# Multi-Agent Reinforcement Learning for Modeling, Simulating, and Optimizing Energy Markets

**Matan Levy**[1], **Itay Segev**[2], **Alexander Tuisov**[2], **Sarah Keren**[2]

[1] Faculty of Electrical Engineering and Computing, Technion - Israel Institute of Technology
[2] Faculty of Computer Science, Technion - Israel Institute of Technology

## Abstract

The objective of this study is to advance the optimization of hybrid electricity markets using *multi-agent reinforcement learning* (MARL). The transition from centralized systems to public–private models introduces significant challenges, including the emergence of independent market players and the increasing integration of renewable energy sources (RESs). These challenges are further intensified by rapidly shifting demand patterns, driven both by energy-intensive data centers and AI inference workloads, as well as by political and societal instabilities.

To address these complexities, we develop a formal model of market participants' behavior and propose a MARL-based framework for optimizing system operator strategies. This framework incorporates dynamic pricing and dispatch scheduling to minimize operational costs, maintain grid stability, and align market incentives. We also present a new, adaptable simulation environment compatible with state-of-the-art MARL methods. Empirical evaluations in increasingly complex scenarios demonstrate the effectiveness of our approach in capturing the dynamic and decentralized nature of modern electricity markets.

## Introduction

Our aim is to harness ***multi-agent reinforcement learning*** (MARL) to optimize modern electricity markets as they transition from historically centralized, government-controlled systems to complex public–private hybrids driven by the growing penetration of *renewable energy sources* (RESs) and by the widely available AI-enabled decision-making methods. These new technologies have blurred the distinction between the traditional production, consumption, and storage roles of different system components, enabling a single entity to assume multiple roles. As a result, large-scale electricity generators, system operators, and utility companies are no longer the sole decision-makers. Instead, the new technologies have induced a decentralized architecture in which AI-enabled grid-edge agents can decide not only the amount of energy to consume at a given time based on market signals but also on when and how much energy to trade. The resulting challenges are amplified by rapidly shifting demand patterns, intensified by energy-hungry data centers and AI inference workloads, and by unstable political and societal conditions.



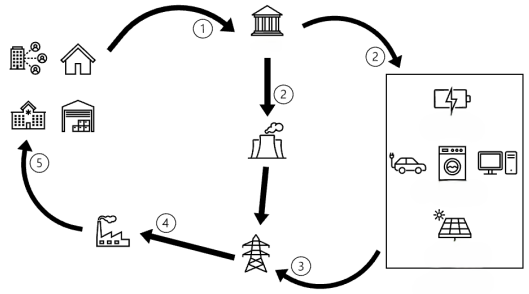

Figure 1: The day-ahead control cycle. At the start of the cycle, the SO sets a dispatch plan for its controlled generators. Then, every 30-minutes: (1) the SO receives realized demand for the current time step (2) the SO monitors the dispatch directives and, if prices are not fixed in advance, posts real-time buy/sell tariffs (3) grid-edge agents buy/sell power (4) if needed, peaker reserves are dispatched or curtailment is performed (5) balanced power flows to consumers.

These factors underscore the urgency of developing robust, adaptive market-optimization strategies while simultaneously posing substantial computational challenges amid a highly uncertain operational and regulatory environment[1].

**Example 1** *To demonstrate the challenges in managing current energy markets, consider the day-ahead market depicted in Figure 1 in which the* ***system operator*** **(SO)** *aims to optimize electricity generation based on forecasted demand, generation costs, and grid constraints. The resulting decisions, made 24 hours in advance, specify the amount of electricity to be produced (dispatch), the prices, and the allocation of reserve capacity, i.e., the ability to generate additional power at short notice, often at high environmental costs, in the event of generation failures or unexpected demand surges. In real time, the SO must continuously balance supply and demand by activating reserves or curtailing generation as needed. In addition, depending on regulation and pricing regimes, it may adjust prices dynamically.*

[1]This work is a result of insights gained from an ongoing collaboration with NOGA Ltd.- Israel's Independent System Operator.

Adapting the day-ahead market to today's energy systems requires accounting for the variability and limited controllability of increasingly heterogeneous ***grid-edge agents***, or **GEAgents**, particularly those with local generation and storage capabilities. For example, a household with a photovoltaic (PV) unit and a battery can autonomously optimize its energy storage policy, learning when to store energy, when to consume it, and when to trade with the grid to maximize economic benefits. While these distributed resources can enhance efficiency and resilience by shaving peaks, supplying energy, and reducing the amount of centrally dispatched generation required, such behaviors are focused on optimizing individual utility and introduce significant uncertainty into aggregate demand forecasts. As a result, they can destabilize the system, especially under sudden shifts in consumption or generation patterns.

To address these challenges, the SO must adjust the electricity production plan, or *dispatch*, and feed-in and sell prices to influence market participants and align their behavior with grid operational objectives. Additionally, it manages reserves and peaking power plants, which can be activated to address unmet demand and ensure system stability. The challenge is thus one of cost optimization while satisfying the demand in the presence of strategic market players.

Reinforcement Learning (RL) and, in particular, Multi-Agent Reinforcement Learning (MARL), are well-suited for modeling modern energy systems and networks, which are inherently multi-agent in nature. These systems consist of diverse, distributed, and strategically autonomous entities, such as grid-edge devices, utility companies, system operators, and market participants, that pursue different objectives, interact over shared physical and economic infrastructures, and respond to evolving conditions, market prices, and regulatory constraints. MARL offers a natural framework for capturing this decentralized and interactive structure and enables agents to learn adaptive long-term policies, coordinate under uncertainty, and reason about both cooperative and competitive dynamics by interacting with one another and with the environment (Zhu et al. 2023). In addition, unlike centralized optimization approaches, MARL supports real-time adaptation at the agent level, making it especially suitable for the dynamic and distributed nature of modern energy systems.

Thus, despite its complexity and the fact that it is neither the simplest approach to implement nor to analyze, and that the current MARL literature still lacks efficient, general-purpose implementations, its close alignment with the core characteristics of these dynamic markets makes it the most appropriate modeling choice. Moreover, MARL's ability to simulate emergent behavior and to explore decentralized strategies makes it a powerful tool for designing and analyzing resilient, efficient, and adaptive energy systems.

We make three key contributions. First, we offer a general MARL model that captures the incentives and rational decision-making of independent participants in energy markets. Leveraging these models, we then focus on studying the optimization problem of the system operator (SO) under various assumptions, revealing how each setting shapes optimal dispatch and pricing policies. Finally, we offer a novel configurable, open-source grid simulator, we call `Energy-Net`, that supports diverse topologies and uncertainty patterns. Using the day-ahead market described in Example 1, we demonstrate via experiments across increasingly complex settings how RL-driven agents can jointly optimise participant and SO strategies, highlighting the promise of MARL for modern energy-market design.

## Background and Related Work

*Reinforcement Learning* (RL) is a learning paradigm where an agent learns to optimize its behavior by interacting with an environment and receiving rewards or penalties for its actions (Sutton and Barto 2018). *Multi-agent Reinforcement Learning* (MARL) extends RL to scenarios involving multiple agents that concurrently learn and make decisions within a shared or partially shared environment (Albrecht, Christianos, and Schäfer 2024). Each agent aims to maximize its own utility (typically measured as accumulated reward), but its actions can influence both its own outcomes and the outcomes of other agents, leading to complex emergent behaviors and the need for coordination and cooperation.

The most common MARL model is the Multi-agent Markov Decision Process (MAMDP) (also known as *Stochastic Game* or *Markov Game*) (Shapley 1953; Littman 1994) defined as a tuple $\langle \mathcal{S}, \mathcal{A} = \{\mathcal{A}_i\}_{i=1}^n, \mathcal{T}, \mathcal{R} = \{\mathcal{R}_i\}_{i=1}^n, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the *joint action space* with $\mathcal{A}_i$ as the $i^{th}$ agent action space s.t. $a \triangleq (a_1, a_2, \ldots, a_n)$ for $a \in \mathcal{A}$, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function $\mathcal{T}(s', a, s)$ such that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A} : \sum_{s' \in \mathcal{S}} \mathcal{T}(s, a, s') = 1$, $\mathcal{R}$ is the *joint reward function* with $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as the $i^{th}$ agent reward function, and $\gamma \in [0, 1)$ is the discount factor.

A solution is a *joint policy* $\pi \triangleq (\pi_1, \ldots, \pi_n)$ associating each agent with policy $\pi_i : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$ that specifies the probability of agent $i$ taking an action at a given state. The value (utility) function $V_i^\pi(s)$ denotes the expected cumulative discounted reward agent $i$ receives when starting in state $s$ and following $\pi$ thereafter. The action-value function or Q-value $Q_i^\pi(s, a)$ quantifies the expected value when performing $a$ in $s$, and following $\pi$ thereafter. This general definition captures a variety of interactions and relationships that can exist between agents in collaborative, competitive, and mixed-incentive MARL settings.

The suitability of MARL for modeling energy systems and networks has been the focus of are several previous frameworks. In fact, MARL is relevant to the three challenges Keren et al. (2024) outline for optimizing energy network: optimizing grid-edge agents (Shen et al. 2022), maintaining stability of power systems (Gao, Wang, and Yu 2021; Li et al. 2023), and for maintaining energy markets (Zhu et al. 2023; Li et al. 2023; Charbonnier, Morstyn, and McCulloch 2022; Jang et al. 2023).

However, applications of RL and MARL in energy markets often assume a single, all-knowing controller optimizing the entire system. In such formulations, a central agent directly controls all generation and storage decisions using global information and perfect foresight, an assumption that is unattainable in practice. These centralized op-

timization models can yield system-level insights but cannot capture the strategic, profit-driven behavior of individual market participants (Harder, Weidlich, and Staudt 2023; Perera and Kamalaruban 2021). Moreover, as modern grids grow more heterogeneous and stochastic with high renewable penetration, a monolithic control scheme becomes impractical (Wolgast and Nieße 2023). Recent studies emphasize that managing numerous distributed resources under uncertainty requires moving beyond one-size-fits-all control toward more decentralized decision-making structures (Michailidis, Michailidis, and Kosmatopoulos 2025; Ahlqvist, Holmberg, and Tangerås 2022).

On the other end of the spectrum, many RL-based models use a fully decentralized approach in which each market participant (e.g. a storage unit owner or consumer) acts independently. In these formulations, multiple RL agents learn their own policies (for bidding, charging, discharging, etc.) based on price signals or local observations, without a central coordinator explicitly optimizing the whole system (Werner and Kumar 2023; Zhang et al. 2024). This bottom-up approach reflects competitive markets by giving each market player its own profit-maximizing RL agent (Guan et al. 2015; Vázquez-Canteli and Nagy 2019; Qiu, Nguyen, and Crow 2015). However, purely decentralized models typically assume the market rules or prices are exogenous or fixed (Zhu et al. 2023; Perera and Kamalaruban 2021). In our model, the SO acts as an active participant and directly shapes the market dynamics. Related efforts on dynamic dispatch and end-to-end RL in energy systems include (Yang et al. 2021; Zhang et al. 2019), and comprehensive overviews of RL for power systems can be found in (Ginzburg-Ganz et al. 2024).

From an algorithmic standpoint, our challenge is to create a general MARL formulation that captures the complex dynamics brought by the existence of multiple agents in the system. Specifically, we aim to address the strategic and tightly coupled interplay between a central SO and price-responsive market participants. Most prior approaches either employ fully centralized optimization, which ignores competitive behavior, or simulate independent agents interacting with static SO actions. Consequently, hybrid markets, with a dynamic, learning-enabled SO and autonomous market agents, remain underexplored.

In our framework, the SO continually adjusts dispatch and pricing signals, while market agents react strategically to maximize their own profit. Capturing this two-way interaction is essential for realistic market modeling, yet most RL studies to date have only touched on limited aspects of this SO-agent feedback loop (Harder, Weidlich, and Staudt 2023; Navon et al. 2024). The scarcity of work in this hybrid paradigm highlights the potential of our approach for integrating a central coordinator's adaptive decisions with the learning-based responses of individual market players and for optimizing modern energy systems.

## Energy Market Dynamics

Energy markets are applied at different scales and levels of temporal granularity and are managed by a system operator (SO) that is tasked with ensuring the reliable, secure, and efficient operation of the electricity grid and its associated markets (in liberalized markets, this role is often performed by an Independent System Operator (ISO)).

In a typical *day-ahead market*, the SO predicts the following day's power demand (electricity consumption) and issues a *dispatch*, a production schedule, while considering operational constraints and generation costs (see Example 1). In addition to the generation of the predicted, or *nominal* demand, the SO also manages the *reserve*, which sets a backup production capability for each time step.

In real-time, the SO is tasked with continuously maintaining a balance between demand and supply. If there is a surplus, energy is discharged, or *curtailed*. If production determined by the dispatch is not enough to cover the *realized demand*, reserves, which are more flexible but also more expensive and polluting, are deployed. Producers are then compensated based on the System Marginal Price (SMP) mechanism, calculated as the marginal cost of producing the final unit of energy required to satisfy system demand, based on the least-cost dispatch solution. In this work, we abstract the dispatch details (i.e., distribution of generation load among producers) and consider only the total amount and cost of power produced at each timestamp (see the appendix for details on SMP computation).

Historically, energy markets comprised three principal components: power producers (e.g., power plants), power consumers (industrial and residential), and the system operator, responsible for market management and coordination. The producers typically used conventional coal-based generation and were either units under the full control of the operator, or independent units that participated in the market but were regulated and bound by production agreements made for different temporal horizons.

Recent power-market reforms have introduced independent grid-edge agents (GEAgents), from private utilities to smart homes, alongside traditional producers and consumers. In these markets, each GEAgent operates a **Production-Consumption-Storage unit** (**PCS-unit**), which may produce (e.g., via PV), consume (e.g., via electrical appliances), and store (e.g., via a battery) energy (and any combination of these three capacities).

Unlike traditional controlled producers, these GEAgents are not legally required to adhere to dispatch instructions and may buy from or sell energy to the grid at will to maximize their profits, overlooking the instability this may cause. Therefore, modeling the GEAgents' behaviors and strategies is essential for the SO's planning. Since we assume GEAgents are rational, the natural way for the SO to align player incentives with stability constrains and efficiency objectives is via pricing.

Thus, with the aim of minimizing total costs for the SO (thus the taxpayers) while satisfying the supply and demand balance, the dispatch, denoted $\Delta_t$, and selling and feed-in prices, denoted $\xi_t$ and $\phi_t$, respectively, for each time $t$, are the primary tools for market control. In what follows, we analyze the SO 's optimization problem under increasingly complex market regimes, highlighting how dispatch and pricing jointly control system stability and efficiency. We note that due to space constraints, we provide the key

details of each setting and refer the reader to the appendix for the complete formulations.

In the basic setting, the SO receives at the beginning of each episode the nominal production and reserve capabilities and costs for market participants, as well as the demand for all time steps in the horizon $T$. Based on this information and on the operational constraints, it determines the scheduled $\Delta_t$ and prices $\xi_t(\cdot)$, $\phi_t(\cdot)$ for all timestamps $t \in [T]$ to minimize total costs. Formally,

$$\min C^{\text{total}} \;=\; \min\left[C^{\text{dispatch}} \;+\; \sum_{t=1}^{T} C_t^{\text{online}}\right]$$

(Deterministic SO Objective)

where $C^{\text{dispatch}}$ is the total dispatch cost for the complete episode, and $C_t^{\text{online}}$ is the online cost (including reserve cost) for time $t$.

Since all information in this deterministic model is given in advance, the GEAgent can also compute its policy at the beginning of each episode and decide how much power to buy from ($P_t^b$), and sell to ($P_t^s$) the grid at every timestamp $t$ to maximize its total revenue under its operational constraints. Formally:

$$\max \sum_{t=1}^{T}\Big(\phi_t\, P_t^s - \xi_t\, P_t^b\Big)$$

(Deterministic GEAgent Objective)

In a stochastic extension of this setting, we account for the inability to exactly predict demand and production. In this case, it may be possible to estimate these distributions from historical data and observations using machine learning methods to improve decision-making under these forms of uncertainty (de Vilmarest, Nowotarski, and Ziel 2024). In this setting, the objectives of the SO and GEAgents are replaced by an expectation-based optimization.

**Accounting for Strategic Demand:** In modern energy systems, demand is not only stochastic but also strategic since GEAgents can intelligently manage the operation of devices and energy resources, in response to system-level signals. This *demand (load) flexibility* is reshaping energy markets by introducing new ways to contribute to their efficient and stable operation (Charbonnier, Morstyn, and McCulloch 2022; Zhu et al. 2023). However, this shift also introduces challenges such as increased system complexity, uncertainty in demand forecasting, and the need for regulatory mechanisms to ensure fair and reliable participation.

In this extended setting, the SO needs to determine $\Delta_t$, $\xi_t$ and $\phi_t$ for each $t$ according to the demand $D_t$ at time $t$ while accounting for the GEAgents ability to sell, buy, and store power. From the perspective of the GEAgent, the price signals are exogenous signals set by the SO, but they depend on the GEAgents' sales $P_t^s$ and purchases $P_t^b$ and other variables. This coupling results in a feedback mechanism where the player's actions influence the prices, and the prices, in turn affect the player's actions. This introduces a game-theoretic dimension where the GEAgents' decisions are influenced by the SO 's pricing strategy and vice versa.

Formally, the GEAgent's input includes all the parameters that were relevant for the deterministic and stochastic settings, including the expected local demand $l_t$ and production $g_t$ at time $t$. A key difference is that, depending on regulation, the selling price $\xi_t$ and feed-in prices $\phi_t$ can be set either in advance or dynamically, in response to the market state. The objective of the GEAgent is now:

$$\max_{P_t^b, P_t^s} \mathbb{E}_{l_t, g_t}\left[\sum_{t=1}^{T}\big(\phi_t(P_t^s, P_t^b, \ldots) - \xi_t(P_t^s, P_t^b, \ldots)\big)\right]$$

(Strategic Player Objective)

As in the stochastic settings, the SO receives at the beginning of each episode (day) all the information about the GEAgents and the controlled producers and needs to determine the scheduled amount of production $\Delta_t$ for each time step. However, it is crucial to distinguish between two components of the demand. The **nominal demand** refers to the exogenous, inelastic portion of load that remains unaffected by local control strategies, real-time market incentives, or variations in renewable generation. In contrast **flexible (or strategic) demand**, refers to the portion of demand that can be adjusted in time, quantity, or pattern in response to external signals, such as price changes, grid conditions, or availability of renewable energy.

Since the SO cannot loyally model the demand without considering the strategic nature of the GEAgents, optimization methods that are appropriate for deterministic and stochastic settings won't work here. Thus, we propose a model of market participants as RL agents.

## The Energy Market as MARL

In modeling modern power systems using MARL, it is essential to account for multiple interacting perspectives. These include the physical constraints of the grid (e.g., stability limits), agent-level decision processes under partial observability, and the heterogeneity of demand profiles encompassing both nominal and flexible demand. Effective models must also incorporate market and pricing signals that influence agent behavior, and the temporal-spatial scalability required for real-world deployment. While these considerations are crucial for realistically and robustly capturing decentralized control strategies in complex energy environments, they also pose significant challenges to preserving the underlying Markovian structure that traditional agent-based decision models rely on.

Through the lens of RL, the SO aims to learn an optimal policy that balances overall system efficiency with the mitigation of risk, such as insufficient power supply and grid instability. Simultaneously, GEAgents seek to maximize their individual utility in response to market signals, subject to their own operational constraints and preferences. We formally model this decentralized setting as a Multi-Agent Markov Decision Process (MAMDP) (defined above), involving two types of agents: the SO, and the GEAgents.

A key characteristic of the setting we aim to model is that while the states, actions, and rewards are relatively straightforward to define, it is challenging to model the joint transition function: the next system state and its stability depend on the actions performed by all agents. This is a challenging feature for MARL solution approaches.

Our model for the SO's includes the following.

- **State Space $\mathcal{S}$:** At every time step $t$, typically representing a half-hour interval, the system state is associated with a vector $s_t \in \mathcal{S}$ that specifies operational factors that may affect decision-making. For the SO, this includes the system-level demand forecast $\hat{D}_t$ for the specified horizon, the realized demand $D_t$ for the current time step, supply capacities, storage states, etc. It may also include factors that indicate the stability state of the system, for example, whether the supply-demand balance is violated.
- **Action Space $\mathcal{A}$:** The SO's actions include the dispatch directives $\Delta_t$ and setting the sell prices $\xi_t(\cdot)$ and buy prices $\phi_t(\cdot)$ for each time step. In real-time, the SO also activates reserves and curtails power if needed, but since we assume these actions are dictated by the state and dispatch action, they are not modeled explicitly.

  We support two types of pricing mechanisms. In a day-ahead pricing regime, the SO makes the prices public at $t = 0$ while in an online pricing setting, the SO can dynamically set prices in response to market signals. We discuss the characteristics of several mechanisms, including the benefits of quadratic pricing, in the next section.
- **Reward Function $\mathcal{R}$:** At each time step $t$, the reward is set by the dispatch cost $C_t^{\text{dispatch}}$ and online adjustment cost $C_t^{\text{online}}$. To expedite learning, we shape the reward by including a weighted mismatch penalty, where $\eta_o$ is the weight associated with energy curtailment (due to overproduction) and $\eta_u$ is the weight associated with power shortage.

$$\begin{aligned}\mathcal{R}_t(\eta_o, \eta_u) = &-\left(C_t^{\text{dispatch}} + C_t^{\text{online}}\right) \\ &- \eta_o \left[\max\{0,\ \Delta_t - D_t\}\right]^2 \\ &- \eta_u \left[\max\{0,\ D_t - \Delta_t\}\right]^2.\end{aligned}$$

The model of the GEAgents includes:

- **State Space $\mathcal{S}$:** Each GEAgent is associated with a PCS-unit for which the state includes its local information (e.g., state-of-charge) as well as the price signals advertised by the SO.
- **Action Space $\mathcal{A}$:** Modern GEAgents have significant decision-making autonomy, allowing them to choose how much energy to store, consume, or sell based on their local goals, capabilities, and constraints. We assume the GEAgent sees the current prices and local state at the start of each iteration before deciding how to act. If generation and consumption are non-controllable (e.g., corresponding to PV-based generation and consumption from operating critical appliances), these become exogenous values governed by stochastic processes and are part of the state. In this case, the only decisions are the charge and discharge actions, which may have stochastic effects.
- **Reward Function $\mathcal{R}$:** For each GEAgent $i$, the step-wise reward is the net revenue obtained by trading with the grid: $\mathcal{R}_t^i = \phi_t\left(P_t^s, P_t^b\right) - \xi_t\left(P_t^s, P_t^b\right)$. Maximizing the sum of $\mathcal{R}_t^i$ over the horizon is equivalent to the strategic GEAgent objective defined in the previous section.

**Joint Transition Function $\mathcal{T}$:** By modeling joint actions, a MAMDP allows each agent's choice (including how GEAgents respond to prices or storage opportunities) to influence the next state. As mentioned above, the difficulty of modeling the transition function is a key challenge in modeling and using MARL for the problems we aim to address.

In general, the transition function can be decoupled into two parts. The physical dynamics capture the dynamics of the electrical network. For example, when a charge or discharge action is performed, the battery dynamics must obey its physical constraints. In contrast, market dynamics capture the interactions between agents. These strategic decisions create a coupled system where each agent's payoff depends on the actions of others.

In principle, the Markov transition function must capture all aspects of the dynamics, but writing a closed-form for these different layers is hopeless. Instead, we create the `Energy-Net` simulator (presented later on) to maintain the physics and book-keeping, and enable *learning* directly from roll-outs. This can side-step the need for explicit modeling of the complex dynamics and allows extracting value functions and policies using methods such as deep neural networks, rather than from first principles.

**Episode:** As is typical in the day-ahead market (Figure 1), at the beginning of each episode (time step $t = 0$) the SO receives the predicted demand $\hat{D}_t$ for the next episode (e.g., $48$ half-hour intervals). It also receives the production and reserve capacities of its controlled units, the prices of each generated unit, and other information that might be relevant (i.e., weather forecast, special events, etc.). If day-ahead pricing is applied, the SO sets and advertises the $\xi_t(\cdot)$ and feed-in tariff $\phi_t(\cdot)$ for the whole episode. Otherwise, online pricing is applied. This iterative process continues until the end of the planning horizon.

**Evaluation:** While RL agents are typically evaluated based on the accumulated expected reward, in the model we propose this may be insufficient. Thus, while we may use pricing to bias learning toward a proxy objective, our actual objective is typically expressed in terms of system stability, efficiency, and sustainability, which may require careful design of the evaluation criteria. For example, if the objective is to minimize the amount of energy generated by reserves in real time, we can use a cost-based reward function to train the agents, but perform evaluation based on the total *amount* of reserve power that is activated.

## Solution Approaches

The MARL formulation we propose captures the strategic interactions that typify modern hybrid power systems. In this section, we discuss solution approaches that can be adopted by the market participants. Importantly, even if our main challenge is in computing optimal market management approaches for the SO, we must equip the GEAgents with the strongest policies to guarantee the SO can predict their response to different price signals.

In principle, the deterministic and stochastic formulations of the market dynamics can be solved using state-space and

dynamic programming methods, respectively (see the appendix for details). Even if distributions are not fully known, it may be possible to learn them from data. Nevertheless, such methods are not appropriate to our problem, which is inherently challenging due to the agents' ability to strategically adapt their behavior and due to the dual-action learning structure, which operates across different time frames.

A specific challenge is that pricing may be dynamic and set at every time step, while the dispatch (planned generation) action $\Delta_t$ for each time step $t$ is typically decided at the beginning of each episode. This temporal disparity adds a layer of complexity, as the reward for an action may be reflected only at the end of the episode. Moreover, determining the dispatch sequence involves generating a time series output that must account for dynamic market conditions and the behavior of market participants. A further complication arises from the interdependence of the agents' decisions; dispatch decisions are influenced by the market agents' responses to price signals, while optimal pricing strategies depend on real-time outcomes.

Since our model represents a sequential and highly non-linear multi-agent interaction (SO first, GEAgent second, repeatedly), we iteratively train each of the policies for continuous control in an online RL regime. If the agents' policies converge, it is toward a *practical* equilibrium in function-approximation space rather than a formal Nash point. There are several abstractions that can be used to facilitate computation. One example is to make the problem easier by abstracting the dispatch optimization, which we denote as **dispatch abstraction**. In this simplified model the ISO only has control over the prices, and we assume that the ISO production $\Delta_t$ is fixed to be equal to the predicted $\hat{D}_t$.

**Quadratic Pricing:** We employ two pricing regimes, online and day-ahead tariffs. One way to allow the SO to influence consumption and injection patterns is to impose price curvature by *quadratic pricing*. Following (Papadaskalopoulos and Strbac 2015), we use a superlinear surcharge on purchases and a sublinear bonus on feed-in:

$$\xi_t = \alpha_0 + \alpha_1 P_t^b + \alpha_2 [P_t^b]^2, \quad \phi_t = \beta_0 + \beta_1 P_t^s + \beta_2 \sqrt{P_t^s},$$

The superlinear term steepens the marginal purchase price, thereby discouraging demand spikes and reducing reliance on peaker reserves, while the sublinear feed-in adjustment tempers incentives for excessive injections, promoting smoother system operation(see our online appendix [2] for the full details).

## The `Energy-Net` Simulator

In spite of a variety of simulators that currently exist (Pigott et al. 2022; Moriyama 2018; Vázquez-Canteli et al. 2019; Marot 2021), there is no current framework that allows modeling the complex structure we want to account for and that is designed to work with off-the-shelf RL and MARL methods. We therefore develop a novel simulator, `Energy-Net` we can be used to examine various solution approaches.

`Energy-Net` is a modular, discrete–time simulator representing a flexible and adaptable environment, and can be used to accommodate different system configurations. At the core of the design of the software is a decoupling between the physical dynamics of the electrical system and the strategic agents, i.e., it is built around a strict *physics–agent split*. A physical core advances loads, renewables, batteries, and reserves, while the SO and GEAgents interact only through a gym-style `step()` interface (Farama-Foundation 2023). This design (i) lets us plug in any off-the-shelf RL algorithm without touching the power-system code, (ii) isolates market rules in a single controller module, and (iii) ensures that learned policies can affect the grid *only* via explicit levers, such as prices and dispatch tweaks, thus preserving physical realism while streamlining experimentation.

Building on our formal setting, we use `Energy-Net` to instantiate the day-ahead electricity market. A single simulation episode therefore comprises $T$ uniform intervals of length $\Delta t$ (in our experiments $T$=48 and $\Delta t$ = 30min). At each step $t \in \{1, \ldots, T\}$ the environment reveals the current forecast and grid state to the agents, applies their actions, propagates the physical dynamics, and returns next-state observations and rewards through the standard `step` interface (see the appendix for the full details and codebase).

## Case Study - A Day-Ahead Market as MARL

To demonstrate how MARL can be used to model and optimize energy markets, we use our proposed model to optimize the policy of the SO in the simplified day-ahead market described in Example 1. For this, we employ our `Energy-Net` simulator together with state-of-the-art RL algorithms.

**Setup** We evaluate our MARL formulation and pricing schemes described above under a variety of scenarios. Thanks to the fact that `Energy-Net` cleanly separates physical dynamics from agent logic, we are able to stage the empirical study in escalating phases of coordination for the SO and GEAgents.

First, `Baseline` represents a non-adaptable SO for which the dispatch scheduled is set according to the mean of the predicted demand. In `ISO-Dispatch`, we trained and evaluated the SO in isolation; all GEAgents were disabled, so the operator optimised its dispatch $\Delta_t$ under a stochastic yet *non-strategic* demand profile. Next, we enabled a set of PCS-units with a fixed, pre-defined charging policy and retrained the SO, thereby quantifying the benefit of price coordination when storage is present but *non-adaptive*. We examined this setting with two pricing mechanisms: *online linear*, denoted `ISO-L`, and *quadratic*, denoted `ISO-Q`. We then allowed *both* agents to learn concurrently: the SO tunes its real-time dispatch and tariffs, while the PCS-unit adapts its behavior to these market signals. In settings `Joint-Storage-L` and `Joint-Storage-Q` we examined the online and linear pricing, respectively, for a storage-only GEAgent, while in `Joint-PCS-L` and `Joint-PCS-Q`, we added local production and consump-

[2] Codebase and online appendix: https://github.com/CLAIR-LAB-TECHNION/energy-net

tion capabilities (see the appendix for full details). For each episode, we sample the *realized* demand from a Gaussian noise induced predicted demand for each time step $t$, and, when relevant, the realized load and production for the PCS-units. We ran each training phase for $40$ iterations with $4800$ time steps each ($1000$ days) and evaluated it over $20$ episodes. The non-strategic demand (i.e., the non-shiftable power load that comes from appliances that do not change their consumption patterns according to market signals) is described as a stochastic process with Gaussian distribution $D_t \sim \mathcal{N}(7200, 10)$ in MWh. All settings were run using the same demand pattern and performance parameters with resources of 10 cores of Intel(R) Xeon(R) CPU E5-2683 v4 @ 2.10 GHz and 1 × NVIDIA GeForce GPU (12 GB).

**Results** Our focus is on evaluating the SO's ability to operate in the presence of strategic GEAgents by examining (i) how GEAgents affect the dispatch plan, specifically whether the SO leverages renewable output to displace conventional generation, and (ii) the resulting change in reserve usage. Importantly, as discussed above, during training the agents' rewards reflect operational preferences (e.g., the SO is penalized more for energy obtained from the more polluting, quick-to-operate reserves, than for power generated as part of the dispatch plan). However, to enable evaluation independent of these shaped rewards, we report domain-grounded metrics, namely the quantities of energy drawn by each source[3].

The effect of the GEAgents on the market is depicted in Figure 2. Here, each column corresponds to each of the settings described above and shows the aggregate power from the three sources: *Dispatch* (light blue) is the average power generated as part of the dispatch plan, *PCS sold* (green) is the average power bought from the PCS, and *Reserve* is the average reserve activated. While we examined several RL methods, we only show here the key results for agents trained using the TD3 algorithm (Fujimoto, Hoof, and Meger 2018) as implemented by (Raffin et al. 2021).

As demonstrated, the transition from a fixed policy (`Baseline`) to an RL-based SO (`ISO-Dispatch`) increases the use of reserves due to the stochasticity and learning-induced uncertainty introduced into the SO's policy. In settings `ISO-L` and `ISO-Q`, we observe the effect that GEAgents with storage capabilities and fixed charge/discharge policies have on demand patterns. Results demonstrate that incorporating the quadratic pricing approach in `ISO-Q` substantially reduces reserve usage, but increases total (nominal) demand due to storage activity. In contrast, online pricing in `ISO-L` reduces nominal demand but increases reliance on reserves. This trade-off persists in `Joint-Storage-L` and `Joint-Storage-Q`, where GEAgents are adaptive to market signals and slightly further increase overall demand. Finally, `Joint-PCS-L` and `Joint-PCS-Q` demonstrate that local demand and generation from GEAgents reduce both their participation in the market and, under quadratic pricing, the total demand.

In summary, the results demonstrate several key effects

[3]Due to space constraints, we present only key findings here and defer full results to the online appendix.

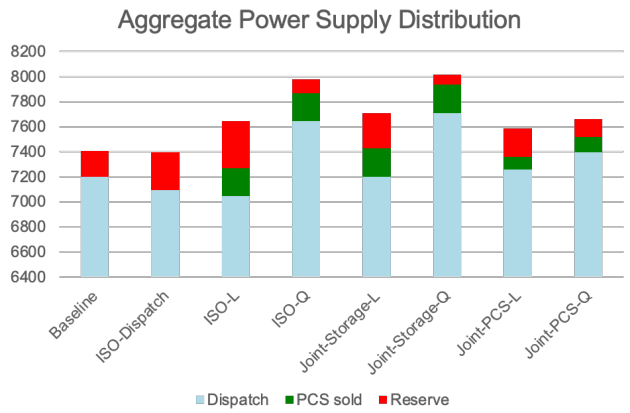


Figure 2: Power supply distribution results with TD3: Dispatch, PCS sold (power sold to the grid by the PCS units), and activated Reserve.

introduced by incorporating GEAgents with PCS-units into the energy market. While these are promising first steps in showing how MARL can be used to analyze and optimize such complex systems, the findings also highlight important challenges, particularly the difficulty faced by the ISO in adapting its dispatch policy and pricing schemes to fully leverage the power supplied by GEAgents. Thus, although we have demonstrated the use of general off-the-shelf RL approaches to represent the learning agents, extending the RL literature with methods tailored to the unique requirements of these systems may lead to deeper understanding and improved optimization outcomes.

## Conclusion

We demonstrate the benefit of modeling modern power systems using MARL, where physical grid constraints, market signals, and heterogeneous agent behaviors interact through tightly coupled feedback loops. Our framework captures both nominal and flexible demand, enabling realistic and robust evaluation of decentralized control strategies and pricing mechanisms through a new simulation environment we developed. The results show that strategically coordinated ISO policies coupled with strategic grid-edge agents can reduce reserve requirements and lower carbon intensity.

Together with these achievements, developing RL methods that are robust to uncertainty, capable of reasoning about risk, and explicitly aware of system-level constraints will be essential to ensure reliable real-world deployment. The fragility of current RL methods was demonstrated in our experiments, where modest forecasting errors may lead to supply shortfalls or excessive generation. Another challenge lies in scaling the approach to operational grids, which will require hierarchical or federated MARL architectures and hardware-in-the-loop testing. Finally, while algorithmic coordination can reduce reserve usage and lower tariffs, the distribution of benefits is unlikely to be uniform; ensuring fairness and transparency is essential to achieving beneficial real-world impact.

## Acknowledgments

This work was supported by NOGA Ltd.- Israel's Independent System Operator. We gratefully acknowledge our collaborators at NOGA for their valuable discussions and insightful feedback, which inspired and guided the development of this work and its future directions.

## References

Ahlqvist, V.; Holmberg, P.; and Tangerås, T. 2022. A survey comparing centralized and decentralized electricity markets. *Energy Strategy Reviews*.

Albrecht, S. V.; Christianos, F.; and Schäfer, L. 2024. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press.

Charbonnier, F.; Morstyn, T.; and McCulloch, M. D. 2022. Coordination of resources at the edge of the electricity grid: Systematic review and taxonomy. *Applied Energy*.

de Vilmarest, T.; Nowotarski, J.; and Ziel, F. 2024. Adaptive Probabilistic Forecasting of Electricity (Net-)Load. *IEEE Transactions on Power Systems*.

Farama-Foundation, T. 2023. Gymnasium: A standard API for reinforcement learning environments. https://github.com/Farama-Foundation/Gymnasium. Accessed: 1 August 2025.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research. PMLR.

Gao, Y.; Wang, W.; and Yu, N. 2021. Consensus Multi-Agent Reinforcement Learning for Volt-VAR Control in Power Distribution Networks. *IEEE Transactions on Smart Grid*.

Ginzburg-Ganz, E.; Segev, I.; Balabanov, A.; Segev, E.; Kaully Naveh, S.; Machlev, R.; Belikov, J.; Katzir, L.; Keren, S.; and Levron, Y. 2024. Reinforcement Learning Model-Based and Model-Free Paradigms for Optimal Control Problems in Power Systems: Comprehensive Review and Future Directions. *Energies*.

Guan, C.; Wang, Y.; Lin, X.; Nazarian, S.; and Pedram, M. 2015. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*.

Harder, N.; Weidlich, A.; and Staudt, P. 2023. Finding individual strategies for storage units in electricity market models using deep reinforcement learning. *Energy Informatics*.

Jang, D.; Spangher, L.; Srivastava, T.; Yan, L.; and Spanos, C. 2023. Personalized Federated Hypernetworks for Multi-Task Reinforcement Learning in Microgrid Energy Demand Response. In *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*.

Keren, S.; Essayeh, C.; Albrecht, S. V.; and Morstyn, T. 2024. Multi-agent reinforcement learning for energy networks: computational challenges, progress and open problems. *arXiv preprint arXiv:2404.15583*.

Li, S.; Cao, D.; Hu, W.; Huang, Q.; Chen, Z.; and Blaabjerg, F. 2023. Multi-energy management of interconnected multi-microgrid system using multi-agent deep reinforcement learning. *Journal of Modern Power Systems and Clean Energy*.

Littman, M. L. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the 11$^{th}$ International Conference on Machine Learning (ICML)*.

Marot, A. e. a. 2021. Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*.

Michailidis, P.; Michailidis, I.; and Kosmatopoulos, E. 2025. Reinforcement Learning for Optimizing Renewable Energy Utilization in Buildings: A Review on Applications and Innovations. *Energies*.

Moriyama, T. e. a. 2018. Reinforcement learning testbed for power-consumption optimization. In *Methods and Applications for Modeling and Simulation of Complex Systems: 18th Asia Simulation Conference (AsiaSim)*. Springer.

Navon, A.; Belikov, J.; Orda, A.; and Levron, Y. 2024. On the Stability of Strategic Energy Storage Operation in Wholesale Electricity Markets. *arXiv preprint arXiv:2402.02428*.

Papadaskalopoulos, D.; and Strbac, G. 2015. Nonlinear and randomized pricing for distributed management of flexible loads. *IEEE Transactions on Smart Grid*.

Perera, A.; and Kamalaruban, P. 2021. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*.

Pigott, A.; Crozier, C.; Baker, K.; and Nagy, Z. 2022. GridLearn: Multiagent reinforcement learning for grid-aware building energy management. *Electric Power Systems Research*.

Qiu, X.; Nguyen, T. A.; and Crow, M. L. 2015. Heterogeneous energy storage optimization for microgrids. *IEEE Transactions on Smart Grid*.

Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; and Dormann, N. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. https://github.com/DLR-RM/stable-baselines3.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the national academy of sciences*.

Shen, R.; Zhong, S.; Wen, X.; An, Q.; Zheng, R.; Li, Y.; and Zhao, J. 2022. Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. *Applied Energy*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Vázquez-Canteli, J. R.; Kämpf, J.; Henze, G.; and Nagy, Z. 2019. CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning. Association for Computing Machinery.

Vázquez-Canteli, J. R.; and Nagy, Z. 2019. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy*.

Werner, L.; and Kumar, P. 2023. Multi-market Energy Optimization with Renewables via Reinforcement Learning. *arXiv preprint arXiv:2306.08147*.

Wolgast, T.; and Nieße, A. 2023. Approximating Energy Market Clearing and Bidding With Model-Based Reinforcement Learning. *arXiv preprint arXiv:2303.01772*.

Yang, T.; Zhao, L.; Li, W.; and Zomaya, A. Y. 2021. Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. *Energy*.

Zhang, B.; Hu, W.; Cao, D.; Huang, Q.; Chen, Z.; and Blaabjerg, F. 2019. Deep reinforcement learning–based approach for optimizing energy conversion in integrated electrical and heating system with renewable energy. *Energy conversion and management*.

Zhang, Y.; Robu, V.; Cremers, S.; Norbu, S.; Couraud, B.; Andoni, M.; Flynn, D.; and Poor, H. V. 2024. Modelling the Formation of Peer-to-Peer Trading Coalitions and Prosumer Participation Incentives in Transactive Energy Communities. *Applied Energy*.

Zhu, Z.; Hu, Z.; Chan, K. W.; Bu, S.; Zhou, B.; and Xia, S. 2023. Reinforcement learning in deregulated energy market: A comprehensive review. *Applied Energy*.