ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

Adaptive Reinforcement Learning Frameworks for Dynamic Tax Policy Decision Support

Dhanaraj Sathiri
Independent Researcher
dhanrajsathiri@gmail.com

Abstract

Tax policy aims to stabilize the economy and provide basic public services to meet the needs of the domestic economy. However, it is dynamically implemented by governments to regulate economic fluctuations. Tax revenue, affected by many internal and external factors, is difficult for governments to predict. Tax noncompliance and evasion also hinder the effectiveness of tax policy. Consequently, the formulation of tax policy is difficult and needs to be based on a predictive model that can provide reliable decision support. Reinforcement learning (RL) is a branch of machine learning that forms policy through reward-driven interaction with an environment. By setting the control problem of tax policy into an RL framework, reinforcement learning can realize the adaptive and autonomous optimization of tax policy.

Tax policy needs to improve the return on taxation while pursuing other economic goals and maintaining the stability of taxation in order to stimulate compliance and avoid deformation of the tax base. Therefore, it is necessary to ensure that the improvement of taxation return does not affect investment incentives, both domestic and foreign. For public systems, such as social security and education, the rational planning of short-term expenditure and its cycle through a counter-cyclical stance of tax policy in line with economic needs within the overall revenue and expenditure plan.

Keywords : Economic modeling; decision support systems; tax policy; reinforcement learning; simulation Anyone can send a message to this channel.

1. Introduction

The paper proposes a smart decision support system for dynamic economic policymaking that dynamically adjusts the input space using reinforcement learning. The proposed research design quantifies the additional requirements introduced by economic policy dynamics and uncertainty, particularly stakeholders' constraints, through a comprehensive framework that aims at creating excessive tax revenue for the entire tax system.

Dynamic economic policy optimization in uncertain environments requires decision-support tools that fulfil the specific requirements induced by such conditions. For tax policy optimization, these requirements include (i) modelling the tax system as a policy search problem, (ii) enabling policy evaluation with a flexible simulator that accounts for social compliance constraints, (iii) measuring policy performance with respect to decision-makers' objectives, (iv) controlling the trade-off between economic exploration and exploitation, (v) testing robustness to economic shocks and uncertainty, (vi) assuring generalization beyond the used data, and (vii) offering a layout that joins governance, transparency, and privacy concerns. These features, together with the proposed exploration strategy, allow incorporating the dynamic nature of tax policies into the optimization process and better coping with uncertainty. Such a system would be very useful also for central banks in identifying risks to the achievement of their policy objectives.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

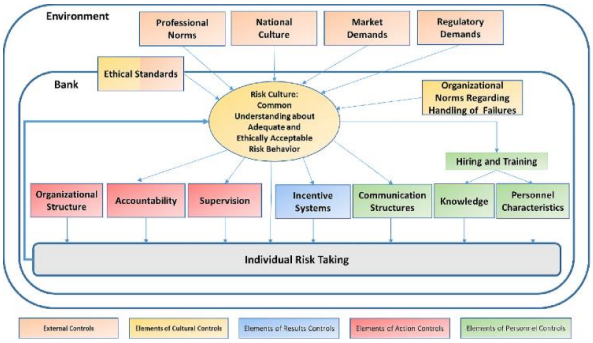


Fig 1: Banks' risk culture and management control systems

1.1. Background and Significance

Modern economies are developing rapidly, and fiscal policy must change simultaneously with these developments. A well-structured tax system makes an undeniable contribution to the achievement of government objectives: redistributing wealth in society, taking measures against inflation, avoiding a higher level of unemployment, maintaining lower prices, promoting balanced foreign trade, encouraging production and increasing investments. An important component of the fiscal policy of the government, an important source of income is the tax revenue obtained from corporate taxes, income tax, personal income tax, consumption tax, VAT, and capital gains taxes. Tax revenues are, however, dynamic and subject to change. Shocks also occur in the economy from time to time because of natural, international or external fiscal policy considerations, and political influences, all of which make it necessary to take effective measures to rectify any decline through policy change.

Policy making is a highly complex task. Tax policy decision makers have to have information about many uncertain elements affecting tax revenues. They have to find ways to operate and implement policies that can result in desired outcomes. There is very often a heavy pressure on governments to introduce policies that are not sustainable from a medium and longer term perspective. They do not have the luxury to see beyond the next election. With the help of a properly designed smart decision support system, macro-economists and tax risk specialists can take the pressure away from tax policy makers and allow them to examine the long term sustainability of policy alternatives over relevant time horizons.

1.2. Research design

Existing methods for optimizing tax policy are based on stochastic optimization, differential games, and direct policy search. Yet it remains challenging to discover direct reinforcement learning architectures that are able to handle the complexity of public tax policy. Three requirements distinguish the dynamics of tax policy and set limits for proper decision-making: objectives of tax systems are multiple and often conflicting; tax policy is responsive to shifting environments, while revenue is affected not only by the level and structure of rates, but also by financial responsibility, compliance, and tax morale of taxpayers; the public choice perspective takes into account the behavior of the political stakeholders involved in the design, selection, enforcement, and decision. Two research questions can be formulated. First, which decision support system for dynamic tax policy exists and what are its main components? Second, how does the integrated decision support for tax policy optimization benefit from reinforcement-learning processes compared with existing economic models?

Reinforcement learning plays an important role in the proposed framework. It provides an algorithm that models the decision-making process as a trial-and-error process, exploring a given environment, learning from the consequences of its actions, and, while doing so, improving its ability to make better decisions. This makes reinforcement learning especially useful for problems for which a precise mathematical model does not exist or is difficult to derive. To discover a direct reinforcement-learning architecture model capable of handling the dynamics of tax policy optimization, existing RL concepts are explored and three major architectural lines in RL are identified: model-free safe and constrained reinforcement-learning methods, model-based methods for environment simulation and indirect policy search, and methods aimed at direct policy search in Markov decision processes.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

2. Theoretical Foundations of Reinforcement Learning in Policy Optimization

Dynamic tax policy optimization can naturally be framed as policy optimization and learning in a sequential decision making context. When the system is relatively well understood and accurate analytical models either exist or can be developed, reinforcement learning methods naturally lend themselves to a policy search approach. However, the absence of an accurate simulator, or the scarce availability of past decision data, necessitate exploring a model-free approach, where a policy is learned directly from the reward signal. A policy learning approach is also appealing in the case of safe or constrained reinforcement learning due to the presence of safety and fairness constraints that govern the acceptable space of action.

Reinforcement learning can also be employed to find dynamic tax policies in conditions where a rich database of economic data exists, but a simulator of the economy is lacking. In these cases, model-based reinforcement learning appears very suitable, permitting the development of an environment capable of testing the effect of a multitude of policies in a variety of different scenarios of the economy. Safety and constraints remain important elements to consider, especially when the policies are being created for actual deployment. The focus must be placed on these aspects when building the simulated environment.

Equation 1: Value functions and Bellman equations (derived step-by-step)

$$\pi(a \mid s) = \Pr(A_t = a \mid S_t = s)$$

State-value function

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid S_t = s]$$

Action-value function

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[G_t \mid S_t = s, A_t = a]$$

These match the paper's "value-based methods… estimate the value of state-action pairs (Q-learning)."

Smart Decision Support Systems …

Start with the definition:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \, r_{t+k} \mid S_t = s\right]$$

Split the first reward term:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[r_t + \sum_{k=1}^{\infty} \gamma^k \, r_{t+k} \mid S_t = s\right]$$

Factor out one γ from the remaining sum:

$$\sum_{k=1}^{\infty} \gamma^k \, r_{t+k} = \gamma \sum_{j=0}^{\infty} \gamma^j \, r_{t+1+j} = \gamma G_{t+1}$$

So:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[r_t + \gamma G_{t+1} \mid S_t = s]$$

But $\mathbb{E}_{\pi}[G_{t+1} \mid S_{t+1} = s'] = V^{\pi}(s')$. Therefore:

$$\boxed{V^{\pi}(s) = \mathbb{E}_{\pi}[r_t + \gamma V^{\pi}(S_{t+1}) \mid S_t = s]}$$

Expanding the expectation over actions and transitions:

$$\boxed{V^{\pi}(s) = \sum_{a} \pi(a \mid s) \sum_{s'} P(s' \mid s,a)(r(s,a,s') + \gamma V^{\pi}(s'))}$$

2.1. Markov Decision Processes and Policy Optimization

A reinforcement learning (RL) policy optimization model is presented. The method can be adapted to support tax policy decision makers in dynamically changing environments. In an RL setting, a policy-search approach generates a mapping

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

of the underlying state-space representation that describes a dynamic environment with uncertainties. The learning and inferencing algorithms work without a model that describes the legal tax policy and without scaling of the space of simulated taxes environments. Tax policies have dynamic objectives, not least because of changing domination policies in the scrutinized economy. Development gains often suffer from not being carefully balanced. Riding a boom, the fan of lowering taxes for whatever budge has often bitten its user, hence an automatic stabilizer in times of boom and jelly back in times of recession may help sustain equilibrium without additional collaterate gains. Involving citizen, firms, and other groups in the shaping and controlling of taxes that swallow and digest public money is preferable to government insuring and controlling all public services. Compliance is further based on the believed fairness of taxes. Here taxes are a collective promise, keeping order and living together safe.

In public policy, causal uncertainty makes is essential to develop decisions that are safe. Reinforcement learning allows to impose safety during exploration by using a model capable of simulating the effects of acting with the policy into situations where training data is missing. Because of the large scales of the play of Articial Intelligent, it is also essential to put boundary constraints on possible policy. In tax policy, the political actors and analysis are generally known. The comprehension of how to report, whose aspects of supporting and control are mainly in discussed with national, local micro and macro mainly been using. Saving oring by group of the external unknown shocks is always catastrophic in the long-run.

2.2. Value-Based and Policy-Based Methods

Methodologically, reinforcement learning is often organized into two complementary categories: value-based methods that focus on learning to estimate the value of state-action pairs (Q-learning) and policy-based methods that focus on learning an optimal action-selection policy directly (policy gradient). Value-based methods estimate the expected long-term return of performing action a from state s and subsequently following policy p. These estimates then update the action-selection policy according to the principle of maximum expected return. (Because the maximization can be difficult—if not impossible—in complex continuous and high-dimensional environments, “soft” policies are used in practice that learn to assign higher probabilities to more rewarding actions.) Since value functions contain information about the environment that can be used to identify good policies, when the value function can be accurately approximated from “off-policy” experiences the is often treated as a testing ground for evaluating the performance of agent training strategies.

Despite these appealing features, value-based methods have limitations. First, they often require substantial amounts of training data to learn even an approximate value function of reasonable accuracy. In particular, if the environment has multiple goals that are unequally likely to occur (e.g. high-stakes rare events), a sampling-based approach for learning the value function will suffer from the curse of dimensionality. In such cases, there is a compelling case for methods (whether in the form of actor-critic frameworks) that learn an optimal policy directly and more efficiently. Second, value-based methods concentrate on learning the expected return of policies, but exploring policies that are just good enough is often more important than identifying the optimal policy. Some measures of performance more concretely and directly related to the policies themselves, such as expected log probability of receiving rewards, are therefore more natural to optimize.

2.3. Exploration-Exploitation Trade-offs in Economic Policy Contexts

Exploration-exploitation trade-offs are fundamental to reinforcement learning (RL) and are the most discussed and studied aspect within the economics community. In conventional RL applications, exploration is achieved by agent’s randomness, which leads to collecting suboptimal reward and potentially risk-exposure. Nevertheless, considering that the agent is modeling a real-world environment, exploration may be risky and unacceptably costly or may even lead to catastrophes, crash, or collapse of the environment, particularly in domains such as investment, monetary policy, tax policy and many other econ-related domains. It is therefore important for the agent in these

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

settings to perform a safe exploration of the environment in order to increase its expected future reward without taking unwanted, excessive risks. Consequently, the agent should carefully balance the advantage of exploration (reward) against the disadvantage (risk). More formally, by joint modeling of risk and reward, the objective of the agent can be altered into maximizing the expected reward while staying sufficiently risk-loving. There already exist some safe RL algorithms from machine learning that accept such joint-risk-and-reward-setting extensions, background concepts, state-of-the-art models as well as implementations of RL applied to econ-related domain can be found. With respect to tax-policy-constrained setting, the government takes decisions about the level of tax-rates and welfare-transfers. Risky investments by households and businesses in such RL set-up may be either actively encouraged or passively tolerated depending on the unease with taking risks within society. Consequently, the degree of risk-aversion parameter in the Households and businesses in the model then jointly determines the level of uncertainty within the equilibrium path.

3. Tax Policy Dynamics and Decision Support Requirements

Three main factors that must be taken into account for tax policies in a dynamic context are explored here. First, the objective of tax policies is not clearly defined, and therefore, policymakers decision support systems should be enhanced with multiple criteria based on the policy stakeholders requirements. Second, large fluctuations can occur in tax revenue of countries or states, and the economic environment or economy state should be also considered for tax policies decision support systems. Third, constraints imposed by some of the stakeholders (e.g., taxpayers compliance and economic agents reactions) influences more the tax policies than the desires of the tax administrator. Consequently, decision support systems capable of representing the policies objectives of all stakeholders and their constraints can help avoiding non-compliance with tax policies.

The objective of a tax policy can be defined as the guiding principle used to manage the current structure and mechanisms of a tax system in order to promote the best possible performance in a dynamic environment. Similarly to a business firm objective, the priority of a tax policy is usually implied and the tax system is taken as a tool or a means to achieve certain goals. However, a formulation of the joint objectives of tax policies is still missing. Consequently, models of policymakers decision support system do not represent the tax policies objectives explicitly and often include only the preferences of the tax administrator in terms of supporting the tax system performance. The own performance of the tax policies is also neglected. Therefore, Decision Support Systems capable of reflecting the preferences of all policy stakeholders, allowing a multi-criteria evaluation of supporting and constrained decision policies, should be developed.

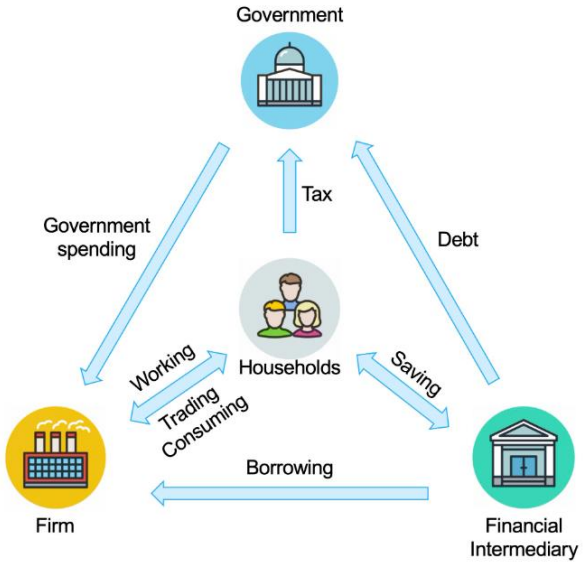


Fig 2: TaxAI A Dynamic Economic

3.1. Objectives of Tax Policy

Tax policy should increase the government's revenue for the funding of the public sector. The priorities in the public budget reflected by public expenditures make it necessary

ISSN: 3067-4166

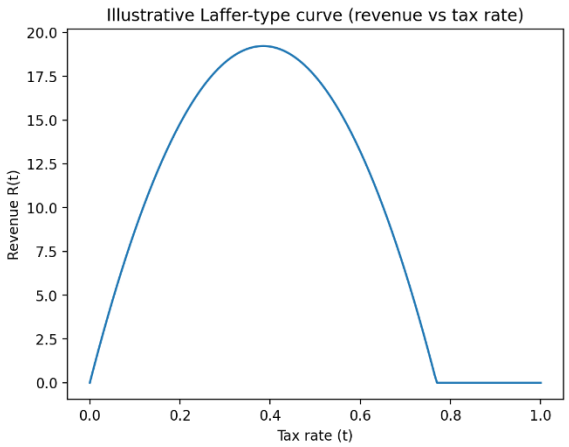
AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

that the income from taxes match carefully the different requirements. Resources should not be taken away from the private sector without a good reason, as they can otherwise not be used at their most effective. However, reducing taxes can lead to fewer public services. Therefore, tax policy comprises a stable long-term view, not short-term tax cuts to boost the economy. The tax authorities must treat taxpayers fairly and thus avoid unfair competition. But tax policy must also provide a stable environment that encourages businesses and individuals to invest. To succeed at this particular balancing act, a well-designed tax policy requires an understanding of changing economic realities.

Tax policy needs a watchdog to confront nasty surprises whenever the unexpected occurs, fuelling calls for tax cuts or tax increases. In general, the demand for new spending and lower taxes during good times is greater than in bad times. Many governments simultaneously face pressure to reduce deficits and fund expensive new demands, and their budgets often forecast surpluses. The temptation is to cut taxes, yet the long-term harmful consequences of elections in government surpluses justify resisting this clamour.



3.2. Dynamic Environments and Uncertainty in Tax Revenue

Tax systems directly impact economic development and growth. Among the various objectives of tax policy, the stability and predictive capability of government revenue are essential. However, rapid social changes can exceed a government's ability to respond effectively. Hence, changes in tax policy can introduce sudden shocks in tax revenue. These shocks are difficult to properly anticipate and plan for because governments frequently lack adequate forecasting models. Furthermore, methods based on smoothing extremes, such as developing ideal filters, time series decomposition, and higher-order condition trends, have failed to predict extreme changes in recent decades.

Therefore, the capacity of tax policy to prevent the emergence of economic shocks in revenue is not always guaranteed. A careful inspection of the objectives of tax policy shows that other social factors may need to be considered or integrated into the design of tax policy in order to maintain tax revenue within predefined limits. This integration can be realized by applying a more general framework capable of incorporating governing rules or requirements as major components within the design model. Such requirements may refer to maintaining government revenue within confident levels, avoiding extreme changes in tax burden, preserving taxpayers' income levels and purchasing power, and preventing economic fluctuations beyond predefined limits. These aspects are more easily captured with the help of simulation models than simple regression-type forecasting econometric models.

3.3. Stakeholder Constraints and Compliance Considerations

Systems embedding a smart decision support for policy-makers may benefit from being visible, interpretable, and valuable for stakeholders' interests. Regulatory bodies, enterprises and citizens interact with taxes, and any smoothing involving taxes will impact each relationship in different ways. Hence, tax policies, supported by advice of predictive systems should feature a holistic approach that takes into account the compliance in stakeholders' relationships. The outputs of the model-free smart decision system explore the effect of just some taxes for economic development, but citizens' relationship should be also analyzed. In this sense, it is surprising that tools are not available to evaluate the effect of tax policies with poor tax

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

compliance (especially from entrepreneurs, associated in organized groups). The diffusion of the model by mitigation of a poor tax compliance should be used as additional evaluation for economic development. To test constraints related to political-relational economy rules, enforcement rates should approach distinct level of penalty.

Political and social aspects (in tax compliance perspective) should impact the regulation of the predictive systems and with data available in different economic and historical backgrounds, the policies will be improved generation after generation. The backward relation predicted with proper data could also be used in the development of the training data set of predictive policy-support systems.

4. Reinforcement Learning Architectures for Policy Optimization

Methodological challenges associated with dynamic tax policy formulation can be addressed using various types of reinforcement learning (RL) approaches. RL techniques that require no prior modelling of the process being controlled may be particularly suitable for users with limited expertise in quantitative policy modelling. However, general model-free policy-search methods cannot be applied directly to discrete-action problems with large action spaces, which limits their use for tax policy optimisation, an area in which strategies are required for changing many tax rates simultaneously.

Methodologies that seek to automate and facilitate the dynamic optimisation of sensitive policies incorporating diverse stakeholders with challenging, frequently conflicting, demands can benefit from the use of simulated scenarios to generate large datasets. Data-driven approaches such as supervised ML demand considerable effort from the policy-design community (for feature engineering and data preparation) and may not provide satisfactory solutions outside their training domains. The ability of deep learning (DL) techniques to learn from and efficiently generalise solutions to complex problems can be harnessed using an imitation-learning framework. A second direction explores the possibility of combining ideas from model-based RL and constrained policy optimisation. Economic theory offers a sound basis for simulating the impact of any given policy in a dynamic environment containing multiple economic agents, which is likely to prove invaluable for policy-support applications.

Equation 2: Optimality equations and Q-learning update (step-by-step)

$$V^*(s) = \max_{\pi} V^{\pi}(s), Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

Take the best possible future action at the next state:

$$\boxed{Q^*(s,a) = \sum_{s'} P(s' \mid s,a)(r(s,a,s') + \gamma \max_{a'} Q^*(s',a'))}$$

If we had samples (s_t, a_t, r_t, s_{t+1}), then:

$$Q^*(s_t,a_t) \approx r_t + \gamma \max_{a'} Q(s_{t+1},a')$$

Define the **TD target**:

$$y_t = r_t + \gamma \max_{a'} Q(s_{t+1},a')$$

Define the **TD error**:

$$\delta_t = y_t - Q(s_t,a_t)$$

Update (stochastic approximation):

$$\boxed{Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\, \delta_t}$$

i.e.

$$\boxed{Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_t + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)]}$$

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

4.1. Model-Free Approaches for Policy Search

When deploying RL to optimize tax policy, the process considers tax collection to be regulation of social behavior. Control theory models the regulated system, whose state represents changes in the economy, such as economic surpluses. Policy simulation defines the simulated system behavior, while reinforcement learning defines the tool and rewards for governing the regulated system. The process is separation of concerns: it considers tax collection as regulation; separates simulated system behavior from the policy optimization; and defines optimal policies using algorithms that learn based on rewards. Simple longitudinal relations, graphical models, STL, CBN, and model-free RL outline the other three directions.

The majority of RL algorithms used in active learning and control policy search are model-free, requiring neither a model of the environment nor the description of the reward function. Instead, an agent explores the environment to maximize the cumulative reward for a certain time horizon or to discover a goal that can be reached as quickly as possible. A learning agent must either explore unknown environments or exploit the knowledge already gained, and the balance between the two actions is known as the exploration-exploitation dilemma. This dilemma is not exclusive to active learning, but is inherent in any dynamic environment, such as optimal tax policy in an uncertain and dynamic economy. Tax revenues, a source of funds for public goods and services, are not guaranteed, and the tax system is overloaded because it also comprises redistributive aspects. Therefore, the tax system affects and regulates the entire economy, and is impacted by economic dynamics.

4.2. Model-Based Techniques for Policy Simulation

Exploratory simulation studies are widely recognized for their ability to assess the potential effects of any number of alternative policies on socioeconomic systems. Within the context of tax policy design and evaluation, they are routinely employed by international organizations such as the International Monetary Fund and the World Bank. Such studies ultimately reflect the top-down perspective of a representative agency. However, while they are instrumental in providing preliminary information about the potential effects of tax policy as well as raising public awareness about the issues at stake, they should not serve as a direct input into policy decisions. Indeed, the resulting scenarios do not represent predicted outcomes since they cannot account for the actual (often strategic) behavioral responses of taxpayers and tax administers. For reliable predictions about the future, a bottom-up perspective—one that directly considers the response behavior of the various stakeholders involved in tax systems—is essential.

Reinforcement learning techniques can help provide such a perspective by enabling the exploration of representative, and potentially optimal, decision pathways through the vast and uncertain space of possible policies. When utilization of exploration data generated by independent agents is permissible—or such exploration is deemed feasible—policy evaluation can be undertaken using value-based reinforcement learning techniques. However, demand for information about untested policies or for predictions in scenarios of crisis or change necessitates a model of the tax system's operation. Here, exploratory training of a model—be it a surrogate model of the environment or a model of the system itself—when combined with policy search conducted using the learned model, can help mitigate concerns about solution optimality, complexity, feasibility, and trust.

4.3. Safe and Constrained Reinforcement Learning in Public Policy

Prioritizing natural disaster risk mitigation, citizen safety, and poverty alleviation makes tax revenue (contribution) not only economically relevant but also politically salient in any country. Supplementing revenue stabilizes the economy in the wake of natural disasters, giving governments fiscal space to respond effectively and subsidy schemes that support citizens and the economy during crises such as the COVID-19 pandemic. In this context, major features of contribution demand and supply need to be carefully aligned, with the public always ultimately paying the price. Not having a system in place or the political will to ensure such responsiveness invariably comes with high economic costs. While courage and political will are often in short supply, the potential of political and economic leadership can be augmented through suitable policy-support tools.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

Two fundamental aspects must be taken into account when searching for an optimal tax policy space: ensuring that the simulation environment satisfies important desired properties and providing an intelligent policy-searching engine. Public policy is subject to limitations and constraints that need to be considered during the search for an optimal response. With the natural environment already pushing the economy towards crisis and high poverty levels in many countries, public policy toward crisis-proofing the economy on a tight budget becomes a key issue.

5. Data, Features, and Simulation Environments for Tax Policy

Real-tax data from national and international agencies provide numerical inputs. Interactions with the local population and business community help identify qualitative factors that may predict compliance behavior and affect tax-revenue stability, such as greater regional stability in the context of an armed conflict or the development of infrastructures that allow for easier and safer transportation and reduce tax-collection costs.

Tax simulation environments should be built to generate the numerical features required for the learning agent. These features include variations in tax rates, unit rates, exemptions and discounts. State-action-revenue intervals allowing for a fine-tuning of the data require design scenarios. The simulation environment should allow assessing the stability of the series generated by these features under deterministic and stochastic conditions and, finally, the generalization capabilities external to the parameters used in the generation process.

Evaluation, robustness, and generalization

Evaluation metrics must focus on the capacity of the developed policy to achieve its defined objective, such as the maximization of tax revenue. Robustness can be assessed through the evaluation of the policy in the presence of economic shocks, forecast uncertainties or noise disturbances introduced in the data. Generalization capabilities can be evaluated by applying the learned policy to situations that differ from those used in the training process.

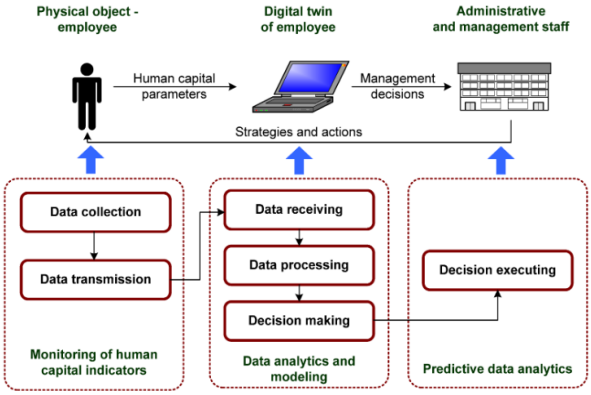


Fig 3: Data, Features, and Simulation Environments

5.1. Data Sources and Preprocessing

Systematic analysis of the requirements for a RL-based approach to the dynamic optimization of tax policy indicates that a supporting decision support system must operate over a two-dimensional feature space. The first dimension consists of a temporal series, describing to different granularities both the previous evolution of the tax system and the corresponding reactions of the tax revenue to specific reforms. The second dimension concerns the interaction between the tax system and the economic cycles. Both pieces of information must be adequately described in terms of suitable inputs for the algorithms and features must be developed accordingly.

The data sources covering the previous evolution of the considered territories are different administrative databases collected and maintained by Eurostat and the Italian Ministry of Finance. Missing indicators were carefully reconstructed following the same methodology used to build Eurostat databases. National account series for the European Union, the Euro area, and the EFTA countries were downloaded

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

from the European Commission's Directorate-General for Economic and Financial Affairs (DGAECF) and publicly available national datasets. The database concerning the Italian tax system shows: main tax rates for the central government and the local authorities, revenues, and composition in terms of sources. Statistics covering the distribution of the given revenues were obtained from Ministry of Finance databases. Tax revenues in the remaining territories are tax rate and tax base assumptions, along with estimates of tax revenue elasticities with respect to the economic cycle, used to check the plausibility of tax revenue.

5.2. State Representation for Tax Systems

The design of a reinforcement-learning policy optimization environment requires the definition of a state space tailored specifically to the tax policy decision problem within the chosen application domain. Five factors are generally regarded as critical influencers of tax-system performance. Tax revenues are the most obvious, directly determining the extent of public consumption and public investment expenditures, as well as driving the budgetary constraint for private-sector savings. Economic growth (or lack thereof) has clear implications for various budget items, influencing their volume relative to GDP. Cyclical divergences from potential GDP are crucial for understanding labour-market absorption, as are the economic cyclical and growth outlooks when they deviate from potential.

The other two critical state variables, which are less frequently part of the explanatory feature set than needed to satisfy generally accepted modelling and forecasting requirements, are more directly linked to the responsiveness of the private sector to fiscal policy. Tax-system complexity is often cited as a hindrance to economic development, while compliance is self-evidently desirable. The more complex, higher-rate and more time-consuming any tax is (or becomes), the more likely taxpayers will seek to evade it and/or employ less productive resources in tax avoidance."

5.3. Environment Design and Scenario Generation

Previous sections have demonstrated that Smart Decision Support Systems are valuable and feasible tools for the dynamic optimization of public policies with uncertain outcomes. Tax policies, however, present unique challenges for their design, data requirements, and practical deployment. To assist with tax policy optimization, the important design choices for the data, state representation, and environment are now described. Finally, the implemented Smart Decision Support System illustrates how these components may be combined to create an operational environment.

As with any model-free reinforcement-learning approach, the data must cover a sufficiently wide variety of optimal and suboptimal policies for the agent to learn a useful policy at a reasonable speed. Since most public policies are long-term commitments with uncertain outcomes, a systematic exploration of scenarios or situations would take an extremely long time. For public taxation policies, however, most tax agency databases around the world provide sufficient long-range data histories of government collections and distributions over the economy, making possible a different approach to scenario exploration. The data can be used to create a set of situations that train the underlying policy-search model and encompass a wide range of situations and economic conditions, including recessions, deep recessions, slight and deep recoveries, bubbles, stock crashes, oil shocks, and growth stagflation.

6. Evaluation, Robustness, and Generalization

Policy performance falls short of social and political aims, so a systematic process with sound metrics is essential. Reliable evaluations consider external shocks and stochastic disturbances, and out-of-sample performance underpins trust in learned, value-based tax policies.

Any reinforcement learning (RL)-based policy-optimization method is only as good as its evaluation metrics. While RL employs a reward signal, many aspects of a tax system cannot easily be encapsulated in a single number: fairness

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

and equity, for instance, are notoriously hard to quantify, and some parameter points may simply be unacceptable from an ethical or political point of view—no matter what total welfare is, or appears to be. The danger of proposing tax regimes that sound good merely in terms of the choicest of the objective-function parameters can be avoided only by evaluating the proposed RB-IPS over a large number of judiciously selected aspects. An "empirical check" against plausibility, equity and fairness remains essential, despite the methodological framework being novel.

The empirically governed checking of strategies has earned some reputation in other areas of decision support, especially in finance; therefore, it is prudent to inherit a thorough treatment of the design, control and checking of such financial algorithms. The bad news is that any realistic policy must somehow deal with how suddenly a social system can change from stability to instability and vice versa—specifically what ranges of speed are permissible. Also the fact that domain checks can be inherently uncertain is worrying: for example, a purely mathematical model may assert that the density of taxes at some time is large for all times, but experience might show that drops are merely slow. Nevertheless, such hedging or robustness aspects should be tested wherever possible, along with performance and searchability.

Equation 3: Policy gradient (policy-based methods) derived step-by-step

Let the policy be differentiable with parameters θ: $\pi_\theta(a \mid s)$.

Define:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \, r_t\right]$$

where $\tau$is a trajectory.

Trajectory probability:

$$p_\theta(\tau) = p(s_0)\prod_{t\geq 0} \pi_\theta\,(a_t \mid s_t)\, P(s_{t+1} \mid s_t, a_t)$$

Only $\pi_\theta$depends on θ, so:

$$\nabla_\theta \log p_\theta(\tau) = \sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

Now:

$$\nabla_\theta J(\theta) = \nabla_\theta \int p_\theta(\tau)\, R(\tau)\, d\tau = \int \nabla_\theta p_\theta(\tau)\, R(\tau)\, d\tau$$

Use $\nabla p = p\nabla \log p$:

$$\begin{aligned}\nabla_\theta J(\theta) &= \int p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau)\, R(\tau)\, d\tau \\ &= \mathbb{E}_{\tau\sim p_\theta}[\nabla_\theta \log p_\theta(\tau)\, R(\tau)]\end{aligned}$$

Substitute the sum expression:

$$\boxed{\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, R(\tau)\right]}$$

To reduce variance, replace $R(\tau)$with the **reward-to-go** from time t:

$$G_t = \sum_{k=0}^{\infty} \gamma^k \, r_{t+k}$$

So:

$$\boxed{\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, G_t\right]}$$

And in practice use an advantage function $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$:

$$\boxed{\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t\geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, A^\pi(s_t, a_t)\right]}$$

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

6.1. Metrics for Policy Performance

The performance of tax policy can be evaluated with respect to established economic criteria. It must generate sufficient revenue to finance public expenditure and support the macroeconomic environment. For high-income economies, the main source of tax revenue is the personal income tax. Consequently, its performance can be measured in terms of the yield and its responsiveness to economic activity, computed with respect to real gross domestic product. For emerging economies, the corporate income tax also serves as a criterion. Yields should generally increase over time and they should be estimable. A stable tax system reduces economic uncertainty and enhances steady growth. Finally, a stable tax policy reduces uncertainty and enhances faithfulness. Objective baselines can be established by computing the policy performance of near-optimal controllers using dummy reinforcement learning agents. The tax policy should approach these baselines or out-perform them on out-of-sample data.

These established economic criteria can thus serve to advise governments and controllers navigating economic downturns. The time profile of each economic criterion defines a simulated policy scenario and learned reinforcement-learning models capable of producing the policy profiles can be deployed as real-time economic controllers. Transparency and interpretability design principles embedded into Artificial Intelligence models provide an open governance mechanism for policy decision support. Ultimately, simulated policy performance helps to achieve these end-goals—policy relaxation. The degree of loading or unloading detected in real time provides simulated policy scenarios.

6.2. Robustness to Economic Shocks and Uncertainty

Policies that affect investment and consumption are often induced by temporary changes in market conditions. Past economic crises exemplified this volatility, necessitating the generation of policies that anticipate potential changes in fundamental variables like growth, inflation, or external capital flows. An adaptive tax policy is robust against unforeseen changes in the economic environment and ensures the sustainability of the government budget, both in the deterministic system and in the stochastic simulations with random shocks in the model. In drawing the tax-path, the government takes into consideration the capital market's reaction function as well as future expectations; here, an optimal tax policy would restrict longer deviations from the balanced growth path.

One of the factors generating bias in the fiscal policy rules is the uncertainty associated with the forecast of the fundamentals of the economy. If there is an large uncertainty band surrounding the dynamics of the economy, this band will induce agents in the economy to adopt more conservative behavior since they will react more to worse-than-expected outcomes. The choice of tax and expenditure multipliers are a way to introduce the uncertainty risk.

6.3. Cross-Validation and Out-of-Sample Testing

Validation of reinforcement learning-guided tax policy using exploratory simulations is performed in two complementary ways. First, performance on tax revenue targets in out-of-sample scenarios not included in the training data is reported to indicate generalization capacity. Second, diagonal cross-validation is used to assess capability in a yet different tax environment. Starting from a tax system optimally tuned to the conditions of one real-world jurisdiction, out-of-sample optimal policies are then determined for a second jurisdiction, inclusive of testing on the original training environment. Such an exploration-confined validation procedure captures the setup typically found in applications of supervised machine learning. The focus remains on demonstrating the operationality of the proposed approach and less on rigorous in-sample performance evaluation.

The need for carefully investigated out-of-sample testing arises from the concern that the explored reinforcement learning-based tax revenue prediction model may be overfitted to the dynamics of a specific setting. While built-in regularization helps address such risk, the development of a model-free tax policy recommendation system for use across very different economic environments—unsupervised, fully exploratory reinforcement learning—is an active area of interest. Typically, exploration of one environment does not guarantee the ability to accurately act

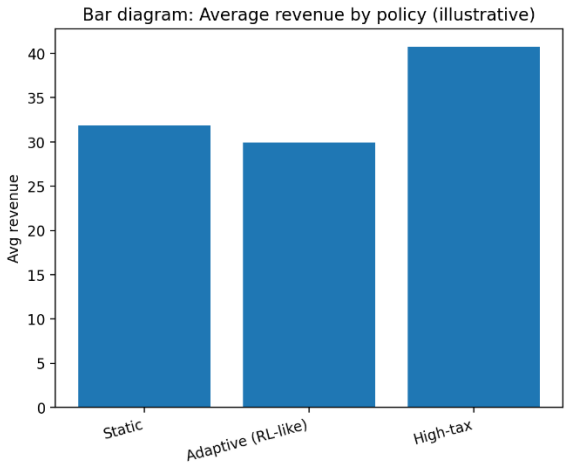
ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

in others, and this issue is investigated here via simple diagonal cross-validation.



7. Practical Considerations for Implementation

To meet the principles of good governance, any decision-support tool requires an evaluation of its appropriateness and risk for society. Public policy issues rely heavily on the constructive channels between analysts and decision-makers, and the effect of AI models on such a relationship should be considered cautiously. Consideration of privacy and security implications, compliance with applicable laws and regulations, and dependence on a principled governance framework ensures that these systems are developed and applied responsibly. In the face of the potential for even greater system complexity, governance of the systems should balance transparency and privacy interests and account for the growing regulatory requirements regarding AI. Analysts operating decision support tools built on rules of thumb apply experience-based interpretation to the AI-generated outputs for ultimate policy decisions. Similarly, policy explainability is critical for building trust and confidence in the use of AI in this context.

The value of a decision-support tool does not solely reside in the final AI model—the development process is likely to expose insights that improve the understanding of the decision problem and inform policy choices, even in the absence of AI. Finally, while modern AI technologies are likely only to be applied in a narrow space of the economy, mistakes have the potential to create severe ramifications elsewhere, including AI in government decision-making. AI may become the core of government decision-making as memory and computer resources expand. However, the very complexity and secrecy of most modern algorithms—and the difficulty of checking their final predictions—have the potential to undermine trust and close the channels of analysts' intuition.

7.1. Governance, Transparency, and Accountability

The development of Smart Decision Support Systems for Dynamic Tax Policy Development requires the involvement of stakeholders both in the training of the models and during the subsequent evaluations. Although enabling autonomous learning by interactive agents with qualitative simulation is a possible route, regulation requires an established audit trail for policy changes, as well as due process for modifying rules that are in place. As a result, the Smart Decision Support Systems are envisaged as assistance tools for public service employees rather than automatic agents with autonomous power to modify policies without human or institutional oversight.

Transparency and accountability are indispensable in all aspects of the data and model development as well as during the application of the trained models. Any data set used for training, evaluation and testing needs to be disclosed, as do the procedures and operations derived from the original data sets to obtain state and action feature representations. All model hyperparameters and learning outcomes also require public disclosure. Such transparency enables external validation of model performance and exploration of vulnerabilities that may compromise the correctness of the trained models in delivering the objectives set by the policy makers.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

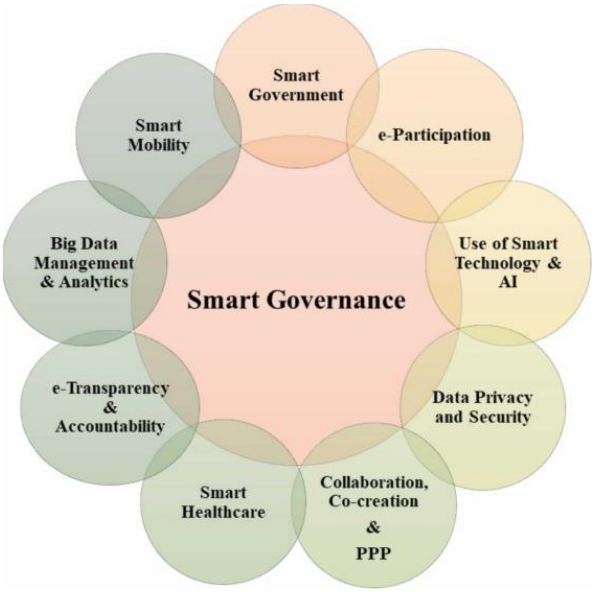


Fig 4: Governance, Transparency, and Accountability

7.2. Privacy, Security, and Ethical Implications

Governance risk related to privacy, security, and ethical implications of the systems has to be a critical aspect when introducing any new technology or approach to decision making in government context in relation to artificial intelligence. The implementation of AI technologies has thus to safeguard ethically, legally, and socially sensitive information. It must take into account that privacy protection control mechanisms and processes could have a significant impact and serve as a prevention method for possible privacy invasion cases. Local and national laws on privacy protection, such as Regulation 2016/679 of the European Parliament and of the Council, which is getting a follow-up on a horizontal basis on AI-generated personal data, are very strict on this matter. Security protection from cybercrime must also have a major role in any development project on AI that supports or integrates technology in every program and decision-making of private and public entities. In this regard, a dedicated governance function is also necessary to pay specific attention to the links between innovation, security of information and systems, and the consequent promptness of requirements in order to overcome potential threats. Detecting biases, fairness, and discrimination is another ethical implication that now has to be introduced in the AI decision-making process before any rollout of applications/solutions supporting the expected new technology. Every decision supported by AI technology must be intended as a precautionary principle that may require verification of the absence of bias.

7.3. Interpretable AI and Policy Explainability

In the realm of artificial intelligence (AI), the quest for interpretability has gained momentum. Increasingly, stakeholders—ranging from developers and business leaders to policymakers and the public—demand transparent AI systems internally as well as clear explanations of their outputs. For example, AI-powered chatbots should provide coherent reasons supporting their responses. Consequently, black-box AI systems, which yield predictions without revealing the underlying rationale, face scrutiny and rejection. ChatGPT exemplifies such tension: millions appreciate its functionality, while it remains excluded from various applications because users cannot ascertain how it generates responses.

Integrating explainability into AI systems poses distinct and rigorously studied challenges across all domains. For instance, many voice assistants accelerate a driver's navigation but lack interpretable AI models to clarify assistance. Developers and users cannot ascertain whether a voice assistant possesses reliable knowledge about the road and its environment. In the context of Reinforcement Learning (RL), policies trained through black-box models can generate good or bad results depending on the complexity of the learned environment. Furthermore, robust RL models should identify and filter dangerous or impossible policies from a stage before making decisions and determining their outputs. Reinforcement learning policy optimization lacks a transparent interpretable AI model to unveil the supporting rationale behind selected policies and the produced recommendations. Authorities ultimately remain responsible for fulfilling societal and

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

environmental conditions by carefully weighing the suggestions generated by the RL model.

8. Conclusion

The research demonstrates how game and economic theory provide distinct methods for generating Markov Decision Process environments. By applying model-free reinforcement learning techniques to discover tax policies in simulated settings, implications for evaluation robustness and out-of-sample performance can guide the practical deployment of support systems in real macroeconomic environments. Delegating the search for other real-world prosecutor and defense attorneys\u2019 policies to dry run agents places such decisions at the disposal of the public administration or institutions that oversee it, taking care to anticipate potential abuse risks. Nevertheless, these systems are never a substitute for human decision-making; at least some members of a population should remain free and capable of acting in conditions of ethical privacy and identity security.

Tax data from the Americas were used to train models that generate the Markov Decision Process for short-term revenue scenario policy simulations. A simplified model-free reinforcement learning framework was established, embodying tax bureaucrats. The optimized revenue-maximization policy proved robust to negative GDP shocks, mock Keynesian stimulations, and increasing tax evasion levels. Scenarios simulation affirmed that maintaining lower tax rates increases the resiliency of public finances. A supervised model predicts policy performance concerning the probability that investors will adjust GDP expectations when deciding to invest or withdraw in the next period. Building a supervised system based on model-free reinforcement-learning outputs broadens its application range to explore policies that can reduce unemployment rates.

8.1. Simulated Policy Scenarios

Numerical analyses applying RL techniques to data from the United Kingdom, the United States, and Hebei, China, demonstrate how tailored tax policies can enhance national welfare objectives by accounting for variations in economic environments. These policy optimization findings were subsequently applied to a simulated environment in which RL methodologies were leveraged to produce a policy-support system. Methods that allow more effective exploration—emphasis on value-based learning, enhanced realism through integrated modeling for policy simulation, and incorporation of risk constraints into the RL framework—serve to advance overall robustness and generality of the solution-support technology. The resulting system offers a structured basis for learning policy configurations that can guide decision makers toward performance-enhancing outcomes within their preferred objectives.

To maintain realism, application areas span provinces rather than countries, where data sets can be synthesized rather than relying on past observations—an approach especially suited for testing tax system designs that have never been simultaneously explored. Four application domains are considered: Hebei province in China, the United Kingdom, and the United States of America, with the number of VAT brackets representing a critical design choice in the first three, together with the possibility of near-zero corporate tax; and the United Kingdom, where tourism enters profit equations for the first time.

8.2. Real-World Applications and Lessons Learned

In practice, RL-based decision support systems for dynamic tax policy optimization offer several advantages over existing tools. First, they enable the rapid generation of smart dynamic policy rules adapted to a given tax system. Second, they ensure that simulation environments are business-cycle-consistent, easily adaptable to other environments, and large enough to assess policy robustness. Third, they cover a wide set of scenarios, including revenue targets or compliance constraints, and allow for the assessment of rules' feasibility, robustness, and downside risks. Fourth, established dynamic economic problem formulations enforce the optimality of policy derived via popular RL training algorithms.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

Applying this integrated framework in the Czech Republic produced a set of optimally operating automatic stabilizers. Shrewd rules could be derived, though without leading to risk-pushing revenue changes. Moreover, the corresponding state representation lent itself to the calculation of feasible and risk-sensitive tax policy responses to large adverse – and even pandemic-like – economic shocks, highlighting the Czech Republic's above-average vulnerability to such events. Empirically validating an advanced RL-derived automatic stabilizer policy set – one based on the Czech tax–benefit system – offered additional insights into using RL for tax policy optimization within the Czech Republic and more widely.

9. References

[1] Agarwal, A., Kakade, S., Lee, J. D., & Mahajan, G. (2020). Optimality and approximation guarantees for policy gradient methods in reinforcement learning. *Journal of Machine Learning Research, 21*(98), 1–76.

[2] Athey, S., Bayati, M., Doudchenko, N., Imbens, G. W., & Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association, 116*(536), 1716–1730.

[3] Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives, 31*(2), 3–32.

[4] Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica, 89*(1), 133–161.

[5] Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. *Journal of Machine Learning Research, 18*(1), 2633–2680.

[6] Bai, Y., Basu, S., & Sun, X. (2023). Safe reinforcement learning for constrained decision-making: A survey. *ACM Computing Surveys, 55*(12), 1–37.

[7] Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In A. Prieditis & S. Russell (Eds.), *Proceedings of the 12th International Conference on Machine Learning* (pp. 30–37). Morgan Kaufmann.

[8] Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM, 64*(7), 58–65.

[9] Bertsekas, D. P. (2019). *Reinforcement learning and optimal control*. Athena Scientific.

[10] Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science, 66*(3), 1025–1044.

[11] Bertsimas, D., & Tsitsiklis, J. N. (1997). *Neuro-dynamic programming*. Athena Scientific.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

[12] Besley, T., & Persson, T. (2014). Why do developing countries tax so little? *Journal of Economic Perspectives*, *28*(4), 99–120.

[13] Bhandari, J., & Russo, D. (2019). A contextually adaptive method for constrained bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 3062–3069.

[14] Blanchard, O., & Summers, L. H. (2017). Rethinking stabilization policy. *IMF Economic Review*, *65*(1), 1–35.

[15] Borkar, V. S. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research*, *27*(2), 294–311.

[16] Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

[17] Cai, Q., Yang, Z., Wang, Z., & He, Z. (2019). Exploring under-appreciated challenges in offline reinforcement learning. *arXiv preprint arXiv:1909.05833*.

[18] Campbell, J. Y. (2018). Financial decisions and markets: A course in asset pricing. *Annual Review of Financial Economics*, *10*, 1–25.

[19] Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and high-frequency trading*. Cambridge University Press.

[19] Castro, P. S., Dabney, W., & Rowland, M. (2018). Distributional reinforcement learning: A review. *arXiv preprint arXiv:1806.06923*.

[20] Chen, L., Hallak, A., & Mannor, S. (2020). Learning a mixture of policies for offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(4), 3365–3372.

[21] Chow, Y., & Ghavamzadeh, M. (2014). Algorithms for CVaR optimization in MDPs. *Advances in Neural Information Processing Systems*, *27*, 3509–3517.

[22] Chow, Y., Tamar, A., Mannor, S., & Pavone, M. (2017). Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, *18*(1), 6070–6120.

[23] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*, 4299–4307.

[24] Cobbe, K., Klimov, O., Hesse, C., Kim, T., & Schulman, J. (2020). Leveraging procedural generation to benchmark reinforcement learning. *Proceedings of the 37th International Conference on Machine Learning*, 2048–2056.

ISSN: 3067-4166

AMERICAN DATA SCIENCE JOURNAL FOR ADVANCED COMPUTATIONS

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 14 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 18

[25] Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.

[26] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness. *arXiv preprint arXiv:1808.00023*.

[27] D'Amour, A., Heller, K., Moldovan, D., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research, 23*(226), 1–61.

[28] Dearden, R., Friedman, N., & Russell, S. (1998). Bayesian Q-learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 761–768.

[29] Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2017). Learning and policy search in stochastic dynamical systems with Bayesian neural networks. *Proceedings of the International Conference on Learning Representations*.

[30] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[31] Duflo, E. (2020). Field experiments and the practice of policy. *American Economic Review, 110*(7), 1952–1973.

[32] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

[33] Efroni, Y., Mannor, S., & Pirotta, M. (2021). Exploration-exploitation in constrained MDPs. *Advances in Neural Information Processing Systems, 34*, 21250–21262.

[34] Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *Administrative Law Review, 72*(4), 1–36.