computation

MDPI

Article

Extending Q-Learning for Economic Modelling: A Design Framework with Equilibrium Benchmarks

Jorge Moya Velasco [1,*], Jorge Soria Ruiz-Ogarrio [2], Pedro Caja Meri [1] and Silvia Álvarez-Santás [3]

[1] Departamento de Economía y Empresa, Facultad de Derecho, Empresa y Ciencias Políticas, Universidad Cardenal Herrera-CEU, CEU Universities, Carrer Lluís Vives 1, 46115 Alfara del Patriarca, Spain; pedcame@uchceu.es
[2] Departamento de Economía, Facultad de Ciencias Económicas y Empresariales, Universidad de Navarra, Edificio Amigos, c/Universidad 1, 31009 Pamplona, Spain; jsoriar@unav.es
[3] Departamento de Economía Financiera y Contabilidad, Facultad de Ciencias de la Economía y de la Empresa, Universidad Rey Juan Carlos, Paseo de los Artilleros, s/n, 28032 Madrid, Spain; silvia.alvarez@urjc.es
* Correspondence: jorge.moyavelasco@uchceu.es

Abstract

This paper proposes a methodological architecture to integrate Q-learning into economic modelling systematically. It addresses a common gap: the lack of a shared framework linking economic foundations to Reinforcement Learning components. Rather than introducing a new algorithm, it specifies and reports how preferences, frictions, information structures, and time horizons map to the reward function, discount factor, and learning environment design. Equilibrium outcomes serve as benchmarks for comparing learned policies, not as imposed axioms. This approach interprets learning dynamics through standard economic categories and enables comparability across studies. The architecture organizes models along explicit dimensions: behavioural preferences, institutional frictions, economic environment class, information structure, learning and exploration mechanisms, and evaluation metrics. A simulation illustrates how variations in frictions, risk attitudes, and intertemporal preferences affect learned policies, their stability, and their relationship to static benchmarks. The paper aims to promote the cumulative use of Reinforcement Learning in applied economics by providing a general specification that improves interpretability, comparability, and reproducibility, turning deviations from theoretical equilibria into measurable diagnostics that refine economic fundamentals.

Keywords: Q-learning; reinforcement learning; computational economics; economic modelling; benchmarks; methodology

Check for updates

Academic Editor: Shengkun Xie

Received: 13 January 2026
Revised: 6 February 2026
Accepted: 11 February 2026
Published: 14 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Economic Theory has built a solid corpus through simplified models that make it possible to obtain equilibrium solutions and conduct comparative statics [1,2]. These models, although analytically powerful, rely on highly restrictive characterizations of the economic agent, modelled as a very constrained homo oeconomicus, poorly aligned with the behaviour of a real economic agent [3]. To overcome these limitations, there arises a need to treat economic decision-making as a dynamic process, more closely reflecting what reality shows. This is where Reinforcement Learning (RL) enters as an area of Machine Learning that allows agents to learn optimal policies through repeated interaction. RL provides a natural language to model adaptation, trial-and-error, and feedback, connecting

with classical bounded rationality [4] and enabling equilibrium to be reformulated as an emergent outcome of a dynamic process.

The use of RL in applied Economics, implemented through computational algorithms, has generated three fundamental contributions that explain its rapid expansion. First, it allows us to address problems for which there are no proper analytical solutions, characteristic of markets or strategic environments that are extremely complex. This makes it possible to provide operational rules obtained through the learning process inherent to this type of solution. Second, it has opened the possibility of studying strategic adaptation as a dynamic process, allowing the observation of adjustment paths, historical dependencies, and out-of-equilibrium patterns that remained invisible in traditional models. Finally, it has opened a computational pathway that enables the testing of the economic foundations provided by Economic Theory, treating equilibrium as a possible outcome to which learning may or may not converge.

The observation of these three main contributions reveals the conceptual implication of economic equilibrium in all of them. In Economics, the notion of equilibrium is a central issue, whose result is a fixed-point solution that appears as an optimum, as a best response to an optimization process subject to constraints. However, in RL models, results are generated by an updating process, and equilibria, when they appear, are endpoints of a dynamic process, not solutions previously determined by the optimization programme of Economic Theory. This difference is key and does not eliminate the analytical role of theoretically grounded equilibrium but rather shifts it from a prior axiomatic assumption to a benchmark that can be contrasted through explicit decision dynamics.

However, RL implementations in Economics operate as isolated exercises to solve each specific case. That is, state, action, and reward representations are defined case by case, without an architecture that links these components to the economic foundations of Economic Theory. This hinders interpretation, comparability, and the accumulation of evidence within a given context. Moreover, it prevents an easy return to economic foundations to adjust or refine their general postulates. This is where the proposal developed in this work connects, by establishing an interpretative framework that allows integrating RL results in Economics into the language and corpus of Economic Theory, enabling the advancement of the foundations of economic knowledge.

1.1. Methodological Limitations in the Use of Reinforcement Learning in Applied Economic Models

The growing use of RL in applied economic models allows for a closer examination of how these approaches are implemented in practice. This development has brought to light a few recurrent methodological limitations in their use, which constitute precisely the object of improvement of this paper. Beyond the diversity of existing approaches, patterns of model design can be found in the literature that hinder both comparability across studies and the systematic economic interpretation of results [5–7].

In particular, a problem arises from the implicit incorporation of economic assumptions into algorithmic design choices. Central economic assumptions, such as preferences, constraints, information structures, or notions of agents' rationality, are often encoded in implementation decisions specific to the algorithm. The discretization of the state space, the definition of the action set, the design of the reward function, the time horizon, or the discount factor are typically presented as technical choices, without being explicitly linked to the economic concepts they are meant to represent. This lack of correspondence between the algorithmic model and its economic reference generates interpretative difficulties, as algorithmically similar models may represent substantively different economic environments, while different designs may reflect equivalent economic assumptions without this being evident [6,8].

A second limitation concerns the omission of explicit reporting of the learned policy. In many applications of Q-learning, the analysis focuses on aggregate magnitudes such as prices, profits, or costs, without characterizing the policy learned by the agent. From an economic perspective, however, the policy constitutes the central object of analysis. When it is not made explicit, the results are not reintegrated into the underlying theoretical framework, and learning operates as an instrumental mechanism disconnected from the economic theory it is intended to inform. The lack of policy-level reporting prevents the identification of regions of inaction, decision thresholds, historical dependence, or the coexistence of qualitatively distinct policies that generate similar aggregate outcomes [6,7].

Finally, it can be observed that a significant share of existing studies uses economic benchmarks as ex post numerical points of comparison, without integrating them systematically into the analysis of the learning process. This practice complicates the interpretation of observed deviations, as it remains unclear whether they stem from learning frictions, model design choices, or intrinsic economic behaviour. Consequently, both cross-study comparison and the accumulation of knowledge in this area are severely constrained [5–7].

These limitations should not be interpreted as isolated deficiencies of individual works, but rather as the manifestation of a more general methodological shortcoming, which can be summarized as the absence of an explicit interface between economic concepts and the design choices inherent to RL. In this context, this article proposes addressing this gap by offering a taxonomy and a reporting scheme that make these decisions explicit and enable a systematic economic interpretation grounded in Economic Theory.

In this article, the notion of a taxonomy does not refer to a classification of algorithms, solution methods, or equilibrium concepts. Rather, it refers to a structural map between economic assumptions and RL design decisions. The object of classification is not the learning algorithm itself, but the mode of modelling through which economic hypotheses are incorporated into a RL framework.

The proposed taxonomy is intended to make explicit how core economic assumptions, such as preferences, constraints, information structures, and notions of rationality, are operationalized through design choices including state representation, action space definition, reward specification, discounting, and temporal structure. These assumptions constitute the fundamental building blocks of standard economic models, as they determine agents' behaviour and the nature of equilibrium outcomes. In RL-based models, however, such assumptions are often embedded implicitly in implementation choices that are presented as technical details. By organizing these decisions within a common framework, the taxonomy provides an explicit interface between RL models and the conceptual categories of Economic Theory.

The purpose of this taxonomy is therefore methodological rather than algorithmic. It does not aim to improve predictive accuracy, convergence speed, or empirical performance. Instead, it seeks to discipline model specification so that learned policies can be interpreted as economically meaningful objects, compared across studies, and evaluated against theoretical benchmarks. In this sense, the taxonomy functions as a protocol for economic interpretation and reporting, rather than as a substitute for existing learning methods or solution concepts [9].

1.2. Delimitation of Comparable Frameworks

The purpose of this subsection is not to offer an exhaustive literature review, which is provided later, but rather to define the methodological frameworks conceptually similar to the one discussed here, clarifying the specific contribution of the proposed framework and its limitation in relation to the lack of a taxonomy that would allow for a proper transfer from the field of Economics. While these approaches share formal or computational elements,

they differ substantially in their objectives and the type of methodological contribution they offer, which allows them to be grouped into the following sets.

1.2.1. Learning in Games (Game Theory)

A first strand of related literature corresponds to learning in games within Game Theory. This body of work studies learning dynamics such as fictitious play, regret minimization, no-regret learning, or evolutionary dynamics as analytical tools to characterize convergence properties, stability, and equilibrium selection [10–12]. Within this framework, learning rules are introduced to analyze whether and how agents' behaviour converges to equilibrium concepts that are defined independently of the learning process itself. Learning does not constitute an object of study in its own right but rather appears as a means to validate or refine theoretical equilibrium analysis. As a consequence, algorithmic design choices are typically dictated by the equilibrium concept under consideration, and no explicit methodological layer is developed to relate learning architectures to underlying economic assumptions, which lies outside the scope of this literature.

1.2.2. Agent-Based Models

Agent-based models provide a flexible framework to represent interactions among heterogeneous agents in economic environments that can be described with different degrees of complexity [13–15]. These approaches have been widely used to study phenomena such as market dynamics, coordination failures, financial instability, labour markets, and macroeconomic adjustment processes, allowing researchers to explore emergent outcomes generated by specified behavioural rules. Agent-based models are highly diverse and adaptable, but they differ from standard economic models in that they generally do not require decision processes to be formulated as explicit sequential optimization problems. Solutions are obtained through computational procedures that simulate adaptive behaviour and complex interaction patterns. However, decision processes are not formalized as policies in the RL sense, nor are economic assumptions systematically disentangled from algorithmic implementation choices. As a result, comparability across models typically relies on contextual or narrative descriptions, rather than on a shared methodological structure that would allow results to be interpreted and contrasted at a common analytical level.

1.2.3. Computational Experiments and Reproducibility Frameworks

Another body of contributions focuses on computational experiments and reproducibility frameworks in Economics. This literature addresses issues related to transparency, replication, and robustness of computational results, proposing standards for data availability, code sharing, and experimental documentation [16–18]. These efforts play a crucial role in strengthening empirical validation and credibility in computational economic research. However, their primary concern lies with computational reproducibility and verification, rather than with the economic interpretation of learning algorithms. They do not explicitly address the linkage between RL design decisions and economic concepts, nor do they provide guidance on how learned policies should be specified and reported as objects of economic analysis.

1.2.4. Ad Hoc Applications of Reinforcement Learning in Economics

A large and growing body of work consists of ad hoc applications of RL to specific economic environments, typically modelled as Markov Decision Processes or stochastic games, with model design adapted to each context. These applications have demonstrated their ability to generate valuable results in areas such as pricing, strategic competition, or dynamic resource allocation [7,19,20]. Nevertheless, the economic assumptions underlying these models are commonly encoded implicitly in the architecture of the learning algorithm.

This makes it difficult to interpret learned policies in economic terms and severely limits comparability across studies, as similar algorithmic designs may correspond to different economic hypotheses, while different designs may embody equivalent assumptions without this being transparent. Consequently, the integration of these partial results into the broader corpus of economic knowledge remains limited, given the absence of protocols that allow them to be systematically contrasted with theoretical equilibrium concepts or other interpretative benchmarks.

Although these approaches differ substantially in their objectives, analytical focus, and methodological tools, they share a common limitation from a methodological standpoint: none of them provides an explicit structural map that links core economic assumptions to RL design decisions in a way that enables systematic interpretation, comparability across studies, and cumulative knowledge building. A review of the proposals within these frameworks reveals no explicit protocol for operating at this methodological level, one that would make economic assumptions transparent through structured mapping with RL design decisions and standardize the reporting of learned policies as economically significant objects. The proposal developed in this article is intended to complement these approaches by operating at a distinct methodological level. Its objective is not to introduce new algorithms or to compete in terms of empirical performance, but to provide a methodological framework that allows RL design decisions to be explicitly articulated, organized, and interpreted from an economic perspective. By offering a taxonomy of such decisions and an associated reporting scheme, the proposed framework aims to improve the economic interpretability of learned policies and to facilitate the systematic accumulation of knowledge in Economics.

Building on the limitations identified in Section 1.1 and the delimitation of existing approaches presented in Section 1.2, the following sections introduce the methodological framework proposed in this article. This framework provides an explicit interface between RL design choices and the categories of Economic Theory. To this end, an architecture is specified that disciplines RL-based models so that their solutions can be framed within economically meaningful categories, allowing equilibria emerging from learning dynamics to be contrasted with theoretical benchmarks. Rather than proposing a new algorithm, the objective is to establish a protocol for specification and economic interpretation that makes results comparable, interpretable, and accumulative across studies.

In this context, the article addresses the following research questions:

1. How can RL design decisions be systematically linked to economic concepts so that economic assumptions are made explicit rather than implicitly encoded in algorithmic choices?;
2. How can learned policies in RL–based economic models be specified and reported in a way that enables economic interpretation and comparability across studies?;
3. How can equilibrium concepts from Economic Theory be used as interpretative benchmarks in learning-based economic models, without being imposed as axiomatic solutions?

The first two questions are central to the methodological and interpretative objectives of the paper, while the third follows as a consequence of the proposed framework.

The remainder of the article includes, in the following section, a literature review related to the purpose of this work, followed by the presentation of the design framework, offering the general specification of an economic Q-learning model and the families of models that fit within the proposed architecture. A demonstrative simulation is then described, and finally, the main patterns are reported and their implications discussed in conclusions.

2. Literature Review

The application of RL and related computational methods in Economics has expanded significantly in recent years. A growing body of work applies learning algorithms to problems involving strategic interaction, dynamic pricing, market design, agent-based simulations, and policy analysis. These contributions demonstrate the potential of learning-based approaches to address environments that are analytically complex, dynamic, or characterized by incomplete information. Across these diverse applications, RL is typically employed as a Machine Learning (ML) tool that generates decision rules through iterative interaction with a dynamic and stochastic environment [21]. Models differ in their economic context and algorithmic implementation, but they commonly define states, actions, and rewards in a way that allows agents to adapt over time. In many cases, learning outcomes are evaluated using aggregate indicators such as prices, profits, convergence properties, or welfare measures.

Essentially, the method evaluates sequences of actions [22], whose evaluations are used to improve subsequent action decisions [23]. One of its main advantages is the ability to develop optimal behaviour autonomously and independently, without human intervention, based on a simple modelling framework. Q-learning is a popular algorithm widely used in RL [24] which, together with other related algorithms [25,26], has become widespread [23,27] due to its capacity to shift analysis from static conditions toward adaptive trajectories [28]. Equilibrium, when it appears, ceases to be a postulate and becomes a possible convergence outcome, or a benchmark against which learned dynamics can be contrasted [29,30].

There exist broad literature reviews explaining learning techniques [31] and their use in Economics, Game Theory, and Finance [32,33], which are valuable for situating application domains and the computational meaning of the technique [34,35]. Regarding its use in Economics, Q-learning has proven effective in addressing, among others, optimization challenges in finance [28,36,37], trading [38–40], pricing and resource allocation [41–44], and problems involving simulation or interaction among agents and within markets [45–47]. Despite their diversity, these approaches share a common methodological limitation because the core economic assumptions are frequently embedded implicitly in algorithmic design choices. As a result, learned policies are often treated as technical outputs rather than as economic objects, and deviations from theoretical benchmarks are reported without a systematic economic interpretation. Outside of a limited number of contributions, often presented in conference settings [48,49], there is no established line of work exploring conceptual extensions of RL aimed at constructing a systematic and general framework for economic contexts such as the one proposed here.

In studies examining social dilemmas and repeated games under noise, extensions aimed at stabilizing early learning have been proposed [50]. There are also contributions systematizing connections between learning and computational approaches [51]. However, these and similar approaches do not connect with the general economic framework and show that, as strategic complexity increases, it becomes necessary to work with standards for specifying this economic approach. The literature on Q-learning emphasizes the value of algorithmic neutrality [24,52], leaving aside the structural connection between this computational modelling and its economic content. This lack of an explicit mapping between economic modelling operations and RL design decisions limits comparability across studies and hinders cumulative knowledge building. Without a shared specification and reporting discipline, it becomes difficult to assess whether differences in outcomes reflect substantive economic assumptions or merely implementation choices. The present article addresses this gap by proposing a methodological architecture that makes this correspondence explicit

and operational, thereby enabling systematic interpretation, comparison, and extension of learning-based economic models.

Based on this diagnosis, the following section develops the methodological core of the proposal. Rather than extending existing algorithms, the focus is on establishing a disciplined specification and reporting architecture that makes economic assumptions explicit and comparable across learning-based models.

3. Economic Q-Learning Design Framework

Before presenting the proposed framework, we outline the typical operations involved in the modelling activity performed by Economic Theory in its formalization process. Here, formalization does not refer to the characterization of economic objects such as consumers, producers, firms, or markets, but rather to the set of structural operations that make it possible to analyze the behaviour of these objects, their interactions, and the resulting outcomes.

Economic models first specify an evaluation criterion through which outcomes are assessed, such as utility, profit, cost, or welfare. Once this criterion is defined, Economic Theory specifies the set of feasible actions associated with each type of agent's decision, together with the constraints and frictions that restrict choice. These include income and budget constraints, technological limitations, and institutional rules, which may generate adjustment costs or capacity limits for the agents under consideration.

Subsequently, Economic Theory characterizes the information and temporal structure under which decisions are made, including, for instance, the degree of uncertainty faced by agents or the presence of discounting over time. Decision-making under these assumptions is then embedded in an interaction structure, which may range from individual optimization problems to strategic interaction, aggregation mechanisms, or dynamic systems.

Finally, economic outcomes are evaluated relative to interpretative benchmarks, such as optimal policies, equilibrium allocations, steady states, or counterfactual reference points.

RL implements these same types of structural operations in algorithmic form. Objectives are encoded through reward functions, feasible actions through action sets, and constraints and frictions through state variables and transition dynamics. Information structures are represented via the agent's observation space, while temporal considerations are incorporated through discount factors and episode definitions. Interaction structures can be modelled using different learning architectures, such as single-agent or multi-agent settings.

This conceptual correspondence makes RL a potentially valuable tool for Economic Theory. However, in most applied economic implementations, this correspondence remains implicit: economic assumptions are absorbed into algorithmic design choices without being systematically articulated in a way that connects both domains.

The framework proposed in this article is built precisely around this deficiency. Its purpose is to make explicit the structural correspondence between economic modelling operations and RL design choices. In this way, it provides a methodological approach for specifying, interpreting, and comparing RL-based economic models, rendering their results intelligible as economic objects and facilitating their evaluation within the standard analytical parameters of Economic Theory and Economic Policy.

From this perspective, Economic Theory formalizes economic behaviour through a limited set of structural modelling operations that recur across different fields and applications:

1. Specification of objectives, defining how outcomes are evaluated, such as utility, profit, cost, or welfare;
2. Definition of feasible actions and constraints, determining what agents can do given technologies, budgets, institutional rules, and frictions;

3. Representation of information structures, specifying what is observed, when it is observed, and under what uncertainty;
4. Temporal organization of decisions, including horizons, discounting, irreversibility, and stopping conditions;
5. Embedding of behaviour within interaction structures, ranging from individual decision problems to strategic interaction, aggregation, or market-level dynamics;
6. Evaluation relative to benchmarks, such as optimal policies, equilibrium allocations, steady states, or counterfactual reference points.

RL implements these same operations through algorithmic components. Objectives are encoded via reward functions; feasible actions through action sets; constraints and frictions through state variables and transition dynamics; information structures through the agent's observation space; temporal considerations through discount factors and episode definitions; interaction structures through single-agent or multi-agent learning architectures; and benchmarks through externally defined reference policies or equilibrium concepts.

The proposed framework is organized around this correspondence, making explicit how economic modelling operations are translated into RL design choices.

From this point onward, an architecture is formalized to delimit how the agent's valuation and frictions should be separated and declared to structure a return function. In this way, the information and metrics obtained from learned policies can be interpreted as economic rules that are comparable and contrastable against equilibrium benchmarks.

In its standard formulation, Q-learning is framed within a Markov decision process, where the agent observes a state s_t, chooses an action a_t, receives an immediate return r_{t+1}, and transitions to a new state s_{t+1}. The central object is the function $Q(s, a)$, which approximates the expected value of taking action a in state s and continuing with a policy that maximizes future returns. The typical incremental update is expressed through a Bellman-type rule, with learning rate α and discount factor γ:

$$Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') \right], \tag{1}$$

where the term in brackets acts as a temporal target that combines the immediate return and the best estimated future value [24,29,30]. Naturally, extensions can be introduced depending on parameters such as dimensionality or the environment [21,32,53–55], although here we present the minimal basic Q-learning model.

Implementing the algorithm requires specifying an exploration rule, for example ε-greedy, which controls the trade-off between exploiting the policy induced by Q and continuing to generate information to improve it. This dimension is not a minor technical detail in the economic context addressed in this work, because it affects model stability, convergence speed, and the persistence of results. From a methodological standpoint, it is essential to translate the model reward into an economically interpretable objective where the factor γ can be treated as a representation of time preferences and not merely as a technical parameter [24,30].

From here onward, we present the set of possible decisions or design dimensions that must necessarily be considered when constructing an economic model based on Q-learning. In practice, these dimensions correspond to the core objects of the RL framework, including state, actions, objective signal, discounting, and exploration, but formulated in terms of standard Economic Theory models, expressed as preferences, constraints, frictions, market structures, or information structures. These main modelling dimensions are presented below.

3.1. Design Dimensions

3.1.1. Agent Preferences and Behavioural Foundations

In Economics, agents' behaviour is modelled in relation to objectives, typically in the form of a utility function subject to a set of constraints. The analogous operational concept in Q-learning models is a criterion guided by an updating process that operates simply through the evaluation of a generic reward that does not inherently require structure. Payment outcomes guide future actions that correct past errors. However, if everything is reduced to payment evaluation, equilibria will emerge detached from economic interpretation, forcing the exercise to be repeated for each new proposal. For this reason, the first design dimension consists of fixing which preference components are to be represented and how they enter the agent's objective.

To proceed along this line, at least three families of elements must be addressed. The first refers to risk preferences. In economic applications, it is common to distinguish agents with different degrees of risk aversion, which can be represented through transformations of the immediate reward. This implies an explicit formulation of utility over payments, profits, or received consumption. The methodological point is not to choose a particular functional form, but to require the model to make explicit whether the agent maximizes expected payments, expected utility, or combinations that may incorporate penalties.

Second, intertemporal preferences must be made explicit. In Q-learning, the discount factor is usually treated as a hyperparameter, as noted above, but in Economics it must become an object interpreted in terms of patience, time preference for adjustment, or costs associated with waiting. In this sense, model design must establish how this factor, now considered an agent preference, is interpreted in relation to time. It should also be specified whether it differs across agents and whether it depends on the state or type of interaction. This decision is especially relevant in negotiation-based environments, where waiting costs and agents' ability to wait determine their capacity to reach different outcomes.

Finally, in connection with behavioural foundations, any other relevant factors, such as reputational consequences, reciprocity, or inequality aversion, should be considered if pertinent to the economic phenomenon under study. In general terms, these diverse elements may be represented as part of the agent's objective function or as state variables affecting expected returns. The proposed rule is that, if an additional behavioural dimension of this kind is introduced, it must be made explicit whether it resides in the objective, the state, or the interaction mechanism, adapting the model accordingly.

3.1.2. Frictions and Institutional Factors

The second major dimension must incorporate costs in the form of frictions, together with institutional elements that alter incentives and that, in Applied Economics, are decisive in explaining deviations from equilibrium predictions, whether in Economic Theory or Game Theory.

Here it is useful to distinguish between frictions that affect action implementation and those referring to adjustment processes. The former includes technological, regulatory, or contractual constraints that limit the set of available actions in certain states. For example, this includes production capacity limits, budget constraints, market operation rules such as price floors, or compliance costs. In a Q-learning model, these frictions are incorporated by restricting the action space or altering environmental dynamics.

Regarding the second type of frictions, these do not refer to penalties from prolonged negotiations, which should be included above as intertemporal preferences, but to costs penalizing changes or delays. These include adjustment costs, search costs, or transaction costs. The methodological point is to avoid introducing them as arbitrary penalties and instead treat them as a structured part of the objective, making explicit the components

related to utility or profit and those based on costs or frictions. This separation is essential for correct economic interpretation.

Finally, there exists an institutional level that cannot be ignored, referring to rules determining action or interaction protocols, such as contract forms, commitment enforcement, observability of actions, or agreement closure mechanisms. In many contexts, the economic game is not defined solely by a payoff function, but also by institutional structures regulating sequence, process, or available information. Therefore, the proposed design must explicitly declare the institutional protocol, as it conditions both incentives and learning dynamics.

3.1.3. Classes of Economic Environments

The third dimension involved in this article's purpose refers to the class of environment represented by the model. Without enumerating applications or algorithms, we identify families of environments for which state design, actions, and objectives change systematically.

First, it is necessary to distinguish competitive market conditions, where the agent responds to signals essentially in the form of prices or quantities. A different class comprises oligopolistic environments, characterized by strategic competition, where returns depend on other agents' actions and on whether interaction is repeated. Bilateral negotiation or contract environments, involving sequencing, time costs, and agreement breakdown, must also be delimited here. Auction-based environments should also be specified, where outcomes depend on allocation rules involving payments. This category further includes search and matching environments, where relevant decisions include accepting, rejecting, waiting, or searching, and where the state incorporates partial information and expectations.

The value of this classification may seem basic, but its inclusion is necessary to explain which learning assumptions apply in each case. Moreover, it allows connecting learned outcomes with known theoretical benchmarks, such as competitive equilibrium or Nash equilibrium, without requiring dynamics to necessarily reproduce those benchmarks.

3.1.4. Information Structure and State Representation

This dimension concerns information structure. As is well known, in Economics, available information determines plausible expectations. In a Q-learning model, it determines the state and information content on which learning occurs. Here, it is necessary to delimit which variables are observable by an agent in each period, such as prices, past actions, public signals, outcomes, or own action histories. It must also be considered whether observation is perfect or noisy, and whether noise is exogenous or institutionally determined. If reputational components must enter the state, such as compliance history or quality signals, they should be included. Finally, it may be specified whether the state representation is strictly Markovian or summarizes history, as often occurs in economic environments.

This dimension is also decisive for interpretation, as changing what the agent can observe determines the nature of the learning process, its stability, and its relation to equilibrium predictions based on full information.

3.1.5. Learning and Exploration Mechanism

It is not appropriate to provide a complete description of relevant realities without introducing one referring to the learning mechanism itself. Although the article focuses on basic Q-learning models, even within this framework there are important decisions to consider, such as action-space granularity, discretization of continuous variables, heterogeneity in learning parameters, and agent independence.

Exploration is not a minor technical detail in economic contexts, where it can be interpreted as innovation, experimentation, or strategic testing, and has direct effects

on convergence or persistence of off-equilibrium results. This delimits patterns such as regularity, rigidity, or dispersion. Likewise, allowing heterogeneity in learning or exploration rates, reflecting differences in adaptability, attention, or learning capacity, introduces a behavioural factor with direct economic interpretation.

In multi-agent environments, it must be made explicit whether learning is modelled independently, where each agent treats the environment as given, or whether components reflecting anticipation or response to other agents are incorporated. Declaring these assumptions affects economic interpretation of outcomes.

3.1.6. Evaluation Metrics and Comparison with Equilibrium Benchmarks

In Economics, it is not sufficient to observe algorithm convergence; it is also necessary to check whether learned outcomes approximate theoretical benchmarks and to specify what constitutes a relevant result. From an economic perspective, it is relevant to detect whether outcomes are efficient, whether welfare losses arise, whether stability or excessive volatility emerges, and whether stabilization occurs at a stationary state or in cycles.

To incorporate this dimension, three evaluation planes are proposed. The first is dynamic, measuring learning speed, stability, and exploration sensitivity. The second is economic, referring to efficiency, possible surpluses, distributions, and rigidity or dispersion. Finally, an interpretative plane is needed to examine the relationship with theoretical equilibrium. Specifically, it seeks to identify the extent to which equilibrium acts as an attractor and when dynamic outcomes reveal the importance of frictions or limits relative to that equilibrium. This aspect is crucial, especially in evaluations conducted in spaces where outcomes do not sharply delimit equilibria, whose non-attainment seriously compromises overall assessment and model validity.

3.2. Economic Specification of the Objective and Its Translation into Q-Learning

As seen above, a key methodological point is to clearly separate two levels that, in much of the applied literature, tend to be mixed. On the one hand, there is the algorithm, which defines a general rule for updating action values in a sequential decision problem. On the other hand, there is the economic specification. It is therefore necessary to establish an explicit correspondence between both levels, so that links can be drawn between the components and outcomes of the learning process and the terms of the underlying economic foundations [24,30].

In an economic model, the agent decides under constraints and with an objective that is interpreted as utility or net profit, often affected by institutional frictions and adjustment costs. To reflect this structure in Q-learning, we propose that the immediate return driving the update be expressed as an economically interpretable object, separating a benefit-utility component from a cost-friction component. A general formulation, sufficiently flexible for different contexts, is to write the return as

$$r_{i,t+1} = u_i(\pi_{i,t+1}) - c_i(s_t, a_t; \phi_i), \tag{2}$$

where $\pi_{i,t+1}$ denotes the relevant economic payoff of the period (profit, consumption, surplus, or contractual outcome), $u_i(\cdot)$ encodes the agent's valuation, allowing for heterogeneity in risk attitudes, and $c_i(\cdot)$ collects state, and action, dependent costs and frictions. These include adjustment costs, transaction costs, delays, regulatory penalties, or information processing costs discussed above. This formulation does not introduce a new algorithm; rather, it constitutes an accounting discipline that identifies which part of the return responds to preference fundamentals, and which part reflects frictions and institutions. This facilitates comparison across models because it transparently fixes which economic assumptions have been incorporated into the objective [24].

Definition 1. *The term $r_{i,t+1}$ denotes the agent's immediate return at $t+1$. In this article, it is interpreted as an economically meaningful object, typically expressed as utility or net profit, constructed from a valuation component $u_i(\pi_{i,t+1})$ and a cost/friction component $c_i(s_t, a_t; \phi_i)$. The return explicitly separates economic valuation from penalties associated with constraints, institutional frictions, or adjustment costs.*

The interpretation of discounting also requires an explicit treatment. This interpretation connects with the tradition of discounted value typical of economic analysis [56,57]. Thus, the discount factor γ is associated with patience, understood as the agent's cost of waiting. Of course, the model allows for heterogeneity across agents.

Remark 1. *The factor γ_i is interpreted, when explicitly stated in the model specification, as agent i time preference parameter, that is, as an operational representation of patience or impatience. In applied implementations, it may also be treated as a technical parameter for stability and effective horizon. This distinction is methodologically relevant, because it determines whether γ_i should be calibrated as a preference parameter rather than merely as an algorithmic hyperparameter.*

Under this specification, the Q-learning update preserves its form, but the signal driving it is now anchored in an explicit economic objective. For agent i, the update can be written as

$$Q_{i,t+1}(s_t, a_t) = (1 - \alpha_i) Q_{i,t}(s_t, a_t) + \alpha_i [r_{i,t+1} + \gamma_i \max_{a'} Q_{i,t}(s_{t+1}, a')], \tag{3}$$

with $r_{i,t+1}$ defined in (2). Equation (3) shows where the desired economic interpretation appears: not in the algebraic structure of the update, but in the definition of the immediate return, in the interpretation of discounting, and in the effective constraints on the action space. This relationship between discounting, return, and Bellman equations directly connects with the standard formulation of dynamic economic problems, where discounting appears as a preference parameter and the solution is characterized recursively [30].

The methodological advantage of this approach is twofold. First, the model becomes legible in economic terms, because frictions are not absorbed into an opaque or generic reward, and discounting is not reduced to a hyperparameter without interpretation. Second, comparison with equilibrium benchmarks becomes operational, because the return is defined in units of utility or net profit, and learned outcomes can be evaluated as convergence toward a theoretical reference. In what follows, the proposed architecture organizes precisely these decisions, with objective, frictions, environment, information, learning mechanism, and metrics, as explicit design dimensions [24,30].

3.3. Operational Specification of Implementation and Parameter Reporting

The proposed framework can be understood both as a decision template and as a reporting template. Table 1 summarizes these elements as a minimum reporting list establishing correspondences across dimensions so that an economic Q-learning model is interpretable, replicable, and comparable with results from other applications.

Table 1. Minimum reporting template for economic Q-learning models.

Dimension	What Must Be Specified	Economic Interpretation	Typical Options (Illustrative)
Agent objective and payoff primitive	Definition of the period payoff $\pi_{i,t+1}$ and its mapping into the RL return	What the agent is effectively maximizing in economic terms	Profit, net utility, surplus, contract outcome
Return decomposition	Explicit separation of valuation and frictions in $r_{i,t+1} = u_i(\pi_{i,t+1}) - c_i(s_t, a_t; \phi_i)$	Distinguishes preferences from institutional or adjustment effects	Risk curvature in $u_i(\cdot)$, menu costs, transaction costs, regulatory penalties in $c_i(\cdot)$
Risk attitudes	Functional form or transformation encoding risk sensitivity (if included)	Heterogeneity in risk aversion, variability penalization	Concavity in u_i, variance penalty, downside-sensitive terms
Time preferences	Interpretation and value of γ_i, and whether heterogeneous	Patience, impatience, cost of waiting, effective horizon	Constant γ, heterogeneous γ_i, state-dependent discount (if justified)
Feasible actions	Definition of the action space $A(s)$ and constraints	Technological, budgetary, regulatory feasibility	Capacity constraints, price grids, production bounds
Environment class	Market/interaction family and protocol	Defines the institution generating incentives and timing	Competitive market, oligopoly, bargaining, auction, search/matching
Timing and protocol	Sequence of moves, stopping rules, update frequency	Institutional microstructure and strategic timing	Simultaneous vs. sequential moves, negotiation rounds, clearing rule
Information structure	What is observed at decision time; noise and observability	Available information, signals, and inference burden	Full observability, partial observability, public signals, noisy prices
State representation S	Variables constituting s_t; Markov justification or history summary	What the agent knows operationally when acting	Current price/quantity, history summaries, reputation states
Multi-agent specification	Number of agents, symmetry/asymmetry, interaction topology	Strategic interdependence and non-stationarity	Independent learners, shared environment, heterogeneous types
Learning parameters	α_i, exploration rule, update schedule	Learning speed, experimentation, behavioural interpretation	ε-greedy, decaying ε, fixed α, decaying α
Convergence and stability criteria	What counts as stability; stopping conditions	Learning steady state as economic outcome	Stationary policy, cycles, persistent heterogeneity, convergence rate
Economic evaluation metrics	Welfare/efficiency measures and distributional outcomes	Economic meaning of learned outcomes	Surplus, profits, price dispersion, rigidity, volatility
Benchmark comparison	Which theoretical benchmark is used (if any)	Equilibrium as reference, not as imposed solution	Nash/Cournot, competitive equilibrium, bargaining solution, auction benchmark
Reproducibility	Seeds, code availability, parameter disclosure	Replicability and auditability of claims	Fixed random seeds, full parameter table, code repository

The table does not aim to restrict the space of models, but rather to prevent substantively relevant decisions for economic interpretation from being concealed or absorbed by technical parameters. By requiring the immediate return to be expressed as an explicit economic object and discounting to be declared in interpretative terms, the framework allows the dynamics to be read as an adaptive decision procedure comparable to standard equilibrium references. In this way, equilibrium no longer appears as a starting assumption, as in classical approaches to economic and game theory, but as a reference point for

comparisons that allow both improvement of foundations and identification of specific features in frictions or behavioural dimensions that alter effective comparative statics.

On the other hand, the applicability of the framework depends on the design being operational in concrete implementations. Given the methodological purpose of the article, the appropriate implementation level is one that illustrates the architecture with minimal and reproducible ingredients. However, in economic models, design decisions are not merely technical, as they condition the type of behaviour emerging from the process and therefore affect economic interpretation. Essentially, the main requirements when carrying out this exercise can be summarized as studying how continuous variables are discretized and how state representations are constructed. Additionally, one must define the feasible action space and determine which criteria are used to stabilize learning dynamics.

The above elements are sufficient for tabular Q-learning versions that converge to the optimum under standard conditions and appropriate learning rates [30,58]. However, it should be noted that with other functional approximations (e.g., DQN), stability is not guaranteed, reinforcing the need to report parameters and stopping criteria. Discretization of prices, quantities, types, or signals must be reported as part of the design, since grid granularity defines the available decision set and thus the approximation degree to the underlying economic problem. Similarly, state representation must be fully described, specifying which variables are observable, how memory elements are incorporated when interactions depend on history, and which transformations or normalizations are applied. The framework therefore recommends explicitly stating learning parameters such as rates, exploration rules and decay, stopping criteria, and the number of simulated periods or episodes. Likewise, randomness should be controlled through fixed seeds to ensure reproducibility and to separate economic effects from computational noise.

To clarify how the proposed methodological architecture can be applied in practice, the sequence of modelling decisions that an economist would follow to systematically implement an RL-based economic model under the proposed framework is illustrated. If, for example, the economist wants to model a price problem with demand uncertainty and adjustment frictions, applying the proposed architecture requires the following decisions to be explicit and reportable:

1. The economic objective must be specified. The researcher defines whether the agent maximizes benefits, expected utility, welfare, or another evaluation criterion. Under the proposed framework, this objective is explicitly encoded in the reward function, with a clear distinction between gross benefits and the costs or penalties introduced by institutional frictions;
2. The feasible set of actions must be declared. In economic terms, this corresponds to permissible prices or quantities. In the learning model, this choice determines the action space and its discretization, which is reported as an economic modelling decision and not as a purely technical choice;
3. Constraints and frictions must be identified. Adjustment costs, capacity limits, or institutional rigidities are introduced as explicit components of the state space and transition dynamics. This ensures that frictions affecting economic behaviour are not implicitly absorbed into the algorithmic adjustment parameters;
4. Next, the information and time structure inherent to decision-making must be established. The researcher specifies what the agent observes at the time of the choice and how future outcomes are valued. These assumptions are assigned to the observation space and the discount factor, and the interpretation of the discount is explicitly stated as economic time preference where appropriate;
5. Subsequently, the interaction structure must be defined. The framework allows the researcher to determine whether the problem is single-agent or multi-agent, and

whether strategic interaction exists. This choice determines the learning architecture but does not alter the reporting discipline imposed by the framework;

6. Finally, an interpretive reference point must be specified. Instead of imposing itself as a solution, the static or dynamic equilibrium reference point is introduced as a contrasting element that allows for the evaluation of learned policies and the interpretation of the economic implications of possible deviations. Following this sequence, the proposed architecture organizes the transition from economic modelling assumptions to RL design options. The result is that learned policies, stability properties, and deviations from theoretical equilibria can be reported as objects of economic learning.

This section does not aim to impose an adjustment strategy, but to delimit elements that allow distinguishing between results reflecting the economic structure of the environment and those depending on implementation choices or exploration regimes. Similarly, when scenarios are compared, it is advisable to declare whether changes are introduced along a single design dimension or across several simultaneously, to preserve causal interpretation in sensitivity analyses.

Finally, we stress that these simulations can be used as controlled exercises to determine how modifications in preferences or frictions, holding everything else constant, alter learning trajectories and their relationship with the benchmark implied by economic fundamentals.

In summary, the previous sections have delimited a disciplined transition between economic foundations, like preferences, frictions, institutional factors and information, and the components of a Q-learning model. This makes it possible to document design decisions often left implicit and, at the same time, to interpret learned outcomes against equilibrium references without imposing them as starting assumptions. To show that the framework is operational, the next section introduces a deliberately simple and reproducible demonstrative simulation. Its function is not to offer a concrete application, but to isolate the effect of a modelling dimension, such as a friction or a preference feature, on learning dynamics and their relation to a theoretical benchmark.

4. Methodology: Demonstrative Simulation Design

Because the objective of this article is methodological rather than empirical, the design choices adopted in the simulation are justified according to criteria that differ from those typically used in applied or calibrated economic models. In this context, design decisions are not selected to achieve empirical realism or numerical optimality, but to preserve economic meaning, enable diagnostic interpretation, and make explicit the correspondence between economic modelling operations and RL design choices.

First, discretization levels are chosen to maintain interpretability of states and actions, rather than to approximate a continuous solution with high numerical precision. Coarser discretizations allow learned policies to be represented and analyzed as economically meaningful decision rules, making thresholds, regions of inaction, or qualitative policy differences directly observable. Second, parameter values are selected to isolate and illustrate specific modelling assumptions, rather than to match empirical magnitudes. This facilitates the identification of how preferences, frictions, or temporal assumptions are encoded into the learning process. Third, benchmark definitions are introduced as interpretative reference points, consistent with the role played by equilibrium concepts in Economic Theory, allowing deviations generated by learning dynamics to be evaluated in a structured manner.

Under this perspective, alternative discretizations, parameterizations, or benchmark choices could be adopted without altering the methodological role of the example. Such variations would not change the structural mapping highlighted by the framework, but

could be explored in future work to address substantive economic questions. The simulation should therefore be understood as a controlled methodological exercise designed to demonstrate the interpretative and comparative capabilities of the proposed framework, rather than as a test of empirical validity or predictive performance.

4.1. Environment and Institutional Protocol

The demonstrative simulation presented here has the methodological purpose of showing how the proposed architecture translates concepts from economic fundamentals into an operational Q-learning design, in a way that is both reproducible and economically interpretable.

The environment is defined in discrete time. In each period t, an agent observes a state s_t, chooses an action $a_t \in A$, and the environment generates an economic payoff defined as π_{t+1} and a return given by r_{t+1}, updating the state to the next period. State evolution follows a Markovian transition conditional on (s_t, a_t):

$$s_{t+1} \sim P(\cdot \mid s_t, a_t), \tag{4}$$

and the period economic payoff is determined by an institutional rule and an exogenous shock:

$$\pi_{t+1} = \Pi(s_t, a_t, \varepsilon_{t+1}), \tag{5}$$

where ε_{t+1} captures uncertainty, for example due to fluctuations in demand. The institutional friction of interest in this example is introduced as an explicit adjustment cost associated with changes in actions across periods, so that the environment reflects, in a stylized way, that operational decisions face implementation costs and not only instantaneous incentives.

4.2. State and Action Spaces

The example is particularized as a discrete price-setting problem. The action is the price chosen in each period, $a_t \equiv p_t$, with a finite set of possible prices:

$$A = \left\{ p^1, \ldots, p^K \right\} \tag{6}$$

The state collects sufficient economic information for the decision to be operational and for the adjustment friction to be explicit. In its minimal version, it is defined as the pair formed by a demand state and the previous price:

$$s_t = (z_t, p_{t-1}), \tag{7}$$

where z_t represents a demand factor, or alternatively a market regime, which can be modelled as a discrete variable or as a continuous variable discretized on a grid. For the simulation, the price space A is discretized into a uniform grid of 20 points, while the demand factor z is divided into 10 levels. This granularity is selected to balance approximation accuracy with computational efficiency and stable convergence. This representation makes transparent which information conditions the decision and allows disciplined incorporation of observability restrictions, such as noisy observation of z_t, if desired.

4.3. Economic Return Specification and Parameterization

The period economic payoff is defined as a standard operating profit. For constant marginal cost mc and demand $q(\cdot)$ depending on price and demand state, it is written as

$$\pi_{t+1} = (p_t - mc)\, q(p_t, z_t) \tag{8}$$

Demand can be kept generic at the conceptual level, but for simulation a parsimonious form preserving monotonicity and non-negativity is adopted. A standard choice is, for example, $q(p,z) = \max\{0,\ z - bp\}$, with $b > 0$, together with discretization of z when necessary. The institutional friction is incorporated as a menu-cost type adjustment cost that penalizes price changes:

$$c(s_t, a_t; \phi) = \kappa\, 1\{p_t \neq p_{t-1}\}, \tag{9}$$

where $\kappa \geq 0$ measures friction intensity and ϕ denotes the vector of environment parameters. The immediate return driving learning is defined following the general structure proposed in Section 3, separating valuation and friction:

$$r_{t+1} = u(\pi_{t+1}) - c(s_t, a_t; \phi). \tag{10}$$

In the demonstrative simulation, $u(\cdot)$ is used to introduce, when desired, heterogeneity in risk attitudes. For example, a baseline case is $u(\pi) = \pi$, corresponding to risk neutrality, while a common alternative is a parametrized concave transformation (e.g., Constant Relative Risk Aversion-CRRA) when π is non-negative by construction.

As can be seen, the goal of the example is not to exhaust possible specifications, but to show how the framework forces explicit declaration of which part of the return corresponds to economic valuation and which part corresponds to frictions and institutional constraints.

4.4. Learning Procedure and Evaluation

Under the previous specification, learning is implemented through tabular Q-learning on the discrete space $S \times A$. For each state–action pair, the agent updates action values according to the standard rule

$$Q_{t+1}(s_t, a_t) = (1 - \alpha)\, Q_t(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a' \in A} Q_t(s_{t+1}, a') \right], \tag{11}$$

where $\alpha \in (0, 1]$ is the learning rate and $\gamma \in [0, 1)$ is the discount factor, interpreted as discussed in Section 3. The exploration policy is implemented using an ε-greedy rule: with probability $1 - \varepsilon$, the agent chooses the action that maximizes $Q_t(s_t, \cdot)$, and with probability ε explores alternative actions in A.

To ensure reproducibility, random seeds, grid sizes, number of episodes or periods, and stopping criteria are reported. Specifically, convergence is defined when the maximum variation in the Q-value matrix between two successive episodes falls below a threshold of 0.0001 over a window of 100 episodes, or when the resulting optimal policy remains unchanged over the same period. All reported results correspond to averages over 50 independent runs with fixed random seeds, ensuring that the rigidity and convergence patterns described are robust properties of the model and not artefacts of simulation noise.

Evaluation is organized around metrics connecting the learned outcome with its economic interpretation. First, convergence and stability are reported in terms of stationary policies, stabilization speed, and presence of cycles or persistent heterogeneity. Second, directly interpretable economic metrics are reported, such as average profit levels, price volatility, and a price rigidity measure defined as the frequency of changes $\mathbb{P}(p_t \neq p_{t-1})$ under the learned policy. Third, a benchmark reference is introduced for each demand state, defined as the action that maximizes static payoff without adjustment frictions:

$$p^*(z) \in \arg\max_{p \in A} (p - mc)\, q(p, z), \tag{12}$$

and the learned policy is compared with this reference in terms of action distance, efficiency loss, or systematic deviations induced by friction κ or by curvature in $u(\cdot)$. The sensitivity exercise design keeps all other elements constant and varies one dimension at a time, typically κ or a risk parameter in $u(\cdot)$, to isolate the effect of institutional frictions or behavioural heterogeneity on learning dynamics and on the relationship with the benchmark.

5. Results

We summarize the results of the demonstrative simulation introduced above. This section does not aim to establish an RL-based empirical finding or present an applied case, but rather to demonstrate how the proposed architecture generates a reporting protocol that allows the learning dynamics to be read through categories integrated into the classical narrative of Economics, thereby allowing comparison with the equilibrium benchmarks proposed by Economic Theory. The contribution therefore lies in the type of evidence that can be extracted from the model and in its interpretability. That is, the learned policy is described as an economic object, as a decision rule under frictions, and the metrics are formulated in interpretable units such as profit, volatility, or the mentioned rigidity. For its part, as explained, the static benchmark is used as a form of contrast, not as an imposed solution.

Results should not be interpreted solely in terms of convergence or numerical performance. Under the proposed framework, the learning process itself constitutes an economically meaningful object. Specifically, the learned policy allows us to characterize how agents adjust their decisions across states during the learning phase, revealing patterns such as regions of inaction, discrete adjustment thresholds, and path-dependent responses. These features are not visible when attention is restricted to aggregate outcomes or terminal values but emerge clearly once policies are explicitly reported and interpreted as economic decision rules.

A central contribution of the framework is to discipline the comparison between learned outcomes and theoretical benchmarks. This is achieved by aligning benchmark comparisons with the structure of the learned policy, rather than with aggregate outcome levels alone. Rather than treating equilibrium solutions as numerical reference points evaluated ex post, the framework allows deviations to be interpreted structurally. In the simulation, deviations from the benchmark equilibrium are not uniform across the state space but concentrated in specific regions associated with learning frictions and adjustment costs. This makes it possible to distinguish between transient deviations arising during exploration and persistent deviations that reflect economically meaningful constraints embedded in the model design and therefore interpretable within standard economic categories of frictions and adjustment costs.

An additional insight revealed by the framework concerns the non-uniqueness of aggregate outcomes. The simulation shows that similar aggregate indicators, such as average prices or profits, may be generated by qualitatively different learned policies. In this sense, policies that differ in the timing and magnitude of adjustments across states can produce comparable averages while implying markedly different dynamic responses to shocks or deviations. From an economic perspective, this implies distinct adjustment mechanisms and strategic responses, despite producing comparable summary statistics. This observation highlights a limitation of result reporting based exclusively on aggregates and illustrates how policy-level analysis is necessary to recover the underlying economic behaviour.

To further illustrate the interpretative advantages of the proposed reporting framework, Figures 1 and 2 offer visual representations of the policy-level relationships that underpin the analysis developed in this section. These figures do not introduce additional simulation results, nor are they based on a separate calibration exercise. Rather, they

provide stylized depictions constructed from average and representative values of the model's variables, with the aim of making explicit how the proposed framework enables economically meaningful relationships to be visualized and interpreted.

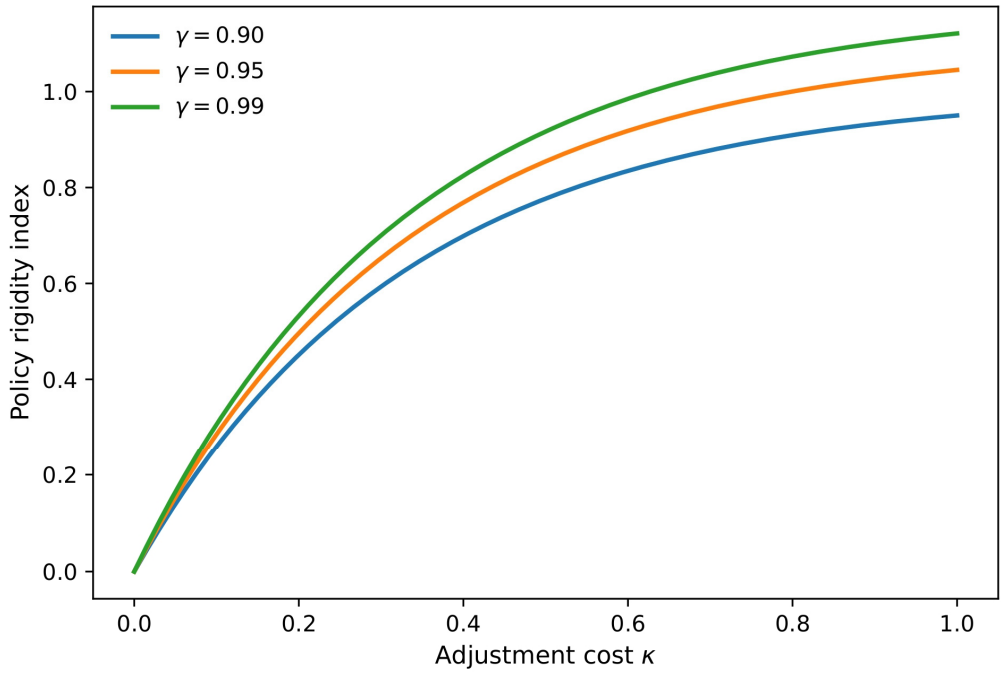


Figure 1. Interaction between adjustment costs and time preferences. The figure is illustrative and not based on calibrated data. Its purpose is to visually clarify the structural implication of the methodological framework.

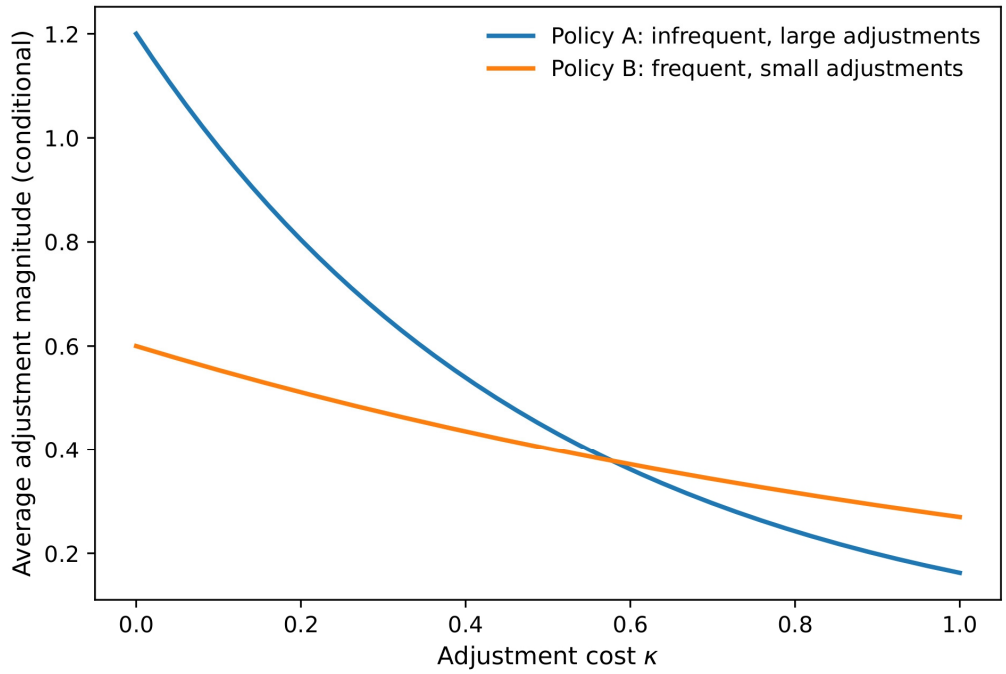


Figure 2. Distinct adjustment regimes behind similar aggregates. The figure is illustrative and not based on calibrated data. Its purpose is to visually clarify the structural implication of the methodological framework.

Figure 1 illustrates the interaction between adjustment costs and time preferences as they are encoded in the learning-based model. The figure shows how higher adjustment costs translate into greater policy rigidity, and how this effect is amplified by stronger time

preferences (higher discount factors). Although the precise functional form is illustrative, the figure reflects the structural role played by these dimensions in the model: adjustment costs discipline the frequency and size of policy changes, while time preferences shape the persistence of learned decision rules.

Figure 2 illustrates a central insight emphasized by the proposed framework: the non-uniqueness of aggregate outcomes. The figure contrasts two stylized adjustment regimes that generate similar aggregate indicators but differ markedly in the timing and magnitude of adjustments. One regime is characterized by infrequent but large adjustments, while the other exhibits more frequent but smaller adjustments. From an economic perspective, these regimes correspond to distinct adjustment mechanisms and dynamic responses to shocks, despite producing comparable averages.

For these reasons, the simulation should be understood as a diagnostic exercise rather than as a validation test of a specific economic hypothesis. Its role is to demonstrate how the proposed framework uncovers economically relevant dimensions of learning dynamics that would otherwise remain hidden. The objective is not to claim general empirical validity, but to show how design choices, benchmarks, and policy representations interact in shaping economic interpretations.

To maintain traceability, a baseline configuration of the environment and the learning procedure is defined, and sensitivity exercises are developed that vary one dimension at a time. In particular, it is recalled that the adjustment cost κ is used as an explicit institutional friction and, when appropriate, the specification of $u(\cdot)$ is varied, or the discount parameter γ when it is interpreted as the described time preference. This proposed organization thus makes it possible to separate institutional effects from behavioural effects, and makes visible, in operational terms, the link between economic assumptions and learning dynamics.

Table 2 summarizes the reporting protocol and the one-factor sensitivity design used in this section to interpret, in economic terms, the learned policy and its implications in the demonstrative simulation.

Table 2. Reporting protocol and sensitivity design for the demonstrative simulation.

Item	What Is Reported (Object)	How It Is Constructed	What It Diagnoses Economically	Benchmark/Reference
$\mu(s)$ over the state grid	Extract the greedy policy from $Q(s,a)$ after stabilization; represent as mapping $s_t = (z_t, p_{t-1}) \mapsto p_t$	Decision rule implied by the model, dependence on history, presence of inaction regions or thresholds	Compared to static benchmark $p^*(z)$ state-by-state	$\mu(s)$ over the state grid
Stability/learning outcome	Policy stability and learning stability	Report stabilization criterion (policy unchanged over a window; or ΔQ below tolerance); document possible cycles/alternation	Whether outcomes are stationary, cyclical, or path-dependent under the declared frictions and preferences	Stability is not assumed; it is an outcome assessed against a stated criterion
Economic performance (gross vs. net)	π_{t+1} and r_{t+1} summaries	Compute gross payoff π and net return $r = u(\pi) - c(\cdot)$; report them separately	Separates operational performance from institutional/adjustment burden, prevents "reward opacity"	Static payoff under $p^*(z)$ without adjustment cost provides a clean reference

Table 2. *Cont.*

Item	What Is Reported (Object)	How It Is Constructed	What It Diagnoses Economically	Benchmark/Reference
Price rigidity/adjustment frequency	$\mathbb{P}(p_t \neq p_{t-1})$	Frequency of action changes under the learned policy, conditional and unconditional on z	Operational rigidity induced by adjustment frictions; identifies where adjustment becomes costly	Compared across κ values and against the $\kappa = 0$ baseline
Dispersion/variability of decisions	Dispersion of p_t and state-conditional variation	Measure variability of actions across time or across demand states; report conditional patterns	Diagnoses whether the learned policy yields stable rules, switching, or heterogeneity across regimes	Interpreted relative to benchmark mapping $p^*(z)$ and to the baseline case
Benchmark deviation	Distance between $\mu(s)$ and $p^*(z)$	For each z, compute deviation of chosen action from $p^*(z)$; aggregate by states visited	Quantifies how frictions/preferences shift effective comparative statics away from the static solution	$p^*(z) \in \arg\max_{p \in \mathcal{A}}(p - mc)q(p, z)$
Sensitivity: institutional friction	Variation in κ	One-factor sensitivity holding all else constant	Isolates institutional/implementation effects on policy and stability	Baseline $\kappa = 0$ and incremental $\kappa > 0$ comparisons
Sensitivity: risk preferences	Alternative $u(\cdot)$ or risk parameter	Compare linear valuation vs. concave valuation under identical environment	Isolates preference-based differences from environmental change; keeps interpretation within the objective	Baseline $u(\pi) = \pi$ as reference specification
Sensitivity: time preferences	Variation in γ (only when interpreted as patience)	Keep algorithmic settings fixed; interpret γ explicitly as preference when used	Distinguishes economic time preference from purely technical discount tuning	Baseline γ declared; alternative γ scenarios reported with the interpretation stated

5.1. Learned Policy, Stability, and Traceability of the Outcome

The first result to be reported is the policy induced by learning. Given that the state is defined as $s_t = (z_t, p_{t-1})$, the learned policy can be presented as a correspondence $p_t = \mu(z_t, p_{t-1})$ over the state grid, which provides an immediate reading of the role of history (through p_{t-1}) and of the demand state (through z_t). In a methodological exercise, this object is more informative than observing isolated metrics, because it allows one to observe directly how frictions and preferences translate into decision rules, and to observe whether learning induces path dependence or adjustment thresholds.

The second result refers to stability. Instead of treating convergence as an event or a binary fact, the operational criteria are reported:

1. Stabilization of policy μ under a defined tolerance;
2. Evolution of the average magnitude of Q updates;
3. Persistence of cycles or alternations of actions in recurrent states.

This reporting format avoids extrapolating theoretical guarantees beyond their domain and allows the dynamics to be described in terms that the economic narrative can interpret as stable rules, persistent fluctuations, or sensitivity to incomplete information and exploration.

Finally, the traceability of the policy is interpreted with respect to the explicit economic components of the return. Given that the return is defined as $r_{t+1} = u(\pi_{t+1}) - c(s_t, a_t; \phi)$, the economic reading of the learned policy in this case depends on whether deviations from the benchmark are explained by the friction component $c(\cdot)$, by the curvature of

$u(\cdot)$, or by the weight on the future controlled by γ. This decomposition is precisely the methodological advantage of the proposed architecture: the learned outcome is not reduced to the functioning of the network or to algorithmic convergence but is presented as a consequence of the declared economic assumptions, to which it refers.

5.2. Comparison with the Static Benchmark and Interpretable Economic Metrics

In addition to dynamic metrics, the gross payoff π and the net return $r = u(\pi) - c$, should be reported separately, as well as a possible welfare measure, defined as the sum of returns. This facilitates normative interpretation and above all comparison across scenarios. The explicit separation in reporting between the gross payoff and the net return makes it possible to identify what part of the detected inefficiency stems from market dynamics and what part is a direct consequence of the modelled institutional friction. This distinction is key to avoiding biases in the assessment of economic welfare.

The central comparison is built with the static reference defined in (12), $p^*(z) \in \arg\max_{p \in \mathcal{A}}(p - mc)q(p, z)$, which operates as a benchmark by demand state when adjustment costs are ignored. The purpose is not to evaluate accuracy in predictive terms, but to quantify to what extent the learned policy coincides with the reference in the absence of frictions, and how it departs from it when $\kappa > 0$ is introduced or when the valuation $u(\cdot)$ incorporates some sensitivity to risk. As can be seen, the interpretive field refers back to familiar Economic Theory, which is the main objective of such exercises when one wants to frame them within economic discourse.

Reporting is organized around three metrics with direct economic meaning. First, average performance in terms of π and r, distinguishing the operating payoff and the net return in the presence of frictions. Second, reporting should include the price rigidity metric discussed above, defined as the frequency of change $\mathbb{P}(p_t \neq p_{t-1})$. This allows one to measure transparently the disciplined effect of the adjustment cost. Finally, measures of dispersion or volatility of the action p_t should be introduced, capturing whether the policy induces a pattern of smooth adaptation, discrete jumps, or alternations, and in which states of the environment they appear.

In this framework, the evidence of interest is not about global quantitative improvement, but about qualitative structure: if the learned policy replicates the benchmark when friction is null or negligible, learning is consistent with the equilibrium prediction as a plausible approximation under that characterization. If, upon introducing frictions, the policy departs from $p^*(z)$ systematically and that departure is explained by the term $c(\cdot)$ or by the valuation $u(\cdot)$, the divergence is interpreted as an informative economic result, not as a methodological anomaly. Again, comparison with the benchmark is used as a structural contrast to identify which economic assumptions shift effective comparative statics when the decision is treated as learning.

5.3. Sensitivity by Dimensions of the Architecture: Frictions, Risk, and Time

The methodological value of the demonstration is consolidated when one dimension is varied at a time, holding constant the remaining elements of the environment and the algorithm. First, κ is varied as an institutional friction. This exercise makes it possible to document how the adjustment cost alters the frequency of changes and the distance from the benchmark, and whether it induces path dependence in the operational sense that the learned optimal action depends on the previous price even when the demand state is the same. In reporting terms, the relevant result is whether the friction generates inaction regions in the state space, that is, adjustment thresholds or intermittent adjustment patterns, because these objects connect directly with economic interpretations of rigidity and implementation costs.

Second, when a non-linear valuation $u(\cdot)$ is activated, the sensitivity exercise is organized to show that the difference from the baseline case is not due to changes in the environment, but to changes in the agent's objective function, that is, in the way outcomes are evaluated. Here the methodological requirement is not to confuse risk aversion with mere technical penalization: the term capturing risk sensitivity must be declared within $u(\cdot)$ and must not be absorbed into an ad hoc reward adjustment. The result to be reported, therefore, must be the explanation of how this dimension modifies the learned policy and the dispersion metrics of actions and payoffs, and to what extent it alters the relationship with the benchmark.

Finally, the discount γ can be considered a time-preference parameter, when declared as such, or a technical parameter defining an effective horizon. Sensitivity with respect to γ acquires economic meaning only under the first reading. It is essential to distinguish that, whereas for algorithmic stability a baseline technical discount value is maintained, in economic sensitivity exercises this parameter is calibrated exclusively to model the agent's time preference, thereby allowing the observed variations in the policy to be attributed to the decision-maker's patience and not to numerical instabilities. Therefore, when heterogeneity in γ or variations of γ are explored, reporting must keep the adopted interpretation clear and separate the economic discussion of patience and agents' cost of waiting from numerical considerations of stability or learning speed.

5.4. Scope and Methodological Contribution

Ultimately, comparison with a static benchmark is now a contrast to identify which dimensions shift comparative statics when the decision is modelled as a form of learning. This proposed procedure thus also offers a tool for theory: it makes it possible to verify when classical conditions sustain equilibrium and when specific parameters reconfigure it. Thus, computation does not replace theory but complements it with a systematic exploration of its margins of validity. The contribution is thus located in the specification architecture and in the reporting discipline, which separates economic return, frictions, and the reading of discounting, and produces comparable objects across models, policy, stability, and economic metrics, under controlled variations in assumptions. The reporting protocol and explicit parameter declaration guarantee reproducibility and the critical condition for results to be auditable and cumulative in the literature.

The environment employed is deliberately parsimonious and discretized, and the use of tabular Q-learning responds to the demonstrative purpose and interpretive clarity. Of course, extensions to multi-agent environments, continuous spaces, or functional approximations can be made, but the interest lies in maintaining the explicit link between economic assumptions and the design of the learning process described here.

Overall, the exercises in the previous subsections show that the proposed architecture not only produces results that are legible in economic terms, but also organizes systematically the relationship between assumptions, learning dynamics, and comparison with equilibrium references.

After illustrating how learning is interpreted on the basis of economic fundamentals, the next section synthesizes the future scope that this type of modelling has for the future development of computational Economics based on this line.

6. Discussion

The results reported allow us to revisit the broader claims made in the Introduction regarding comparability, cumulative knowledge, and the refinement of Economic Theory. The demonstrative simulation shows that these claims can be operationalized once RL models are disciplined by an explicit reporting architecture. As shown in Section 5, policies

yielding similar average prices or profits can differ markedly in how adjustments are implemented across the state space. Particularly, some learned policies concentrate price adjustments in a narrow set of states associated with large deviations from the benchmark, while others distribute smaller adjustments more evenly across states. These differences imply distinct economic behaviours, such as higher rigidity with occasional sharp corrections versus smoother but more persistent adjustment paths, despite producing comparable aggregate outcomes [5–7].

This policy-level distinction is consistent with evidence from the price-adjustment/menu-cost literature showing that rigidity can coexist with episodic, state-dependent corrections, so aggregate statistics alone may mask economically different adjustment mechanisms [59]. By making these policy-level differences observable, the proposed framework establishes a common analytical basis for comparing learning-based economic models beyond numerical performance. This directly addresses the problem identified in the Introduction: aggregate convergence alone is insufficient to assess economic behaviour when learning dynamics and frictions are present.

Taken together, the reported results show that learning-based models can generate economically distinct adjustment regimes that are observationally equivalent at the aggregate level. By making the learned policy explicit and contrasting it with a benchmark, the framework reveals whether deviations arise from exploration dynamics, institutional frictions, or preference curvature [60]. This diagnostic reading also connects with learning-in-games results showing that Q-learning dynamics can converge to stable behaviours that need not be characterized by aggregate outcomes alone, reinforcing the need for policy-level objects and stability reporting [60]. This diagnostic capability is precisely what enables meaningful comparison across models and supports the cumulative use of RL in economic analysis. This distinction would be obscured under standard aggregate reporting, but becomes explicit once policies are treated as primary economic objects.

Relative to existing applications of RL in Economics, which typically emphasize aggregate outcomes or convergence properties, the proposed framework shifts the focus toward policy-level structure as the primary object of economic interpretation. This perspective aligns with, but goes beyond, strands of the literature on learning in games, agent-based modelling, and ad hoc RL applications, where learned behaviour is often evaluated ex post without a unified reporting discipline [5–7].

This discussion should likewise draw attention to the limitations of the present contribution, and a clear one in this regard is that the illustrative simulation is deliberately parsimonious and does not aim to provide empirical validation or assess predictive performance. Moreover, the framework does not eliminate modelling discretion but disciplines it by requiring that economic assumptions be declared explicitly and traced through learning outcomes.

These considerations clarify the scope of the contribution and prepare the ground for Section 7, where the implications of this methodological approach for future research in computational Economics are summarized.

7. Conclusions

This article has developed a methodological proposal for the use of Q-learning in economic modelling and in business systems, aimed at offering a specification and reporting architecture that establishes, in operational terms, which elements and dimensions must be implemented for a learning model to be interpretable in economic terms, comparable across studies, and assessable against theoretical benchmarks.

First, we conclude that the modelling approach we present now allows comparison with Economic Theory to recover its true analytical role, where equilibrium ceases to

operate as an assumption and instead functions as a contrastable benchmark, unlike what currently happens in most approaches to economic and business problems addressed through RL.

Consistent with the above, the models that incorporate the dynamics offered here facilitate the economic interpretation of equilibrium and do so as an approximation to the results of theory. The proposed system makes it possible to understand that, when dynamics converge to stationary rules different from those predicted by Economic Theory, this should not be interpreted as a failure of the procedure, but as structural evidence about which frictions, information restrictions, or behavioural margins shift effective comparative statics.

In addition, it has been shown that the framework proposed here makes it possible to turn the reading of economic modelling using Q-learning into a reproducible procedure, since a reporting protocol is provided that produces comparable objects, learned policy, stability, economic metrics, and distance to benchmarks, under controlled variations in design dimensions.

The demonstrative simulation allows us to draw several conclusions that go beyond the paper's initial methodological motivation. As shown in Section 5, reporting learned policies as economic objects reveals dimensions of behaviour that are not captured by aggregate outcomes alone. Consequently, policies that generate similar prices or profits may in fact correspond to distinct economic regimes, differing in the timing and magnitude of adjustments across states and in their dynamic responses to shocks or deviations.

The results also demonstrate that equilibrium benchmarks can be used as interpretative devices rather than imposed solutions. When deviations from benchmarks are analyzed structurally, they provide information about the role of learning frictions and institutional constraints embedded in the model design. This allows results obtained under different parameterizations or environments to be compared within a shared economic language.

Taken together, these findings support the idea that the framework developed in this article resolves the methodological tension generated by the gap between learning algorithms and the semantic content of economic decision-making. By explicitly linking the learning mechanism's tuning rules to their economic meaning, the proposed architecture clarifies where economic assumptions enter the model and how learned outcomes are shaped. Rather than modifying the algorithms to improve interpretability, this contribution aims to restore semantic discipline at the specification level, allowing the learning dynamics to be interpreted in terms of economically significant objects, not merely as computational results.

The demonstrative simulation provided shows that the architecture can be implemented in RL models with minimal ingredients without losing interpretive capacity. By fixing a parsimonious and discretized environment, it becomes possible to facilitate the entry of economic interpretation in the definition of payoff and return, in the encoding of institutional frictions, and in the interpretation of discounting. The system proposed in this work illustrates how changes in one of the proposed dimensions alter the learned policy and its relationship with a benchmark. In this way, it promotes a cumulative use of learning techniques that, compared to current collections of isolated cases that are difficult to interpret and compare, favours the construction of a family of models integrated into Economic Theory.

The proposed framework is designed to be broad enough to accommodate standard equilibrium environments, where convergence is expected, but it is also valid for environments where learning may not settle on the equilibrium prediction. That is, the proposal is transferable to more complex economic scenarios and provides guidelines for working in multi-agent contexts with behavioural heterogeneity, provided that the corre-

spondence described here between economic assumptions and the design of the learning process is maintained. An immediate agenda will consist of applying this architecture to canonical modelling, such as oligopolies, bargaining, or auctions, to systematically assess how frictions reconfigure the relationship between learned outcomes and equilibrium references. This makes it possible to generate interpretive and transferable comparisons across domains.

Finally, it is noted that the current use of Q-learning in Economics, which advances at an accelerated pace imposed by computational capacity and pressure to solve applied problems, is compatible with a formulation such as the one proposed here, which simultaneously contributes to the development of the conceptual corpus. This reduces the risk of accumulating technical solutions disconnected from one another and misplaced with respect to economic fundamentals. In this way, it becomes possible for rapid computational progress not to erode theoretical coherence, but to contribute to reformulating it, subjecting assumptions to dynamic contrast and generating cumulative evidence on the role of frictions, information, and behaviour in economic outcomes.

Although the contribution of this article is methodological rather than prescriptive, the proposed framework has direct implications for applied economic analysis. By making learned policies explicit and economically interpretable, the framework allows researchers and practitioners to assess how learning-based models respond to institutional constraints, shocks, and policy parameters.

This is particularly relevant in contexts where RL models are used to inform pricing rules, market regulation, or strategic decision-making, as the framework enables policymakers and analysts to distinguish between outcomes driven by structural frictions and those arising from transient learning dynamics. In this sense, the contribution of the framework lies in improving the reliability and interpretability of learning-based analyses before they are used in real-world decision-making or policy evaluation.

With all of this, we move from isolated exercises to a common language that integrates computation and theory, opening an agenda for richer and more comparable models in Economics and business. In addition, thinking about future research on this proposal, we also propose extending the model to other alternatives, such as Deep Q-Network (DQN) architectures, which combine the Q-learning algorithm with deep neural networks to approximate the Q function. The same direction can also be explored for models that change or do not change the learning regime (on-policy/off-policy), or the way the Q function is updated (as in State-Action-Reward-State-Action—SARSA), or forms that directly parameterize policies (as in policy-gradient and actor–critic).

Author Contributions: Conceptualization, J.M.V., J.S.R.-O., P.C.M. and S.Á.-S.; methodology, J.M.V. and J.S.R.-O.; software, J.M.V. and J.S.R.-O.; formal analysis, J.M.V. and J.S.R.-O.; investigation, J.M.V., J.S.R.-O., P.C.M. and S.Á.-S.; writing—original draft preparation, J.M.V., P.C.M. and S.Á.-S.; writing—review and editing, J.M.V., J.S.R.-O., P.C.M. and S.Á.-S.; supervision, J.M.V. and P.C.M.; project administration, J.M.V. and S.Á.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CRRA	Constant Relative Risk Aversion
DQN	Deep Q-Network
ML	Machine Learning
RL	Reinforcement Learning
SARSA	State-Action-Reward-State-Action

References

1. Arrow, K.J.; Debreu, G. Existence of an equilibrium for a competitive economy. *Econometrica* **1954**, *22*, 265–290. [CrossRef]
2. Debreu, G. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*; Yale University Press: New Haven, CT, USA, 1959.
3. Friedman, M. The methodology of positive economics. In *Essays in Positive Economics*; University of Chicago Press: Chicago, IL, USA, 1953; pp. 3–43.
4. Simon, H.A. A behavioural model of rational choice. *Q. J. Econ.* **1955**, *69*, 99–118. [CrossRef]
5. Tesauro, G.; Kephart, J.O. Pricing in agent economies using multi-agent Q-learning. *Auton. Agents Multi-Agent Syst.* **2002**, *5*, 289–304. [CrossRef]
6. Klein, T. Autonomous algorithmic collusion: Q-learning under sequential pricing. *RAND J. Econ.* **2021**, *52*, 538–558. [CrossRef]
7. Calvano, E.; Calzolari, G.; Denicolò, V.; Pastorello, S. Artificial intelligence, algorithmic pricing, and collusion. *Am. Econ. Rev.* **2020**, *110*, 3267–3297. [CrossRef]
8. Alamdar, P.F.; Seifi, A. A deep Q-learning approach to optimize ordering and dynamic pricing decisions with strategic customers. *Int. J. Prod. Econ.* **2024**, *267*, 109154. [CrossRef]
9. Graf, P.; Emami, P. Three pathways to neurosymbolic reinforcement learning with interpretable model and policy networks. *arXiv* **2024**, arXiv:2402.05307. [CrossRef]
10. Wu, M.; Amin, S.; Ozdaglar, A. Convergence and stability of coupled belief–strategy learning dynamics in continuous games. *Math. Oper. Res.* **2025**, *50*, 459–481. [CrossRef]
11. Che, C.; Tian, J. Game theory: Concepts, applications, and insights from operations research. *J. Comput. Technol. Appl. Math.* **2024**, *1*, 53–59. [CrossRef]
12. Swenson, B.; Murray, R.; Kar, S. On best-response dynamics in potential games. *SIAM J. Control Optim.* **2018**, *56*, 2734–2767. [CrossRef]
13. Axtell, R.L.; Farmer, J.D. Agent-based modeling in economics and finance: Past, present, and future. *J. Econ. Lit.* **2025**, *63*, 197–287. [CrossRef]
14. Farmer, D. Quantitative agent-based models: A promising alternative for macroeconomics. *Oxf. Rev. Econ. Policy* **2025**, *41*, 571–590. [CrossRef]
15. Neugart, M.; Westerhoff, F. Editorial—Special issue on "Advancing agent-based economics". *Jahrbücher Natl. Stat.* **2024**, *244*, 289–291. [CrossRef]
16. Vilhuber, L. Reproducibility and transparency versus privacy and confidentiality: Reflections from a data editor. *J. Econom.* **2023**, *235*, 2285–2294. [CrossRef]
17. Dreber, A.; Johannesson, M. A framework for evaluating reproducibility and replicability in economics. *Econ. Inq.* **2025**, *63*, 338–356. [CrossRef]
18. Ankel-Peters, J.; Fiala, N.; Neubauer, F. Do economists replicate? *J. Econ. Behav. Organ.* **2023**, *212*, 219–232. [CrossRef]
19. Graf, C.; Zobernig, V.; Schmidt, J.; Klöckl, C. Computational performance of deep reinforcement learning to find Nash equilibria. *Comput. Econ.* **2024**, *63*, 529–576. [CrossRef]
20. Hady, M.A.; Hu, S.; Pratama, M.; Cao, Z.; Kowalczyk, R. Multi-agent reinforcement learning for resources allocation optimization: A survey. *Artif. Intell. Rev.* **2025**, *58*, 354. [CrossRef]
21. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
22. Ritto, T.G.; Beregi, S.; Barton, D.A. Reinforcement learning and approximate Bayesian computation (RL-ABC) for model selection and parameter calibration of time-varying systems. *Mech. Syst. Signal Process.* **2023**, *200*, 110458. [CrossRef]
23. Lou, Z.; Zhang, W.; Bao, X. Adaptive multi-layer reward based Q-learning for source tracing in diffusive molecular communications environment with obstacles. *Nano Commun. Netw.* **2023**, *38*, 100478. [CrossRef]
24. Clifton, J.; Laber, E. Q-learning: Theory and applications. *Annu. Rev. Stat. Its Appl.* **2020**, *7*, 279–301. [CrossRef]
25. Seyyedabbasi, A.; Aliyev, R.; Kiani, F.; Gulle, M.U.; Basyildiz, H.; Shah, M.A. Hybrid algorithms based on combining reinforcement learning and metaheuristic methods to solve global optimization problems. *Knowl.-Based Syst.* **2021**, *223*, 107044. [CrossRef]

26. Shakya, A.K.; Pillai, G.; Chakrabarty, S. Reinforcement learning algorithms: A brief survey. *Expert Syst. Appl.* **2023**, *231*, 120495. [CrossRef]
27. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
28. Tao, X.R.; Pan, Q.K.; Sang, H.Y.; Gao, L.; Yang, A.L.; Rong, M. Nondominated sorting genetic algorithm-II with Q-learning for the distributed permutation flowshop rescheduling problem. *Knowl.-Based Syst.* **2023**, *278*, 110880. [CrossRef]
29. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
30. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
31. Talaei Khoei, T.; Ould Slimane, H.; Kaabouch, N. Deep learning: Systematic review, models, challenges, and research directions. *Neural Comput. Appl.* **2023**, *35*, 23103–23124. [CrossRef]
32. Mosavi, A.; Faghan, Y.; Ghamisi, P.; Duan, P.; Ardabili, S.F.; Salwana, E.; Band, S.S. Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics* **2020**, *8*, 1640. [CrossRef]
33. Maliar, L.; Maliar, S.; Winant, P. Deep learning for solving dynamic economic models. *J. Monet. Econ.* **2021**, *122*, 67–78. [CrossRef]
34. Charpentier, A.; Élie, R.; Remlinger, C. Reinforcement learning in economics and finance. *Comput. Econ.* **2023**, *62*, 425–462. [CrossRef]
35. Nosratabadi, S.; Mosavi, A.; Duan, P.; Ghamisi, P.; Filip, F.; Band, S.S.; Reuter, U.; Gama, J.; Gandomi, A.H. Data science in economics: Comprehensive review of advanced machine learning and deep learning methods. *Mathematics* **2020**, *8*, 1799. [CrossRef]
36. Jiang, Z.; Xu, D.; Liang, J. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv* **2017**, arXiv:1706.10059. [CrossRef]
37. Singh, V.; Chen, S.-S.; Singhania, M.; Nanavati, B.; Kar, A.K.; Gupta, A. How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100094. [CrossRef]
38. Jiang, W. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
39. Salehpour, A.; Samadzamini, K. Machine learning applications in algorithmic trading: A comprehensive systematic review. *Int. J. Educ. Manag. Eng.* **2023**, *13*, 41. [CrossRef]
40. Shavandi, A.; Khedmati, M. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Syst. Appl.* **2022**, *208*, 118124. [CrossRef]
41. Kastius, A.; Schlosser, R. Dynamic pricing under competition using reinforcement learning. *J. Revenue Pricing Manag.* **2021**, *21*, 50–63. [CrossRef]
42. Zhou, Q.; Yang, Y.; Fu, S. Deep reinforcement learning approach for solving joint pricing and inventory problem with reference price effects. *Expert Syst. Appl.* **2022**, *195*, 116564. [CrossRef]
43. Nomura, Y.; Liu, Z.; Nishi, T. Deep reinforcement learning for dynamic pricing and ordering policies in perishable inventory management. *Appl. Sci.* **2025**, *15*, 2421. [CrossRef]
44. Kåge, L.; Milić, V.; Andersson, M.; Wallén, M. Reinforcement learning applications in water resource management. *Front. Water* **2025**, *7*, 1537868. [CrossRef]
45. Ding, Q.; Jahanshahi, H.; Wang, Y.; Bekiros, S.; Alassafi, M.O. Optimal reinforcement learning-based control algorithm for a class of nonlinear macroeconomic systems. *Mathematics* **2022**, *10*, 499. [CrossRef]
46. Soria, J.; Moya, J.; Mohazab, A. Optimal mining in proof-of-work blockchain protocols. *Financ. Res. Lett.* **2023**, *53*, 103610. [CrossRef]
47. May, R.; Huang, P. A multi-agent reinforcement learning approach for investigating and optimising peer-to-peer prosumer energy markets. *Appl. Energy* **2023**, *334*, 120705. [CrossRef]
48. Brusatin, S.; Padoan, T.; Coletta, A.; Delli Gatti, D.; Glielmo, A. Simulating the economic impact of rationality through reinforcement learning and agent-based modelling. In *Proceedings of the 5th ACM International Conference on AI in Finance*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 159–167. [CrossRef]
49. Chen, R.; Zhang, Z. From Individual Learning to Market Equilibrium: Correcting Structural and Parametric Biases in RL Simulations. *arXiv* **2025**, arXiv:2507.18229. [CrossRef]
50. Bereby-Meyer, Y.; Roth, A.E. The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation. *Am. Econ. Rev.* **2006**, *96*, 1029–1042. [CrossRef]
51. Lauriére, M.; Perrin, S.; Pérolat, J.; Girgin, S.; Muller, P.; Élie, R.; Geist, M.; Pietquin, O. Learning in mean field games: A survey. *arXiv* **2024**, arXiv:2205.12944. [CrossRef]
52. Taherdoost, H. Deep learning and neural networks: Decision-making implications. *Symmetry* **2023**, *15*, 1723. [CrossRef]
53. Rummery, G.A.; Niranjan, M. *On-Line Q-Learning Using Connectionist Systems*; Technical Report CUED/F-INFENG/TR 166; University of Cambridge: Cambridge, UK, 1994.

54. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. [CrossRef]
55. Konda, V.R.; Tsitsiklis, J.N. On actor–critic algorithms. *SIAM J. Control Optim.* **2003**, *42*, 1143–1166. [CrossRef]
56. Samuelson, P.A. A note on measurement of utility. *Rev. Econ. Stud.* **1937**, *4*, 155–161. [CrossRef]
57. Koopmans, T.C. Stationary ordinal utility and impatience. *Econometrica* **1960**, *28*, 287–309. [CrossRef]
58. Melo, F.S. *Convergence of Q-Learning: A Simple Proof*; Technical Report; ISR, Instituto Superior Técnico: Lisboa, Portugal, 2001.
59. Nakamura, E.; Steinsson, J. Five Facts about Prices: A Reevaluation of Menu Cost Models. *Q. J. Econ.* **2008**, *123*, 1415–1464. [CrossRef]
60. Leslie, D.S.; Collins, E.J. Individual Q-Learning in Normal Form Games. *SIAM J. Control Optim.* **2005**, *44*, 495–514. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.