$ (months). For $\delta = 10^{-6}$, $t_d = 10$ (months). Both are considerably lower than the desired time horizon ($T = 60$ months). In simple terms, the practical example of this section comes to show that there exist environments where:

> *Market equilibrium prices will irrefutably lead to the depletion of the resource, under optimal harvesting.*

Another takeaway from the results of this section is that computing the optimal harvesting strategy analytically is quite challenging, as there is no closed form and the λ_t values of Theorem 2.1 can only be computed recursively via repeated partial derivation (see the proof in the supplement for more details). Similarly, preventing the potential depletion of the resource requires the policy-maker to at the very least have knowledge of the maximum and minimum budgets over the time horizon T, which typically will not be the case in practice. This motivates the use of DRL over the analytical approach, as we also advocate in the Introduction.

2.4 Simulation Settings: Fishery & Market

2.4.1 Resources. We simulated two scenarios, one with more plentiful resources, and a *scarce resources* scenario, using the findings of [24] as a guide on the selection of the proper 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). [9] For the majority of the simulations we used $M_s = 0.8$, while for the *scarce resources* scenario we used $M_s = 0.45$.[10] 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), $\forall r \in \mathcal{R}$.

2.4.2 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., $c_{n,r,t}(\cdot) = 0$, $\forall n \in \mathcal{N}$, $\forall t$.

2.4.3 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 [72].

3 AGENT ARCHITECTURE

3.1 Decentralised Multi-Agent Deep RL

We consider a *decentralized* multi-agent reinforcement learning scenario in a partially observable general-sum stochastic game (e.g., [47, 62]). 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. More formally, let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of agents, and $\mathcal{M}$ be an N-player, partially observable stochastic 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 $\boldsymbol{a} = \times_{\forall n \in \mathcal{N}} a^n$, where $a^n \in \mathcal{A}^n$, the joint action. The states change according to a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \rightarrow \Delta(\mathcal{S})$, 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 $\sigma_t \in \mathcal{S}$ and joint action $\boldsymbol{a}_t$. The latter is given by the reward function $r^n : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \rightarrow \mathbb{R}$. Finally, each agent learns a policy $\pi^n : O^n \rightarrow \Delta(\mathcal{A}^n)$ 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 $V_{\boldsymbol{\pi}}^n(\sigma_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^n(\sigma_t, \boldsymbol{a}_t) | \boldsymbol{a}_t \sim \boldsymbol{\pi}_t, \sigma_{t+1} \sim \mathcal{T}(\sigma_t, \boldsymbol{a}_t)\right]$, where γ is the discount factor and σ_0 is the initial state.

3.1.1 Learning Algorithm. The policies for all learning agents (harvesters and the policymaker)[11] are trained using the Proximal Policy Optimization (PPO) algorithm [61]. PPO was chosen because it avoids large policy updates, ensuring a smoother training, and avoiding catastrophic failures. Please see the supplement for implementation details and hyperparameters.

[8] $D^x(\cdot)$ denotes the xth function composition of $D(\cdot)$ with itself.

[9] For example, for $M_s = 1$ the resource will not get depleted, even if all agents harvest at maximum effort. Yet, coordination is far from trivial; see [24].

[10] In the simulations of [24], for $N = 8$ and $M_s \leq 0.4$, the agents failed to find a sustainable strategy, and always depleted the resource.

[11] The buyers are not learning agents; see Section 3.5.

3.2 Harvesters' Architecture

Each harvester's input (observation) 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., $\sum_{r\in\mathcal{R}} u_{n,r,t}(\varepsilon_{n,r,t}, s_{r,t})$.

3.3 Policymaker Architecture

The input of the policymaker 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 harvesters for all resources, $\boldsymbol{s}_t$ is the current stock level of each resource, $\boldsymbol{\beta}_t$ is the budgets of the current set of buyers (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.

3.3.1 Multi-objective Optimization. The policymaker's reward is the weighted average of the desired objectives, specifically:

$$w_h \frac{1}{|\mathcal{N}|} \sum_{n\in\mathcal{N}} \boldsymbol{u}_{n,t}(\cdot) + w_b \frac{1}{|\mathcal{B}|} \sum_{b\in\mathcal{B}} \boldsymbol{u}_{b,t}(\cdot) + \\ w_s \min_{r\in\mathcal{R}} \left(\min(s_{r,t} - S_r^{eq}, 0)\right) + w_f Fair(\mathbf{x}) \tag{9}$$

where $w_h, w_b, w_s, w_f \in [0, 1]$ 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 utilities, $\sum_{b\in\mathcal{B}} \boldsymbol{u}_{b,t}(\cdot)$), the sustainability objective (defined in this work as the maximum negative deviation from the equilibrium stock, $\min_{r\in\mathcal{R}} \left(\min(s_{r,t} - S_r^{eq}, 0)\right)$), and the fairness objective ($Fair(\mathbf{x})$). Given the broad literature on fairness, we evaluated three different well-established fairness indices: the Jain index [42], the Gini coefficient [35], and the Atkinson index [4].[12] 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*.

3.4 Robustness

3.4.1 Valuation Obfuscation. 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 [46] (see also [10]). 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 [8] 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 [2] 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 ground truth, pioneered by [49] but in fact dating back to the works of Marquis de Condorcet, more than two centuries ago.

3.4.2 Effort Misestimation. We also test the robustness of the policymaker against harvesters that misestimate their exerted effort, by adding uniform noise $\mathcal{U}(0, y)$ to the effort values. This can be considered as a form of robustness against sim-to-real gap. In simple terms, agents would attempt to fish with some effort ε, but due to hardware implementation mismatching, partial observability, noisy environments etc., they might experience perturbations in the actual effort exerted, now $\varepsilon' = \varepsilon + \mathcal{U}(0, y)$.

3.5 Buyers' Architecture

The buyers are not learning agents; rather, they maximize their utility. To allocate resources to buyers, we solve a constraint optimization program (see supplement) that assigns each buyer a bundle based on their budget constraints, aiming to maximize the social welfare of the buyers.

3.6 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 2.4.3) 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 supplement).

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 we describe in the supplement).

We evaluated the "vanilla" policymaker ($w_h = w_b = w_s = w_f = 1$), and four extreme cases where we optimize only one objective,

[12] For brevity, we only report results on the Jain index. Similar results were obtained for the other indices (see supplement).

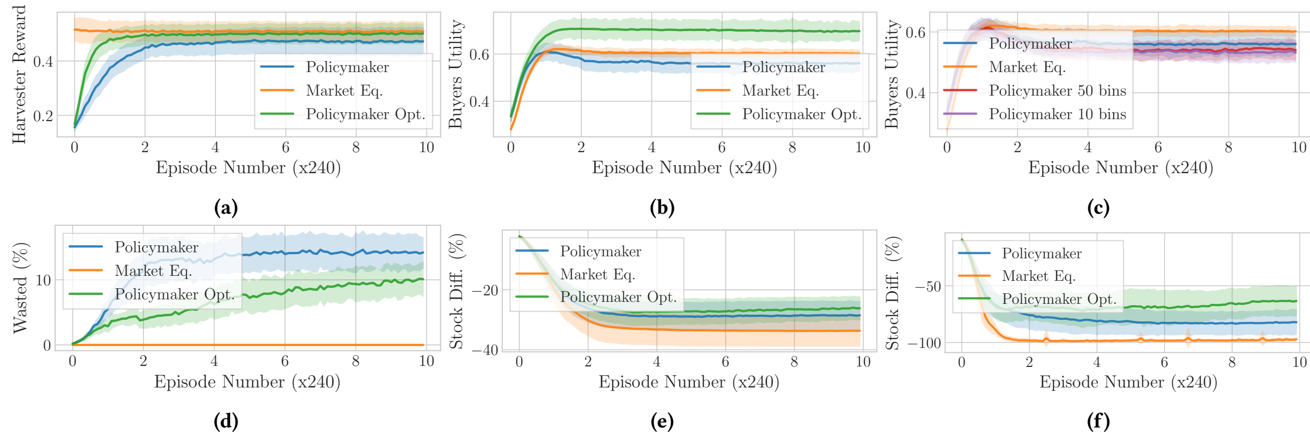

Figure 1: Evolution of several metrics over the number of training episodes. The orange line is the baseline (market equilibrium prices). The blue line refers to the vanilla policymaker where each objective in the reward function has the same weight (see Section 4). The green line refers to the policymaker 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 1d, 1e, and 1f we set $w_s = 1$ and the rest of the weights to 0). The red and purple lines in 1c refer to a policymaker with obfuscated valuations (see Section 3.4.1). In 1f we have a scarce resource setting (see Section 2.4 and supplement). Shaded areas represent one standard deviation.

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 [38, 39]. In this work, we opted to use the Student's T-test [67] due to it's robustness [19]; p-values can be found in the supplement. 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 the supplement for details and the *source code*.

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

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 ($100(X_{\textbf{policymaker}} - Y_{\textbf{market eq.}})/Y_{\textbf{market eq.}}$), 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 three columns refer to a vanilla policymaker with obfuscated valuations (valuations split into 10 bins), noisy efforts (10% uniform noise), and noisy efforts and valuations (10% uniform noise), see Section 3.4.1.

	Policymaker							
	Vanilla	$w_h = 1$	$w_b = 1$	$w_s = 1$	$w_f = 1$	$\approx v$	$\approx \phi$	$\approx v$ & ϕ
Harvesters' SW	-7.44	-1.74	-72.91	-31.37	-34.14	-9.71	-7.08	-9.24
Buyers' SW	-7.01	-24.71	15.42	1.23	2.88	-11.51	-5.19	-10
Stock Difference[14]	-15.3	-2.64	-10.58	-21.83	-12.99	-21.73	-19.59	-24.58
Harvesters' Fairness	-0.61	-0.05	-0.64	-0.72	-0.14	-1.04	-0.53	-0.77
Buyers' Fairness	-0.12	-0.18	-0.05	-0.09	-0.07	-0.31	-0.09	-0.31

achieving a fair allocation (Jain index ≥ 0.98[13]). This is particularly important, as the MEP are 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* (see Section 4.5).

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

[13]The higher the better; Jain index of 1 indicates a fair allocation.

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.

4.3 Robustness

Buyers' valuations are hard to elicit, while harvesters might experience perturbations to the actual effort exerted (see Section 3.4). To evaluate this, we report results on noisy buyers' valuations (see Section 3.4.1), split into 50 and 10 bins (third from the last column of Table 1, and Figure 1c; see supplement 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. Similar results were achieved for noisy efforts, and both noisy efforts and noisy valuations (last two columns of Table 1, and supplement).

4.4 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 versions of the policymaker achieve a fair allocation (Jain index ≥ 0.98,[13] see supplement 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.

4.5 Sustainability

We measure sustainability as the maximum negative deviation from the equilibrium stock (see Section 3.3.1). The introduced policymaker results in the emergence of *significantly and consistently more sustainable harvesting strategies*. Figure 1e shows that the MEP maintain a 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 2.4). 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, 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.[14]

4.6 Wasted Resources and Leftover Budget

Figure 1d shows the percentage of wasted resources (harvested resources that remain unsold). 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 supplement), 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).

4.7 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 ($p^p_{r,t} - p^{ME}_{r,t} / p^{ME}_{r,t}$, where $p^p_{r,t}$ ($p^{ME}_{r,t}$) 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 Equation 9: $-w_i|p^p_{r,t} - p^{ME}_{r,t}|$, where w_i is a hyperparameter.[15] 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 ($\leq$ 5%), medium (5% – 20%), and high (> 20%) intervention. The vast majority of the instances ($\approx$ 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

[14]Note that the stock difference has negative values (negative deviation from the equilibrium stock) thus, in this metric, large negative numbers denote better performance.
[15]Of course, one can use a more involved function of the prices.

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

[1] Bharat Adsul, Ch Sobhan Babu, Jugal Garg, Ruta Mehta, and Milind Sohoni. 2010. Nash equilibria in Fisher market. In *International Symposium on Algorithmic Game Theory*. Springer, 30–41.

[2] Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A Voudouris. 2021. Distortion in Social Choice Problems: The First 15 Years and Beyond. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*.

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

[4] Anthony B Atkinson. 1970. On the measurement of inequality. *Journal of Economic Theory* 2, 3 (1970), 244–263. https://doi.org/10.1016/0022-0531(70)90039-6

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

[6] William J Baumol. 1972. On taxation and the control of externalities. *The American Economic Review* 62, 3 (1972), 307–322.

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

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

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

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

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

[12] Krishnendu Chatterjee, Rupak Majumdar, and Marcin Jurdziński. 2004. On Nash equilibria in stochastic games. In *International workshop on computer science logic*. Springer, 26–40.

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

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

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

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

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

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

[19] Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. 2019. A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. *arXiv preprint arXiv:1904.06979* (2019).

[20] Thomas D Crocker. 1966. The structuring of atmospheric pollution control systems. *The economics of air pollution* 61 (1966), 81–84.

[21] JH Dales. 1968. Pollution, property & prices; an essay in policy-making and economics. (1968).

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

[23] Panayiotis Danassis. 2022. *Scalable Multi-agent Coordination and Resource Sharing*. Ph.D. Dissertation. École Polytechnique Fédérale de Lausanne (EPFL), Lausanne. https://infoscience.epfl.ch/record/291383?ln=en https://infoscience.epfl.ch/record/291383?ln=en.

[24] Panayiotis Danassis, Zeki Doruk Erden, and Boi Faltings. 2021. Improved Cooperation by Exploiting a Common Signal. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 395–403.

[25] Panayiotis Danassis, Zeki Doruk Erden, and Boi Faltings. 2022. Exploiting Environmental Signals to Enable Policy Correlation in Large-scale Decentralized Systems. *Autonomous Agents and Multi-Agent Systems* (2022). https://doi.org/10.1007/s10458-021-09541-7 https://doi.org/10.1007/s10458-021-09541-7.

[26] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. 2022. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991* (2022).

[27] Xiaotie Deng, Yuhao Li, David Henry Mguni, Jun Wang, and Yaodong Yang. 2021. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *arXiv preprint arXiv:2109.01795* (2021).

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

[29] Paul Duetting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. 2019. Optimal Auctions through Deep Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 1706–1715. http://proceedings.mlr.press/v97/duetting19a.html

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

[31] EU. 2022. EU Emissions Trading System. https://ec.europa.eu/clima/eu-action/eu-emissions-trading-system-eu-ets_en [Online; accessed 7-February-2022].

[32] EU. 2022. EU Green Taxation. https://ec.europa.eu/taxation_customs/green-taxation-0_en [Online; accessed 7-February-2022].

[33] Irving Fisher. 1892. *Mathematical Investigations in the Theory of Value and Prices*. Ph.D. Dissertation. Yale University.

[34] Peter H Friesen. 1979. The Arrow-Debreu model extended to financial markets. *Econometrica: Journal of the Econometric Society* (1979), 689–707.

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

[36] Denizalp Goktas, Jiayi Zhao, and Amy Greenwald. 2022. Zero-Sum Stochastic Stackelberg Games. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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

[38] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep Reinforcement Learning That Matters. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669/16677

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

[40] CJ Himmelberg, Thiruvenkatachari Parthasarathy, TES Raghavan, and FS Van Vleck. 1976. Existence of p-equilibrium and optimal stationary strategies in stochastic games. *Proc. Amer. Math. Soc.* 60, 1 (1976), 245–251.

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

[42] Raj Jain, Dah-Ming Chiu, and W. Hawe. 1998. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems. *CoRR* cs.NI/9809099 (1998). http://arxiv.org/abs/cs.NI/9809099

[43] Michael Bradley Johanson, Edward Hughes, Finbarr Timbers, and Joel Z. Leibo. 2022. Emergent Bartering Behaviour in Multi-Agent Reinforcement Learning. https://doi.org/10.48550/ARXIV.2205.06760

[44] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

[45] Jean-Jacques Laffont. 2008. Externalities. *New Palgrave Dictionary of Economics* (2008).

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

[47] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (New Brunswick, NJ, USA) *(ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163.

[48] Michael L. Littman and Peter Stone. 2002. *Implicit Negotiation in Repeated Games*. Springer Berlin Heidelberg, Berlin, Heidelberg, 393–404. https://doi.org/10.1007/3-540-45448-9_29

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

[50] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

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

[52] Denis Nekipelov, Vasilis Syrgkanis, and Eva Tardos. 2015. Econometrics for learning agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 1–18.

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

[54] David C. Parkes and Michael P. Wellman. 2015. Economic reasoning and artificial intelligence. *Science* 349, 6245 (2015), 267–272. https://doi.org/10.1126/science.

aaa8403 arXiv:https://science.sciencemag.org/content/349/6245/267.full.pdf
[55] Michael Pickhardt. 2006. Fifty Years after Samuelson's "The Pure Theory of Public Expenditure": What are we Left With? *Journal of the History of Economic Thought* 28, 4 (2006), 439–460.
[56] Arthur Cecil Pigou. 1924. *The Economics of Welfare.* Macmillan.
[57] Alan Randall. 1983. The problem of market failure. *Natural Resources Journal* 23, 1 (1983), 131–148.
[58] Ariel Rubinstein and Carl-johann Dalgaard. 1998. *Modeling bounded rationality.* MIT press.
[59] Paul A Samuelson. 1954. The pure theory of public expenditure. *The review of economics and statistics* (1954), 387–389.
[60] Paul A Samuelson. 1995. Diagrammatic exposition of a theory of public expenditure. In *Essential readings in economics.* Springer, 159–171.
[61] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347 http://arxiv.org/abs/1707.06347
[62] L. S. Shapley. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences* 39, 10 (1953), 1095–1100. https://doi.org/10.1073/pnas.39.10.1095
[63] Weiran Shen, Pingzhong Tang, and Song Zuo. 2019. Automated mechanism design via neural networks. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems.* 215–223.
[64] Adam Smith. 1791. *An Inquiry into the Nature and Causes of the Wealth of Nations.* Vol. 1. Librito Mondi.
[65] Matthew J Sobel. 1971. Noncooperative stochastic games. *The Annals of Mathematical Statistics* 42, 6 (1971), 1930–1935.
[66] Eilon Solan and Nicolas Vieille. 2002. Correlated equilibrium in stochastic games. *Games and Economic Behavior* 38, 2 (2002), 362–399.
[67] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.
[68] Richard James Sweeney, Robert D Tollison, and Thomas D Willett. 1974. Market failure, the common-pool problem, and ocean resource exploitation. *The Journal of Law and Economics* 17, 1 (1974), 179–192.
[69] Éva Tardos. 2019. Learning and Efficiency of Outcomes in Games, Seminar Slides. (2019).
[70] Nelson Vadori, Leo Ardon, Sumitra Ganesh, Thomas Spooner, Selim Amrouni, Jared Vann, Mengda Xu, Zeyu Zheng, Tucker Balch, and Manuela Veloso. 2022. Towards Multi-Agent Reinforcement Learning driven Over-The-Counter Market Simulations. *arXiv preprint arXiv:2210.07184* (2022).
[71] Leon Walras. 1874. *Éléments d'Économie Politique Pure.*
[72] Michel Wedel and Wagner A Kamakura. 2000. *Market segmentation: Conceptual and methodological foundations.* Springer Science & Business Media.
[73] Wikipedia. 2022. Clean Air Act (United States) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Clean_Air_Act_(United_States)&oldid=1069225593 [Online; accessed 7-February-2022].
[74] WPR. 2018. USDA Plans To Buy $50M In Milk To Reduce Surplus. https://www.wpr.org/usda-plans-buy-50m-milk-reduce-surplus [Online; accessed 7-February-2022].
[75] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to Incentivize Other Learning Agents. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 15208–15219. https://proceedings.neurips.cc/paper/2020/file/ad7ed5d47b9baceb12045a929e7e2f66-Paper.pdf
[76] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 2022. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances* 8, 18 (2022), eabk2607. https://doi.org/10.1126/sciadv.abk2607 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abk2607