Energy and AI 23 (2026) 100665

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Assessing long-term electricity market design for ambitious decarbonization targets using multi-agent reinforcement learning

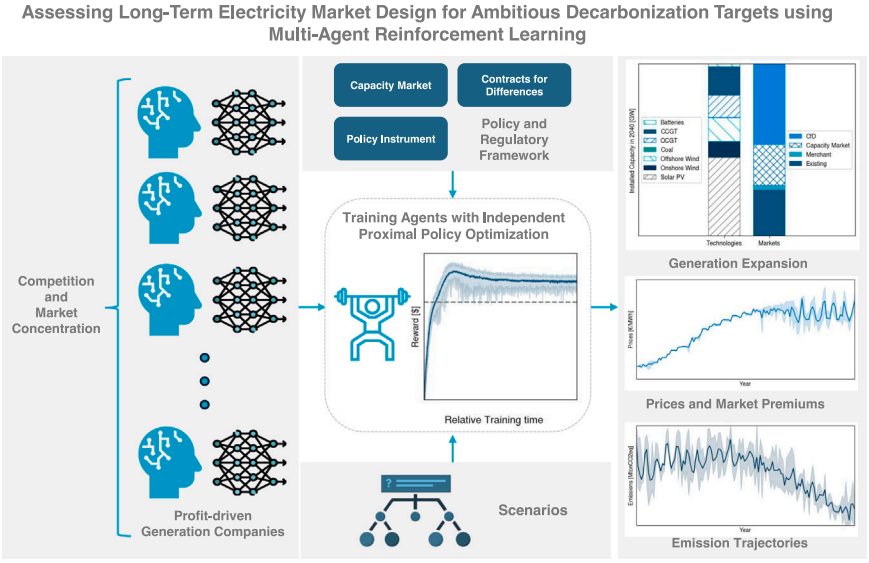
Javier Gonzalez-Ruiz [a,c,d,*], Carlos Rodriguez-Pardo [b,c,d], Iacopo Savelli [d,e], Alice Di Bella [a,c,d], Massimo Tavoni [b,c,d]

[a] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy
[b] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy
[c] CMCC Foundation- Euro-Mediterranean Center on Climate Change, Via Marco Biagi 5, Lecce, 73100, Italy
[d] RFF-CMCC European Institute on Economics and the Environment, Via Bergognone 34, Milan, 20144, Italy
[e] Centre for Research on Geography, Resources, Environment, Energy and Networks (GREEN), Bocconi University, Via Roberto Sarfatti, Milan, 20136, Italy

HIGHLIGHTS

- This work develops a reinforcement learning model for long-term electricity markets.
- Multi-agent Independent Proximal Policy Optimization is used.
- The model enables assessment of simultaneous policy and market interactions.
- An extensive hyperparameter search is applied to the competitive environment.
- Tests in a stylized Italian system under deep decarbonization scenarios.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Multi-agent reinforcement learning
Agent-based modeling
Electricity markets
Energy transition
Energy policy
Capacity markets
Contracts for difference

ABSTRACT

Electricity systems are key to transforming today's society into a carbon-free economy. Long-term electricity market mechanisms, including auctions, support schemes, and other policy instruments, are critical in shaping the electricity generation mix. In light of the need for more advanced tools to support policymakers and other stakeholders in designing, testing, and evaluating long-term markets, this work presents a multi-agent reinforcement learning model capable of capturing the key features of decarbonizing energy systems. Profit-maximizing generation companies make investment decisions in the wholesale electricity market, responding to system needs, competitive dynamics, and policy signals. The model employs independent proximal policy optimization, which was selected for suitability to the decentralized and competitive environment.

* Correspondence to: Via Bergognone, 34, 20144 Milano, Italy.
E-mail address: javier.gonzalez@cmcc.it (J. Gonzalez-Ruiz).

https://doi.org/10.1016/j.egyai.2025.100665
Received 26 May 2025; Received in revised form 15 September 2025; Accepted 12 December 2025
Available online 18 December 2025

Nevertheless, given the inherent challenges of independent learning in multi-agent settings, an extensive hyperparameter search ensures that decentralized training yields market outcomes consistent with competitive behavior. The model is applied to a stylized version of the Italian electricity system and tested under varying levels of competition, market designs, and policy scenarios. Results highlight the critical role of market design for decarbonizing the electricity sector and avoiding price volatility. The proposed framework allows assessing long-term electricity markets in which multiple policy and market mechanisms interact simultaneously, with market participants responding and adapting to decarbonization pathways.

1. Introduction

One of the key actors of a climate transition is electricity generation, given its relevance and its central role as an enabler of decarbonization across other sectors [1]. In scenarios aligned with the Paris Agreement reviewed by the Intergovernmental Panel on Climate Change (IPCC), electricity generation is projected to reach net-negative carbon emissions before 2050, thus enabling the electrification of all end-use sectors.

In this context, most scenarios informing policy-making are predominantly generated by central planning models, which have limited capacity to represent the wide set of policies needed to ensure the green transition. Most notably, given that wholesale electricity systems are predominantly structured as markets featuring active participation from private merchant actors alongside regulated institutions (including transmission companies), electricity market design is crucial for aligning the evolution of energy systems with climate objectives [2]. In parallel to their relevance for the energy transition, electricity systems and markets have come under increased scrutiny due to their vulnerabilities and questions regarding resilience in the face of large-scale crises, such as the disruptions triggered by the war in Ukraine [3].

Amid these concerns, a key policy challenge lies in the design of long-term electricity markets. These markets, including auctions, support schemes, and contracting mechanisms, aim to incentivize adequate investment in generation and flexibility resources. Their goals range from de-risking capital investment and ensuring resource adequacy, to shielding consumers from price volatility and ensuring affordability [4–6]. Long-term markets complement short-term markets (e.g., day-ahead, intraday, and balancing markets), which focus on efficiently dispatching and utilizing existing assets [7]. Regarding long-term markets, two dominant paradigms have emerged: Energy-only Markets (EoMs) and Capacity Remuneration Mechanisms (CRMs) [4]. However, the evolving policy landscape has prompted a surge in reform proposals. These range from enhanced long-term contracting for renewable energy sources in the form of Contract for Differences (CfD) [8–11], to the introduction of mandatory contracting schemes [12], and the development of new financial instruments aimed at shielding consumers from price volatility [3]. A more far-reaching approach to transforming electricity markets is gaining traction alongside these incremental changes: hybrid electricity market models [13–16], frameworks in which governments assume a stronger role in guiding and selecting large-scale investments in a competition for the market. In contrast, decentralized markets continue to govern short-term operations via competition in the market. Although intriguing and attractive from a theoretical standpoint, such arrangements lack quantitative evaluations in the Literature and are still far from real-life implementations.

Considering this broad spectrum of potential market architectures, each with far-reaching implications beyond the electricity sector, modeling tools are essential for informed decision-making [17], where Agent-based and partial equilibrium models have traditionally dominated this space. However, these approaches may face limitations in capturing the dynamics of transformative policy scenarios or exploring radically new institutional designs. More specifically, the explicit modeling of auctions and their system-wide interactions and complexities, a defining characteristic of modern electricity markets that will become increasingly significant as supporting and long-term mechanisms gain prominence, is often overlooked or only partially addressed in the existing literature.

Multi-Agent Reinforcement Learning (MARL) offers a promising modeling alternative that addresses these shortcomings, providing a highly flexible modeling framework that can complement existing approaches by simulating adaptive behaviors in complex, multi-agent systems. Nevertheless, the application of MARL to electricity market modeling is still in its early stages and requires further development to become a robust tool for policy and technical assessment. This work concentrates on enabling MARL models as a comprehensive tool for long-term electricity market assessments, with contributions expanded as follows:

- We develop an open-source multi-agent environment that models the long-term electricity market, extending current implementations that concentrate on short-term markets. In its current version, the model allows for investing in generation assets through stylized capacity and contract for difference markets, aside from merchant investments. Moreover, integrating other incentive instruments and accommodating different market designs can be easily extended.
- This work includes a detailed process for a hyperparameter search in the context of proximal policy optimization applied to the electricity markets. Efforts in this area showcase the potential and difficulties of applying MARL to competitive environments.
- By integrating a long-term market environment with a MARL training pipeline, the first application to the best of our knowledge, this work provides a versatile framework for evaluating ambitious decarbonization strategies in electricity systems. The approach offers near-unlimited flexibility in electricity market design. Four key advantages distinguish this modeling framework from existing methods in the literature:
  - First, the model explicitly incorporates auction mechanisms, which are central to many long-term market designs and are commonly used to facilitate and de-risk investment in energy infrastructure.
  - Second, it supports comparing multiple market instances and policy layers, such as carbon taxes, within a unified framework, enabling a systematic evaluation of the effectiveness and potential redundancy of various policy instruments.
  - Third, agents in the model manage portfolios that include both new investments and existing assets, allowing the analysis of possible divergent incentives and behaviors of incumbent versus entrant actors.
  - Fourth, the model captures the impact of market competition on outcomes, a critical feature in wholesale electricity systems that often exhibit comparatively high market concentration.
- The code for this framework is open-source, available via a public repository[1].

The rest of this article is organized as follows. Section 2 presents the related work on modeling practices for electricity markets. Section 3.1 introduces the complete MARL implementation. Next, Section 4 describes the training setup and the process for hyperparameter selection. After, Section 5 presents market results obtained with the MARL model. Finally, Section 6 concludes.

[1] https://github.com/jjgonzalez2491/MARLEY_V1

2. Related work

This work addresses a research gap by applying Multi-Agent Reinforcement Learning to long-term electricity markets, where strategic decisions regarding investment in utility-scale generation assets are crucial. To establish the foundation for this contribution, this section presents related literature on two main fronts: first, highlighting current practices in long-term electricity market analysis, and second, examining the flexibility, opportunities, and limitations identified in applying MARL to a similar field. As such, the first two subsections introduce Agent-Based and Partial-Equilibrium models for long-term electricity markets, the two primary approaches that have addressed pressing concerns related to decarbonization efforts and have examined investment strategies in detail. From this baseline, applications of Multi-Agent Reinforcement Learning to short-term electricity markets are discussed, drawing connections to long-term market modeling where applicable.

2.1. Agent-based models

In Agent-Based models, electricity markets are constructed and represented via the detailed depiction of decisions and interactions between market players, actors, and policy-makers. In this category, EmLab-Generation, an Agent-Based model designed for long-term electricity market analysis [18], has been used to analyze the impact of capacity markets in a system with a growing share of renewable energy sources. Specifically, studies show the effectiveness of the scheme in comparison to strategic reserves for ensuring reliability via the promotion of low-cost peak generation units [19]. Similarly, authors in [20] demonstrate that capacity markets that allocate long-term commitments to investments, instead of annual contracts, are preferable when ensuring security of supply. In [21], EmLab is enhanced with short-term modules, showcasing the potential of demand response and energy storage systems to outright replace or compete against capacity markets. Finally, given the possible phase-out of support schemes for renewable technologies in Germany and the Netherlands, scenarios demonstrate significant reductions in new capacity additions of low-carbon technologies and price increases [22].

Expanding on the previous framework, Brain-Energy has been developed to increase agent heterogeneity and include institutional agents, such as governments and regulators, enabling endogenous decision-making regarding key policies in the energy sector [23,24]. Analysis of the UK, Germany, and Italy transition pathways using Brain-Energy shows that historic-path dependence in investment choices can displace low-carbon investments in scenarios with weak regulatory frameworks [25]. Furthermore, in [24], authors argue that, in scenarios with heterogeneous agents and higher capital requirements, more aggressive policy action to promote decarbonization efforts is necessary to achieve environmental targets.

Authors in [26] have developed the EMIS-AS model by implementing a richer representation of electricity systems and market sessions. EMIS-AS learning capabilities, implemented via Kalman Filters, are particularly relevant and aim to forecast key parameters for the investment profitability assessments. Using the previous framework, several market designs are tested when pursuing clean energy targets ranging between 45% and 100% by 2030 [27]. Among other insights, authors find that energy-only and capacity markets can achieve clean energy targets while maintaining operational constraints. Moreover, carbon pricing is the most effective mechanism for reaching the first wave of renewable penetration, while stacking several mechanisms to promote the energy transition in scenarios with more aggressive targets demonstrates only marginal benefits. Expanding these ideas, the EMIS-AS model enabled a detailed comparison between capacity markets and operating reserve demand curves [28].

This modeling category has also included risk metrics in the agent decision-making formulation. In [26], risk-averse agents compare projects and technologies through a utility function, implicitly assigning higher discount rates to comparatively larger investments. More broadly, authors in [29] implement and compare three risk metrics in an agent-based model: Value-at-Risk, Mean-Variance, and Adjusted discount rate. Finally, it is worth mentioning the family of short-term Agent-Based models, such as AMIRIS, that complement the long-term perspective with decisions regarding the dispatch of power plants, in addition to interactions between a broader set of agents and actors in short-term markets [30,31].

2.2. Partial equilibrium and Bi-level optimization models

Partial Equilibrium and/or Bi-level optimization for electricity markets refer to model formulations where hierarchical optimization problems represent a market and are generally resolved in a Nash equilibrium. In [32], a comprehensive overview of the different types of formulations and their application to electricity markets is presented.

In this modeling category, particular attention has been given to long-term market design, and most relevantly, the impact of Capacity Markets in the presence of risk-averse agents. Initially, [33] introduced an equilibrium model capable of representing both Capacity Markets and Strategic Reserves. Results showcased comparative benefits in the former policy option, considering the additional incentives it provides for actors to participate in the rest of the market. Extending the model, [34] presented a market formulation that includes risk-adverse agents and a solution algorithm based on the Alternating Direction Method of Multipliers. Similar to the previous case, when compared to an Energy-only framework, the capacity market improved the system costs while maintaining the necessary reserve margins. On this note, authors in [35], using a two-stage stochastic program for capacity expansion, argue that capacity mechanisms have an asymmetric effect on the risk profiles of generation technologies, which promotes the integration of low-capital, high-variable-cost technologies. Regarding the impact of demand response on capacity market design, analysis carried out in [36] using a stochastic non-cooperative capacity planning model with risk-averse investors demonstrated that these schemes partially mitigate the effect of risk aversion on social welfare.

Aside from Capacity Market design, authors in [37] design a mechanism capable of aligning consumer preferences with potential investors for strategic reserves by implementing two bi-level optimization problems, the first representing an Energy-only market, and a second in which retailers present their willingness to pay for extra levels of reliability to be provided by an insurer of last resort. Regarding the cost of capital on systems with high shares of wind and solar technologies, the two-stage stochastic equilibrium model developed in [38] complements standard market models by including hedge providers, thus being able to quantify the impact of market and price volatility in the cost of capital for particular technologies.

Lastly, equilibrium models have also been utilized to understand the implications of incomplete risk-trading schemes in electricity markets. By considering the system conditions leading to the contingencies faced by ERCOT during February 2021, in [39], the authors promote, among other measures, a shift toward mandatory contracting obligation on retailers. To directly understand the impact of missing markets in the presence of risk-averse preferences, in [40] three models for the power system are compared: a risk-neutral expansion problem, a risk-averse and missing risk market scenario, in which both supply and demand formulation are risk-averse agents but their optimization problems are solved independently, and a risk-averse expansion problem. By following a similar framework, the work in [41] aims to support the design of decarbonization policies when considering risk-averse agents in incomplete electricity markets.

2.3. Multi-agent reinforcement learning and electricity markets

Reinforcement Learning (RL) is a particular subset of Machine Learning techniques in which agents are trained to perform specific tasks by performing actions in their environment, having received the corresponding feedback from these interactions [42,43]. Recent advancements have enabled agents to be parametrized using Neural Networks, extending the reach of RL to areas where a flexible, adaptable, and scalable nonlinear optimizer is advantageous [43,44].

A natural extension of RL that models the interaction of several agents within the same environment, trained to perform tasks with shared or competing objectives, is denoted Multi-Agent Reinforcement Learning (MARL) [45]. MARL has been extensively applied to various subjects, including the energy sector and policy-making [46]. Particularly relevant to this work are applications where, beginning with naive and unstructured environments characterized by agent competition, neural networks have outperformed both humans and traditional algorithms in solving highly complex tasks [47–49]. In the electricity sector, RL and MARL have been applied across the electricity value chain [50–52]. Among these applications, the most prominent focus has been on analyzing bidding and participation strategies in short-term electricity markets, moving beyond the conventional paradigm in which agents submit bids approximating their marginal production costs. In this context, MARL represents a technique to improve current Agent-based models by filling the gap with partial-equilibrium setups. Specifically, MARL takes advantage of the flexibility that has characterized agent-based models, but extends it to setups where partial-equilibrium models have been used to understand efficient and competitive market interactions.

Starting from [53], authors apply a Deep Policy Gradient algorithm to model price-bidding strategies of single-plant Generation Companies (GENCO) in a short-term market. Tests showcased that the proposed RL algorithm reached a Nash Equilibrium in not-congested cases compared to a simplified partial-equilibrium model. Extending the previous work, in [54], a Deep Deterministic Policy Gradient (DDPG) algorithm is applied to a market and network configuration resembling the previous implementation. In this case, the system improves computational and market performance compared with other RL algorithms, namely Q-learning and Deep-Q network. Compared with the Mathematical Programs with Equilibrium Constraints solution, it obtained higher profits for the single-plant GENCOs. Similarly, the work in [55] uses a DDPG to analyze a system in which GENCO presents quadratic cost functions, parametrized by their bidding strategy. Most relevantly, results highlight the importance of hyperparameter selection, as increasing the discount factor during training exhibited increments in the profit earned by agents from the market, thus deviating from the equilibrium obtained in the analytic solution. Shifting from the independent learning presented thus far, where each agent is trained concurrently without a centralized architecture, in [56], the Multi-Agent DDPG (MADDPG) is implemented to reach the Nash Equilibrium in a day-ahead market, concentrating again on price-bidding strategies. Authors find that the centralized training and decentralized execution framework in the algorithm improved computational efficiency compared to the independent learning methods. Continuing with DDPG, in [57], the authors conduct an extensive hyperparameter search, enabling the independent learning scheme to reach a Bertrand Equilibrium in a stylized system. Furthermore, a Mean-Field algorithm based on DDPG is applied to analyze trading strategies in peer-to-peer (P2P) energy markets [58], a setup where prosumers have limited access to information for their decision making. Results demonstrate agents can learn efficient strategies to bid prices and quantities in the double-auctions from the P2P market, while the Mean-Field algorithms outperform other MARL options. Deviating from models dedicated to price-bidding strategies, a Soft-Actor Critic algorithm is proposed to bid prices and quantities jointly in a short-term market [59]. Although the framework is not applied to a multi-agent context, the authors find the computational and economic benefits of the coupled evaluation in bidding for short-term markets.

Shifting from DDPG, authors in [60] develop the model ASSUME, an agent-based approach focused on short-term markets and structured upon the multi-agent Twin-Delayed DDPG. By following a centralized training and decentralized execution paradigm, authors can extend their framework to hundreds of agents, thus showcasing prices that resemble real market prices from the German electricity market. The model includes an additional incentive that penalizes inefficient actions undertaken by agents, complementing the pure economic profits as the training reward utilized in most of the work presented thus far. In [61], the previous model's performance is compared with its bi-level optimization problem counterpart, showcasing limits in the RL approach while achieving the theoretical equilibrium, but also demonstrating the potential in the modeling framework in terms of flexibility and scalability to more complex setups. Similarly, by applying explainable artificial intelligence techniques, authors analyze the bidding strategies obtained by the ASSUME model [62]. In particular, agents are willing to increase their strategic bids in cases where they are potential price setters. At the same time, this incentive is reduced when it is more likely that their bids will be lower than the marginal market price.

In parallel to the work focused on short-term market analysis, other works have placed efforts extending MARL toward related topics. Worth noting are the SustainGym environments [63], which have been designed to enable a standardized comparison of RL and MARL algorithms in systems from the energy and environmental sectors. Complementary, in a stylized model, authors in [64] propose the inclusion of an active regulator/central planner as part of the learning agents in a MARL market simulation, enabling policy design that endogenously responds to strategic behavior from market actors, and vice-versa.

3. Multi-agent long-term electricity market reinforcement learning environment

This section describes the implementation of the Long-term Electricity Market in the Multi-Agent Reinforcement Learning framework. Section 3.1 introduces the long-term electricity market environment. Next, Section 3.2 describes the MARL scheme applied to the long-term electricity market environment. As explained throughout this section, maintaining a rich and complex representation of a long-term electricity market that enables policy assessments is a key goal of this work. As a result, design elements and features in the environment have become intrinsically and inevitably connected with the algorithm selection. Thus, the generalization of this environment to other MARL algorithms remains a compelling area for future research.

3.1. Long-term electricity market environment

The Long-term Electricity Market Environment, schematized in Fig. 1, is described via its two main structural components, starting from the electricity market design and moving toward translating such market to the MARL framework. Following this structure, Appendix A goes into further detail on the model description.

3.1.1. Long-term electricity market

The model presented in this work targets mid- and long-term (several years or, at most, a couple of decades) decarbonization analysis in wholesale electricity markets, focusing on utility-scale investment decisions and market mechanisms designed to promote them. To this end, the model represents GENCOs, as shown in the upper part of Fig. 1. GENCOs are assumed to be profit-maximizing entities while subject to system conditions, policy signals, and competition from other market players. GENCOs invest in generation and Energy Storage Systems (ESS) assets to participate in the electricity market and expand their portfolio. The characteristics of generation and storage technologies are detailed in Appendix A.1.3 and Appendix A.1.4. The investment framework

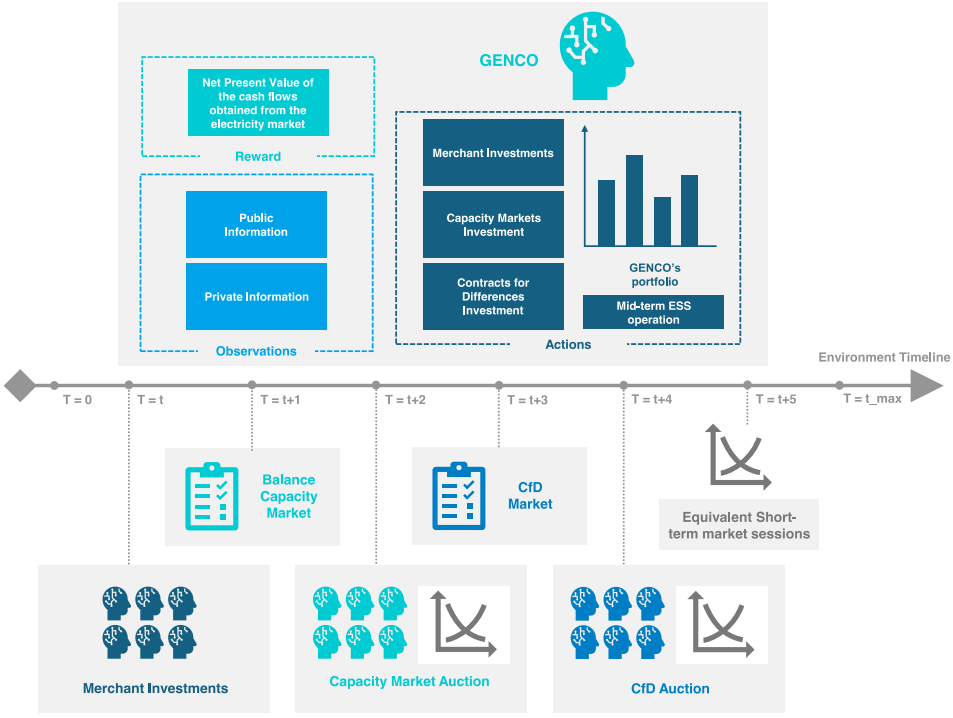


Fig. 1. Structure of the long-term electricity market in the reinforcement learning model. The diagram highlights the interactions between GENCO agents and the market environment. From the market, agents receive observations *(public information, published by entities such as Market Operators and TSOs, and private information, regarding the performance of their portfolio)*, to take actions *(investment decisions)* in the system, with the aim to maximize their profits during the simulation. The figure illustrates one year of market operation, where agents participate in the short-term market at each environment step while making investment decisions annually. Investment decisions occur sequentially through mutually exclusive entry mechanisms *(merchant, Capacity Market, CfD market)*, with the latter two depending on system balance calculations, as detailed in Section 3.1.

assumes unrestricted access to financing with no equity constraints, capital expenditures (CAPEX) incurred during the construction phase, and pre-tax cash flows.

To represent day-ahead markets, the model uses equivalent periods through representative days. Representative days are defined as sequential 24-hour windows, with an hourly resolution, intended to capture daily patterns in electricity systems. To model renewable resource availability, time series projections from [65,66] describe their hourly capacity factor. In the case of electricity demand, projections for the Italian power system are used to describe the yearly profile [67]. Once all necessary time series are obtained, the TSAM Python library [68] transforms them into representative periods, which are then applied to the equivalent short-term market sessions. For each hour in the equivalent short-term sessions, GENCOs participate in the market by presenting bids for their portfolio (energy quantity and offered price). In line with most long-term partial-equilibrium models, the environment assumes that GENCOs present bids for the short-term market that correspond to their marginal production costs. The bids from all GENCOs and resources in the system serve as input for double-sided marginal price auctions, dispatching resources based on their merit order and setting a unique system price.

Regarding the operation of the electricity system, the model adopts a copper plate assumption, similar to the day-ahead market clearing in, for example, France and Germany. Although limited, this approach directly evaluates long-term electricity market design as a tool or potential obstacle toward ambitious decarbonization objectives. Future work will concentrate on improving network modeling within the MARL context, with possible pathways for enhancement detailed in Appendix A.1.5.

As previously stated, GENCOs are also able to invest in ESS assets. However, ESS investments are restricted to technologies with relatively short-duration storage capacities (3–4 h), such as Lithium-ion batteries. In contrast, longer-duration storage technologies remain under incumbent agents' control for operational decisions. In each equivalent short-term market session, GENCOs decide the desired state of charge level of the long-duration ESS for the next period, thus representing inter-period flexibility in the technology. Within the day-ahead markets, both short and long-duration ESSs are operated for efficient system operation, as described in Appendix A.1.4.

Given this selection of generation and storage assets, reduced in comparison to existing planning models but representative of the main trends in the sector, investment can occur via the three available and mutually exclusive channels and markets:

- **Merchant Investments:** Any GENCO can freely decide to place projects of different technologies under construction, with no centralized planning process by the regulator. Once a merchant investment enters operation, the project earns revenues directly from the day-ahead market. From a GENCO's perspective, its plant has no protections against sustained low prices during excess supply situations, which might prove insufficient to justify the investment. However, when scarcity conditions occur, the plant will harness the full-fledged scarcity rents.
- **Contract for Difference Market:** In this market, a regulator *(or central planner)* establishes a penetration target for RES in the system, defined as the share between renewable production and total demand. When GENCOs have built insufficient RES projects to achieve the desired penetration target, an auction takes place to cover the missing renewable share. Winning projects in the auction are awarded a stylized version of a two-way Contract for Difference [11]. From a GENCO perspective, the two-way CfD ensures a fixed price for the total output of the participant plant. Equivalently, the GENCO hedges its project against low-price conditions in the system. In exchange, the GENCO must build the project presented in the auction and withhold the financial obligation associated with the CfD during the contract's lifetime (usually 15 to 25 years).

- **Capacity Market:** In this market, the regulator ensures resource adequacy by guaranteeing sufficient resources to meet peak demand conditions. The Capacity Market is inspired by the Reliability Option framework [4,69,70]. From a GENCO's perspective, participation in the Capacity Market provides a premium linked to the project's contribution to system adequacy, as measured by its capacity credits and the auction allocations. In return, the GENCO must develop the generation project and protect consumers from high-price events through the Reliability Option mechanism.

With this design, the environment concentrates on well-established long-term market arrangements, serving as a baseline that facilitates validation, model comparison, and policy assessments. Exploring other market arrangements, such as those from hybrid design schemes, constitutes an interesting direction for future studies.

Apart from market design, the model includes two additional policy instruments that could help to shape and define scenarios. The first is a Carbon Tax, directly linked to the CO_2 emissions produced by generation technologies, which passes through to consumers via the bids in the short-term market. The second instrument is an exogenous limit for investing in specific technologies. This limit can be enforced at any point during the simulation and applied discriminately to market players and investment channels.

3.1.2. Electricity market as a multi-agent reinforcement learning environment

This section integrates the Long-term Electricity Market into the Gymnasium standard [71], and more specifically, the version of Multi-Agent environments developed by the RLLIB team [72]. Section Appendix A.2 presents further information for the RL Environment.

3.1.2.1. Environment structure. In the Gymnasium standard, environments for Reinforcement Learning use steps to simulate the transitions in the underlying Markov Decision Process. In each step, agents observe the system's state, take action, and receive the corresponding reward. In the context of the market model, these environment steps are directly linked to Equivalent Short-Term Market sessions and the corresponding Representative Periods. Each step is designed to represent a fixed period of operation by condensing interactions into a single 24-hour representation.

During each environment step, the GENCOs' portfolios participate in the Equivalent Short-Term Market, generating market outcomes that serve as the basis for the environment's observations and rewards. However, aside from operating mid-term ESS, agents do not actively make decisions affecting the short-term market: generation assets bid their availability and marginal costs to the day-ahead market, without strategic bidding enabled, and the intra-day charge/discharge cycles of ESS are defined for efficient system operation. Instead, investment decisions are enabled every year, allowing agents to take actions to expand their portfolios. Investments in the different markets occur every two environment steps and in sequence across the year, as shown in the lower panel of Fig. 1. This implementation enables yearly investments in all the markets if system conditions demand it.

3.1.2.2. Reward. In the model, agents are assumed to maximize the net present value of the cash flows obtained from the electricity market. This involves calculating and aggregating revenues and costs from all assets and markets relevant to the agents' portfolio across the simulation. Yet, to be consistent with real company operations, the reward function is designed to be passed to agents step-by-step, as in Eq. (1). The reward is divided into two parts: profits and investment costs. On the one hand, the profits $(P_{\mathrm{M}}, P_{\mathrm{CM}}, P_{\mathrm{CfD}})$ derived from all markets in the system; on the other hand, the investment costs $(IC_{\mathrm{M}}, IC_{\mathrm{CM}}, IC_{\mathrm{CfD}})$ of new portfolio additions, similarly disaggregated across markets. Apart from the previous terms, and to improve clarity in the financial modeling across the formulation, the discounting of cash flows is carried out inside the environment while considering the discount rate as an exogenous parameter. The detailed disaggregation of profits and costs for each market is presented in Appendix A.2.2.

$$r_{\mathrm{t}} = \left(p_m + P_{\mathrm{CM}} + P_{\mathrm{CfD}} - \left(IC_{\mathrm{M}} + IC_{\mathrm{CM}} + IC_{\mathrm{CfD}}\right)\right)\left(\frac{1}{(1+r)^t}\right) \tag{1}$$

3.1.2.3. Action space. In the current implementation, the actions available to GENCO for interacting with their environment are, in principle, a combination of Discrete *(e.g., investment quantities)* and Continuous *(e.g. bidding prices for long-term auctions)* actions. Yet, considering the advantages provided by Action Masking to represent both internal and external constraints for the agents and the system [73], the model adopts a fully multi-discrete action space, increasing the number of discretization steps for continuous variables to enhance the realism of agent strategies, following the recommendations of [74].

In this context, Table A.6 provides a comprehensive overview of the actions available to GENCOs. In particular, Action Masking is used to control which agents have access to which decisions *(e.g., to limit investments in specific technologies)* and when a given action can take place *(e.g., considering that investments are not enabled in every environment step)*. Although flexible for constraint representation, using Multi-discrete actions creates several limitations in the implementation. First, investment limits are regulated solely by technology-specific maximum rates (Tables A.4 and A.5). Second, agents have prior knowledge of auction price caps. Last, the number of available actions (Table A.6) scales with the number of technologies, justifying the choice of a limited yet representative set of options aligned with energy transition trends.

3.1.2.4. Observation space. The design of the observation set available to GENCOs in the model follows three key principles. First, the selected variables should align with real-world market conditions, where agents can access publicly shared system information while specific details remain private and inaccessible to competitors. Second, no internal price forecasting tools are included in the observation set, ensuring that the model retains full autonomy in decision-making. Finally, the selection process adheres to the principle that any information available in the market is also provided to the agents.

Considering this, Table A.7 lists the observations made available to agents across environment steps, subdivided according to their thematic category. The set of observations includes indicative values for the time series modeling demand and variable resource availability, the composition of the energy mix via the assets in operation, individual and system-wide reservoir levels, market information *(such as prices and balances from the Capacity and CfD markets)*, the aggregated reward per technology in these markets, policy-relevant information, and the time associated to the environment step. The framework designed for the GENCO's observation does not depend on the number of agents in the market, as it harnesses aggregated system information, given that the number of agents in each simulation is constant.

3.2. Multi-agent reinforcement learning and competitive environments

To start, the preliminaries for Single-Agent and Multi-Agent RL are provided. Next, details of Single-Agent and Multi-Agent PPO algorithms are presented. From these premises, this section concludes with the arguments used for the algorithm selection.

3.2.1. Single-agent reinforcement learning

In RL, the agent's interactions with the environment are modeled using a Markov Decision Process (MDP). MDPs are defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ corresponds to set defining the state space, $\mathcal{A}$ represents the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition probability between states as a function of the agent's actions, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, which depends on the system state and the agent's actions. For simplicity, the notation assumes a fully observable setting, thus allowing the agent to directly translate states to observations $\mathcal{O} \to \mathcal{S}$ [42,43].

Trajectories in the MDP can be formalized by the succession of states, actions, and rewards, $\tau = (s_0, a_0, r_0, \dots s_t, a_t, r_t, \dots, s_T, a_T, r_T)$, where the subscript t represents time steps, $[0, T]$ denote the initial and terminal steps in the system, and states, actions, and rewards are sampled from the corresponding sets. The system is a finite-horizon discounted Markov Decision Process for cases where T is finite, and rewards are discounted by a $\gamma \in (0, 1]$. Generally, RL algorithms aim to maximize the expectation of rewards, shown in expression (2) by learning a policy π_θ, characterized by trainable parameters θ. In Deep RL, policies are represented with Neural Networks, where θ refers to the set of describing parameters, varying according to the selected architecture.

$$J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]. \tag{2}$$

For optimizing the objective function, auxiliary expressions can be defined. Starting from the concepts of State-Value, in expression (3), and Action-Value functions, in Eq. (4), defined as the expected future reward of being in a specific state and following policy π_θ, in the former, and the expected future reward of selecting a particular action in a specific state, in the latter. The Advantage function A_t, quantifying the reward improvement of taking action a_t in state s_t, in comparison to the expected reward in state s_t, can be obtained as the difference between Action and State Value functions. In Actor-Critic RL algorithms, the Action-Value function is associated with the policy π_θ, while the State-Value function serves as a baseline for performance evaluation.

$$V_{\pi_\theta}(s_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=k}^{T} \gamma^{k-t} r_k \mid s_t\right] \tag{3}$$

$$Q_{\pi_\theta}(s_t, a_t) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=k}^{T} \gamma^{k-t} r_k \mid s_t, a_t\right] \tag{4}$$

Two key assumptions commonly used in RL algorithms to ensure convergence are particularly relevant compared to the multi-agent framework. The first is the Markov Property, describing MDPs in which the future state depends only on the current state and actions. Similarly, the second is the stationarity of transition dynamics, a condition achieved when $\mathcal{P}$ remains constant during training and execution, even if the transitions are described stochastically [42,43,75]. Finally, it is worth mentioning that the framework outlined above operates under risk-neutral assumptions, where the objective is to maximize expected rewards. Theoretical and algorithmic advances have extended reinforcement learning to accommodate risk-averse decision-making [76,77]. Nevertheless, the formal evaluation of explicit risk-averse algorithms in the current framework is left for future work.

3.2.2. Multi-agent reinforcement learning

The framework for Multi-Agent Reinforcement Learning (MARL) can be introduced as a Partially Observable Stochastic Game (POSG) [45]. In particular, the POSG considers a set of N agents, $I = \{1, \dots, i, \dots, N\}$, each with its set of actions $\mathcal{A}_i$, where $\mathcal{A} = \mathcal{A}_1, \dots, \mathcal{A}_N$, interacting within an environment with a finite number of states $\mathcal{S}$. In the system, transition probabilities are defined as $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, a function of states and Actions, an agents harness a reward based on a function $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{R}$. Finally, agents have access to system states via an Observation function $\mathcal{O}_i : \mathcal{A} \times \mathcal{S} \times \mathcal{O}_i$.

Analogous to MDP, agent's trajectories can be formalized by the succession of observations, actions, and rewards, $\tau_i = (o_{i,0}, a_{i,0}, r_{i,0}, \dots o_{i,t}, a_{i,t}, r_{i,t}, \dots, o_{i,T}, a_{i,T}, r_{i,T})$, where the subscript t represents time steps, $[0, T]$ denote the initial and terminal steps in the system, γ is the discount factor, and observations, actions, and rewards are sampled from the corresponding sets and functions of each agent.

Unlike single-agent MDPs, a POSG requires the definition of a solution concept to guide RL algorithms to find the set of joint policies $\pi = \pi_1, \dots, \pi_N$ that aim to achieve the desired objective [45]. The long-term electricity market environment is considered competitive, as agents intend to maximize their profit, with no explicit mechanism for cooperation or communication, apart from standard market interactions, and a solution concept that should approach a Nash Equilibrium.

Aside from the environment category, MARL is subject to additional challenges compared with the Single-Agent framework [45]. The POSG is non-stationary and non-Markovian, as the transition probability function depends on all agents' actions sampled from policies subject to change during training. Moreover, MARL suffers from credit assignment issues, given that in the reward estimation functions, it is difficult to distinguish between a reward variation caused by the agent's actions and a change produced by other factors at play. Additionally, MARL algorithms may converge to suboptimal equilibria in environments governed by equilibrium solution concepts. Finally, MARL implementations and applications may suffer from scalability issues, given the accelerated growth in the number of states, actions, and observations in POSG as a function of the number of agents in the system. Nevertheless, MARL remains a rapidly evolving field, with ongoing advancements in both theoretical foundations and practical applications [44–46,51,52].

3.2.3. Proximal policy optimization and independent multi-agent learning

This section introduces PPO, following the work in [75,78]. Policy Gradient Algorithm aims to obtain a stochastic policy $\pi_\theta(a_t|s_t)$ that maximizes the expected reward of the RL agent, as described in expression (2). To produce such a policy, algorithms of the REINFORCE type, a category within PGM, apply stochastic gradient ascent (SGA) to an expression that seeks to find actions that maximize the expected reward in the given state, as measured by the comparison between the Action-state function *(the Actor)* representing the agent's policy, and the Value-state function *(the Critic)* [42,43,75].

PPO improves upon standard PGM and REINFORCE algorithms by redesigning the objective function to avoid significant updates in the Actor network [78]. The modifications mitigate the risk of converging to a local maximum during training and enable the re-use of the trajectories collected with a given Actor-Critic combination in multiple iterations *(epochs)* of SGA. In particular, PPO proposes a three-part objective function, shown in Eq. (5) and further disaggregated in expressions (6)–(9), where:

$$L(\theta, w) = L^{\text{CLIP}}(\theta) + hL^{\text{entropy}}(\theta) - vL^{\text{VF}}(w) \tag{5}$$

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t\left[\min\left(p_t(\theta)\hat{A}_t, \text{clip}\left(p_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] \tag{6}$$

$$\hat{A}_t = V_t^{target} - V_{w_{old}}(s_t) \tag{7}$$

$$L^{\text{VF}}(w) = \hat{\mathbb{E}}_t\left[\left(V_w(s_t) - V_t^{target}\right)^2\right] \tag{8}$$

$$V_t^{target} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{n-1} r_{t+n-1} + \gamma^n V_{w_{old}}(s_{t+n}) \tag{9}$$

- The terms $L^{\text{CLIP}}(\theta)$ and $L^{\text{entropy}}(\theta)$ guide the updates in the Actor network $\pi_\theta(a_t|s_t)$, while $L^{\text{VF}}(\theta)$ drive updates in the Critic network $V_w(s_t)$;
- θ and w are the parameters of the Actor and Critic networks. Moreover, parameters from the previous algorithm iteration are referred to as θ_{old} and w_{old};
- The main objective function of PPO, $L^{\text{CLIP}}(\theta)$, uses an Advantage estimation to shift the Actor Policy toward actions that maximize the expected reward while controlling for the maximum size in the updates using the combination of the clip and min functions;
- ϵ is a hyperparameter that, in conjunction with the clipping function, limits the ratio between new and old policies to remain in the range $[1-\epsilon, 1+\epsilon]$;

- $L^{\text{entropy}}(\theta)$ is an entropy term that procures the exploration of new strategies by inducing randomness in action selection in the Actor network. For cohesiveness, the entropy term is not expanded, but [75] includes expressions for both continuous and discrete action spaces;
- $p_t(\theta)$ is defined as the variation in actor policies between the algorithm updates, measured by the fraction $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$;
- A_t corresponds to the Advantage function, estimated via the difference between an estimate of the Action-state function, V_t^{target}, denoted as target state value, and the previous version of the Value-state function $V_{w_{old}}(s_t)$;
- $L^{\text{VF}}(\theta)$ drives the updates in the Critic network via minimizing the squared error between the predicted State-value function $V_w(s_t)$, and the estimated target state value V_t^{target};
- V_t^{target} is used as a proxy of the Action-state value function, and is calculated using the accumulated rewards in a given trajectory with length $(t+n)$; and
- h and v are coefficients that balance exploration, via the entropy term, and the weight of the value function loss with respect to the total loss.

Given the PPO objective function, the training cycle starts with initializing the Actor and Critic networks. Using the current version of the Actor Policy, trajectories are collected from the environment until the batch, controlled by the hyperparameter with the same name, is filled. Given these trajectories, auxiliary state-value functions are estimated. Moreover, multiple epochs of SGA are performed using the same batch of collected trajectories, either through randomly selected subsets of trajectories *(mini-batches)* or full-batches of trajectories. Results from the SGA are then used to update the Actor and Critic networks. Last, the process is repeated until convergence, or until the training budget is reached.

To implement IPPO, the single-agent version of the algorithm is extended to all agents in the environment. Starting from the definition of the Actor, $\pi_{i,\theta_i}(a_{i,t}|o_{i,t})$, and Critic networks, $V_{i,w_i}(o_{i,t})$, for agent i. In these definitions, parameters θ_i and w_i are, in principle, independent and different between agents. Both functions are no longer associated with the system state but depend upon the observations $o_{i,t}$ harvested by agents from the environment.

Given this starting point, the PPO objective function, showcased in expression (10) and detailed in (11)–(14), is extended across agents in the environment. In general, the description of the main variables from the single-agent case holds while extending the notation to account for the multiplicity of agents. Similarly, IPPO follows the same principles as the single-agent case. Critic and Actor networks are initialized for each agent. Trajectories are collected using the most up-to-date policies until the batch is complete, and auxiliary reward functions are estimated. The collected batch undergoes multiple SGD updates, independent for each agent, based on the PPO objective function. Networks are updated, and the process repeats iteratively until convergence or until the training budget is reached.

$$L_i(\theta_i, w_i) = L_i^{\text{CLIP}}(\theta_i) + h_i L_i^{\text{entropy}}(\theta_i) - v_i L_i^{\text{VF}}(w_i) \tag{10}$$

$$L_i^{\text{CLIP}}(\theta_i) = \hat{\mathbb{E}}_{i,t}\left[\min\left(p_{i,t}(\theta_i)\hat{A}_{i,t}, \text{clip}\left(p_{i,t}(\theta_i), 1-\epsilon_i, 1+\epsilon_i\right)\hat{A}_{i,t}\right)\right] \tag{11}$$

$$\hat{A}_{i,t} = V_{i,t}^{target} - V_{i,w_{i,old}}(o_{i,t}) \tag{12}$$

$$L_i^{\text{VF}}(w_i) = \hat{\mathbb{E}}_t\left[\left(V_{i,w_i}(o_{i,t}) - V_{i,t}^{target}\right)^2\right] \tag{13}$$

$$V_{i,t}^{target} = r_{i,t} + \gamma_i r_{i,t+1} + \gamma_i^2 r_{i,t+2} + \cdots + \gamma_i^{n-1} r_{i,t+n-1} + \gamma_i^n V_{i,w_{i,old}}(o_{i,t+n}) \tag{14}$$

Considering this baseline, it is worth discussing the implications of Independent Multi-Agent PPO in light of the MARL challenges highlighted in previous Sections. Beginning with the Advantage and Value function target definitions, where expressions (12) and (14) show that the accumulated rewards in the trajectories exclusively guide the evaluation of individual agent action performance. As a result, the algorithm will inherently suffer from Credit Assignment issues, as it is impossible to discern the actual effect of individual actions on the agent's performance beyond the information provided by the observations. Similarly, the estimates produced by value and policy networks are affected by exogenous variability in the environment, from the single-agent perspective, induced by other agents' actions. As demonstrated in [79], this issue can prevent meaningful learning altogether, even in relatively simple environments. Furthermore, IPPO does not explicitly constrain the learning process to reach an equilibrium. Because of that, the algorithm neither ensures that an equilibrium can be reached, nor guarantees that the obtained solution corresponds to a Nash Equilibrium [45]. Moreover, if an equilibrium is achieved, it would result from agents independently adjusting their strategies until significant variations in the Actor and Critic networks are no longer encouraged. In this sense, the algorithm cannot discriminate and/or select between multiple equilibria.

3.2.4. Algorithm selection for the multi-agent long-term electricity market environment

Despite the possible concerns highlighted in the last section, independent learning and PPO (IPPO) have been selected as the model's core *solver* for this work. This section provides the foundations for this selection, first addressing the key issues of algorithm selection, and later delving into training paradigms for multi-agent configurations.

To begin, it is helpful to summarize the market model's key characteristics and desired features. A primary contribution of the MARL approach is its ability to represent auctions as an explicit entry mechanism for market participants. In this context, enabling agents to develop stochastic policies could be advantageous, as it allows for a richer set of learned behaviors and strategies for market interaction. From an implementation standpoint, the market environment is fast and computationally efficient, making it well-suited for parallelized trajectory collection. Moreover, no explicit model is known beforehand to explain market interactions.

This set of characteristics leads to the selection of the family of model-free policy optimization algorithms, in line with the conclusions in [60]. Yet, in this case, the use of multi-discrete actions and the possibility of parallelization given the fast-running environment lean toward PPO [78] and IMPALA [80] algorithms, with the latter being better suited for large-scale RL applications. In contrast, algorithms derived from DDPG, such as MADDPG, are less appropriate for the current context, given their exclusive support for continuous actions [43]. This reasoning is consistent with recommendations from Ray and RLlib [72], the framework currently used for the implementation.

Regarding training paradigms, four main trends have emerged for multi-agent environments [45]. First, independent learning treats each agent as a separate learner using the same algorithm, resulting in fully decentralized training and action execution. Second, centralized training with decentralized execution uses shared value functions (e.g., system-wide action/state value functions) to address non-stationarity and credit assignment [79]. Third, mean-field methods enable agents to interact with a limited subset of peers, modeling local interactions while preserving decentralized training and execution [81]. Finally, population-based approaches train a single policy across multiple agents, leveraging diverse training conditions to enhance robustness [48].

For the model introduced in Section 3.1, independent learning emerges as the most suitable learning architecture. To start, GENCOs in the wholesale electricity market are competitors, where explicit information and strategy sharing are discouraged and/or strictly prohibited. Moreover, independent learning aligns well with a key characteristic of electricity markets: repeated market interactions. This repetition allows agents to refine their strategies continuously and, in some cases,

even develop tacit collaboration with competitors. Additionally, certain actions, such as bids in specific auction types, remain private and are never disclosed to other players. Furthermore, independent learning facilitates experimentation with heterogeneous and misaligned agent objectives.

On the contrary, alternative paradigms could face substantial limitations. Training a fully shared value function across dozens of agents, each with multidimensional discrete action spaces, poses severe scalability challenges that may hinder effective learning. Furthermore, shared-value function methods commonly assume a shared-reward function [79], a characteristic of cooperative environments that cannot be replicated for long-term electricity markets. Additionally, while effective in peer-to-peer energy trading scenarios [58], mean-field approaches are less relevant in wholesale markets where explicit coalition formation is excluded by design. Similarly, population-based methods, which augment the independent learning paradigm across multiple instances, remain out of reach of most academic research, given their extreme computational requirements [48].

These arguments strongly support independent learning as the most appropriate choice under limited computational resources. While population or league-based training may offer superior robustness when resources are abundant, independent learning provides a realistic, scalable, and conceptually faithful representation of competition in wholesale electricity markets.

Complementing the discussion related to algorithm selection and the training paradigm, multi-agent PPO has shown strong empirical performance in various competitive and cooperative scenarios [49,82]. Specifically, empirical studies have argued that the PPO objective function, designed to shield policy gradient methods from large and destructive policy updates, has also proven effective in mitigating the non-stationary conditions of multi-agent settings [49]. More specifically, these empirical results have shown that the PPO hyperparameters can be adjusted to trade off sample efficiency for training stability beyond what is customarily done in single-agent environments, which, in turn, appears to be the main reason behind successful multi-agent applications.

Overall, the previous reasons motivate the selection of IPPO for this work. Nonetheless, acknowledging the challenges associated with IPPO in competitive environments, further analyses are carried out before delving into market outcomes. Section 4 first evaluates the algorithm's performance, discusses hyperparameter selection, and compares the equilibria emerging from different parameter choices. Furthermore, extensive testing is conducted in Section 5, highlighting expected and unexpected behaviors in the solutions from an electricity market perspective, serving as complementary validation of the modeling approach.

4. Implementation details, training, and hyperparameter selection

This Section details the implementation, describes the training, and presents the hyperparameter Selection applicable to the current context.

4.1. Implementation details

The implementation is based on the RLLIB libraries, and particularly, the Python environment uses Python 3.9, Ray 2.4.3, PyTorch 2.1.2, and CUDA 11.8. For training, single computing nodes in a supercomputing system were used. Each node comprises two Intel Xeon Platinum 8360Y (36 cores), two NVIDIA A100 GPUs, and 800 GB of RAM. During training, 69 parallel environments were used for sampling, only one GPU was employed for SGA, and 50 GB of RAM were allocated. Job scheduling was carried out using the LSF system, with reproducibility scripts shared as part of the repository.

4.2. Training

Extensive tests are carried out to analyze the training behavior and serve as an input for the hyperparameter Selection discussed in the next Section. For these tests, two environment configurations are used: an EoM, denoted as environment A, and a system with all investment channels enabled, denoted as environment B. In both, 16 agents compete in the market; 8 incumbents with all technologies available for investments, seven entrants, each with one technology available, and a last agent dedicated to operating its mid-term hydro-reservoirs. The Maximum investment per technology is set to 4GW, a relatively high value considering the system conditions, thus increasing competitive pressure in the market. Apart from these differences, both environments abide by the description and input parameters presented in Appendix A.

With these environments, 58 tests are carried out starting from the recommendations of [49], showing performance improvements using conservative hyperparameter setups for multi-agent training. From this starting point, the runs aim to evaluate the contribution of the most relevant parameters in the algorithm: Clipping parameter ϵ, Entropy Coefficient h, Batch Size, and Actor and Critic Network Architectures. Regarding the latter, testing is conducted using Multi-Layer Perceptrons (MLPs) and Long Short-Term Memory (LSTM) configurations. Furthermore, a training budget of 25 h of wall time is set for all tests. Last, no hyperparameter scheduling is applied to ensure the training process represents algorithm behavior. Following these guidelines, Table 1 summarizes the training tests, while Tables B.8 and B.9 provide complementary information.

Showcasing the training behavior, Fig. 2 presents the evolution of aggregated reward as a function of sampled environment steps in four selected tests. Several consistent trends can be identified during training. First, training begins with large negative rewards, driven by excessive merchant investments beyond system requirements. This oversupply results in persistently low prices, limiting the profits agents can extract from the market. From this initial condition, agents gradually reduce their investments until aggregate profits turn positive. At that point, they typically continue reducing investments, increasing the profitability of their installed assets, until scarcity conditions emerge. In the last stages of training, agents refine their strategies, filtering profitable from unprofitable technologies for investment. Figures B.16 to B.19 further describe the training behaviors of test cases.

4.3. Hyperparameter selection

A Hyperparameter Selection using a complementary set of metrics to assess performance is carried out using the testing runs presented previously. The first of the metrics is the aggregated system reward, interpreted as the net present value of the profit accumulated by all agents in the simulation and given by the sum of expression (2) for all agents. It has been observed that a system with well-trained agents would exhibit a relatively low, but positive, aggregated reward. This is because, while individual agents try to maximize their profits, they should be limited by market interactions with their competitors. Moreover, periods with excessive scarcity events, those behind large profit spikes in the simulation, would also open the opportunity for additional investments in the system. Finally, none of the agents are obliged to participate in market sessions. Consequently, if profits harnessed from the system are consistently negative, agents can stop investment altogether, thus mitigating their losses.

Nonetheless, a relatively low aggregated reward could stem from various factors and conditions, such as inefficient/random agent behavior or uneven distribution of profits across agents. As a result, relying solely on the aggregated reward could be insufficient to assess the hyperparameter configurations properly. Therefore, three additional metrics are proposed to facilitate comparison and evaluation across runs:

Table 1
Summary of tests conducted during the hyperparameter search. According to the order presented in the testing category, the parameter of interest is increased *(e.g. for Batch Size tests, M.6 configuration has a higher Batch Size than M.5)*. For individual hyperparameter tests, the corresponding code combines the environment and the set of hyperparameters, where A indicates the Energy-only-Market and B the Capacity Market environments *(e.g. AM.11 indicates the test in the EoM environment)*.

Code	Network type	Test description	Code	Network type	Test description
M.1, M.2, M.3, M.4	MLP	Increasing clipping factor - ϵ	*L.3, L.4*	LSTM	MLP in tail Medium LSTM Increasing batch size
M.5, M.6, M.2, M.7, M.8, M.9	MLP	Increasing batch size	*L.5, L.6*	LSTM	MLP in tail Long LSTM Increasing batch size
M.10, M.2, M.11	MLP	Increasing entropy coefficient - h	*L.7, L.8*	LSTM	MLP in head Short LSTM Increasing batch size
M.12, M.2, M.13, M.14	MLP	Varying network architecture by increasing the size in hidden layers	*L.9, L.10*	LSTM	MLP in head Medium LSTM Increasing batch size
L.1, L.2	LSTM	MLP in tail Short LSTM Increasing batch size	*L.11, L.12*	LSTM	MLP in head Long LSTM Increasing batch size

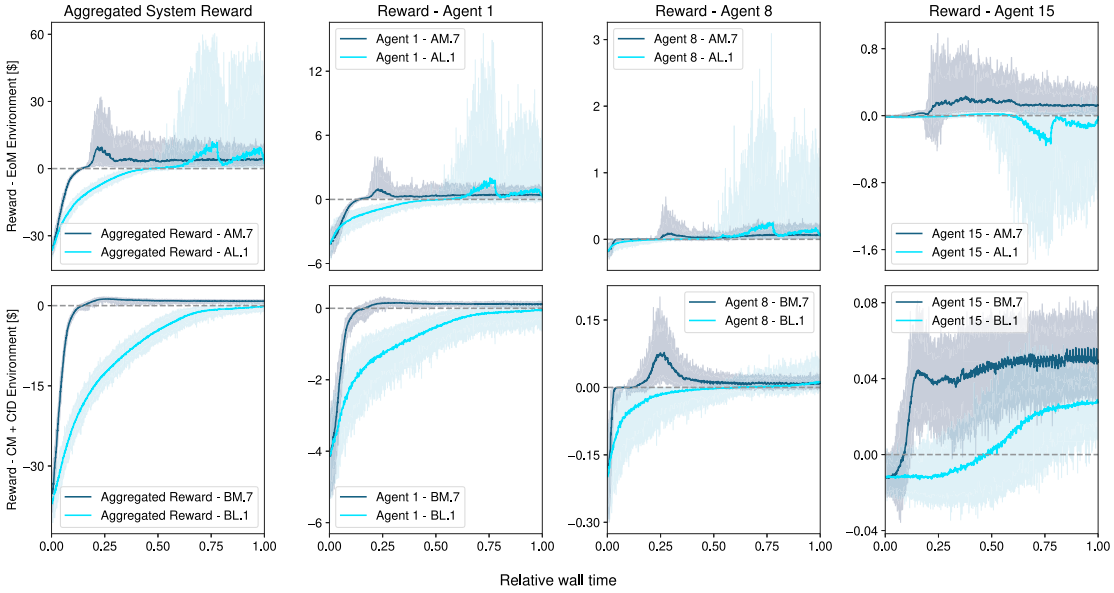


Fig. 2. Evolution of aggregated and individual reward during training for Hyperparameter configurations M7 (*MLP network)* and L1 (*LSTM network)*. The upper graphs present results for the EoM environment, while the lower ones concentrate on the Capacity plus CfD market. The left panels show the aggregate reward for the system. Panels in the second to the fourth columns present the reward evolution during training for Agent 1 *(Incumbent)*, Agent 8 *(Entrant - Solar PV)*, and Agent 15 *(mid-term storage operation)*, respectively. Results are normalized according to the relative wall time used for training. Solid curves indicate average values obtained during sampling, while shaded areas represent minimum and maximum values.

- **Penalty:** Following the auxiliary reward function implemented in [60], a metric, initially conceptualized as a penalty to be used during training, was devised to quantitatively measure the distance between the agent's decisions and the market equilibrium. The penalty aims to represent the net present value that can be obtained by projects participating in the market in a given system condition. This profitability is based on agents' actions and virtual plants and is calculated per agent, market, and technology.
- **HHI index**: In its current version, the environment models relatively homogeneous GENCOs, differentiated only by their existing assets and enabled investments. Thus, it is reasonable to expect that, if training is successful across market participants, most of them should be able to place investments in the system. As a proxy to measure uniformity across agents, the Herfindahl–Hirschman Index (HHI) [83] is calculated using the installed capacity at the end of the simulation.
- **League Ranking:** Inspired by the League-based training used in [48], a league-based evaluation of the hyperparameter configurations is implemented. In the league, market simulations are launched using, for each agent, a random selection of networks *(e.g. Agent 1 uses networks from run M.6, Agent 2 from run L.1, and so on)*. After one episode is completed, the agent's accumulated rewards are recorded. By repeatedly randomizing network selection and running multiple episodes, it is possible to compare the performance of hyperparameter configurations by directly using the agents' reward.

To compare results among runs, market simulations are carried out using the most updated agents for each particular hyperparameter configuration to evaluate system performance, given the lack of stopping criteria in the algorithm. Starting from the aggregated reward, as displayed in Fig. 3, results show that hyperparameter selection and network configuration significantly impact the learning process, system

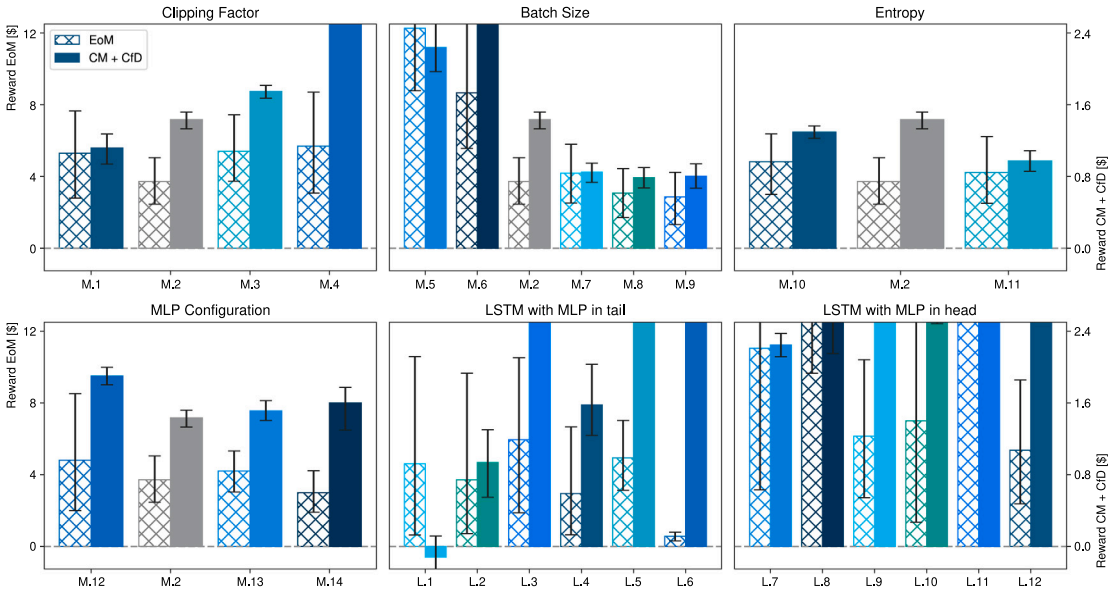


Fig. 3. Average aggregated reward for different hyperparameter configurations in the EoM and CM + CfD environments. Results are obtained using 100 episodes in the environment with the most updated agents' versions. Hatched bars indicate the outcome for the EoM environment, while solid bars represent the CM + CfD environment. Error bars showcased the 10th and 90th percentiles from the 100 episodes. The Y axes in the Figure are adjusted to facilitate comparison among the most relevant hyperparameter configurations.

interactions, and outcomes. Among the hyperparameter selection, it can be noted that increasing the clipping factor while reducing the batch size leads to higher aggregated rewards. In contrast, modifying entropy and MLP configurations results in negligible changes. On the other hand, LSTM-based configurations exhibit greater fluctuations in aggregated reward, likely due to a more unstable learning process, as illustrated in Figures B.16 to B.19.

Fig. 4 presents the application of the additional evaluation metrics to the hyperparameter runs. In the polar plot, runs/curves approaching the unit circle are the best performers. These results confirm the positive impact of increasing the Batch Size in the PPO algorithm and corroborate the erratic behavior of all LSTM configurations tested. Regarding other parameters evaluated, the tests are less conclusive. Yet, a slight advantage, especially in the League Ranking, is observed for hyperparameter configurations with intermediate Clipping Factors and higher Entropy terms. Further details regarding the metrics are provided in Figure B.20, for the Penalty and HHI index, and in Table B.10, for the League Ranking.

Based on previous assessments, an additional ablation study uses the hyperparameter configurations presented in Table 2. Fig. 5 presents a consolidated view of aggregated rewards and complementary evaluation metrics, where comparisons are made exclusively among these relatively well-performing configurations. The results highlight the importance of the Batch Size in the algorithm, as it is the only hyperparameter that differs substantially across configurations. In contrast, other configurations exhibit similar performance, making it challenging to determine an optimal choice based on the selected evaluation criteria. Ultimately, configuration T.1 is chosen for all subsequent tests in this study. This decision is based on its relatively large network, which enhances the expressiveness of agent actions, its higher entropy, which led to a slight performance improvement across all tests by improving exploration, and its intermediate clipping factor. Nonetheless, no significant changes would be expected if any configurations from Table 2, except for T.4, were to be used instead. Importantly, the selected Batch Size enables simulations of up to 40 agents in the tested hardware.

Importantly, the hyperparameter search conducted in this section, the primary source of computational burden in this work, is not required for the practical use of the MARL model. With the selected hyperparameters, training sessions typically reach stability in the main trackable metrics within 20–40 h, depending on the number of agents in the system. Once agents are trained, market simulations can be performed with minimal time and computational requirements, partially offsetting the high training costs. Future work could explore methods that allow a single training setup to adapt effectively to a range of policy parameters *(carbon taxes, auction price caps, and other policy options)*, in addition to other methodological improvements that enhance the computational efficiency in the current framework.

Table 2
Tests and hyperparameters used in the ablation study. For individual tests, hyperparameters not mentioned in the corresponding field are not modified, and the values from Table B.8 are used instead.

Code/Parameter	Clipping factor	Batch size	Entropy	MLP configuration
M.7	0.05	35 328	0.000001	[256-256]
T.1	0.1	35 328	0.01	[512-512]
T.2	0.1	35 328	0.01	[256-256]
T.3	0.1	35 328	0.000001	[512-512]
T.4	0.1	17 664	0.01	[512-512]
T.5	0.05	35 328	0.01	[512-512]

5. Long-term electricity market results

Considering the modeling framework and the hyperparameter search presented previously, this Section applies the MARL model to a system inspired by the Italian electricity system. The analysis aims to showcase the models' capabilities while dynamically representing decarbonization pathways under different market designs, policy scenarios, and competition levels. To this end, the Italian electricity system provides a suitable baseline, given its continued reliance on fossil-fuel-based technologies alongside a relatively high penetration of RES. Nonetheless, the analysis can be readily extended to other systems with comparable characteristics *(energy mix, decarbonization policies, demand growth expectations, among others)*.

For the scenarios, the period between 2020 and 2040 is selected, using the starting conditions regarding installed capacity and the policy

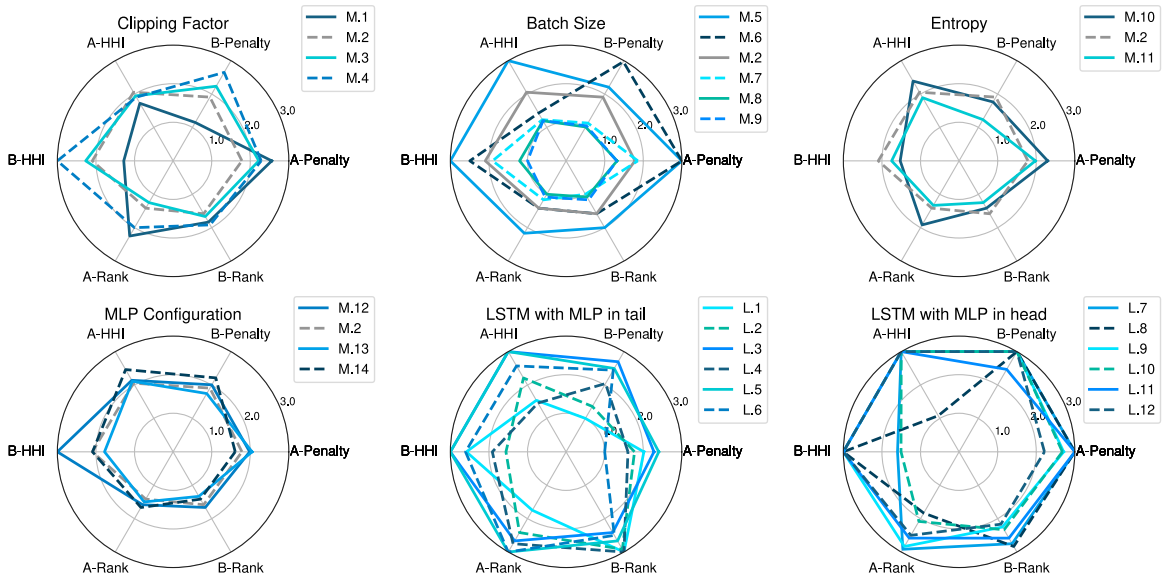


Fig. 4. Average Penalty, HHI index, and League Ranking for hyperparameter configurations in the EoM *(A)* and CM + CfD *(B)* environments. Penalty and HHI index Results are obtained using 100 episodes in the environment with the most updated agents' versions. League Ranking is obtained from the competition set between all hyperparameter configurations per environment, where agents rank according to their overall performance between 1 *(best)* and 26 *(worst)*. To facilitate visualization in the polar plot, all metrics undergo zero-centered median normalization, are clipped between 0 and 2, and are shifted by one unit.

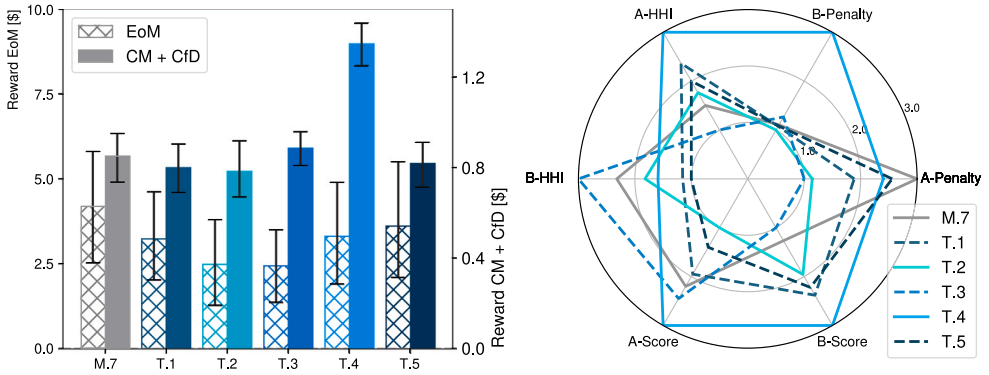


Fig. 5. Average Penalty, HHI index, and League Ranking for hyperparameter configurations from the ablation study in the EoM *(A)* and CM + CfD *(B)* environments. Penalty and HHI index Results are obtained using 100 episodes in the environment with the most updated agents' versions. League Score is obtained from the competition set between the hyperparameter configurations from the ablation study per environment, where agents are scored according to their overall performance between 0 *(best)* and 1 *(worst)*. To facilitate visualization in the polar plot, all metrics undergo zero-centered mean normalization, are clipped between 0 and 2, and are shifted by one unit.

scenarios briefly summarized in Fig. 6. Complementary, demand scenarios are obtained from planning exercises carried out with a central planning energy model (PyPSA-Eur), accounting for the electrification of the energy sector and discounting electricity imports from total consumption. This exercise results in an average demand growth rate of close to 2% for the period under study. Besides these conditions, the system abides by the characteristics introduced in Section 3.1. Importantly, the planning carried out with PyPSA avoids lost-load events.

Initially, 12 scenarios that combine market designs and competition levels are evaluated. On the long-term market design side, four regulatory frameworks are tested; an Energy-only-Market *(EoM)*, a market combining features from the EoM with CfD auctions to meet a RES target *(CfD)*, a Capacity Market to ensure system adequacy *(CM)*, and simultaneous implementation of a Capacity Market and CfD auctions *(CM+CfD)*. Regarding competition levels, the market designs are tested on environments with 8 *(6 incumbents, 2 entrants)*, 16 *(8 incumbents, 8 entrants)*, and 32 *(16 incumbents, 16 entrants)* agents. For each of the previous configurations, existing assets are allocated to incumbent agents, resulting in HHI coefficients of 3000, 1880, and 1000, respectively, thus representing markets with high, moderate, and low levels of market concentration. Furthermore, the decommissioning shown in Fig. 6 is applied uniformly to all incumbents. The maximum investment per technology in the system is maintained constant and divided equally across agents, to allow the number of agents and the emerging competition among them to be driving factors in the simulations. Consequently, the limit for generation technologies and short-term storage, per market and year, is set to 8 GW and 1.6 GW, respectively, which in both cases, as an aggregate, is a comfortable upper bound for the requirements imposed by demand growth. To all previous simulations, a "Moderate Carbon Tax" rising from current levels in the Emission Trading Scheme to around 150€/tCO2 is applied. This carbon tax schedule reflects a rising ambition in decarbonization plans, though not at the levels expected to reach the EU and Italy's

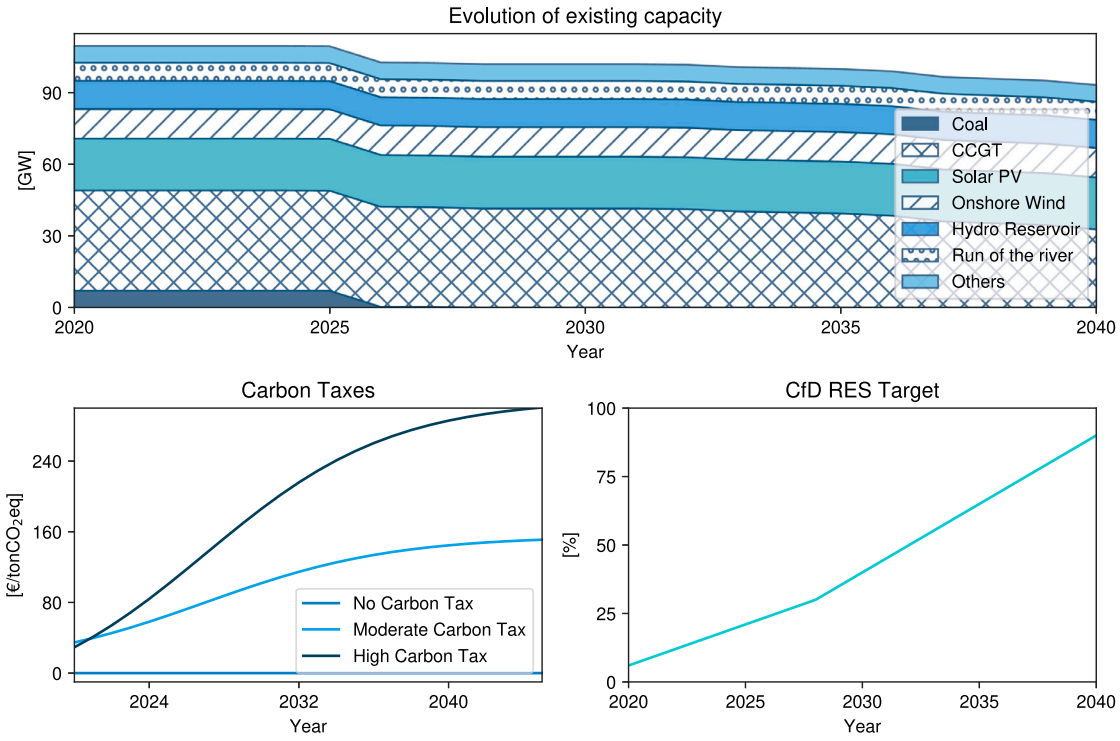


Fig. 6. Existing conditions and main policy assumptions for the study cases. The upper panel showcases the evolution of existing assets, where the key factor is coal power plants decommissioning before 2030. The lower left panel presents the Carbon Tax scenarios, while the lower right one displays the RES penetration target to be achieved with the contracts-for-difference market.

2040 and 2050 climate targets. Thus, we carry out sensitivity with a higher carbon tax schedule. Finally, a computational budget of 16, 20, and 32 h is used for training scenarios with 8, 16, and 32 agents.

To analyze market outcomes, Fig. 7 introduces the installed capacity at the end of the study period, selected as 2040. For the EoM scenarios, expansion is driven by solar, offshore wind, and Open-Cycle Gas Turbines (OCGT), while Combined-Cycle Gas Turbines (CCGT) are limited to a few GWs, and Coal plants are negligible. Except for OCGT and short-term storage, these results are relatively close to the central planning exercise carried out with the PYPSA optimization framework, which has been harmonized by sharing, when possible, input parameters, technical assumptions, and the main energy policy with the current implementation of the MARL model. In general, PyPSA allocates greater capacity to dispatchable technologies such as OCGT and CCGT and storage solutions such as batteries. This outcome is largely attributable to its high temporal resolution, hourly representation across all 8760 h of the year. While the total installed capacity of variable renewable energy sources is broadly comparable across frameworks, PyPSA typically results in a higher share of solar power, reflecting its assumption of progressively declining investment costs for this technology over the modeled period, in addition to the incentives emerging from centralized planning for storage technologies, in contrast to the decentralized market signals in the MARL model. A more detailed discussion of the characteristics of PyPSA and its contrasts with MARL, which account for the observed differences in results, is provided in Appendix D.

The energy outcomes from the MARL simulations change under different market designs. In the case when the CfD market is introduced as the only long-term mechanism (CfD scenario), offshore wind penetration increases considerably, while solar penetration remains relatively constant. Yet, the entry mechanism chosen by agents for solar investments shifts from merchant investment toward long-term auctions. For the case in which the capacity market is the long-term mechanism instead (CM scenario), offshore wind similarly enters the market through these auctions. Still, the main effect of the capacity market is the increase in the share of OCGT in the mix, bringing it closer to the planning model. This similarity can be explained by the fact that the planning model rewards the flexibility of OCGT not through markets, but through the explicit constraint imposed in the system to avoid lost load events. When both long-term auctions are combined (CM + CfD scenario), the composition of the generation mix resembles the case in which only the Capacity Market was in place. However, RES shifts investments from other mechanisms to the long-term CfD auctions. In terms of installed capacity, no significant deviations occurred depending on the number of agents in the simulation.

To capture the dynamics of the transition toward the generation mix previously described, Figs. 8 and 9 present the evolution of CO_2 emissions and electricity prices over the study period. Regarding emissions, we find significantly different outcomes depending on the choice of market designs. In addition to the decommissioning of fossil fuel assets and the competitive costs of renewable technologies, particularly solar PV, we find no significant emission reductions by 2040 in the EoM cases, regardless of the level of competition. In contrast, when the long-term mechanisms of the CfD and CM market designs are introduced, with the subsequent increase in total installed capacity illustrated in Fig. 7 from all technologies, but especially RES, emissions are substantially reduced. Notably, the three market designs that incorporate long-term mechanisms (CfD, CM, and CM+CfD) produce similar outcomes in terms of emissions in 2040, mainly due to the opportunities created by both CfD and Capacity Market auctions for high shares of offshore wind deployment. Nonetheless, emission trajectories are more dynamic, with agents in the CfD market designs delaying merchant investments before 2030, before entering the market in mass once the RES target showcased in 6 becomes more stringent.

In terms of electricity prices, two key trends emerge. First, the EoM design is characterized by frequent scarcity events, particularly toward the end of the simulation period. These events, which lead to price spikes and load curtailment, become more prevalent under lower

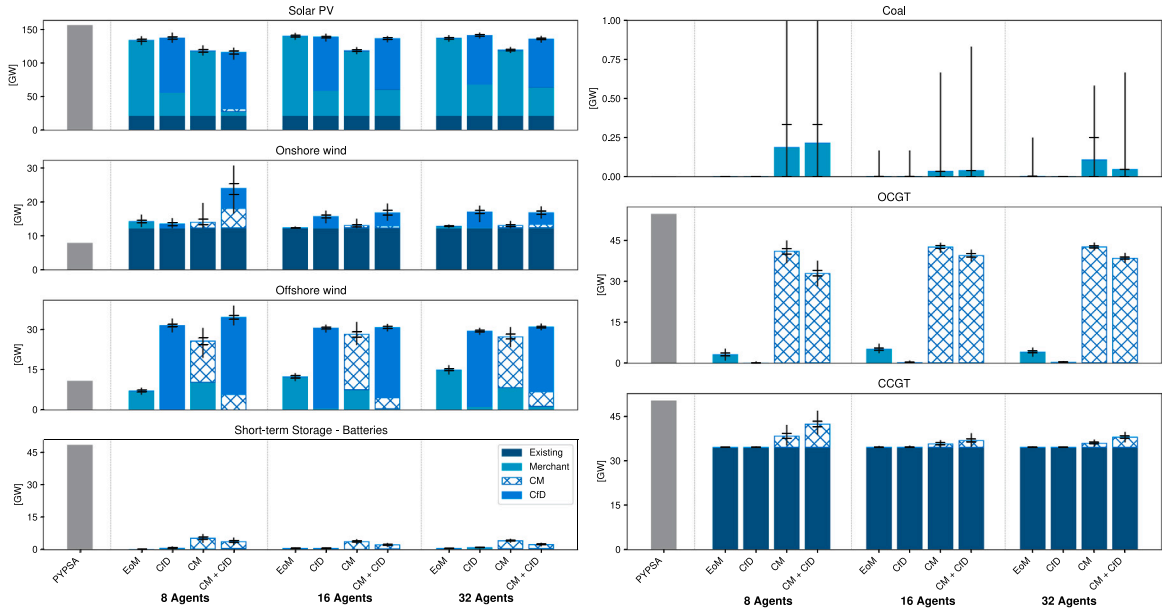


Fig. 7. Installed Capacity of generation and storage technologies in 2040 under different market designs and competition levels. Stacked bars indicate average installed capacities across runs. In each stacked bar, the horizontal markers display the 25th and 75th percentiles, and the vertical markers minimum and maximum values. Columns organize the installed capacities in the four market designs. Furthermore, results are grouped in the column categories, divided by vertical dotted gray lines, according to the number of agents used in the simulation. Different hatching highlights the mechanism used by agents to enter the market. Gray bars, for the corresponding technology, show the output from the central planning exercise carried out in PYPSA. Results from the MARL model are obtained across 200 market simulations using the trained agents.

levels of competition. Although CfD auctions are not explicitly designed to address scarcity, they partially mitigate these events and reduce their frequency. In contrast, Capacity Market designs prevent scarcity events altogether during the simulation. This outcome suggests that the Capacity Market's current design procures more adequate services than strictly necessary in an efficient design, especially considering that such markets are intended to meet a Loss of Load Expectation (LOLE) between 3–4 h per year [84]. Further details on electricity prices in the short-term market for 2040 and the behavior of CfD and CM premiums are presented in Appendix Figure C.23. Crucially, concurrently implementing CfD and CM auctions leads to increased CM premiums. This result is consistent with the exacerbated *Missing Money* problem arising from the additional renewable energy deployment incentivized by CfD schemes.

To further explain the investment behavior of agents, Table 3 presents the aggregated Internal Rate of Return (IRR) and installed capacities by agent type across simulations. In contrast, Table C.11 offers more detailed results disaggregated by technology and market design. A key observation is that, in all scenarios, the aggregate IRR exceeds the 8% discount rate applied in the environment. Nonetheless, the implicit discount rates decreased as the number of agents and the level of competition increased. However, this attenuation effect was not substantial and was less pronounced in the transition from 8 to 16 agents compared to the transition from 16 to 32 agents. These patterns can possibly be attributed to three factors: the emergence of risk-averse strategies during training, limited competition and/or potential collusion induced by the learning algorithm, and the interaction between market design and model input parameters.

Regarding the environment's design, it is essential to highlight that when investments fail to yield profits, agents quickly learn to avoid investing altogether, since there is no explicit penalty for abstaining from market participation. As a result, agents tend to adopt strategies where they limit their investment commitments to opportunities with near-guaranteed profitability. Such behavior, which translates to agents requiring a positive difference between the implicit internal rate of return derived from their investments and the exogenous discount rate used to calculate the net present value of their portfolios, is commonly associated with risk aversion [85]. This occurs despite the environment optimizing a risk-neutral objective per agent, as described in Section 3.2. Regarding the second factor, Section 4 shows that the system quickly converges to an equilibrium, with limited exploration or further improvement in agent strategies. This stagnation may promote implicit collusion, as market outcomes become relatively stable. As discussed in [48], addressing this issue typically requires multiple training restarts to encourage exploration, leaving a stream for future work. Finally, the relatively high endogenous discount rates observed among agents can be linked to active constraints in the model. This is particularly noticeable in the case of solar PV, where investments in 2040 approach the maximum allowable investment level for the study period.

To study the previous hypothesis related to maximum investment quantities, but also to further understand and validate modeling outcomes, a comprehensive set of sensitivities and experiments is carried out. Detailed results are provided in Appendix C, with the following highlighting the main findings from the exercise:

- **Carbon taxes:** Complementing the base case, implemented with the moderate Carbon Tax, the two additional Carbon Tax scenarios *(No Tax and High Carbon Tax)* illustrated in Fig. 6 were tested under different market designs. Additionally, the tests were carried out using 16 agents, representing an intermediate level of competition in the market. Installed capacities for the year 2040 under the Carbon Tax scenarios are presented in Fig. 10. Overall, the penetration levels of most technologies remain relatively consistent across scenarios, suggesting competitive low-carbon investments remain the dominant factor shaping the generation mix in the coming years. However, carbon taxes had a relevant effect for CO_2 emissions reductions, especially in the EoM and CM market designs. Moreover, the high Carbon Tax scenario facilitates increased investment in offshore wind, resulting in a notable reduction in emissions. This finding underscores the importance of analyzing climate policies such as carbon taxes within the

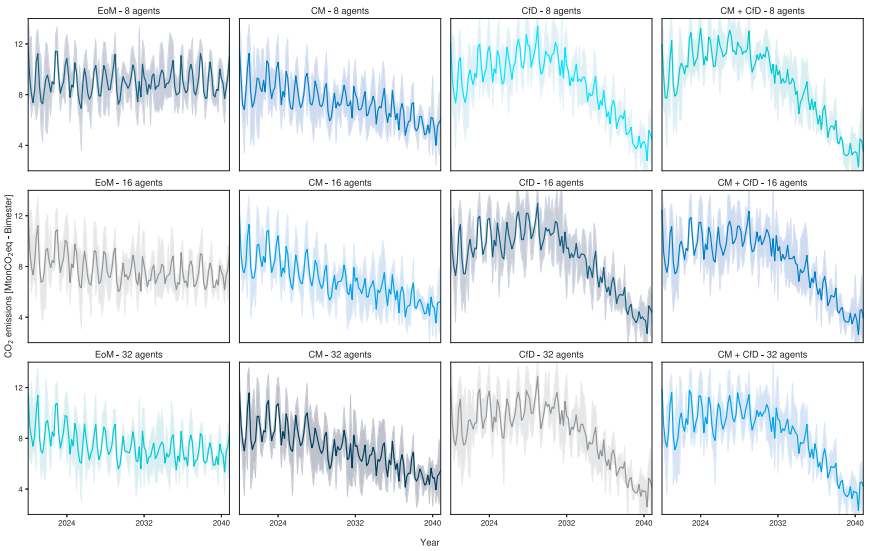


Fig. 8. Bi-monthly CO_2 emissions under different market designs and competition levels. Emissions are obtained using the electricity generated per technology, and its corresponding emission factor. In the plot, rows indicate the number of agents used in the simulation, while columns correspond to the four market designs. Solid lines display average values, while shaded areas are the 25th and 75th percentiles. Results are obtained across 200 market simulations using the trained agents.

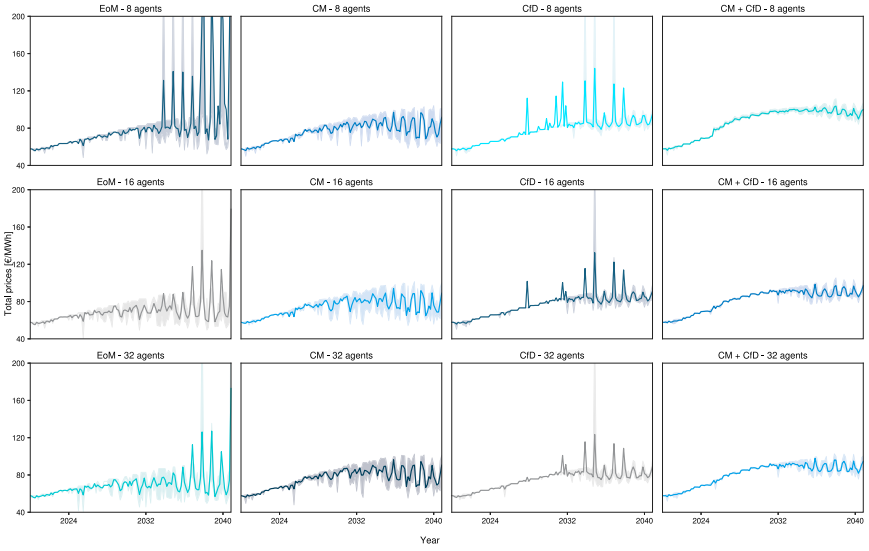


Fig. 9. Total system prices under different market designs and competition levels. Prices are calculated as the net price incurred by demand, including the short-term prices, the premiums from the capacity and contract for difference markets, and any financial settlement derived from these mechanisms. In the plot, rows indicate the number of agents used in the simulation, while columns correspond to the four market designs. Solid lines display average values, while shaded areas are the 25th and 75th percentiles. Results are obtained across 200 market simulations using the trained agents.

Table 3
internal rate of return of investments and share of investments for incumbent and entrant agents under different market designs and competition levels. The internal rate of return is calculated as an aggregate for all investments carried out for the agent during the simulation. The share of investments for agents is calculated using the installed capacity [MW] across generation and storage technologies.

Agent type		Number of agents and market design											
		8				16				32			
		EoM	CfD	CM	CM + CfD	EoM	CfD	CM	CM + CfD	EoM	CfD	CM	CM + CfD
Incumbents	IRR [%]	15.03	16.63	11.47	16.21	11.38	15.51	10.90	14.39	10.59	14.53	11.47	14.11
	Share of investments [%]	72.52	77.34	79.44	76.40	57.65	56.88	54.44	57.94	58.29	56.76	52.29	53.80
Entrants	IRR [%]	14.19	16.47	11.43	17.48	10.69	16.38	10.78	15.38	10.44	15.13	11.28	14.15
	Share of investments [%]	27.48	22.66	20.56	23.60	42.35	43.12	45.56	42.06	41.71	43.24	47.71	46.20

context of complementary policy instruments and market design, which have first-order consequences for the energy sectors and their decarbonization.

- **Maximum investment quantities:** In this sensitivity, the investment cap for solar PV and OCGT technologies doubled compared to the base scenario. The results show an increased share of solar PV in the mix, partially replacing offshore wind investments, especially within CfD auctions. Although this shift leads to a slight reduction in overall IRR, solar PV projects still yield high returns, indicating that agents are still maximizing the share of solar PV eligible for auction participation, which remains highly profitable. These results indicate that all the factors hypothesized to explain the difference between the IRR and the exogenous discount rate have an effect in explaining the observed market results.
- **Market design:** Variations in auction price caps and market targets for CM and CfD markets are tested and compared to the base scenario *(CM + CfD market environment with a 16-agent configuration, also used for the discount rate and robustness analysis)*. On the one hand, the results demonstrate that the model is less sensitive to auction prices than market targets, indicating that market dynamics and competition are expected to be the primary drivers of system outcomes. Yet, differences are observed for CfD auctions when the uptake of the RES target accelerates, denoting a tendency in the trained agents to approach the price-cap in the auctions in such cases. On the other hand, increasing the targets of these mechanisms leads to over-investment in the system with subsequent trade-offs in total system prices. Notably, changes in the adequacy target for the capacity market had limited effects on emissions, while higher RES penetration levels in CfD auctions yielded additional mitigation potential. Moreover, faster uptakes in the RES target yield higher emission reduction potential, especially around the year 2030, but also increase the pressure on system prices and reduce incentives for merchant investments. These findings underscore the critical need for adaptive decision-making in market design, as the current implementation maintains a fixed market structure. Thus, the analysis also showcases the opportunity to model complementary incentives for short-term storage and other flexibility sources, which maintained a limited participation through the capacity market across sensitivities, remaining the largest differentiating factor with the central-planning model.
- **Discount rates:** A range of exogenous discount rates, uniformly applied across all agents in the system, was tested, leading to three main trends. First, higher discount rates resulted in a rapid shift away from merchant investments, with the subsequent utilization of long-term auctions to meet demand requirements. However, the higher auction utilization and increased prices obtained through these mechanisms resulted in substantially higher system costs. Finally, low discount rate scenarios aligned with rapid decarbonization trends through faster and higher RES investments, with notable emission reduction potentials around 2030. These results further validate the model and reiterate the critical importance of low financing costs for achieving net-zero targets, as highlighted by [86,87].
- **Robustness analysis:** Using a base scenario, the MARL training setup was repeated multiple times. Across sessions, the main training metrics and key market variables, such as system prices, emissions, and total installed capacity, remained largely unchanged. However, some degree of substitutability was observed in the technologies selected by agents. From a methodological perspective, this assessment highlights that for in-depth scenario analysis, retraining sessions serve as an additional measure to evaluate the type of solution obtained by the model. From a market perspective, the emergence of different technology portfolios achieving similar aggregate results, even in a relatively static and risk-neutral setting, underscores the opportunity to enhance active planning approaches where the market design considers not only a purely risk-neutral economic perspective but also incorporates differentiated metrics such as resilience to shocks, diversification levels, and other system characteristics to promote and differentiate portfolios.

6. Conclusions and future work

With the advent of MARL, recent advances in algorithmic development and increased hardware availability, new frontiers in modeling capabilities are opening across a wide range of applications. Building on these advancements and drawing from previous efforts applying MARL in the electricity sector, this work offers new steps toward applying MARL methodologies to long-term electricity market design. This work helps close the gap between agent-based and partial-equilibrium modeling paradigms, two of the most common methods currently used for long-term market design and assessment. In this sense, the MARL model introduced, implemented, and tested presents an alternative tool for policymakers to test, evaluate, and refine, or even radically redesign, regulatory frameworks in response to the challenges of the energy transition.

The main results, and their sensitivities, highlighted important caveats regarding the role of long-term markets in stringent decarbonization pathways. Long-term electricity markets provided a pathway for the system to achieve investment levels sufficient to meet demand requirements while maintaining consistent decarbonization trends. This feature remained consistent even under policy misalignment (for example, in the absence of a carbon tax) or during market pressure (as in scenarios with elevated discount rates). Moreover, a comparison with a centralized planning model indicates that current market structures provide insufficient incentives for flexibility technologies such as short-term storage. However, while the large-scale deployment of storage solutions is expected to generate system-wide benefits, their remuneration within the MARL market framework remains underdeveloped, a market failure commonly observed in real-world systems. This gap underscores the potential value of establishing formal long-term flexibility markets, thus informing regulatory decisions on this issue.

A key strength of MARL in the context of electricity market design is the modeling flexibility that it would offer to practitioners, market actors, and eventually, to policymakers. The trained agents demonstrate the ability to respond to various regulatory configurations while acting and competing against other agents in the system. Particularly noteworthy is the simultaneous evaluation of multiple long-term expansion mechanisms, which require agents to assess investment profitability under market conditions that influence short- and long-term cash flows. More importantly, the model outcome is obtained with minimum assumptions, emerging almost endogenously from the training and model configuration.

This flexibility, intrinsic to the MARL framework, enables exploring traditionally hard-to-model features in electricity systems, such as generator contracting strategies or bilateral agreements between utilities. Crucially, this flexibility could also be extended to future market designs for which no historical precedent or proxy for comparison exists. This includes hybrid market systems, where agents operate competitively within a centrally planned context that imposes objectives such as system adequacy or emissions reductions. As such, the proposed modeling setup is well-suited to future research efforts to bridge the gap between centralized planning and decentralized market-based decision-making, a key factor for achieving decarbonization targets in energy sectors worldwide.

Nonetheless, the use of independent learning and Proximal Policy Optimization, an effective method for multi-agent competitive environments, requires careful consideration. In particular, appropriate

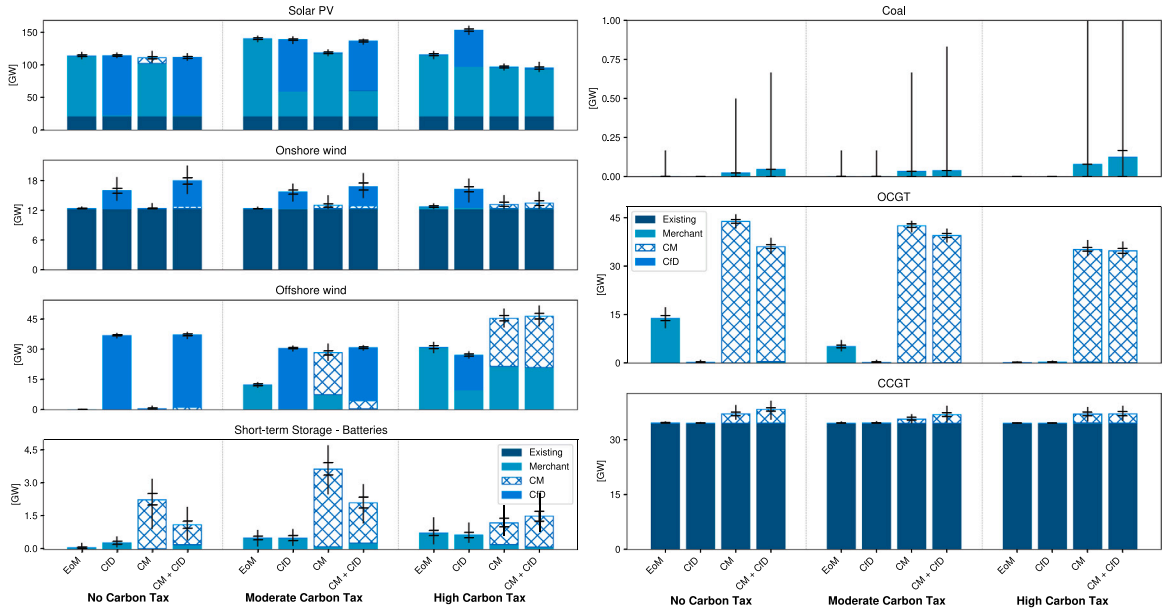


Fig. 10. Installed Capacity of generation and storage technologies in the year 2040 in simulations considering 16 agents and different carbon tax scenarios. Stacked bars indicate average installed capacities across runs. In each stacked bar, the horizontal markers display the 25th and 75th percentiles, and the vertical markers the minimum and maximum values. Columns organize the installed capacities in the four market designs. Furthermore, results are grouped in the column categories, divided by vertical dotted gray lines, according to the carbon tax used in the simulation. Moreover, hatching highlights the mechanism used by agents to enter the market. Results from the MARL model are obtained across 200 market simulations using the trained agents.

hyperparameter selection and robustness analysis proved essential for enhancing the quality of the solutions. More importantly, the potential improvements that MARL offers within this modeling framework remain constrained by the inherent limitations of agent-based models. First, the model assumes rational decision-making by agents, which may not reflect real-world behavior. Second, the analytical scope is limited, excluding interactions with additional power sector stakeholders and other economic segments, such as the financing sector. Additionally, the model does not account for transformational and structural changes beyond the current market structure and electricity system organization, a significant limitation given the long-term nature of the analysis. These constraints, along with the broader limitations typical of energy and power sector models, are not directly addressed through the application of MARL.

Future research could explore MARL applications incorporating a broader and more diverse set of electric market agents and stakeholders. Notably, this includes prosumers, driven by digitalization and democratization trends, who are expected to play a central role in future energy systems by integrating supply and demand-side incentives into their decision-making. More broadly, the MARL framework can be extended to include active regulatory agents in the system. As demonstrated in prior studies, such a regulator can contribute to the strategic design of electricity markets in response to the strategic behavior of market participants. In particular, the regulator could be modeled as controlling key policy levers such as market caps and carbon pricing, and, specific to the model presented in this work, the targets and objectives of long-term auction mechanisms to pursue system-level goals.

However, further development is required before this modeling approach can match the performance and realism of established electricity market models in the academic literature. Notably, the implementation lacks several key technical constraints essential for comprehensive market analysis, including unit commitment constraints, transmission system modeling, and a detailed representation of system flexibility requirements and associated markets. These aspects will be addressed in future work. In addition, while risk-averse behavior has emerged organically under specific configurations in the current setup, explicitly modeling agents' risk preferences through dedicated reinforcement learning algorithms designed to optimize risk-sensitive metrics would provide greater control and insight into strategic behavior under uncertainty. From an algorithmic standpoint, improvements are also needed to enhance the applicability of MARL in competitive settings. Large-scale implementations of MARL at the research forefront remain beyond the reach of most academic studies, even this one, which benefited from access to supercomputing resources. Advancements in MARL algorithms that improve convergence, facilitate exploration of alternative equilibria, and expand the diversity of agent responses in competitive environments would significantly strengthen this research.

CRediT authorship contribution statement

Javier Gonzalez-Ruiz: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Carlos Rodriguez-Pardo:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Iacopo Savelli:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Alice Di Bella:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Massimo Tavoni:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and Grammarly in the writing process to improve the readability and language of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article

Funding

Javier Gonzalez-Ruiz, Carlos Rodriguez-Pardo, and Massimo Tavoni acknowledge support from the European Research Council, ERC grant agreement number 101044703 (EUNICE) CUP D87G22000340006.

Alice Di Bella acknowledges funding from European Union PNRR - Missione 4–Componente 2–Avviso 341 del 15/03/2022 - Next Generation EU, in the framework of the project GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018 – CUP C83C22000890001). Iacopo Savelli received funding from the European Union's Horizon Europe programme under the Marie Skłodowska-Curie grant agreement number 101148367.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.egyai.2025.100665.

Data availability

The code for this framework is open-source, available via the following public repository: https://github.com/jjgonzalez2491/MARLEY_V1.

References

[1] Change IPOC. Working group III contribution to the sixth assessment report of the intergovernmental panel on climate change. 2022.

[2] Newbery DM, Pollitt MG, Ritz RA, Strielkowski W. Market design for a high-renewables European electricity system. Renew Sustain Energy Rev 2018;91:695–707. http://dx.doi.org/10.1016/j.rser.2018.04.025, MAG ID: 2748776435.

[3] Batlle C, Schittekatte T, Knittel CR. Power price crisis in the EU: Unveiling current policy responses and proposing a balanced regulatory remedy. SSRN Electron J 2022. http://dx.doi.org/10.2139/ssrn.4044848, URL https://www.ssrn.com/abstract=4044848.

[4] Cramton P, Ockenfels A, Stoft S. Capacity market fundamentals. Econ Energy & Environ Policy 2013;2(2). http://dx.doi.org/10.5547/2160-5890.2.2.2, URL http://www.iaee.org/en/publications/eeeparticle.aspx?id=46.

[5] Newbery D. Missing money and missing markets: Reliability, capacity auctions and interconnectors. Energy Policy 2016;94:401–10. http://dx.doi.org/10.1016/j.enpol.2015.10.028, URL https://linkinghub.elsevier.com/retrieve/pii/S0301421515301555.

[6] Oren SS. Generation adequacy via call options obligations: Safe passage to the promised land. Electr J 2005;18(9):28–42. http://dx.doi.org/10.1016/j.tej.2005.10.003, URL https://linkinghub.elsevier.com/retrieve/pii/S1040619005001193.

[7] of Energy UD, Government U. Benefits of demand response in electricity markets and recommendations for achieving them. a report to the united sates congress pursuant to section 1252 of the energy policy act of 2005. 2006, URL https://www.energy.gov/sites/prod/files/oeprod/DocumentsandMedia/DOE_Benefits_of_Demand_Response_in_Electricity_Markets_and_Recommendations_for_Achieving_Them_Report_to_Congress.pdf.

[8] Zachmann G, Hirth L, Heussaff C, Schlecht I, Mühlenpfordt J, Eicke As. The design of the European electricity market, current proposals and ways ahead. 2023, URL https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740094/IPOL_STU(2023)740094_EN.pdf.

[9] Newbery D. Efficient renewable electricity support: Designing an incentive-compatible support scheme. Energy J 2023;44(3). http://dx.doi.org/10.5547/01956574.44.3.dnew, URL https://www.iaee.org/en/publications/ejarticle.aspx?id=4000.

[10] Schlecht I, Maurer C, Hirth L. Financial contracts for differences: The problems with conventional cfds in electricity markets and how forward contracts can help solve them. Energy Policy 2024;186:113981. http://dx.doi.org/10.1016/j.enpol.2024.113981, URL https://linkinghub.elsevier.com/retrieve/pii/S0301421524000016.

[11] Parliament E. Improving the design of the EU electricity market - briefing (EU legislation in progress). 2024.

[12] Wolak FA. Long-term resource adequacy in wholesale electricity markets with significant intermittent renewables. Environ Energy Policy Economy 2022;3:155–220. http://dx.doi.org/10.1086/717221, URL https://www.journals.uchicago.edu/doi/10.1086/717221.

[13] Roques F, Finon D. Adapting electricity markets to decarbonisation and security of supply objectives: Toward a hybrid regime? Energy Policy 2017;105:584–96. http://dx.doi.org/10.1016/j.enpol.2017.02.035, URL https://linkinghub.elsevier.com/retrieve/pii/S0301421517301106.

[14] Joskow PL. From hierarchies to markets and partially back again in electricity: responding to decarbonization and security of supply goals. J Institutional Econ 2022;18(2):313–29. http://dx.doi.org/10.1017/S1744137421000400, URL https://www.cambridge.org/core/product/identifier/S1744137421000400/type/journal_article.

[15] Corneli S. A prism-based configuration market for rapid, low cost, and reliable electric sector decarbonization. 2020, URL https://media.rff.org/documents/corneli-prism-markets-for-rapid_decarbonization-final_word_version.pdf.

[16] Keppler JH, Quemin S, Saguan M. Why the sustainable provision of low-carbon electricity needs hybrid markets. Energy Policy 2022;171:113273. http://dx.doi.org/10.1016/j.enpol.2022.113273, URL https://linkinghub.elsevier.com/retrieve/pii/S030142152200492X.

[17] Bublitz A, Keles D, Zimmermann F, Fraunholz C, Fichtner W. A survey on electricity market design: Insights from theory and real-world implementations of capacity remuneration mechanisms. Energy Econ 2019;80:1059–78. http://dx.doi.org/10.1016/j.eneco.2019.01.030, MAG ID: 2787190816.

[18] Chappin EJ, Laurens Jde Vries, de Vries L, Richstein JC, Bhagwat P, Iychettira KK, Khan S. Simulating climate and energy policy with agent-based modelling: The energy modelling laboratory (emlab). Environ Model Softw 2017;96:421–31. http://dx.doi.org/10.1016/j.envsoft.2017.07.009, MAG ID: 2745419972 S2ID: 40af0ea1698a159b65a782023b972c9cf83239ed.

[19] Bhagwat PC, Iychettira KK, Richstein JC, Chappin EJ, De Vries LJ. The effectiveness of capacity markets in the presence of a high portfolio share of renewable energy sources. Util Policy 2017;48:76–91. http://dx.doi.org/10.1016/j.jup.2017.09.003, URL https://linkinghub.elsevier.com/retrieve/pii/S0957178716300406.

[20] Bhagwat P, Marcheselli A, Richstein JC, Chappin EJ, Laurens Jde Vries, de Vries L. An analysis of a forward capacity market with long-term contracts. Energy Policy 2017;111:255–67. http://dx.doi.org/10.1016/j.enpol.2017.09.037, MAG ID: 2761923291.

[21] Khan ASM, Verzijlbergh RA, Sakinci OC, De Vries LJ. How do demand response and electrical energy storage affect (the need for) a capacity market? Appl Energy 2018;214:39–62. http://dx.doi.org/10.1016/j.apenergy.2018.01.057, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261918300680.

[22] Marc Melliger, Marc Melliger, Chappin EJ. Phasing out support schemes for renewables in neighbouring countries: An agent-based model with investment preferences. Appl Energy 2022;305:117959. http://dx.doi.org/10.1016/j.apenergy.2021.117959, MAG ID: 3202699514.

[23] Barazza E, Strachan N. The impact of heterogeneous market players with bounded-rationality on the electricity sector low-carbon transition. Energy Policy 2020;138:111274. http://dx.doi.org/10.1016/j.enpol.2020.111274, MAG ID: 3001239696.

[24] Barazza E, Strachan N. The co-evolution of climate policy and investments in electricity markets: Simulating agent dynamics in UK, german and Italian electricity sectors. Energy Res Soc Sci 2020;65:101458. http://dx.doi.org/10.1016/j.erss.2020.101458, URL https://linkinghub.elsevier.com/retrieve/pii/S2214629620300359.

[25] Barazza E, Strachan N. The key role of historic path-dependency and competitor imitation on the electricity sector low-carbon transition. Energy Strat Rev 2021;33:100588. http://dx.doi.org/10.1016/j.esr.2020.100588, MAG ID: 3112687038.

[26] Anwar MB, Muhammad Bashar Anwar, Stephen G, Dalvi S, Frew B, Ericson S, Brown M, O'Malley M. Modeling investment decisions from heterogeneous firms under imperfect information and risk in wholesale electricity markets. Appl Energy 2022;306:117908. http://dx.doi.org/10.1016/j.apenergy.2021.117908, MAG ID: 3206562533.

[27] Frew B, Bashar Anwar M, Dalvi S, Brooks A. The interaction of wholesale electricity market structures under futures with decarbonization policy goals: A complexity conundrum. Appl Energy 2023;339:120952. http://dx.doi.org/10.1016/j.apenergy.2023.120952, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261923003161.

[28] Anwar MB, Guo N, Sun Y, Frew B. Can wholesale electricity markets achieve resource adequacy and high clean energy generation targets in the presence of self-interested actors? Appl Energy 2024;359:122774. http://dx.doi.org/10.1016/j.apenergy.2024.122774, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261924001570.

[29] Yang J, Fuss S, Johansson DJ, Azar C. Investment dynamics in the energy sector under carbon price uncertainty and risk aversion. Energy Clim Chang 2023;4:100110. http://dx.doi.org/10.1016/j.egycc.2023.100110, URL https://linkinghub.elsevier.com/retrieve/pii/S266627872300017X.

[30] Deissenroth M, Klein M, Nienhaus K, Reeg M. Assessing the plurality of actors and policy interactions: Agent-based modelling of renewable energy market integration. Complexity 2017;2017:1–24. http://dx.doi.org/10.1155/2017/7494313, URL https://www.hindawi.com/journals/complexity/2017/7494313/.

[31] Schimeczek C, Nienhaus K, Frey U, Sperber E, Sarfarazi S, Nitsch F, Kochems J, Ghazi AAE. AMIRIS: Agent-based market model for the investigationof renewable and integrated energy systems. J Open Source Softw 2023;8(84):5041. http:

//dx.doi.org/10.21105/joss.05041, URL https://joss.theoj.org/papers/10.21105/joss.05041.
[32] Gabriel SA, Conejo AJ, Fuller JD, Hobbs BF, Ruiz C. Complementarity modeling in energy markets, Softcover reprint of the hardcover 1st edition 2013. International series in operations research & management science, New York Heidelberg Dordrecht London: Springer; 2014.
[33] Höschle H, De Jonghe C, Le Cadre H, Belmans R. Electricity markets for energy, flexibility and availability — Impact of capacity mechanisms on the remuneration of generation technologies. Energy Econ 2017;66:372–83. http://dx.doi.org/10.1016/j.eneco.2017.06.024, URL https://linkinghub.elsevier.com/retrieve/pii/S0140988317302189.
[34] Höschle H, Hlne Le Cadre, Le Cadre H, Smeers Y, Smeers Y, Papavasiliou A, Ronnie Belmans, Ronnie Belmans, Belmans R, Belmans R. An ADMM-based method for computing risk-averse equilibrium in capacity markets. IEEE Trans Power Syst 2018;33(5):4819–30. http://dx.doi.org/10.1109/tpwrs.2018.2807738, MAG ID: 2789989672.
[35] Mays J, Morton DP, O'Neill RP. Asymmetric risk and fuel neutrality in electricity capacity markets. Nat Energy 2019;4(11):948–56. http://dx.doi.org/10.1038/s41560-019-0476-1, URL https://www.nature.com/articles/s41560-019-0476-1.
[36] Kaminski S, Höschle H, Delarue E. Impact of capacity mechanisms and demand elasticity on generation adequacy with risk-averse generation companies. Electr Power Syst Res 2021;199:107369. http://dx.doi.org/10.1016/j.epsr.2021.107369, URL https://linkinghub.elsevier.com/retrieve/pii/S0378779621003503.
[37] Billimoria F, Fele F, Savelli I, Morstyn T, McCulloch M. An insurance mechanism for electricity reliability differentiation under deep decarbonization. Appl Energy 2022;321:119356. http://dx.doi.org/10.1016/j.apenergy.2022.119356, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261922007000.
[38] Mays J, Jenkins JD. Financial risk and resource adequacy in markets with high renewable penetration. IEEE Trans Energy Mark Policy Regul 2023;1(4):523–35. http://dx.doi.org/10.1109/TEMPR.2023.3322531, URL https://ieeexplore.ieee.org/document/10273857/.
[39] Mays J, Craig MT, Kiesling L, Macey JC, Shaffer B, Shu H. Private risk and social resilience in liberalized electricity markets. Joule 2022;6(2):369–80. http://dx.doi.org/10.1016/j.joule.2022.01.004, URL https://linkinghub.elsevier.com/retrieve/pii/S2542435122000368.
[40] Dimanchev E, Gabriel SA, Reichenberg L, Korp° as M. Consequences of the missing risk market problem for power system emissions. Energy Econ 2024;136:107639. http://dx.doi.org/10.1016/j.eneco.2024.107639, URL https://linkinghub.elsevier.com/retrieve/pii/S0140988324003475.
[41] Dimanchev, Gabriel SA, Fleten S-E, Pecci F, Korpås M. Choosing climate policies in a second-best world with incomplete markets: insights from a bilevel power system model. In: MIT CEEPR. vol. 2024–14, 2024, URL https://ceepr.mit.edu/wp-content/uploads/2024/09/MIT-CEEPR-WP-2024-14.pdf.
[42] Sutton RS, Barto A. Reinforcement learning: an introduction, In: Adaptive computation and machine learning, second ed.. Cambridge, Massachusetts London, England: The MIT Press; 2020.
[43] Winder P. Reinforcement learning: industrial applications of intelligent agents. First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly; 2021.
[44] Zuccotto M, Castellini A, Torre DL, Mola L, Farinelli A. Reinforcement learning applications in environmental sustainability: a review. Artif Intell Rev 2024;57(4):88. http://dx.doi.org/10.1007/s10462-024-10706-5, URL https://link.springer.com/10.1007/s10462-024-10706-5.
[45] Albrecht SV, Christianos F, Schäfer L. Multi-Agent reinforcement learning: foundations and modern approaches. MIT Press; 2024, URL https://www.marl-book.com/.
[46] Sven Gronauer, Gronauer S, Klaus Diepold, Diepold K. Multi-agent deep reinforcement learning: a survey. Artif Intell Rev 2021;1–49. http://dx.doi.org/10.1007/s10462-021-09996-w, MAG ID: 3156295478.
[47] OpenAI, Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, Farhi D, Fischer Q, Hashme S, Hesse C, Józefowicz R, Gray S, Olsson C, Pachocki J, Petrov M, Pinto HPdO, Raiman J, Salimans T, Schlatter J, Schneider J, Sidor S, Sutskever I, Tang J, Wolski F, Zhang S. Dota 2 with large scale deep reinforcement learning. 2019, URL http://arxiv.org/abs/1912.06680 arXiv:1912.06680 [cs, stat].
[48] Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, Oh J, Horgan D, Kroiss M, Danihelka I, Huang A, Sifre L, Cai T, Agapiou JP, Jaderberg M, Vezhnevets AS, Leblond R, Pohlen T, Dalibard V, Budden D, Sulsky Y, Molloy J, Paine TL, Gulcehre C, Wang Z, Pfaff T, Wu Y, Ring R, Yogatama D, Wünsch D, McKinney K, Smith O, Schaul T, Lillicrap T, Kavukcuoglu K, Hassabis D, Apps C, Silver D. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 2019;575(7782):350–4. http://dx.doi.org/10.1038/s41586-019-1724-z, URL https://www.nature.com/articles/s41586-019-1724-z.
[49] Yu C, Velu A, Vinitsky E, Gao J, Wang Y, Bayen A, Wu Y. The surprising effectiveness of PPO in cooperative, multi-agent games. 2022, URL http://arxiv.org/abs/2103.01955 arXiv:2103.01955 [cs].
[50] Yang T, Zhao L, Li W, Zomaya AY. Reinforcement learning in sustainable energy and electric systems: a survey. Annu Rev Control 2020;49:145–63. http://dx.doi.org/10.1016/j.arcontrol.2020.03.001, URL https://linkinghub.elsevier.com/retrieve/pii/S1367578820300079.
[51] Kell AJM, McGough S, Forshaw M. Machine learning applications for electricity market agent-based models: A systematic literature review. 2022, URL http://arxiv.org/abs/2206.02196 arXiv:2206.02196 [cs].
[52] Zhu Z, Hu Z, Chan KW, Bu S, Zhou B, Xia S. Reinforcement learning in deregulated energy market: A comprehensive review. Appl Energy 2023;329:120212. http://dx.doi.org/10.1016/j.apenergy.2022.120212, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261922014696.
[53] Ye Y, Qiu D, Li J, Strbac G. Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: A novel multi-agent deep reinforcement learning approach. IEEE Access 2019;7:130515–29. http://dx.doi.org/10.1109/ACCESS.2019.2940005, URL https://ieeexplore.ieee.org/document/8826539/.
[54] Ye Y, Qiu D, Tindemans SH, Sun M, Papadaskalopoulos D, Strbac G. Deep reinforcement learning for strategic bidding in electricity markets. IEEE Trans Smart Grid 2020;11(2):1343–55. http://dx.doi.org/10.1109/tsg.2019.2936142, MAG ID: 2969843367.
[55] Liang Y, Guo C, Ding Z, Hua H. Agent-based modeling in electricity market using deep deterministic policy gradient algorithm. IEEE Trans Power Syst 2020;35(6):4180–92. http://dx.doi.org/10.1109/TPWRS.2020.2999536, URL https://ieeexplore.ieee.org/document/9106862/.
[56] Du Y, Li F, Zandi H, Xue Y. Approximating Nash equilibrium in day-ahead electricity market bidding with multi-agent deep reinforcement learning. J Mod Power Syst Clean Energy 2021;9(3):534–44. http://dx.doi.org/10.35833/MPCE.2020.000502, URL https://ieeexplore.ieee.org/document/9406572.
[57] Graf C, Zobernig V, Schmidt J, Klöckl C. Computational performance of deep reinforcement learning to find Nash equilibria. Comput Econ 2023. http://dx.doi.org/10.1007/s10614-022-10351-6, URL https://link.springer.com/10.1007/s10614-022-10351-6.
[58] Qiu D, Wang J, Dong Z, Wang Y, Strbac G. Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading. IEEE Trans Power Syst 2023;38(5):4853–66. http://dx.doi.org/10.1109/TPWRS.2022.3217922, URL https://ieeexplore.ieee.org/document/9931995/.
[59] Xu H, Wu Q, Jinyu Wen, Zhihong Yang. Joint bidding and pricing for electricity retailers based on multi-task deep reinforcement learning. Int J Electr Power Energy Syst 2022;138. http://dx.doi.org/10.1016/j.ijepes.2021.107897, 107897–107897, MAG ID: 4200379907.
[60] Harder N, Qussous R, Weidlich A. Fit for purpose: Modeling wholesale electricity markets realistically with multi-agent deep reinforcement learning. Energy AI 2023;14:100295. http://dx.doi.org/10.1016/j.egyai.2023.100295, URL https://linkinghub.elsevier.com/retrieve/pii/S2666546823000678.
[61] Harder N, Mitridati L, Pourahmadi F, Weidlich A, Kazempour J. How satisfactory can deep reinforcement learning methods simulate electricity market dynamics? Benchmarking via bi-level optimization. ACM SIGEnergy Energy Informatics Rev 2024;4(4):65–77. http://dx.doi.org/10.1145/3717413.3717419, URL https://dl.acm.org/doi/10.1145/3717413.3717419.
[62] Miskiw KK, Staudt P. Explainable deep reinforcement learning for multi-agent electricity market simulations. In: 2024 20th international conference on the European energy market. EEM, Istanbul, Turkiye: IEEE; 2024, p. 1–9. http://dx.doi.org/10.1109/EEM60825.2024.10608907, URL https://ieeexplore.ieee.org/document/10608907/.
[63] Yeh C, Li V, Datta R, Arroyo J, Christianson N, Zhang C, Chen Y, Hosseini M, Golmohammadi A, Shi Y, Yue Y, Wierman A. SustainGym: reinforcement learning environments for sustainable energy systems.
[64] Renshaw-Whitman C, Zobernig V, Cremer JL, De Vries L. Non-stationarity in multiagent reinforcement learning in electricity market simulation. Electr Power Syst Res 2024;235:110712. http://dx.doi.org/10.1016/j.epsr.2024.110712, URL https://linkinghub.elsevier.com/retrieve/pii/S0378779624005984.
[65] Antonini EGA, Di Bella A, Savelli I, Drouet L, Tavoni M. Weather- and climate-driven power supply and demand time series for power and energy system analyses. Sci Data 2024;11(1):1324. http://dx.doi.org/10.1038/s41597-024-04129-8, URL https://www.nature.com/articles/s41597-024-04129-8.
[66] Antonini EGA, Di Bella A, Drouet L, Savelli I, Tavoni M. Weather- and climate-driven power supply and demand time series for European countries. 2024, URL https://zenodo.org/doi/10.5281/zenodo.13938926.
[67] Di Bella A, Colelli FP. Mitigation strategies can alleviate power system vulnerability to climate change and extreme weather: a case study on the Italian grid. Environ Res: Infrastruct Sustain 2025;5(1):015003. http://dx.doi.org/10.1088/2634-4505/ada308, URL https://iopscience.iop.org/article/10.1088/2634-4505/ada308.
[68] Hoffmann M, Priesmann J, Nolting L, Praktiknjo A, Kotzur L, Stolten D. Typical periods or typical time steps? A multi-model analysis to determine the optimal temporal aggregation for energy system models. Appl Energy 2021;304:117825. http://dx.doi.org/10.1016/j.apenergy.2021.117825, URL https://linkinghub.elsevier.com/retrieve/pii/S0306261921011545.
[69] Cramton P, Stoft S. Colombia firm energy market. In: 2007 40th annual hawaii international conference on system sciences. HICSS'07, Waikoloa, HI, USA: IEEE; 2007, http://dx.doi.org/10.1109/HICSS.2007.133, 124–124URL http://ieeexplore.ieee.org/document/4076621/.

[70] Mastropietro P, Rodilla P, Rivier M, Batlle C. Reliability options: Regulatory recommendations for the next generation of capacity remuneration mechanisms. Energy Policy 2024;185:113959. http://dx.doi.org/10.1016/j.enpol.2023.113959, URL https://linkinghub.elsevier.com/retrieve/pii/S030142152300544X.

[71] Towers M, Terry JK, Kwiatkowski A, Balis JU, De Cola G, Deleu T, Goulão M, Kallinteris A, KG A, Krimmel M, Perez-Vicente R, Pierré A, Schulhoff S, Tai JJ, Shen ATJ, Younis OG. Gymnasium. 2023, http://dx.doi.org/10.5281/ZENODO.8127026, URL https://zenodo.org/record/8127026 Language: en.

[72] Liang E, Liaw R, Moritz P, Nishihara R, Fox R, Goldberg K, Gonzalez JE, Jordan MI, Stoica I. RLlib: Abstractions for distributed reinforcement learning. 2018, URL http://arxiv.org/abs/1712.09381 arXiv:1712.09381 [cs].

[73] Huang S, Ontañón S. A closer look at invalid action masking in policy gradient algorithms. In: The international FLAIRS conference proceedings. vol. 35, 2022, http://dx.doi.org/10.32473/flairs.v35i.130584, URL http://arxiv.org/abs/2006.14171 arXiv:2006.14171 [cs, stat].

[74] Delalleau O, Peter M, Alonso E, Logut A. Discrete and continuous action representation for practical RL in video games. 2019, URL http://arxiv.org/abs/1912.11077 arXiv:1912.11077 [cs, stat].

[75] Bick D. Towards delivering a coherent self-contained explanation of proximal policy optimization (Ph.D. thesis), University of Groningen; 2021.

[76] Bisi L, Santambrogio D, Sandrelli F, Tirinzoni A, Ziebart BD, Restelli M. Risk-averse policy optimization via risk-neutral policy optimization. Artificial Intelligence 2022;311:103765. http://dx.doi.org/10.1016/j.artint.2022.103765, URL https://linkinghub.elsevier.com/retrieve/pii/S0004370222001059.

[77] Bonetti M, Bisi L, Restelli M. Risk-averse optimization of reward-based coherent risk measures. Artificial Intelligence 2023;316:103845. http://dx.doi.org/10.1016/j.artint.2022.103845, URL https://linkinghub.elsevier.com/retrieve/pii/S0004370222001850.

[78] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, URL http://arxiv.org/abs/1707.06347arXiv:1707.06347 [cs].

[79] Lowe R, WU Y, Tamar A, Harb J, Pieter Abbeel O, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Advances in neural information processing systems. vol. 30, Curran Associates, Inc.; 2017, URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html.

[80] Espeholt L, Soyer H, Munos R, Simonyan K, Mnih V, Ward T, Doron Y, Firoiu V, Harley T, Dunning I, Legg S, Kavukcuoglu K. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. 2018, URL http://arxiv.org/abs/1802.01561 arXiv:1802.01561 [cs].

[81] Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J. Mean field multi-agent reinforcement learning. In: Proceedings of the 35th international conference on machine learning. PMLR; 2018, p. 5571–80, URL https://proceedings.mlr.press/v80/yang18d.html.

[82] de Witt CS, Gupta T, Makoviichuk D, Makoviychuk V, Torr PHS, Sun M, Whiteson S. Is independent learning all you need in the StarCraft multi-agent challenge?. 2020, URL http://arxiv.org/abs/2011.09533 arXiv:2011.09533 [cs].

[83] of Justic UD, Commission UFT. Merger guidelines. 2023, URL https://www.justice.gov/atr/2023-merger-guidelines.

[84] of Energy Security & Net Zero D, Government U. Exploring reliability standard metrics in a net zero transition. Tech. rep., 2023, URL https://shorturl.at/6H7kD.

[85] Cochrane JH. Asset pricing. Rev. ed. Princeton, N.J: Princeton University Press; 2005.

[86] Calcaterra M, Aleluia Reis L, Fragkos P, Briera T, De Boer HS, Egli F, Emmerling J, Iyer G, Mittal S, Polzin FHJ, Sanders MWJL, Schmidt TS, Serebriakova A, Steffen B, Van De Ven DJ, Van Vuuren DP, Waidelich P, Tavoni M. Reducing the cost of capital to finance the energy transition in developing countries. Nat Energy 2024;9(10):1241–51. http://dx.doi.org/10.1038/s41560-024-01606-7, URL https://www.nature.com/articles/s41560-024-01606-7.

[87] Schmidt TS, Steffen B, Egli F, Pahle M, Tietjen O, Edenhofer O. Adverse effects of rising interest rates on sustainable energy transitions. Nat Sustain 2019;2(9):879–85. http://dx.doi.org/10.1038/s41893-019-0375-2, URL https://www.nature.com/articles/s41893-019-0375-2.