HEC MONTRÉAL

Fairness Incentives in Response to Unfair Dynamic Pricing

par

Jesse Thibodeau

Golnoosh Farnadi
HEC Montréal
Directrice de recherche

Sciences de la gestion
(Spécialisation Data Science and Business Analytics)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

April 2024

Résumé

L'utilisation de la tarification dynamique par les entreprises cherchant à maximiser leurs profits soulève des préoccupations concernant l'équité de la demande, mesurée par les disparités dans les réponses à la demande des différents groupes de consommateurs face à une stratégie de tarification donnée. Notamment, la tarification dynamique peut entraîner des distributions d'acheteurs qui ne reflètent pas celles de la population sous-jacente, ce qui peut poser problème dans les marchés où une représentation équitable est socialement souhaitable. Pour y remédier, les décideurs politiques pourraient utiliser des outils tels que la taxation et les subventions pour adapter les mécanismes politiques en fonction de leur objectif social. Dans cet article, nous explorons le potentiel des méthodes d'IA pour assister de telles stratégies d'intervention. À cette fin, nous concevons une économie simulée de base, dans laquelle nous introduisons un planificateur social dynamique (SP) pour générer des calendriers de taxation des entreprises visant à inciter les entreprises à adopter des comportements de tarification équitables, et à utiliser le budget fiscal collecté pour subventionner la consommation parmi les groupes sous-représentés. Pour couvrir une gamme de scénarios politiques possibles, nous formulons le problème d'apprentissage de notre planificateur social comme un bandit manchot multi-bras, un bandit manchot contextuel et enfin comme un problème complet d'apprentissage par renforcement (RL), évaluant les résultats en termes de bien-être dans chaque cas. Pour atténuer la difficulté de retenir des taux d'imposition significatifs qui s'appliquent à des tranches moins fréquemment rencontrées, nous introduisons `FairReplayBuffer`, qui assure que notre agent RL échantillonne les expériences de manière uniforme à travers un espace de justesse discrétisé. Nous constatons que, lors de la mise en œuvre d'une politique apprise de taxation et de redistribution, le bien-être social s'améliore par rapport à la base de référence indifférente à l'équité, et se rapproche de celle de la base de référence optimale analytiquement consciente de l'équité pour les paramètres de bandit manchot multi-bras et contextuel, et la dépasse de 13.19% dans le cadre du RL complet.

Mots-clés

Bien-être social, Optimization, Apprentissage par renforcement, Équité, Conception d'incitations

Méthodes de recherche

Apprentissage par renforcement, Simulation, Modélisation des choix discrets

Abstract

The use of dynamic pricing by profit-maximizing firms gives rise to demand fairness concerns, measured by discrepancies in consumer groups' demand responses to a given pricing strategy. Notably, dynamic pricing may result in buyer distributions unreflective of those of the underlying population, which can be problematic in markets where fair representation is socially desirable. To address this, policy makers might leverage tools such as taxation and subsidy to adapt policy mechanisms dependent upon their social objective. In this paper, we explore the potential for AI methods to assist such intervention strategies. To this end, we design a basic simulated economy, wherein we introduce a dynamic social planner (SP) to generate corporate taxation schedules geared to incentivizing firms towards adopting fair pricing behaviours, and to use the collected tax budget to subsidize consumption among underrepresented groups. To cover a range of possible policy scenarios, we formulate our social planner's learning problem as a multi-armed bandit, a contextual bandit and finally as a full reinforcement learning (RL) problem, evaluating welfare outcomes from each case. To alleviate the difficulty in retaining meaningful tax rates that apply to less frequently occurring brackets, we introduce `FairReplayBuffer`, which ensures that our RL agent samples experiences uniformly across a discretized fairness space. We find that, upon deploying a learned tax and redistribution policy, social welfare improves on that of the fairness-agnostic baseline, and approaches that of the analytically optimal fairness-aware baseline for the multi-armed and contextual bandit settings, and surpassing it by 13.19% in the full RL setting.

Keywords

Social welfare, Optimization, Reinforcement learning, Incentive design, Fairness

Research Methods

Reinforcement Learning, Simulation, Discrete Choice Modelling

Contents

List of Tables

List of Figures

List of Acronyms

AI	Artificial Intelligence
ML	Machine Learning
SAC	Soft Actor-Critic
MAB	Multi-armed Bandit
CB	Contextual Bandit
RL	Reinforcement Learning
SWF	Social Welfare
MDP	Markov Decision Process
NLP	Nonlinear Programming
FIFO	First-In-First-Out

Preface

As fundamentally economic beings, we weigh the impacts of innumerable decisions made by ourselves and others on a daily basis. We quantify these impacts through the somewhat abstract notion of *utility*, which we compute, often subconsciously, for any financial or social transaction we make. *Economics* is the study of human behaviour by way of these expressions of utility. Meanwhile, *machine learning* (ML) is the study of algorithmic solution methods for quantifying reality from data. Given the emergence of widespread computational systems in markets, a natural progression for economists is to validate their intuitions with machine learning, thus paving the way for the adoption of data-driven market strategies. While there are clear benefits to exploiting data to improve market efficiencies, realities are seldom perfectly conveyed in data. In this vein, deploying artificial intelligence (AI) systems at large entails the risk that consumption behaviours become governed by market models, instead of the other way around. Furthermore, allowing markets to be governed by AI makes any of its adverse effects particularly difficult to break away from. To mitigate this, can we encode policy systems with human values which nudge the behaviour of problematic market models towards adopting more socially-desirable behaviours? The research I present here aims to answer this question. While my work is a step in the direction of fair policy mechanism design, it is my hope that the rapidly growing research community at the intersection of AI and economics may further address the concerns I touch on.

Acknowledgements

This work is dedicated to my friends, who inspire me every day, and my family, who enabled and encouraged me to take risks.

A sincere thank-you to my supervisor Golnoosh, for whom I am endlessly grateful, and whose open-mindedness allowed this economics undergrad to gain a footing in the AI scene at Mila.

Additional thanks to my labmates and collaborators, without whom I would have many problems and very few solutions.

Finally, I am appreciative of Nobuhiro Kiyotaki for the valuable insight he provided during my undergraduate work in economics, as well as during the early ideation stages of this research.

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google award, MEI award, and the Natural Sciences and Engineering Research Council of Canada (NSERC). My collaborators and I also express gratitude to Compute Canada for their support in providing facilities for our evaluations.

Chapter 1

General Introduction

The landscape formed at the intersection of public policy and computational innovation gives way to complex ethical considerations, where the tools and methodologies of today–notably artificial intelligence (AI)–might play a pivotal role in shaping societal outcomes. In this vein, efficient policy mechanisms become crucial for steering economies towards sustainable growth, addressing market failures as well as ensuring equitable distribution of wealth and resources. To this end, Agrawal, Gans, and Goldfarb (2018) emphasize how AI can be leveraged to enhance economic decision-making processes. When designed and implemented effectively, these mechanisms enable governments to guide economic activity and impact social welfare such that outcomes align with broader objectives of efficiency, productivity and equity. However, the advent of sophisticated AI systems introduces new dimensions to the traditional challenges of policy design. For instance, dynamic pricing, defined in this research as a method employed by profit-maximizing firms to assign prices to goods based on estimates of consumer willingness-to-pay, is an example of a technique enabled by new technology which poses significant challenges to policymakers who aim to design mechanisms promoting desirable social outcomes (Ezrachi 2016). By leveraging vast datasets and machine learning algorithms, businesses can optimize pricing strategies in real-time, responding to shifts in demand, competition, and market conditions with improved precision (Varian 2014). While such practices offer potential benefits in terms of economic efficiency and productivity, as well as responsiveness to consumer needs, they also raise a number of ethical questions with respect to various notions of fairness, including the primary notion explored in this research: *demand fairness*. A price assignment satisfies demand fairness if the resultant distribution of buyers is reflective of the underlying population distribution. A popular profit-maximization strategy, dynamic pricing entails a risk that resultant distributions of buyers do not reflect the diversity of the underlying population, highlighting the need for a nuanced approach to policy design–one that balances economic productivity with the social objective of ensuring fair and equitable access to goods and services, where satisfying these principles is desirable (O'neil 2017). Addressing these challenges necessitates novel discussion and research aimed at understanding and mitigating the potential adverse effects of dynamic pricing. By employing reinforcement learning (RL) techniques at the policy level in response to dynamic pricing at the firm level, this thesis explores the development and deployment of dynamic policy mechanisms that promote *a)* fairer firm pricing behaviours and *b)* more equitable market participation. To design such incentive frameworks, we use the policy tools of taxation and subsidy to align business practices with a social planner's (SP) welfare objectives, defined in terms of the tradeoff between economic productivity and demand fairness (Kahneman and Tversky 1979).

1.1 Economic Importance of Efficient Policy Mechanisms

Efficient policy frameworks are critical in addressing the complexities of modern economic landscapes, particularly as we become faced with ethical and operational challenges exacerbated by profit-maximizing business practices. These frameworks are instrumental in driving economies toward sustainable growth, rectifying undesirable market outcomes, and fostering equity. Designing and implementing such policies such that they are adaptable allows governments to make more impactful decisions to improve social welfare in ways that align with broader goals of productivity and fairness. As we explore the notion of integrating sophisticated AI methods to traditional policy formulation, let us first broadly consider dynamic pricing strategies, which allow businesses to adjust prices based on estimates of consumer willingness-to-pay. While such strategies are not inherently problematic, that is, they should in theory improve economic efficiency, they occasionaly pose ethical dilemmas regarding fairness and accessibility. A natural response is to develop a framework for generating policies which are equally dynamic and adaptable, and which nudge unfair social outcomes towards improvement. In fiscal policy, that is, policy relating to taxation, one central challenge is to deploy incentive frameworks which aid in achieving social welfare objectives all while ensuring that firms remain motivated to produce at satisfactory capacities (Piketty 2014). The crux of the problem of generating incentive frameworks is therefore in balancing improvements in fairness outcomes and other welfare metrics with retaining as much of the economic productivity seen under free market dynamics as possible. For a profit-maximizing firm, any deviation from an incumbent pricing strategy towards a fairer one would inherently be suboptimal, as profit-maximizing firms are not assumed to consider fairness in their business objectives. Therefore, there exists some tradeoff between the productivity and fairness components of welfare, and a good policy mechanism should be able to navigate this by nudging market participants, firms and consumers alike, towards more desirable social behaviours.

1.2 AI in Policy Design

The utilization of AI in policy design already shows promising results across various sectors, highlighting both the potential benefits and challenges of the approach. For instance, in the realm of environmental policy, AI has been employed to optimize energy consumption patterns, reduce emissions, and forecast the environmental impact of different policy scenarios. This is seen in initiatives like smart grid technologies (Farhangi 2009), which adjust electricity distribution in real time to match demand with renewable energy supply, thereby enhancing efficiency and reducing reliance on fossil fuels. In healthcare policy, AI algorithms are used to predict epidemic outbreaks, allocate medical resources more efficiently, and personalize care to improve long-term health outcomes (Bolhasani, Mohseni, and Rahmani 2021). For example, predictive analytics were used during the COVID-19 pandemic to forecast hospitalization rates, helping to optimize the distribution of patients, personnel and medical supplies in anticipation of rising cases (Ting et al. 2020). There is therefore a strong precedent in recent history for using AI systems in areas of high social impact. However, the adoption of AI in policy design is not without its challenges. A key concern is the explainability and interpretability of AI models. State-of-the-art methods typically involve the use of neural networks, which are "black-box" models, that is, models with highly opaque learning procedures. Given this, policies based solely on learning algorithms can be difficult to justify to the public, potentially leading to trust issues (Castelvecchi 2016). Moreover, the data used to train AI models may reflect existing biases, leading to policies that inadvertently perpetuate inequality (Barocas and Selbst 2016). Ensuring data quality, representativeness, and addressing algorithmic bias are critical challenges that must be overcome to fully realize the benefits of AI in policy design.

1.3 The Social Complexities of Dynamic Pricing

Dynamic pricing strategies, enabled by AI and big data analytics, present a compelling case study in the balance between efficiency and equity. Ride-sharing services, such as Uber and Lyft, use dynamic pricing models to adjust fares in real-time based on demand and supply conditions. While this can optimize resource allocation and potentially increase drivers' earnings during peak times, it also raises concerns about affordability and access. During emergencies or high-demand periods in certain geographies, prices can surge to levels that exclude lower-income individuals from accessing these services, highlighting the social equity challenges inherent in dynamic pricing (Chen and Sheldon 2016). Another example can be found in the airline industry, where dynamic pricing algorithms adjust ticket prices in real time based on factors like booking patterns and seat availability. This can lead to significant price disparities for passengers booking the same flight, raising questions about fairness and transparency in pricing practices (Alderighi, Gaggero, and Piga 2016). There exist equally The challenge lies in designing regulatory frameworks and industry standards that can harness the benefits of dynamic pricing while mitigating its potential to exacerbate social inequalities. The approach we explore involves ensuring that pricing algorithms incorporate considerations for equity and access (Kahneman, Knetsch, and Thaler 1986). Engaging with these challenges requires a nuanced understanding of both how dynamic pricing may have harmful distributional affects on buyers, as well as the technical mechanisms guiding the generation of effective policy mechanisms designed to mitigate these.

1.4 Contributions

Our contributions to the body of research revolving around fairness in mechanism design are as follows:

- Presenting a new framework that includes three distinct policy tools aimed at promoting demand fairness in markets with a single product [1];
- Introducing various fairness incentive mechanisms through diverse economic policy models, structured using multi-armed and contextual bandits, as well as a complete reinforcement learning (RL) framework;
- Introducing FairReplayBuffer, a replay buffer tailored for the soft actor-critic (SAC) algorithm, specifically designed to facilitate learning in fairness taxation scenarios;
- Proposing a dual strategy of subsidy and taxation to effectively tackle demand fairness issues and improve social welfare;
- Conducting a detailed evaluation of our proposed framework through a simulation study involving firms, each addressing two distinct consumer behaviors [2].

[1] arXiv paper: `https://export.arxiv.org/abs/2404.14620`

[2] For code, please see my public GitHub repository: `https://github.com/j-thib/rl-fair-pricing/`

1.5 Thesis Structure

Over the remainder of this thesis, we conduct a review of topics in economics and AI relevant to our work (chapter 2). We follow with our complete research article (chapter 3). Finally, we present concluding remarks as well as a discussion on directions for future work (chapter 4).

Chapter 2

Literature Review

The work we present is interdisciplinary, drawing from the fields of economics and computer science to provide insight on how to adapt policy decision-making to a rapidly evolving economic landscape. The following sections aim to provide an overview of topics relevant to this research.

2.1 Economic Background

2.1.1 Welfare Theory

Rooted in the economic and social sciences, welfare theory encapsulates the study of principles aimed at enhancing the well-being of individuals and communities. At its core, it concerns itself with effective resource allocation and, more practically, the mechanisms through which policymakers can achieve optimal levels of welfare for their constituents. Given that welfare can itself be abstractly defined, welfare theory encompasses a range of perspectives, from the utilitarian approach, which seeks the greatest utility for the greatest number, to more equity-focused theories that prioritize the needs of the more vulnerable or underrepresented populations.

A pivotal notion in welfare theory is the concept of social welfare functions, introduced by economists such as Burk (1938) and Samuelson (1956), which define mathematical formalisms for expressing social welfare in terms of individual utilities. These functions aim to represent societal preferences in a way that balances notions of efficiency with notions equity, highlighting the trade-offs inherent in welfare economics. The maximization goal in our research is centered around such tradeoffs.

Welfare theory critically engages with the role of government and public policy in welfare provision. The debate between market-based versus state-led approaches to welfare is a contentious one, with authors like Hayek (1944) advocating for minimal government intervention and others like Rawls (1971) arguing for a more active role for the state in ensuring distributive justice. These works provide the foundation for reasoning as it pertains to degrees of government intervention. Recent developments in welfare theory have been influenced by behavioral economics and the recognition of human irrationality in decision-making. Authors like Kahneman and Tversky (1979) challenge traditional economic models of human behavior, suggesting that welfare assessments must consider psychological well-being alongside material conditions. For instance, fairness perceptions (Malc, Mumel, and Pisnik 2016) can be included as a psychological component of welfare metrics. These perceptions can lead consumers to behave stochastically from a modelling perspective. In conclusion, welfare theory remains a field of active research, continually evolving to address complex questions of how policy makers can promote the well-being of their constituents. Through its interdisciplinary approach, it provides critical insights into the mechanisms of welfare provision and the ethical foundations of social policy.

2.1.2 Single-product Markets

The study of single-product markets occupies a central place in both theoretical and applied economics, particularly in the context of pricing strategies and market simulations. Notable contributions in this area often leverage a blend of economic theory, computational models, and empirical analysis to understand and predict consumer behavior, firm strategies, and market outcomes.

A seminal work by Cournot (1838) laid the foundational framework for understanding oligopolistic competition, including single-product markets, focusing on quantity adjustment rather than pricing. Later, Bertrand (1883) introduced a model that shifted the focus to price competition, foundational for later research in single-product pricing strategies. While these formulations exist within the context of duopoly and oligopoly, we employ notions of profit-maximization in an economy consisting of multiple parallel markets without interoperability, as written by Robinson (1969).

More recently, Tirole (1988) synthesized various models of market behavior and firm strategy, offering insights into pricing strategies in monopolistic and oligopolistic markets. This work emphasizes the importance of understanding market structure and strategic interaction between firms in setting prices.

In the realm of simulations, Agent-Based Models (ABMs) have become increasingly popular for exploring market dynamics in single-product markets. Tesfatsion (2006) provides a comprehensive overview of agent-based computational economics (ACE), illustrating how ABMs can simulate complex interactions between autonomous agents, including consumers and firms, to investigate market phenomena such as price dispersion and market power. We borrow from this body as a justification for our use of RL methods.

Recent advancements in machine learning and data analytics have opened new avenues for research in pricing strategies and market simulations. Works by Davenport et al. (2006) explore how big data and predictive analytics can inform dynamic pricing strategies in single-product markets, emphasizing the importance of data-driven decision-making in competitive environments.

2.1.3 Discrete Choice Modelling

Discrete choice modeling stands as a pivotal method in understanding decision-making processes where individuals choose from a set of distinct alternatives. Originating from the seminal work of McFadden (1972), which earned him the Nobel Prize, discrete choice models have extensively been applied across fields such as transportation, marketing, and environmental economics to analyze consumer preferences and predict choice behavior. These models rest on the random utility maximization (RUM) theory, positing that the choice made by an individual among a finite set of alternatives is influenced by the utility that the individual derives from each option. We borrow this notion to formulate the consumer choice as one of whether to participate in a market.

Over the years, the scope of discrete choice modeling has broadened, incorporating more complex structures and phenomena such as heterogeneity in preferences, the role of social influence, and the impact of information on decision-making. Work by Train (2009) has been instrumental in elucidating these advanced models, offering a comprehensive guide to estimations and applications of mixed logit models that allow for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. While these notions are not all modelled explicitly in this work, we nonetheless include some preference heterogeneity, as well as temporal noise which can serve as a proxy for unobserved correlations.

2.1.4 Public Finance and Taxation Theory

The study of public finance, particularly the theory of taxation and subsidy allocation, encompasses a vital area of economic inquiry, exploring how governments raise revenue and allocate resources to achieve social and economic objectives. Central to this field is the theory of taxation, which delves into the mechanisms, principles, and effects of tax policies on economic behavior and societal welfare.

Pioneering contributions by Smith (1776) introduced the principles of taxation (i.e. equity, certainty, convenience, and efficiency) that continue to prove foundational in contemporary tax policy discussions. Later, the work of Pigou (1920) on the concept of externalities laid the groundwork for Pigouvian taxes, highlighting how taxation could correct market failures and optimize social welfare by internalizing external costs. Further, the seminal model by Diamond and Mirrlees (1971) advanced the theory by rigorously analyzing optimal taxation and public goods provision, considering constraints like efficiency and equity. This model underscores the complex trade-offs faced by policymakers in designing tax systems that minimize efficiency losses while achieving redistribution goals.

In addition, more recent literature has increasingly focused on dynamic aspects of taxation, including the work of Saez (2001), who explores the elasticity of taxable income and its implications for optimal income tax rates. This body of research emphasizes the importance of understanding behavioral responses to taxation and the role of administrative efficiency in tax policy. Work by Zheng et al. (2020) titled *The AI Economist* uses the Saez tax formula as a theoretical baseline upon which they improve with the use or RL methods. The allocation of subsidies as a complementary tool for addressing market failures and achieving redistribution has been scrutinized. The interplay between taxation and subsidy policies is critical in forming a coherent public finance strategy to promote economic efficiency and equity. While *The AI Economist* uses a basic uniform trax redistribution scheme, we propose a learned scheme by which taxes are redistributed according to consumer group membership.

2.2 AI Background

We provide a broad overview of the AI concepts and methods used in this research, with a particular focus on reinforcement learning. We begin with a broad description of reinforcement learning, and follow with a discussion on multi-armed bandits and contextual bandits. Following, we describe a class of solution methods called policy gradient methods, and elaborate on soft actor-critic (SAC), the specific algorithm used in our work.

2.2.1 Reinforcement Learning: A Broad Overview

There are three main machine learning paradigms, each of which geared towards solving a particular class of problems. For problems involving making predictions from labeled data, one would typically use supervised learning methods. For tasks requiring one to extract meaningful properties from unstructured data, unsupervised learning methods be-

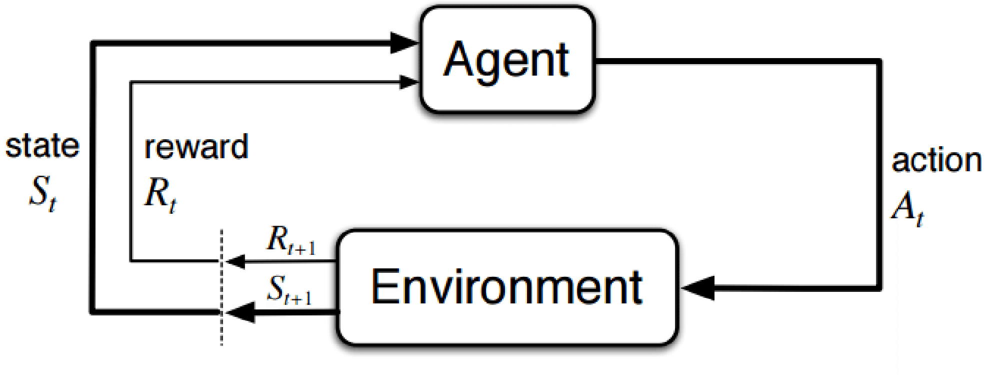


Figure 2.1: The agent-environment interaction in a Markov decision process.

come appropriate. Finally, for tasks involving decision-making in sequential processes, reinforcement learning is used (Sutton and Barto 2018). At its core, RL exists as an approach for learning decision-making and control from *experience*. Specifically, a successful RL agent is one that has learned through trial and error (environment interactions) to take actions which maximize some notion of longterm reward. The mathematical expression governing the actions taken by an RL agent in a given state is called a *policy*. The goal of an RL agent is to learn an optimal policy through trial and error. Reinforcement learning problems can be mathematically formalized as Markov decision processes, which are a classic formalization of sequential decision making, where actions influence not just immediate rewards, but also subsequent situations, or states, and through those future rewards (Sutton and Barto 2018). The sequential element of the problem is therefore apparent in the occasional need for a learning agent to understand the tradeoff between delayed and immediate rewards. MDP's are defined in terms of a state $\mathscr{S}$, an action $\mathscr{A}$, a reward signal $\mathscr{R}$, a transition dynamics function p, and a discount factor γ. In many cases, one might also define a starting state $\mathscr{S}_0$. Figure 2.1, demonstrates how an agent in state S_t may take an action A_t, which will bring it to state S_{t+1} where it collects reward R_{t+1}. The state satisfies the Markov property, which is a general assumption for environment dynamics in RL. This means that the future state depends only on the current state. Thus, the state transition probabilities $p : \mathscr{S} \times \mathscr{S} \times \mathscr{A} \rightarrow [0,1]$ are a function of three arguments,

$$p(s^{'} \mid s,a) \doteq \Pr\{S_t = s^{'} \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathscr{R}} p(s^{'}, r \mid s,a).$$

Having understood the mathematical formalisms allowing one to define and solve a sequential decision making problem, let us examine a simple first class of solution methods: Bandit algorithms.

2.2.2 Bandits

We discuss a class of *action-value* solution methods called bandit algorithms. We define and elaborate on multi-armed bandits, as well as their extension to associative search tasks, also known as contextual bandits.

Multi-armed Bandits

A multi-armed bandit (MAB) problem is simpler than a full reinforcement learning problem, as the optimal agent's action is nonassociative, that is, it doesn't depend on any observations of the current state. Let us consider a setting wherein a learning agent must choose the best action out of *k* options. In a *k*-armed bandit problem, the learner might algorithmically try every available action, while recording the reward collected from having taken each. After sampling actions a sufficient number of times, the agent will compute an increasingly accurate estimate of the "value" of a given action. The rule by which these action-value estimates may be updated is denoted

$$NewEstimate \leftarrow OldEstimate + StepSize[Target - OldEstimate],$$

where $[Target - OldEstimate]$ is an error in the estimate, which we aim to reduce by approaching *Target* by an amount determined by $StepSize$. The learning agent may then *exploit* these value estimates to inform its decision making process. For instance, if the action-values $Q_t(a)$ are known, then a perfectly reasonable policy would be to always select an action A_t such that

$$A_t \doteq \arg\max_a Q_t(a).$$

This policy is a *greedy* action selection method, as it exploits the information it has gathered on action values to consistently choose the action with the highest average reward. However, this gives way to an important consideration: that perhaps there are untried actions that may yield a higher reward, but would rarely be visited due to the greedy policy. For this reason, the importance of the notion of *exploration* becomes apparent. As a learning agent aims to maximize its reward by exploiting the knowledge it has gathered through sampled actions, it should also be allowed to experiment with new actions which may lead the agent to adopt a new policy involving a previously untried action. The tradeoff between exploration and exploitation is a subject of research on its own, and its importance is equal across all solution methods. A simple bandit algorithm may be implemented as follows:

Algorithme 1 : A simple bandit algorithm

```
Input : $\varepsilon$ (probability of exploration)
Initialize : For $a = 1$ to $k$:
    $Q(a) \leftarrow 0$
    $N(a) \leftarrow 0$
Loop forever
    $A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \text{ (breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
    $R \leftarrow \text{bandit}(A)$
    $N(A) \leftarrow N(A) + 1$
    $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}[R - Q(A)]$
```

This ε-greedy algorithm balances exploration and exploitation by selecting a random action with probability $\varepsilon \in [0,1]$. When an action is selected, it increments $N(A)$, which represents the number of times an action A was selected. While this algorithm isn't directly employed in our research, it provides a valuable intuition on how bandits can be used in discrete choice settings.

Contextual Bandits

We can extend the MAB setting described above by adding an associative search component, also known as "context", to the agent's learning problem. Contextual bandits differ from a full RL problem in that an action taken in the current state only affects the immediate reward. However, they are an extension to the *k*-armed bandit problem because they learn a policy that tells the agent which action to take. To conceptualize this further, let us build off the *k*-armed problem faced above. In this new instance, the estimated action values will be conditioned on the current state, so that our action will be a function of some *context vector* $\mathscr{X}$ which contains the observations made by our agent. This is essentially the same as solving multiple state-dependent bandit problems. Thus, contextual bandits are the intermediate step between a MAB and a full RL problem.

2.2.3 Policy Gradient Methods

Thus far, we have discussed how an agent might learn the value of an action in a given state by iteratively updating reward estimates. These are known as *action-value* methods. Policy gradient methods, on the other hand, are a class of algorithms in reinforcement learning that optimize policy directly. These methods parameterize the policy as a probability distribution over actions, dependent on the state and policy parameters. For discrete actions, the policy might use a softmax function, and for continuous actions, a Gaussian distribution. Policy gradient methods compute the gradient of the expected return using the policy gradient theorem, which is a theoretical advantage of policy gradient methods allowing for stronger convergence guarantees. Notably, the continuity of the policy dependence on its parameters allows policy gradient methods to approximate gradient ascent, which is done by sampling trajectories under the current policy and adjusting parameters in the direction indicated by these samples. There are multiple advantages of policy gradient methods over action-value methods. These include improved convergence speed and stability, as well as efficient handling of continuous action spaces.

2.2.4 Soft Actor-Critic

An example of a policy gradient method that is particularly well-suited for exploration of continuous action spaces is soft actor-critic (SAC), which we employ in our work. SAC is an algorithm distinguished for its use of policy entropy to automatically balance exploration with exploitation. In their work, Haarnoja et al. (2018) consider a general maximum-entropy objective which favors stochastic policies by adding an entropy term in the reward function, denoting the expected entropy over $\rho_\pi(\mathbf{s}_t)$, which represents the state marginals of a trajectory distribution induced by a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$. Their objective function is thus formulated:

$$J(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{a}_t,\mathbf{s}_t)\sim\rho_\pi}[r(\mathbf{s}_t,\mathbf{a}_t) + \alpha\mathscr{H}(\pi(\cdot|\mathbf{s}_t))],$$

where the temperature parameter α controls the importance of entropy in the reward and thus determines the policy stochasticity.

Chapter 3

Fairness Incentives in Response to Unfair Dynamic Pricing

Abstract

The use of dynamic pricing by profit-maximizing firms gives rise to demand fairness concerns, measured by discrepancies in consumer groups' demand responses to a given pricing strategy. Notably, dynamic pricing may result in buyer distributions unreflective of those of the underlying population, which can be problematic in markets where fair representation is socially desirable. To address this, policy makers might leverage tools such as taxation and subsidy to adapt policy mechanisms dependent upon their social objective. In this paper, we explore the potential for AI methods to assist such intervention strategies. To this end, we design a basic simulated economy, wherein we introduce a dynamic social planner (SP) to generate corporate taxation schedules geared to incentivizing firms towards adopting fair pricing behaviours, and to use the collected tax budget to subsidize consumption among underrepresented groups. To cover a range of possible policy scenarios, we formulate our social planner's learning problem as a multi-armed bandit, a contextual bandit and finally as a full reinforcement learning (RL) problem, evaluating welfare outcomes from each case. To alleviate the difficulty in retaining meaningful tax

rates that apply to less frequently occurring brackets, we introduce `FairReplayBuffer`, which ensures that our RL agent samples experiences uniformly across a discretized fairness space. We find that, upon deploying a learned tax and redistribution policy, social welfare improves on that of the fairness-agnostic baseline, and approaches that of the analytically optimal fairness-aware baseline for the multi-armed and contextual bandit settings, and surpassing it by 13.19% in the full RL setting.

3.1 Introduction

Firms equipped with modern compute power and vast consumer data logs may enjoy the economic benefit of engaging in the business practice of dynamic (or personalised) pricing. In doing so, profit-maximizing firms are able to charge potential customers, or customer segments, individualized prices based on estimates of their willingness-to-pay. From an efficiency standpoint, dynamic pricing has been shown to increase firm profitability and sales speeds (Schlosser and Boissier 2018). However, the welfare implications of the practice are somewhat unclear, and may occasionally yield socially undesirable outcomes as evidenced in the markets for insurance and housing, among others (Zhu et al. 2023; Betancourt et al. 2022). For instance, while health insurers are known to make extensive use of dynamic pricing, survey data reveal that members of the hispanic population in the U.S. are on average roughly twice as unlikely to have healthcare coverage as members of the black population, who are in turn twice as unlikely as members of the white and asian populations (Martinez 2022). In addition, substantial price differences between consumers may lead to negative perceptions of fairness (Lee, Illia, and Lawson-Body 2011), which can have their own disparate impact on market outlooks with respect to buyer participation, thereby propagating existing disparities. In housing, new constructions are often priced such that they are more appealing to privileged groups, thus contributing to growing wealth gaps and gentrification. Other sectors where buyer distributions should reflect those of the underlying population are the financial services and public transportation.

Dynamic pricing is generally employed by firms to assign prices to consumer segments such that the expected profit derived from each is maximized. Thus, a profit-maximizing firm engaging in dynamic pricing rarely bears any weight to its resulting buyer distribution, which, if not reflective of the underlying population, may be considered unfair. In this research, we address the problem of such imbalanced buyer distributions. Specifically, we examine the issue of dynamic pricing under *demand fairness* (Cohen, Elmachtoub, and Lei 2022), in markets where buyer distributions should be proportional to that of the underlying population. We begin by showing how profit-maximizing price allocations may fail to satisfy demand fairness. We then employ nonlinear programming (NLP) at the firm level to demonstrate how producers willing to consider fairness in their price allocations may do so with often minimal sacrifice to profitability. We call this case *self-regulation*. Building on this, we acknowledge that in many cases, expecting a profit-maximizer to take it upon itself to consider fairness in their price allocations is unreasonable. Thus, we consider how a benevolent social planner (SP) might leverage available policy tools, notably taxation and subsidy, in order to dynamically incentivize market participation among underrepresented consumer groups, while penalizing unfair firm behaviour. To achieve this, we train an agent using reinforcement learning (RL) methods to generate financial incentives aimed at improving demand fairness, i.e. reducing the distributional gaps in buyer populations. In addition, we allow it to learn a simple redistribution scheme whereby consumers are subsidized based on their group membership. We find that social welfare can be improved using taxation aimed at incentivizing firms to adopt fairer behaviours. With the inclusion of subsidy, we find that firm profits and fairness outcomes both increase on average relative to our positive control benchmark.

Throughout our work, we make use of RL methods, which have a growing presence in the body of literature surrounding dynamic pricing and mechanism design. Specifically, we apply soft actor-critic to a variety of mechanism design tasks, from multi-armed bandits, to contextual bandits, and finally to a full RL problem. In addition, RL has been applied to several domains such as modelling markets (Kastius and Schlosser 2021), efficient energy pricing (Lu, Hong, and X. Zhang 2018; Kim et al. 2015) and insurance pricing (Nieuwenhuis, Manstead, and Easterbrook 2019), where RL was shown to be useful in such scenarios where demand for a specific product is either unobserved or misunderstood to the point that it is difficult to uncover via traditional analytical methods. Furthermore, the rise in use cases for AI in industry indicates that RL methods will likely continue to gain popularity for applications to dynamic pricing. Finally, RL methods are scalable and therefore appropriate for the complexity of large consumer bases. Our research proactively addresses the social challenges arising from dynamic pricing, notably via the following contributions:

- Introducing a new framework featuring three distinct policy mechanisms aimed at addressing demand fairness within single-product markets;
- Implementing various fairness incentive mechanisms through a range of economic policy variants, formalized via multi-armed and contextual bandits, as well as a full RL formulation;
- Presenting `FairReplayBuffer`, a replay buffer for the soft actor-critic (SAC) algorithm, designed specifically for learning fairness taxation;
- Proposing a combined approach of subsidy and taxation to effectively address demand fairness issues and enhance social welfare;
- Conducting a comprehensive evaluation of our proposed framework through a simulation study involving firms each addressing two distinct consumer behaviours.

3.2 Related Work

Given the interdisciplinary nature of our work, we conduct a literature review consisting of topics from economics, specifically in the subfields of welfare economics and consumer choice theory, as well as a broad overview of fairness applications in sequential decision making tasks.

3.2.1 Fairness in Dynamic Pricing

The study of welfare within the context of policy design began as a field of economics and gradually found applications in the fields of operations research (Gallego, Topaloglu, et al. 2019) and computer science (Das et al. 2022). Welfare theory covers a broad variety of subtopics, which include definitions of fairness and how these interact given an incumbent policy. For instance, a recent paper by Cohen, Elmachtoub, and Lei (2022) formalizes multiple fairness definitions for dynamic pricing in terms of price, demand, consumer surplus, and no-purchase valuation, while proving analytically that no two of these are simultaneously satisfiable. Our analysis uses the demand fairness definition (Cohen, Elmachtoub, and Lei 2022; Kallus and Zhou 2021) as it most closely measures the disparate impact that dynamic pricing may have on buyer distributions, and which is well motivated by applications in education and healthcare (Kallus and Zhou 2021). In addition, Bertsimas, Farias, and Trichakis (2011) define proportional fairness, a fairness criterion ensuring that the relative welfare improvement among one population subgroup exceed the corresponsing welfare loss among another. Proportional fairness is useful in illustrating the welfare tradeoff between population subgroups under different policies. While the aforementioned works are related through their use of constrained optimization and linear programming, Maestre et al. (2019) use RL to impose fairness using Jain's index, which they treat as a measure of fairness in the price allocations between groups. While we gain inspiration for our consumer demand curves from theirs, they assume that firms will be self-regulating, while we introduce a benevolent SP to generate fairness incentives.

3.2.2 Economics Foundations

In this work, we explore dynamics within single-product markets. Consumer demand behaviours are expressed as purchase probabilities, a formulation commonly applied to evaluate how consumers might react to price fluctuations. The concept of nonlinear consumer preference was first examined in economic theory by Becker (1962) and later formalized in seminal work by McFadden (1972) and continues to prove foundational in consumer choice modelling (Models 2002). In addition, we examine welfare through the lens of fairness, as did by Fleurbaey (2008), who advocates for not only fair outcomes but also fairness in the processes that lead to such outcomes.

3.2.3 Fairness in Sequential Decision Making

Fairness in sequential decision-making is a critical concern as algorithms increasingly influence societal outcomes, in fields such as healthcare (Rajkomar et al. 2018), loan approval (Hu and L. Zhang 2022), and recommender systems (Stratigi et al. 2020). Existing studies have established foundational approaches to fairness and highlighted challenges in ensuring equitable algorithmic decisions over time. Joseph et al. (2016) introduce fairness constraints in bandit algorithms to prevent long-term disadvantages for individuals or groups. Gillen et al. (2018) tackle the implementation of fairness when fairness criteria are undefined, proposing a framework for online learning that adapts to evolving fairness metrics. Yin et al. (2024) explore the delayed impacts of fairness-aware algorithms, revealing how short-term equitable decisions can lead to unfair outcomes in the long run.

Within this body of research, the closest application to our work is the AI Economist (Zheng et al. 2020), where agents interact in a simulated economy by exchanging goods and services, while a SP aims to learn a taxation strategy that improves social welfare, defined as the product of equality and productivity. While their work is effective at showcasing emergent behaviours among consumer-workers under incumbent tax regimes, our focus lies on how an SP can impact societal outcomes by influencing firm objectives. In addition, while the AI Economist uses a fixed uniform tax redistribution policy, we formulate the subsidy as an additional component to the learning problem, as do Abebe, Kleinberg, and Weinberg (2020), who explore approaches for welfare-maximizing subsidy allocation under income shocks. However, while they use *min-sum* and *min-max* formulations, we consider average welfare across consumer groups.

3.3 Problem Formulation and Methodology

In this section, we first define a simulated economy, outlining the consumer dynamics and possible scenarios in which a social planner might interact with firms. We follow with definitions for our firm and social planner learning problems, and finish with a discussion on the solution methods employed.

3.3.1 Consumer Environment

A firm's objective in using dynamic pricing is to charge the maximal price that a prospective consumer is likely willing to pay based on demand estimates. We design a single-product market consisting of distinct consumer groups, wherein members from each group decide whether to purchase a product at a price determined by the firm. Each consumer group i has a unique purchase probability distribution, expressed as a discrete choice model by

$$\mathbb{P}_i(\text{purchase} = 1 \mid p) = [1 + e^{-(b_i + w_i \times p)}]^{-1}, \tag{3.1}$$

where p is the price assigned to the good, and parameters b_i and w_i capture different characteristics of consumer profile i regarding their sensitivity to price fluctuations. Once a price assignment has been made by the firm, we obtain each consumer profile's purchase probability from Equation 3.1. By linearity of expectation, we therefore anticipate that the number of consumers from group i that purchase a product at its given price can be computed by $E[n_i] = N_i \times \mathbb{P}_i(\text{purchase} = 1)$, where N_i is the number of consumers belonging to group i. In our experiments, we approximate these outcomes by simply passing these probabilities into a sequence of N_i Bernoulli trials, such that $n_i \sim \mathscr{B}(N_i, \mathbb{P}_i(\text{purchase} = 1))$, where n_i is a Bernoulli random variable corresponding to the resulting number of consumers from group i who purchased the good. This provides an element of stochasticity to our environment, where purchase outcomes may vary over constant price assignments. For simplicity, we consider the problem of a firm targeting two consumer groups, where one group has a higher sensibility to the price compared to the other one, which makes the consumers of this group underrepresented in the customer base of the firm. The fairness-agnostic firm might aim to maximize its profits with no regard for fairness, while its fairness-aware counterpart might weigh fairness outcomes into its maximization objective. In most real-world applications, we would not expect firms to explicitly self-regulate for fairness, as dynamic pricing is typically used to satisfy classical definitions of utility involving economic productivity (i.e., output, profit-maximization). To mitigate this, we introduce a dynamic social planner that aims to incentivize demand fairness by generating a tax schedule which determines the tax that will be levied on a firm based on their performance in terms of both profitability and fairness. In addition, we allow the social planner to redistribute the tax it collects to consumers as a subsidy to encourage higher market participation among underrepresented groups. Throughout our experiments, we consider demand fairness, and thus are concerned with the gap in purchase probabilities between consumer groups. We formalize our fairness notion as

$$\text{fairness}(p) = \; 1 - \left| \, \mathbb{P}_1(\text{purchase} = 1 \mid p) - \mathbb{P}_2(\text{purchase} = 1 \mid p) \, \right| , \qquad (3.2)$$

so that smaller gaps in demand correspond to higher fairness scores. Given this notion of fairness, the social planner's ultimate objective is to maximize social welfare, denoted by

$$\texttt{swf} = \texttt{profit} \times \texttt{fairness},$$

which illustrates the nonlinear tradeoff between firm productivity and fairness outcomes (Zheng et al. 2020; Bertsimas, Farias, and Trichakis 2012). In Figure 3.1, we illustrate the economic process of the SP agent generating a tax and subsidy mechanism based on the current welfare context. Below it, a firm sets prices such as to maximize profits under the SP's generated mechanism. Finally, the environment responds, whereby consumers decide whether to accept or reject firms' price outputs.

3.3.2 From Policy Preferences to RL Design

In the real world, governments [1] might adopt varying policy plans dependent upon their maximization objective. In this light, our exploration aims to provide an overview of

[1] We use *government*, *policy planner* and *social planner* interchangeably

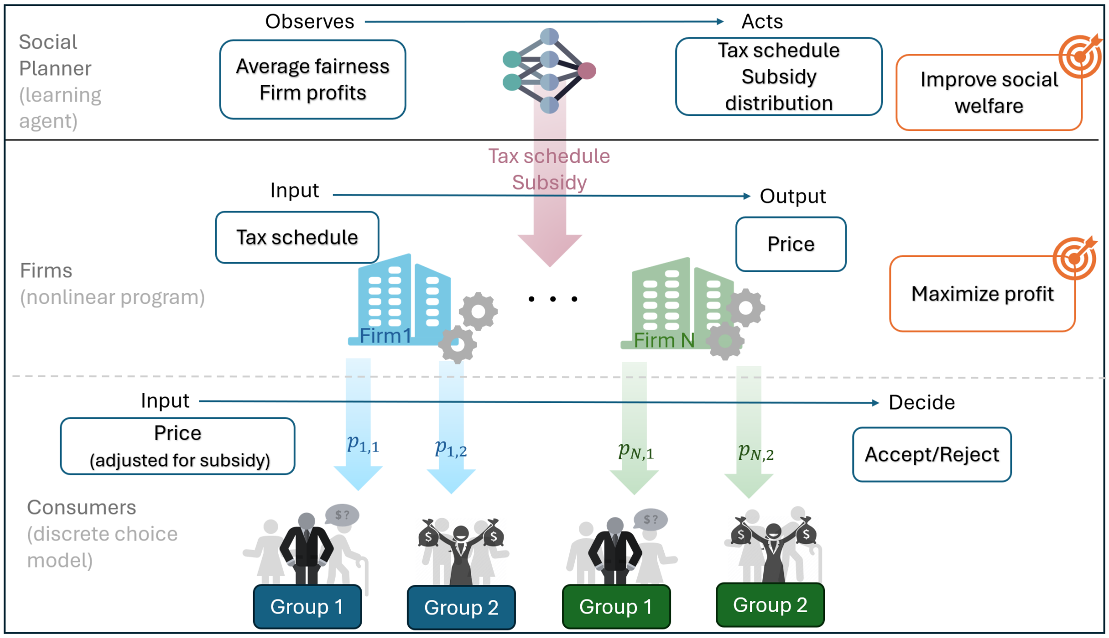


Figure 3.1: Firms are capable of efficiently learning profit-maximizing pricing assignments from consumer demand responses. The social planner (SP) learns these implicitly through firms' fairness and profit scores and designs incentive mechanisms that use taxation and subsidy, pushing firms to minimize the gap in demand responses between consumer groups.

possible policy scenarios, discussing a range of formulations that an incentive mechanism can take on while associating each to different hypothetical use cases. These include (i) a fixed policy mechanism, (ii) an adaptive policy mechanism conditioned on the current economic environment and (iii) an evolving policy framework that allows for frequent and ongoing changes to existing mechanisms. These are formulated as a multi-armed bandit problem, a contextual bandit problem and finally, a full RL problem. Below, we provide a detailed explanation of each policy mechanism.

Fixed Policy Mechanism: Multi-armed Bandit problem

If a policy planner aims to deploy an incentive mechanism to regulate a new market with uncertain or poorly understood dynamics, they might consider designing a tax and subsidy framework based on simulated data or data from economies with analogous market structures. In this scenario, solving a multi-armed bandit problem would be the most appropriate approach, with policy actions tailored to suit the average firms represented in the data.

Contextual Policy Mechanism: Contextual Multi-armed Bandit Problem

If the government was able to experiment with taxation frameworks that are conditioned on current market dynamics, it would want to create a dynamic policy mechanism. This can be formulated as a contextual bandit problem with an SP capable of generating incentives based on the observable economics context.

Evolving Policy Mechanism: RL Problem

Let us further suppose that an SP is at will to make frequent changes to policy frameworks, under the assumption that firms are equally adaptive and will always compute their best response to incumbent policies. Treating this as a full RL problem, a policy planner may adapt taxation and subsidy to maximize their notion of social welfare over time.

3.3.3 Firms

The *firm* component of our problem formulation consists of (i) a nonlinear program (NLP) for the fairness-agnostic firm, and (ii) a modified NLP for the fairness-aware firm.

Fairness-agnostic Firm

Having defined a single-product market environment (subsection 3.3.1), we proceed to formulate the fairness-agnostic firm's problem as the nonlinear program:

$$\begin{aligned} \max_{p} \quad & \sum_{i=1}^{G} n_i p \\ \text{s.t.} \quad & 0 < p \leq p_{max} \end{aligned} \tag{3.3}$$

where G is the number of consumer groups in the market, and n_i is the number of consumers from group i who purchase and is itself a function of firm price p, as described in Equation 3.1. p_{max} may be a value set by firms or governments representing some hypothetical maximum price that can be assigned to the product at hand. Thus, this is a straightforward revenue-maximization problem faced by the firm.

Fairness-aware Firm

In addition, we consider the case where a firm may seek endogenously to maximize some objective which takes the fairness outcomes of their price allocations into account. In similar fashion to Equation 3.3, this self-regulating firm's problem becomes:

$$\begin{aligned} \max_{p} \quad & \sum_{i=1}^{G} n_i p \times \text{fairness} \\ \text{s.t.} \quad & 0 < p \leq p_{max} \end{aligned} \tag{3.4}$$

where the second term in the product enforces demand fairness by penalizing differences in demand between groups, as expressed in Equation 3.2.

3.3.4 Benevolent Social Planner

Given the range of possible preferences for policy makers discussed above, we continue with a mathematical formalism for our dynamic agents, including alternative formulations for our SP agent dependent upon possible policy objectives of government. We first introduce the shared formulations of the SP agent variants, then we outline their differences when introducing the sequential component to the full RL variant.

SP problem formulation

To alleviate the requirement for firms to proactively include fairness considerations in their pricing strategy, we introduce a benevolent SP to incentivize fair firm behaviour by leveraging the policy tools of taxation and subsidization. Specifically, the SP should generate a tax schedule which penalizes the firm based on its *fairness bracket*, and redistributes wealth so as to narrow the distributional gap between market participants. In this scenario, we assume that the SP adopts a revenue-neutral policy; at the end of every tax period, its tax revenue is fully offset by its expenditure, which here takes the form of a consumption subsidy awarded to consumers in proportions determined by their group membership, thereby closing the cycle of wealth in our simulated economy. The SP is formulated as a learning agent that evolves in an environment containing F firms, and each of the firms sets prices for a subset of the population constituted of individuals from both underrepresented and over-represented consumer groups.

For the cases where the SP is a contextual bandit or an RL agent, it observes a context vector from the context space $\mathscr{X}^{SP} = \{f, \phi\}$, where $f \in [0,1]^F$ denotes a vector of individual firm fairness values, and $\phi \in [0, p_{max}]^F$ denotes a vector of firm profits, normalized per customer, and takes actions from a continuous action space

$$\mathscr{A} = \left\{ (A_1, A_2) \mid A_1 \in [0,1]^b, A_2 \in [0,1] \right\},$$

where A_1 is a vector of size b with elements $\tau \in [0,1]$ denoting tax rates for each of b tax brackets, and $A_2 \in [0,1]$, denoting the proportion of the collected tax budget that will be awarded to the underrepresented group as a consumption subsidy in the following period. It follows that the remaining proportion of the tax budget $1 - A_2$ will be distributed to the relatively overrepresented group. Therefore, to determine the effective price faced by a customer from the underrepresented group 1 and majority group 2, we update p via

$$p_1^t := p^t - \frac{A_2 \times \mathscr{B}}{n_1} \quad ; \quad p_2^t := p^t - \frac{(1 - A_2) \times \mathscr{B}}{n_2}, \tag{3.5}$$

where $\mathscr{B} = \sum_{i=1}^{F} \tau_i^{t-1} \times \phi_i^{t-1}$ denotes the total tax revenue generated by F firms in the previous period, and n_1, n_2 denoting the number of consumers belonging to group 1 and 2 respectively.

Social Welfare Over Time

For the multi-armed and contextual bandit policy scenarios discussed in section 3.3.2 and section 3.3.2, we formulate the social planner's problem as one of setting a tax schedule and redistribution scheme that will remain in effect over a time horizon T. Rewards in these settings are therefore computed only at time T. We further assume that the SP is able to compute fairness values by evaluating the buyer distribution resulting from the firm's learned pricing strategy. Consequently, the SP's reward is formulated as

$$\mathscr{R}^{SP} = \frac{1}{F}\sum_{i=1}^{G_j}\sum_{j=1}^{F} n_i p_j (1-\tau_j) \times \text{fairness}_j, \tag{3.6}$$

where F is the number of firms participating in the economy, and G_j is the number of consumer groups addressed by firm j. We note that, with the addition of taxation, the firm's maximization objective becomes $\max_p \sum_{i=1}^{G} n_i p \times (1-\tau)$ because they care about maximizing net profits.

For the RL solutions discussed in section 3.3.2, the SP's objective is to maximize discounted reward over the same horizon. In the RL setting, the SP agent's reward function is denoted

$$\mathscr{R}^{SP}_{\text{RL}} = \sum_{t=1}^{T} \gamma^{t-1} \left(\frac{1}{F} \sum_{j=1}^{F} \sum_{i=1}^{G_j} n_{i,t} p_{j,t} (1-\tau_{j,t}) \times \text{fairness}_{j,t} \right),$$

where $\gamma \in [0,1]$ is a discount factor.

3.3.5 Solution Methods

In this section, we outline the methods employed. The SP agent learns its policy for choosing good incentive mechanisms given the welfare context using soft actor-critic (SAC), an RL algorithm designed to solve Markov decision processes, where there are temporal transition dynamics. S with emphasis on sample efficiency. SAC draws samples from a *replay buffer* to perform policy gradient updates. Upon experimentation, our SAC agent expressed difficulty in generating taxation frameworks which were consistent in retaining meaningful information from lower brackets, which become less common as firms become fairer, leading us to introduce `FairReplayBuffer`, which alleviates this concern.

Soft Actor-Critic

In our study, we employed the soft actor-critic (SAC) algorithm (Haarnoja et al. 2018), an advanced RL technique suited for continuous action spaces. SAC is distinguished by its incorporation of entropy maximization into the reward function, promoting exploration while simultaneously striving for optimal policy development. This approach enhances the algorithm's sample efficiency and stability, making it particularly suited for environments where precise, adaptive control is required. While typically used for sequential decision-making over time, SAC is also appropriate for our bandit setup, as its sample efficiency makes it applicable to public policy where experimenting over long periods entails risk.

FairReplayBuffer

A fundamental challenge in generating meaningful incentives for less frequently seen contexts across the fairness space lies in how our SAC agent performs gradient updates from a traditional First-In-First-Out (FIFO) replay buffer, which would mostly consist of common firm behaviours in training. Due to this, a learning agent in our setting would gradually forget any information gained from infrequently observed regions of the observation space, thereby reducing its effectiveness on out-of-sample fairness brackets.

To address this, we define a replay buffer that uniformly stores examples from across the fairness space, allowing the learning agent to output meaningful tax actions for each region. The algorithm, along with ablations, can be found in Appendix C.

3.4 Experimental Design and Analytical Baselines

In this section, we define our working examples, which are designed to demonstrate the effects that a benevolent SP may have on market dynamics. We begin by defining four firms, each facing varying demand distributions between two consumer groups. Then, we record each firm's optimal pricing strategy and resulting fairness scores under profit-maximizing objectives, with and without self-regulation for fairness, in the absence of policy intervention. We set $p_{max} = 10$. Finally, we allow the SP to dynamically learn a taxation framework with respect to profit and fairness outcomes which will push firms to adopt prices which promote fairness levels approaching those achieved by fairness-aware firms.

3.4.1 Multi-firm Setup

We conduct our experiments in an environment where each of four firms addresses two distinct consumer groups. To describe their behaviours via purchase probability distributions, we apply parameters b and w found in **??** (some of which are borrowed from Maestre et al. (2019)), to Equation 3.1, in order to obtain the curves depicted in Figure 3.2. Using the ground-truth purchase probabilities embedded in our consumer environment, we can analytically derive optimal prices for each firm's objective. For instance, the fairness-agnostic firm's optimal price allocation is obtained by finding

$$\arg\max_{p} \ \mathbb{P}_i(\text{purchase} = 1 \mid p) \times p.$$

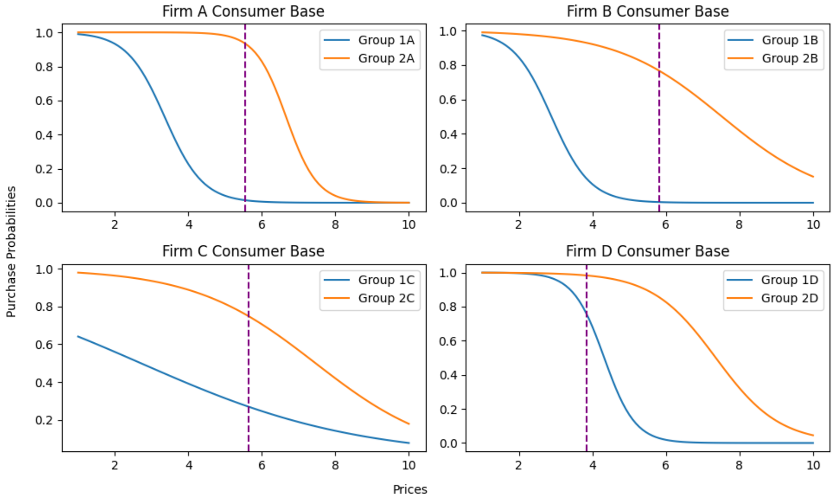


Figure 3.2: Consumer profiles: Each firm serves two consumer groups. For each, group 1 may be considered to have lower tolerance to rising prices than group 2. The vertical purple and green lines represent the analytical profit-maximizing price assigned by each firm in the fairness-agnostic and fairness-aware cases respectively, with the resulting vertical gaps between the orange and blue lines illustrating important discrepancies in purchase probabilities between consumer groups under price allocations associated to both behaviours.

3.4.2 Baselines

Assuming firms are efficient at finding profit-maximizing prices under current market conditions (i.e., demand distributions and incentive schemes), and carrying our working example forward, we would expect fairness-agnostic firms to converge to the prices and fairness values in Table 3.1a in the absence of policy intervention. By solving for the profit-fairness objective, also in the absence of policy intervention, i.e. self-regulation, we find that firms may change their prices significantly while sometimes achieving profits comparable to those in the fairness-agnostic case. These welfare-maximizing results for the fairness-aware firm (Table 3.1b) serve as analytical baselines which can be used as a positive control for comparison with the results from our bandit and RL experiments. Further, referring back to Figure 3.2 provides an alternative visualization of fairness outcomes, where the difference between the orange and blue curves is the gap in demand between consumer profiles resulting from the firms' price assignments. We finally note from Table 3.1a that our initializations for each firm lead them to converge to profit-maximizing prices resulting in a broad coverage of the fairness space. This is a design

Firm$^{\text{agnostic}}$	**A**	**B**	**C**	**D**	**Avg**
f	0.08	0.24	0.52	0.78	0.41
ϕ	2.63	2.24	2.87	3.34	2.77
swf	0.21	0.54	1.49	2.61	**1.14**

(a) Fairness, profit, and corresponding social welfare scores achieved by fairness-agnostic firms A, B, C, and D in the absence of policy intervention.

Firm$^{\text{aware}}$	**A**	**B**	**C**	**D**	**Avg**
f	0.85	0.59	0.66	0.92	0.76
ϕ	2.27	1.65	2.51	3.17	2.40
swf	1.93	0.97	1.66	2.92	**1.82**

(b) Fairness, profit, and corresponding social welfare scores achieved by fairness-aware firms A, B, C, and D, who self-regulate.

Table 3.1: Resulting fairness (f), profit (ϕ), and social welfare (swf) values from analytical optimal price assignments for fairness-agnostic and fairness-aware firms.

choice to demonstrate how optimal decisions from the firms' standpoint can have unforeseeable fairness implications.

3.5 Empirical Evaluation

We deploy a learning agent in the experimental setting defined in section 3.4 using our multi-firm setup. While we initialize firms as described, we allow for a degree of stochastic fluctuations in consumer behaviour to occur, reflecting how demand can change subtley in the short-run. To reflect this, we sample purchase probabilities from a normal distribution around means given by the output of the sigmoid demand curves denoted in Equation 3.1. The curves themselves are parameterized by w_i and b_i, as found in **??**. This indicates that the social planner may experience a different reward from having taken identical actions. This, paired with the Bernoulli trials which determine a consumer group's real purchase outcomes, introduces a degree of noise into our environment, thereby adding complexity to our learning problem. The purchase probability for a consumer group in a given training step is therefore sampled as $\mathbb{P}_{i,j} \sim \mathcal{N}(\mathbb{P}_{i,j}, \sigma^2)$. In our experiments, we set $\sigma = 0.05$, allowing for small fluctuations in demand. We show the best results in terms of swf below. All experiments were run for 20 seeds. Our experiments are designed to answer the following questions:

1. What effects do different policy mechanisms, reflected through RL design choices, have on welfare?
2. How are individual market participants affected by a dynamic policy mechanism?
3. How can RL-generated policies be adopted to assist policy makers and what are the long-term effects of such policy designs?

We train the SP to tax firms based on where they lie along the fairness dimension, which we discretize into 5 brackets, using the multi-armed bandit, contextual bandit, and full RL formulations (see section 3.3). Simultaneously, the SP chooses a proportion of the tax budget that it will re-distribute to the underrepresented consumer group in the next timestep. Our results show that, due to efficient wealth redistribution, the introduction of dynamic policies can approach and even improve upon the global welfare obtained from the analytical optimal case where firms are self-regulating.

3.5.1 RQ1: Welfare effects of various RL methods

Here we evaluate the impact each formulation has on global welfare. Table 3.2 breaks down the average results for the multi-armed bandit, contextual bandit and full RL settings. First, we note that the RL SP yields the best swf results, on-par with the fairness-aware firm baseline, for which results are shown in Table 3.1b, illustrating the effectiveness of the proposed incentive mechanisms. However, while there is a clear pattern of improvement in swf as the problem formulation complexifies, it is worth noting the limitations of each implementation. While a multi-armed bandit SP underperforms relative to other solution methods, it nonetheless significantly improves welfare outcomes compared to profit-maximizing (fairness-agnostic) firms while requiring the least amount of data from the environment. This is suitable, under stationarity assumptions, for new markets with unknown dynamics.

	MAB ($S=0.63$)				CB ($S=0.65$)				Full RL ($S=0.66$)			
Firms	A	B	C	**Avg**	A	B	C	**Avg**	A	B	C	**Avg**
f	0.72	0.66	0.58	0.72±0.04	0.87	0.80	0.58	0.80±0.02	0.84	0.79	0.58	0.79±0.03
ϕ	2.39	1.69	1.67	2.25±0.07	2.35	1.89	1.70	2.23±0.07	2.37	2.26	2.00	2.51±0.11
swf	1.82	0.93	0.89	**1.64±0.14**	2.04	1.51	0.98	**1.78±0.09**	1.99	1.78	1.15	**2.06±0.15**

Table 3.2: Aggregate Results per experimental design. For the full RL, $\gamma = 0.99$. Standard errors are reported only for averages to reduce clutter.

The contextual bandit SP approaches the analytical optimal solution for self-regulating firms. This formulation, while resulting in a more adaptive policy framework, requires more granular environment data in training, which may be difficult to obtain in the real world. Finally, the full RL formulation yields the best results in terms of welfare, achieving improvements relative to the baseline optimum on average. However, it intervenes at higher frequencies, re-defining policy mechanisms at each timestep. In addition to requiring granular training data, this may also reduce its range of applications to highly dynamic or volatile markets with shorter cycles. Globally, our results reveal that it is possible for a dynamic SP to incentive fairness in markets while retaining or even improving economic productivity.

3.5.2 RQ2: Impact on individual firms

We continue with a discussion on the individual impact to welfare bore by firms. Table 3.3 denotes the per-firm welfare comparisons between the analytical baselines in the fairness-agnostic and fairness-aware cases, and firms subjected to the RL SP's learned incentive mechanisms. We note that, while most firms experience improvements in individual welfare, firm **C**'s is significantly reduced. This is likely due to the almost parallel nature of the demand curves of the two groups, and the fact that the fairer price is higher than the profit-maximizing one. Nonetheless, the policy mechanism applied by the social planner RL agent significantly improves welfare outcomes compared to the fairness-agnostic case, and even manages to surpass those in the (unrealistic) case where firms self-regulate for fairness, with a 13.19% improvement in global welfare. We note further that while global welfare improves under this setting, it is impossible for welfare to improve for every individual firm. Firm **C**'s welfare loss indicates that there is indeed no such thing as a free lunch.

Firm	**A**	**B**	**C**	**D**	$\overline{\texttt{swf}}$
$\texttt{swf}_{agnostic}$	0.21	0.54	1.47	2.61	1.21
$\texttt{swf}_{aware}$	1.93	0.77	1.66	2.92	1.82
$\texttt{swf}_{SP}$	1.99	1.78	1.15	3.23	2.06
$\texttt{swf}\%\Delta_{SP}^{aware}$	3.11%	131.17%	-30.72%	10.62%	**13.19**%

Table 3.3: Summary of the welfare effects of the social planner's learned taxation and redistribution scheme compared to analytical baselines. The bottom row denotes the percentage change in welfare between the fairness-aware baseline and the social planner's generated *fairness tax*.

3.5.3 RQ3: Deployment and performance over time

As a high social impact application, it is necessary to study the usability and long-term impact of our framework. This is particularly relevant in the case that a similar solution is used to inform and assist policy makers in mechanism design. We note in Figure 3.3 the emergence of a clear pattern between fairness and taxation, in line to a large degree with human intuition: firms demonstrating fairer behaviours are incentivized by lower tax rates, and this trend persists for each problem formulation (**??**). In addition, each formulation of the SP learns to give the majority of the subsidy budget to the underrepresented consumer group (noted in Table 3.2). For example, a proportion $S = 0.66$ means that for every unit of tax budget the social planner collects, it redistributes 0.66 units to *group 1*, and 0.34 units to *group 2*, thereby reducing their effective price and increasing their market participation. In our evaluation of welfare over time, the full RL SP outperforms both multi-armed and contextual bandit SPs in most timesteps, as shown in Figure 3.3.

3.6 Limitations and Social Impact

There exist opposing schools of thought regarding fiscal policy and policy intervention in general. While one would advocate for intervention for the sake of equality or social progressivity, another would claim that any form of intervention would disrupt the economic processes by which consumption and production are governed, and that any deviation from these processes is inefficient. In our simplified setting, we build an economic subsystem to illustrate the tradeoffs between policy intervention and its absence. Profit-maximizing firms are not likely to be intrinsically motivated to behave fairly with respect to their buyer distributions, especially when said behaviour entails a sacrifice in profitability. We illustrate this in Table 3.1a, where profit-maximizing prices may have unforeseeable fairness outcomes dependent on distributions in demand. To mitigate this, we introduce a third-party social planner whose objective is to generate incentives which nudge firms into considering social welfare. Our experiments ultimately show that net

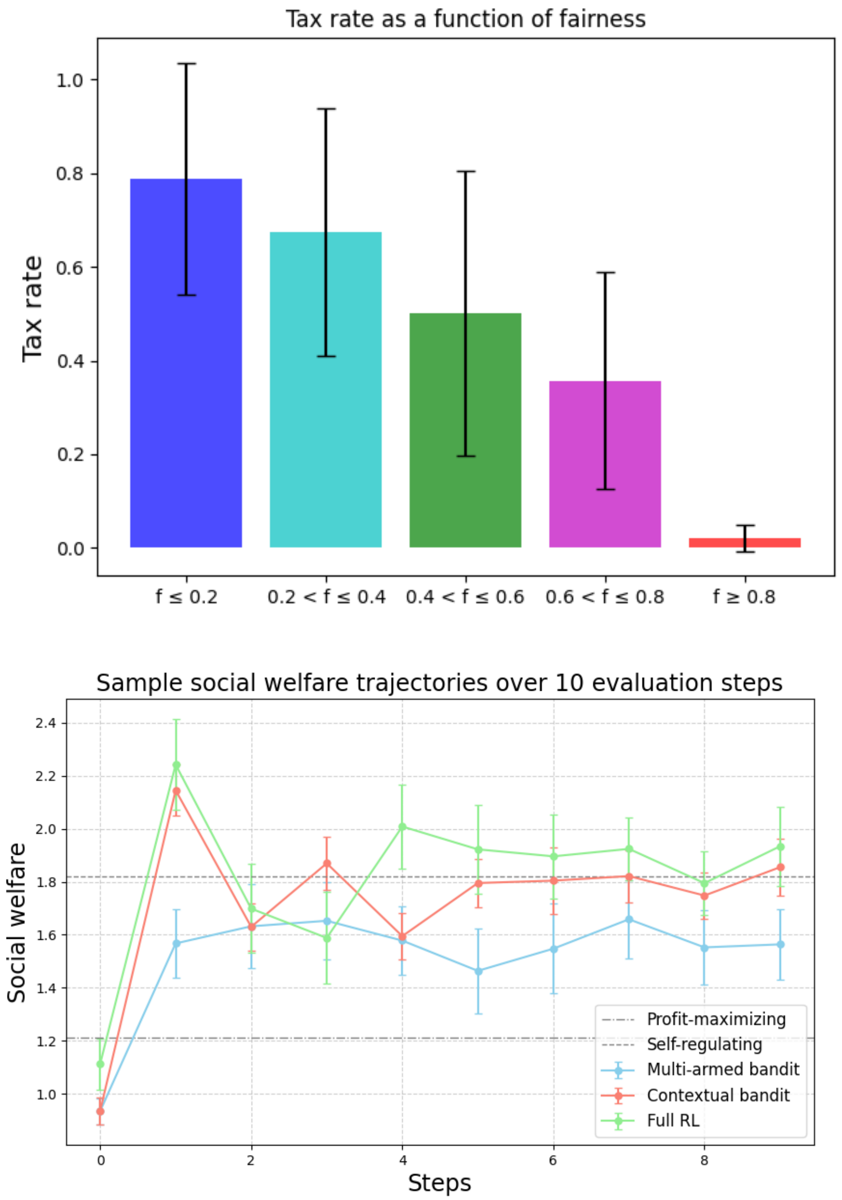


Figure 3.3: *Top*: Tax actions taken by the RL social planner. Reported policy mechanisms record SP actions averaged over 20 seeds. *Bottom*: Social welfare trajectories during evaluation for the SP's learned policy frameworks from multi-armed bandit, contextual bandit, and RL formulations.

improvements in social welfare are achievable with tailored taxation and redistribution schemes learned through a range of RL methods, each with their own applications. We note further that the adjusted prices paid by consumers (post-subsidy) are effective at increasing market participation.

Given that taxation and subsidy have important social implications, deploying AI-based strategies might not be easily accepted by decision-makers or the general public due to interpretability and accountability concerns. While we were able to generate seemingly interpretable tax schedules using RL, the chosen tax brackets themselves might not be easily accepted or interpreted, especially since they are linked with a penalty. Furthermore, while the resulting pattern of tax schedules is consistent across random seeds, the values for each tax bracket varies. Deployment of such a solution therefore requires additional considerations for robustness against distribution shifts and possible poisoning attacks. Additionally, dynamic pricing is becoming more algorithmically driven, thus requiring more studies into how these algorithms might adapt in response to our incentive strategies.

Our research has several simplifying assumptions, making it difficult to utilize this framework in real-world scenarios. For instance, considering only demand fairness might lead to counter-intuitive positive benchmarks, as seen in our examples where analytically-optimal prices are, in some cases, higher than their fairness-agnostic counterpart. Further, it is difficult to conceive of a mechanism for assessing fairness on a per-firm basis in the real world. This would presumably necessitate policy makers to access broader market data, as well as consumer-level data from individual firms, from which they may compute distributions which would be necessary for a fairness-based tax to be enforced. Additionally, our experiments consider a single-product market with no competitive dynamics between firms. While our setting illustrates a real concern for the disparate impacts of dynamic pricing on consumer distributions, introducing consumer choice would increase realism, but would perhaps make achieving improvements in welfare more challenging. Lastly, we focused on demand fairness given how it reflects accessibility to essential goods and services in critical domains such as healthcare, but further research is necessary to explore how our framework can be applied to alternative fairness notions such as price and consumer surplus fairness, which expand to other applications such as loan approval.

In a profit-driven market, our work shows that merely relying on self-regulation is not sufficient, thus requiring interventions from policy makers to ensure equity and equal access to multiple goods and services. While we do not present our framework as the ultimate solution, we believe it can guide future research and inspire exploration at the intersection of AI and society, particularly in the design of dynamic incentive mechanisms, especially in applications where dynamic pricing has high social and economic impacts.

3.7 Conclusion

When used for fairness-agnostic profit-maximization, dynamic pricing can have harmful consequences with respect to equal access. We demonstrate that, by leveraging the policy tools of taxation and subsidy, a dynamic social planner may create global efficiencies which would have otherwise been unattainable, even with locally optimal firm behaviour by self-regulation. This not only reveals that a dynamic third party may be trained to improve fairness outcomes in a broad sense, but may also generate policies which are tailored to specific markets and consumer dynamics, and that these policies may be deployed in ways that improve upon free-market outcomes. While we believe these findings to be relevant for monopolistic markets or geographically-delimited markets with little potential for consumer displacement, we see great value in extending this work to competitive markets, where consumers have more freedom in choosing among different products. Additionally, exploring more realistic and challenging environments with imbalanced consumer distributions is necessary to further understand and analyze the financial and social implications of such frameworks. Finally, extending the framework to use various fairness definitions would be needed to ensure its compatibility with other application domains.

References

Abebe, Rediet, Jon Kleinberg, and S Matthew Weinberg (2020). “Subsidy allocations in the presence of income shocks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 7032–7039.

Becker, Gary S (1962). “Irrational behavior and economic theory”. In: *Journal of political economy* 70.1, pp. 1–13.

Bertsimas, Dimitris, Vivek F Farias, and Nikolaos Trichakis (2011). “The price of fairness”. In: *Operations research* 59.1, pp. 17–31.

— (2012). “On the efficiency-fairness trade-off”. In: *Management Science* 58.12, pp. 2234–2250.

Betancourt, Jose M et al. (2022). *Dynamic price competition: Theory and evidence from airline markets*. Tech. rep. National Bureau of Economic Research.

Cohen, Maxime C, Adam N Elmachtoub, and Xiao Lei (2022). “Price discrimination with fairness constraints”. In: *Management Science* 68.12, pp. 8536–8552.

Das, Shantanu et al. (2022). “Individual fairness in feature-based pricing for monopoly markets”. In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 486–495.

Fleurbaey, Marc (2008). *Fairness, responsibility, and welfare*. OUP Oxford.

Gallego, Guillermo, Huseyin Topaloglu, et al. (2019). *Revenue management and pricing analytics*. Vol. 209. Springer.

Gillen, Stephen et al. (2018). “Online learning with an unknown fairness metric”. In: *Advances in neural information processing systems* 31.

Haarnoja, Tuomas et al. (2018). *Soft Actor-Critic Algorithms and Applications*. DOI: 10.48550/ARXIV.1812.05905. URL: https://arxiv.org/abs/1812.05905.

Hu, Yaowei and Lu Zhang (2022). “Achieving long-term fairness in sequential decision making”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9, pp. 9549–9557.

Joseph, Matthew et al. (2016). “Fairness in learning: Classic and contextual bandits”. In: *Advances in neural information processing systems* 29.

Kallus, Nathan and Angela Zhou (2021). “Fairness, welfare, and equity in personalized pricing”. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 296–314.

Kastius, Alexander and Rainer Schlosser (2021). “Dynamic pricing under competition using reinforcement learning”. In: *Journal of Revenue and Pricing Management*, pp. 1–14.

Kim, Byung-Gook et al. (2015). “Dynamic pricing and energy consumption scheduling with reinforcement learning”. In: *IEEE Transactions on smart grid* 7.5, pp. 2187–2198.

Lee, Simon, Abdou Illia, and Assion Lawson-Body (2011). “Perceived price fairness of dynamic pricing”. In: *Industrial Management & Data Systems*.

Lu, Renzhi, Seung Ho Hong, and Xiongfeng Zhang (2018). “A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach”. In: *Applied energy* 220, pp. 220–230.

Maestre, Roberto et al. (2019). “Reinforcement learning for fair dynamic pricing”. In: *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 1*. Springer, pp. 120–135.

Malc, Domen, Damijan Mumel, and Aleksandra Pisnik (2016). “Exploring price fairness perceptions and their influence on consumer behavior”. In: *Journal of Business Research* 69.9, pp. 3693–3697. ISSN: 0148-2963. DOI: `https://doi.org/10.1016/j.jbusres.2016.03.031`. URL: `https://www.sciencedirect.com/science/article/pii/S0148296316300510`.

Martinez, Michael E. (2022). “QuickStats: Percentage of Uninsured Adults Aged 18–64 Years, by Race and Selected Hispanic Origin Subgroup — National Health Interview Survey, United States, 2020”. In: *MMWR Morb Mortal Wkly Rep 2022* 71.834. DOI: `http://dx.doi.org/10.15585/mmwr.mm7125a3`.

McFadden, Daniel (1972). “Conditional logit analysis of qualitative choice behavior”. In.

Models, Hybrid Choice (2002). “Progress and Challenges”. In: *Marketing Letters* 13.3, pp. 163–175.

Nieuwenhuis, Marlon, Antony SR Manstead, and Matthew J Easterbrook (2019). “Accounting for unequal access to higher education: The role of social identity factors”. In: *Group Processes & Intergroup Relations* 22.3, pp. 371–389.

Pigou, Arthur (1920). *The economics of welfare*. Routledge.

Rajkomar, Alvin et al. (2018). “Ensuring fairness in machine learning to advance health equity”. In: *Annals of internal medicine* 169.12, pp. 866–872.

Schlosser, Rainer and Martin Boissier (2018). “Dynamic pricing under competition on online marketplaces: A data-driven approach”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 705–714.

Stratigi, Maria et al. (2020). “Fair sequential group recommendations”. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. SAC ’20. Brno, Czech Republic: Association for Computing Machinery, pp. 1443–1452. ISBN: 9781450368667. DOI: 10.1145/3341105.3375766. URL: https://doi.org/10.1145/3341105.3375766.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Yin, Tongxin et al. (2024). “Long-term fairness with unknown dynamics”. In: *Advances in Neural Information Processing Systems* 36.

Zheng, Stephan et al. (2020). “The ai economist: Improving equality and productivity with ai-driven tax policies”. In: *arXiv preprint arXiv:2004.13332*.

Zhu, Yushu et al. (2023). “Neoliberalization and inequality: disparities in access to affordable housing in urban Canada 1981–2016”. In: *Housing Studies* 38.10, pp. 1860–1887.

Chapter 4

General Conclusion

The research presented in this thesis explores the complex interplay between dynamic pricing strategies used by firms for profit-maximization and their broader social impacts, particularly concerning fairness and equal access. By investigating the use of policy tools such as taxation and subsidies, this work demonstrates the potential for a dynamic social planner to mitigate the adverse effects of purely profit-driven pricing strategies. Our findings highlight how such policy interventions can lead not only to improved fairness but also to increases in global efficiencies that are difficulty-achieved otherwise, even under unrealistic self-regulated market practices.

Dynamic pricing, while effective for maximizing short-term profits, often neglects the socio-economic disparities it can exacerbate. This oversight can lead to market scenarios where access to goods and services becomes unevenly distributed, favoring certain consumer groups over others. The role of a dynamic social planner, as explored in this thesis, is crucial in such contexts. We show that, by employing a variety of RL algorithms that can adapt and evolve in response to market conditions, a social planner can implement policies that are sensitive to the dynamics of specific markets and the unique needs of disparately impacted consumer groups. Such tailored policies, can significantly outperform the outcomes of unregulated free-market systems. This work not only underscores the potential of integrating advanced machine learning techniques with economic policy design but also highlights the need for continuous research at the intersection of AI and economics. By furthering our understanding of these mechanisms, we can better equip policy systems to guide market behaviors towards more socially desirable outcomes, thereby ensuring that economic and computational advancements contribute positively to all segments of society.

There are several opportunities for building upon this work. While our research focus was on monopolistic or geographically-constrained markets, our promising results invite the extension of this approach to competitive markets. In environments where consumer choice is broader, and the risk of displacement by competitive pricing is significant, the insights gained from a policy-informed framework could prove even more beneficial. Additionally, the exploration of more complex and realistic market environments, notably characterized by imbalanced consumer distributions remains unexamined. Addressing these concerns could provide a clearer picture of how different segments of the population are affected by policy changes and could inform more effective interventions. Further, a promising future direction for this research is the integration of various fairness definitions into the social planning framework. As societal values evolve and new ethical considerations emerge, the flexibility to incorporate diverse definitions of fairness will be key to maintaining the effectiveness of policy interventions. This adaptability will also facilitate the application of our findings across a wider range of domains, improving the generalizability and impact of our work. Finally, one could examine the welfare effects of various policy mechanisms under idiosyncratic shocks to supply or income, leveraging any insights gained to further inform policy design.

Bibliography

Abebe, Rediet, Jon Kleinberg, and S Matthew Weinberg (2020). “Subsidy allocations in the presence of income shocks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 7032–7039.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018). “Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence”. In: *The economics of artificial intelligence: An agenda*. University of Chicago Press, pp. 89–110.

Alderighi, Marco, Alberto A Gaggero, and Claudio A Piga (2016). “The hidden side of dynamic pricing in airline markets”. In.

Barocas, Solon and Andrew D Selbst (2016). “Big data’s disparate impact”. In: *Calif. L. Rev.* 104, p. 671.

Becker, Gary S (1962). “Irrational behavior and economic theory”. In: *Journal of political economy* 70.1, pp. 1–13.

Bertrand, Joseph (1883). “Review of “Theorie mathematique de la richesse sociale” and of “Recherches sur les principles mathematiques de la theorie des richesses.”” In: *Journal de savants* 67, p. 499.

Bertsimas, Dimitris, Vivek F Farias, and Nikolaos Trichakis (2011). “The price of fairness”. In: *Operations research* 59.1, pp. 17–31.

— (2012). “On the efficiency-fairness trade-off”. In: *Management Science* 58.12, pp. 2234–2250.

Betancourt, Jose M et al. (2022). *Dynamic price competition: Theory and evidence from airline markets*. Tech. rep. National Bureau of Economic Research.

Bolhasani, Hamidreza, Maryam Mohseni, and Amir Masoud Rahmani (2021). “Deep learning applications for IoT in health care: A systematic review”. In: *Informatics in Medicine Unlocked* 23, p. 100550.

Burk, Abram (1938). “A reformulation of certain aspects of welfare economics”. In: *The Quarterly Journal of Economics* 52.2, pp. 310–334.

Castelvecchi, Davide (2016). “Can we open the black box of AI?” In: *Nature News* 538.7623, p. 20.

Chen, M Keith and Michael Sheldon (2016). “Dynamic pricing in a labor market: Surge pricing and flexible work on the Uber platform.” In: *Ec* 16, p. 455.

Cohen, Maxime C, Adam N Elmachtoub, and Xiao Lei (2022). “Price discrimination with fairness constraints”. In: *Management Science* 68.12, pp. 8536–8552.

Cournot, Antoine Augustin (1838). *Recherches sur les principes mathématiques de la théorie des richesses*. Vol. 48. L. Hachette.

Das, Shantanu et al. (2022). “Individual fairness in feature-based pricing for monopoly markets”. In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 486–495.

Davenport, Thomas H et al. (2006). “Competing on analytics”. In: *Harvard business review* 84.1, p. 98.

Diamond, Peter A and James A Mirrlees (1971). “Optimal taxation and public production I: Production efficiency”. In: *The American economic review* 61.1, pp. 8–27.

Ezrachi, Ariel (2016). *Virtual competition: The promise and perils of the algorithm-driven economy*. Harvard University Press.

Farhangi, Hassan (2009). “The path of the smart grid”. In: *IEEE power and energy magazine* 8.1, pp. 18–28.

Fleurbaey, Marc (2008). *Fairness, responsibility, and welfare*. OUP Oxford.

Gallego, Guillermo, Huseyin Topaloglu, et al. (2019). *Revenue management and pricing analytics*. Vol. 209. Springer.

Gillen, Stephen et al. (2018). “Online learning with an unknown fairness metric”. In: *Advances in neural information processing systems* 31.

Haarnoja, Tuomas et al. (2018). *Soft Actor-Critic Algorithms and Applications*. DOI: 10.48550/ARXIV.1812.05905. URL: https://arxiv.org/abs/1812.05905.

Hayek, F. (1944). *The Road to Serfdom*. Createspace Independent Publishing Platform. ISBN: 9781522918387. URL: https://books.google.ca/books?id=WHfNjgEACAAJ.

Hu, Yaowei and Lu Zhang (2022). "Achieving long-term fairness in sequential decision making". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9, pp. 9549–9557.

Joseph, Matthew et al. (2016). "Fairness in learning: Classic and contextual bandits". In: *Advances in neural information processing systems* 29.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler (1986). "Fairness and the assumptions of economics". In: *Journal of business*, S285–S300.

Kahneman, Daniel and Amos Tversky (1979). "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, pp. 99–127.

Kallus, Nathan and Angela Zhou (2021). "Fairness, welfare, and equity in personalized pricing". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 296–314.

Kastius, Alexander and Rainer Schlosser (2021). "Dynamic pricing under competition using reinforcement learning". In: *Journal of Revenue and Pricing Management*, pp. 1–14.

Kim, Byung-Gook et al. (2015). "Dynamic pricing and energy consumption scheduling with reinforcement learning". In: *IEEE Transactions on smart grid* 7.5, pp. 2187–2198.

Lee, Simon, Abdou Illia, and Assion Lawson-Body (2011). "Perceived price fairness of dynamic pricing". In: *Industrial Management & Data Systems*.

Lu, Renzhi, Seung Ho Hong, and Xiongfeng Zhang (2018). "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach". In: *Applied energy* 220, pp. 220–230.

Maestre, Roberto et al. (2019). "Reinforcement learning for fair dynamic pricing". In: *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 1*. Springer, pp. 120–135.

Malc, Domen, Damijan Mumel, and Aleksandra Pisnik (2016). "Exploring price fairness perceptions and their influence on consumer behavior". In: *Journal of Business Research* 69.9, pp. 3693–3697. ISSN: 0148-2963. DOI: `https://doi.org/10.1016/j.jbusres.2016.03.031`. URL: `https://www.sciencedirect.com/science/article/pii/S0148296316300510`.

Martinez, Michael E. (2022). "QuickStats: Percentage of Uninsured Adults Aged 18–64 Years, by Race and Selected Hispanic Origin Subgroup — National Health Interview Survey, United States, 2020". In: *MMWR Morb Mortal Wkly Rep 2022* 71.834. DOI: `http://dx.doi.org/10.15585/mmwr.mm7125a3`.

McFadden, Daniel (1972). "Conditional logit analysis of qualitative choice behavior". In.

Models, Hybrid Choice (2002). "Progress and Challenges". In: *Marketing Letters* 13.3, pp. 163–175.

Nieuwenhuis, Marlon, Antony SR Manstead, and Matthew J Easterbrook (2019). "Accounting for unequal access to higher education: The role of social identity factors". In: *Group Processes & Intergroup Relations* 22.3, pp. 371–389.

O'neil, Cathy (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pigou, Arthur (1920). *The economics of welfare*. Routledge.

Piketty, Thomas (2014). *Capital in the twenty-first century*. Harvard University Press.

Rajkomar, Alvin et al. (2018). "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12, pp. 866–872.

Rawls, John (1971). "A theory of justice". In: *Applied ethics*. Routledge, pp. 21–29.

Robinson, Joan (1969). "Price discrimination". In: *The Economics of Imperfect Competition*. Springer, pp. 179–202.

Saez, Emmanuel (2001). "Using elasticities to derive optimal income tax rates". In: *The review of economic studies* 68.1, pp. 205–229.

Samuelson, Paul A (1956). “Social indifference curves”. In: *The quarterly journal of economics* 70.1, pp. 1–22.

Schlosser, Rainer and Martin Boissier (2018). “Dynamic pricing under competition on online marketplaces: A data-driven approach”. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 705–714.

Smith, Adam (1776). “An Inquiry into the Nature and Causes of the Wealth of Nations”. In: *Readings in economic sociology*, pp. 6–17.

Stratigi, Maria et al. (2020). “Fair sequential group recommendations”. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. SAC ’20. Brno, Czech Republic: Association for Computing Machinery, pp. 1443–1452. ISBN: 9781450368667. DOI: 10.1145/3341105.3375766. URL: https://doi.org/10.1145/3341105.3375766.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement learning: An introduction*. MIT press.

Tesfatsion, Leigh (2006). “Agent-based computational economics: A constructive approach to economic theory”. In: *Handbook of computational economics* 2, pp. 831–880.

Ting, Daniel Shu Wei et al. (2020). “Digital technology and COVID-19”. In: *Nature medicine* 26.4, pp. 459–461.

Tirole, Jean (1988). *The theory of industrial organization*. MIT press.

Train, Kenneth E (2009). *Discrete choice methods with simulation*. Cambridge university press.

Varian, Hal R (2014). “Big data: New tricks for econometrics”. In: *Journal of economic perspectives* 28.2, pp. 3–28.

Yin, Tongxin et al. (2024). “Long-term fairness with unknown dynamics”. In: *Advances in Neural Information Processing Systems* 36.

Zheng, Stephan et al. (2020). “The ai economist: Improving equality and productivity with ai-driven tax policies”. In: *arXiv preprint arXiv:2004.13332*.

Zhu, Yushu et al. (2023). “Neoliberalization and inequality: disparities in access to affordable housing in urban Canada 1981–2016”. In: *Housing Studies* 38.10, pp. 1860–1887.

Appendix A – Demand Distribution Parameters

In our experiments, we designed four firms A, B, C and D who each address a customer base whose demand distributions depend on parameters sampled in the table below.

	Firm A		Firm B		Firm C		Firm D	
Params	$g1$	$g2$	$g1$	$g2$	$g1$	$g2$	$g1$	$g2$
w	-1.926	-2.369	-1.9	-0.695	-0.340	-0.600	-2.369	-1.1526
b	6.4757	15.7900	5.4757	5.229	0.9195	4.4757	10.2290	8.4757

Appendix B - Additional Experiments

In addition to the experimental results reported in the main section, we also ran experiments over a variety of reward functions without the use of a subsidy. For instance, including gross profit (pre-tax) into the social planner's reward function made it less concerned for the after-tax income of firms, and more with the overall size of the economy measured by the sum of net-profit and a tax budget. This meant that firms were taxed at very high rates, and while fairness was improved, the profit values reported do not reflect the welfare of individual firms.

	Multi-armed bandit				Contextual bandit				Full RL			
Firms	A	B	C	**Avg**	A	B	C	**Avg**	A	B	C	**Avg**
f	0.76	0.32	0.57	0.60	0.79	0.33	0.56	0.62	0.76	0.46	0.62	0.67
ϕ	2.39	2.16	2.71	2.62	2.37	2.16	2.69	2.63	2.39	1.97	2.52	2.53
swf	1.82	0.69	1.54	**1.57**	1.87	0.71	1.51	**1.63**	1.81	0.91	1.56	**1.70**

Table 1: Aggregate Results w/ Gross Profit

To evaluate firm welfare in the absence of subsidy, the social planner considers net profit in its reward. In this setting, it faces the tradeoff between welfare and taxation, but the tax budget is not reinjected into the economy, making it difficult to draw comparisons with the analytical baselines, where firms are self-regulating and do not pay taxes. Nonetheless, fairness improvements are still achievable.

	Multi-armed bandit				Contextual				Full RL			
Firms	A	B	C	**Avg**	A	B	C	**Avg**	A	B	C	**Avg**
f	0.79	0.33	0.58	0.60	0.62	0.36	0.57	0.58	0.73	0.27	0.55	0.58
ϕ	1.98	1.19	1.35	1.43	2.07	1.42	2.44	2.21	2.30	1.52	2.26	2.24
swf	1.56	0.40	0.79	**0.86**	1.28	0.51	1.39	**1.28**	1.68	0.41	1.25	**1.30**

Table 2: Aggregate Results w/ Net Profit.

Appendix C - Fair Replay Buffer

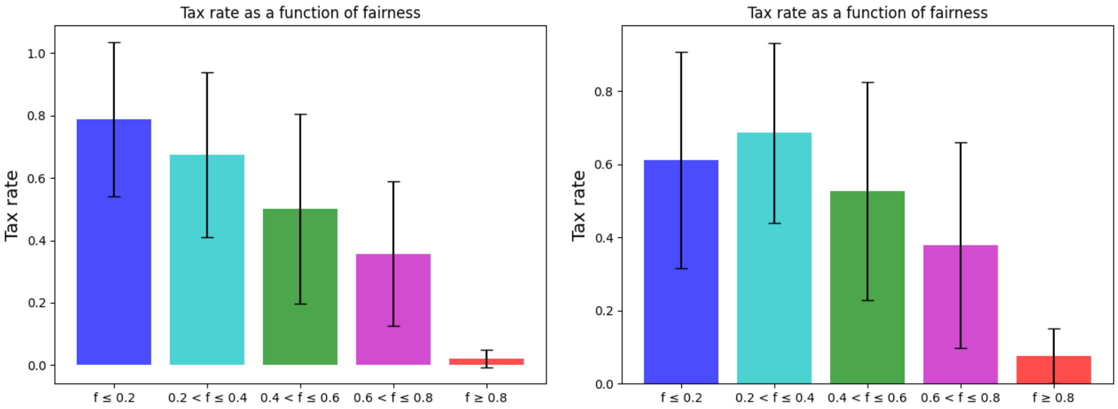
Here, we include details regarding the `FairReplayBuffer`, including the algorithm along with ablations highlighting the differences between the tax schedules learned by the SP agent when using FIFO, and `FairReplayBuffer`.

Algorithm

```
Algorithme 2 : FairReplayBuffer
Input : buffer capacity $|\mathscr{D}|$, $obs$, $action$, $reward$, $done$, $infos$, brackets $\mathscr{I}$, batch
        size $|\mathscr{B}|$
Output : A replay buffer ensuring uniform distribution of experiences across the
         fairness-profit space
Buffer initialization, $\mathscr{D} \leftarrow [\,]$
Storage per bracket initialization, $\mathscr{S} \leftarrow \{\}$
/* Procedure to add experiences                                   */
Procedure ADD($obs, action, reward, done, infos$):
    Determine $i$ based on $obs$
    Append($obs, action, reward, done, infos$) to $\mathscr{D}$
    Append index of new experience to $\mathscr{S}[i]$
/* Procedure to sample experiences                                */
Procedure SAMPLE($|\mathscr{B}|$):
    Sampled batch initialization, $\mathscr{B} \leftarrow [\,]$
    $|b| \leftarrow |\mathscr{B}|/|\mathscr{I}|$
    for $i$ in $\mathscr{I}$ do
        $b \leftarrow$ Sample $(o, a, r, d, info)$ from $\mathscr{S}[i]$ ;
        Add $b$ to $\mathscr{B}$
    end
    Shuffle $\mathscr{B}$
    return $\mathscr{B}$
```

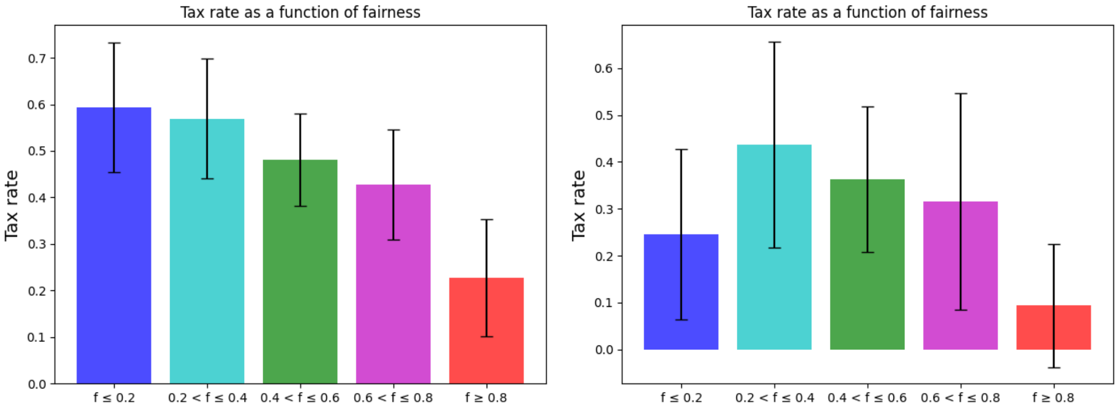
Ablations

We ablate the FairReplayBuffer by comparing it with trials run with a first-in-first-out (FIFO) buffer. The intention behind designing this replay buffer was to generate a tax schedule with intuitive patterns, achieved by ensuring that the learning agent retains information from less frequently-observed fairness brackets.



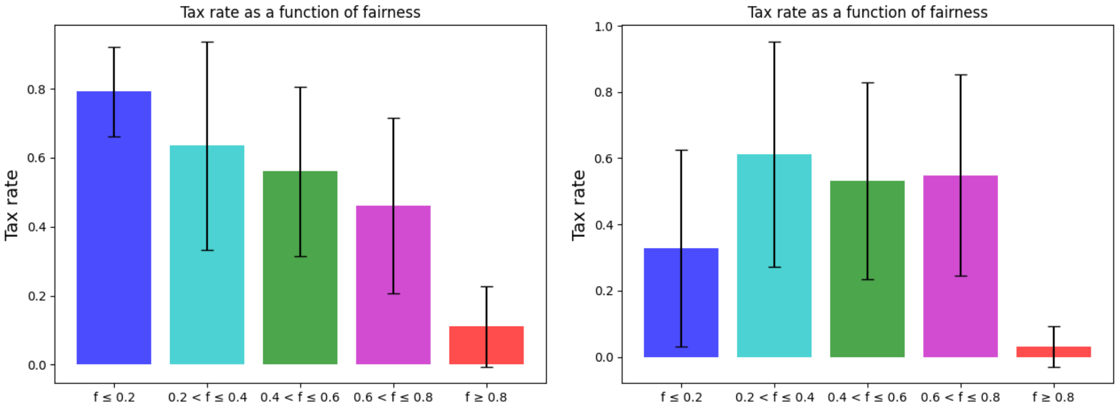
(a) RL, `FairReplayBuffer` (b) RL, `FIFO`

Figure 1: `FairReplayBuffer` vs. FIFO for the RL setting



(a) CB, `FairReplayBuffer` (b) CB, `FIFO`

Figure 2: `FairReplayBuffer` vs. FIFO for the CB setting



(a) MAB, `FairReplayBuffer`

(b) MAB, `FIFO`

Figure 3: `FairReplayBuffer` vs. FIFO for the MAB setting