The emerging threat of artificial intelligence on competition in liberalized electricity markets: A deep Q-network approach

Danial Esmaeili Aliabadi (✉ danial.esmaeili@ufz.de)
Helmholtz Centre for Environmental Research https://orcid.org/0000-0003-2922-2400

Katrina Chan
Helmholtz Centre for Environmental Research UFZ Environmental Engineering and Biotechnology Research Unit: Helmholtz-Zentrum fur Umweltforschung UFZ Themenbereich Umwelt- und Biotechnologie

Research Article

Keywords: Collusion, deep reinforcement learning, day-ahead electricity market, Nash equilibrium

Posted Date: October 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-903041/v1

RESEARCH

The emerging threat of artificial intelligence on competition in liberalized electricity markets: A deep Q-network approach

Danial Esmaeili Aliabadi[*]
and Katrina Chan

[*]Correspondence:
danial.esmaeili@ufz.de
Helmholtz Centre for Environmental Research - UFZ, Permoserstraße 15, 04318 Leipzig, Germany
Full list of author information is available at the end of the article

Abstract

Background: According to sustainable development goals (SDGs), societies should have access to affordable, reliable, and sustainable energy. Deregulated electricity markets have been established to provide affordable electricity for end-users through advertising competition. Although these liberalized markets are expected to serve this purpose, they are far from perfect and are prone to threats, such as collusion. Tacit collusion is a condition, in which power generating companies (GenCos) disrupt the competition by exploiting their market power.

Methods: In this manuscript, a novel deep Q-network (DQN) model is developed, which GenCos can use to determine the bidding strategies to maximize average long-term payoffs using available information. In the presence of collusive equilibria, the results are compared with a conventional Q-learning model that solely relies on past outcomes. With that, this manuscript aims to investigate the impact of emerging DQN models on the establishment of collusive equilibrium in markets with repetitive interactions among players.

Results and Conclusions: The outcomes show that GenCos may be able to collude unintentionally while trying to ameliorate long-term profits. Collusive strategies can lead to exorbitant electric bills for end-users, which is one of the influential factors in energy poverty. Thus, policymakers and market designers should be vigilant regarding the combined effect of information disclosure and autonomous pricing, as new models exploit information more effectively.

Keywords: Collusion; deep reinforcement learning; day-ahead electricity market; Nash equilibrium

Introduction

The 7th united nations' sustainable development goal (SDG7) invites societies to provide affordable, reliable, and sustainable energy for everyone. While access to clean energy is the major concern in many developing countries [40], energy affordability is being emphasized in the developed world [12]. Traditionally, electricity had to be consumed instantly after generation due to the unfavorable economics of electricity storage technologies. Owing to this physical constraint, the electricity industry expanded as vertically integrated monopolies around the globe. Unfortunately, these entities suffered from poor performance and high operation costs, which forced governments to reform the electric power sector. To boost the efficiency of these regulated entities, market designers and policymakers pursue liberalization (i.e., deregulation) that aims to maximize social welfare through promoting competition among self-interested participants. Although market designers expect to witness full competition, it is demonstrated that some electricity markets act more like oligopolies for the following reasons [11]:

- Limited number of generators as a result of high capital investment.
- Network congestion that prevents generators from dispatching power to inaccessible consumers.
- Transmission losses that hinder producers in serving remote consumers.

Oligopolistic markets may incubate collusion that harms open competition among participants. While explicit collusion in electricity markets is prohibited, tacit collusion may still exist in the absence of formal contracts [18]. To achieve a perfectly competitive market, collusion (of any kind) should be eliminated, but it is not a straightforward task for regulators to detect tacit collusion [1, 8, 39]. Heim and Götz [19] study the rising price of reserve power in the German market. The authors conclude that the seemingly collusive behavior is due to the repetitive auctions with the pay-as-bid pricing mechanism.

To make matters worse, antitrust agencies are worried that the autonomous pricing algorithms, often used by suppliers, may learn to collude unintentionally [6, 7]. Algorithmic pricing is common in many markets; for instance, according to Chen et al [9], 500 vendors out of 1,641 on Amazon marketplace benefitted from automated pricing algorithms. Algorithmic pricing attracted more attention after the advent of deep reinforcement learning (DRL) [31].

Thus, in this manuscript, we develop a state-of-the-art algorithm based on a deep Q-network (DQN) model, by which GenCos maximize their long-term profits considering the impact of rivals' decisions. The results of the offered DQN model are compared with an extended version of the conventional Q-learning algorithm [2, 3] in a setting with collusive equilibria.

Literature review

Optimization models, which are employed extensively in formulating the strategic behavior of profit-driven power generating companies (GenCos), often require knowledge about rivals' confidential information and market clearing mechanism [33, 49]. For bi-level models, in particular, the lower level should be free of any binary variables, since it is replaced with the Karush Kuhn-Tucker (KKT) optimality conditions [27]. As such, solving the resulting model may present distorted outcomes, since it lacks variables that capture real-world behaviors including shutdown and start-up [48].

On the other hand, simulation models are considered as alternatives to optimization (and equilibrium) models when underlying problems are intractable for analytical methods to address [16]. Typically, researchers rely on agent-based simulation models in decentralized electricity markets, since it provides sufficient flexibility to investigate the impact of learning on GenCos' strategic behavior. At the forefront of imitating human-like intelligence in agents are model-free reinforcement learning (RL) algorithms [44]: agents learn the optimal set of actions (i.e., optimal policy) with respect to each state, solely by interacting with the environment. In spite of their success in various fields, including operations research, decision, and control theories, RL methods (e.g., Q-learning) suffer from two major drawbacks: the lack of theoretical proof to assure solution optimality, and the curse of dimensionality [5]. As the state space expands, the required memory to store transitions grows exponentially with it. To circumvent the dimensionality curse, *Roth-Erev* learning [13] is developed, which is a streamlined version of RL when a limited number of pure strategies are played by agents. However, Roth-Erev-equipped agents are unable to learn consistent behaviors in complex games, such as the sequential bargaining game [20]. To address the dimensionality challenge, a more recent trend is to estimate the optimal action-selection policy using deep neural networks (DNN).

Artificial neural networks (ANNs) have been used in various fields [4, 36] since the 1950s; nevertheless, the combination of ANNs and RL algorithms together with the ever-increasing computational power and the availability of big data attracted researchers' attention in the field of artificial intelligence (AI). After the preliminary breakthrough of the DQN model in the classic Atari 2600 games [31], *AlphaGo* defeated the world champion of the board game *Go* in 2016 by training DNNs using a combination of the supervised learning and RL [41]. To showcase the competence of DNN models, the classic *Go* game is considered, since previously employed AI search algorithms in other games, such as *Chess*, are ineffective.

Aliabadi et al [2] proposed a Q-learning algorithm with time-dependent parameters showing that agents with the extended learning algorithm can converge to either Nash equilibria or strategies with the same payoff tuple under most parameter combinations. On the other hand, Klein [24] has shown that RL-equipped agents can find collusive equilibria in a simple duopoly setting. In [37, 38] agent-based simulation models with RL-equipped agents are employed to investigate actors' behavior in a common standalone balancing energy market, which is scheduled to be put into practice in 2022. In 2019, a DQN model was developed for the first time to optimize GenCos' bidding strategies in deregulated electricity markets [48]. The same group extended their model in 2020 and compared their results with the conventional Q-learning and bi-level models [49]. Recently, Razmi et al [39] employed supervised learning algorithms to detect collusion in day-ahead markets. This algorithm can be used by independent system operators in markets with limited dynamism. Additionally, Guo et al [17] proposed a data-driven recognition system for bidding objective function using deep inverse reinforcement learning and verify their results using DQN.

In this study, we aim to create a state-of-the-art DQN model that assists generic GenCos to raise and sustain their incomes without using confidential information related to employed technologies (e.g., the unit generation cost), while also taking network constraints into account. The outcomes are then investigated to assess the possibility of players unintentionally engaging in collusive behavior. Although DNN models are applied to a wide variety of problems, to the best of the authors' knowledge, this manuscript is the first to analyze the capacity of the aforementioned models in sustaining collusive behavior in deregulated electricity markets.

Problem definition

In this paper, the strategic bidding problem on a day-ahead market is considered, taking network constraints into account. A typical electric grid is made of interconnected nodes, which function independently. In each node, the produced power by

GenCos is consumed by demand centers, and the excess power flows to the connected nodes through transmission lines. Due to physical limitations, transmission lines are unable to dispatch electricity above a certain threshold. A power network is called "congested" when a thoroughly loaded transmission line reaches its maximum capacity and cannot accommodate further dispatch. Network congestion is managed by penalizing electricity consumption at congested nodes using the locational marginal pricing (LMP) scheme [21]. The independent system operator (ISO), who is responsible for the daily operation of the transmission grid, calculates LMP values at any individual node. These values reflect the minimum additional cost of fulfilling the demand for one additional unit of power (MWh).

Similar to [2, 3], the following assumptions are considered in the presented model:

- Generic GenCos are taken into account; thus, GenCos can utilize various technologies (e.g., biogas power plants, wind turbines).
- To ease modeling, small players (i.e., GenCos and demand centers) in each node are aggregated; therefore, aggregated GenCos are assumed to be influential players, which means they can affect rivals' strategic behavior. This assumption is not disruptive because of the oligopolistic nature of electricity markets [11].

To manage the DAM, the ISO conducts a series of auctions every day, in which GenCos submit their bid prices ($b_i^t \in B_i$) and feasible production capacities (P_i^L and P_i^H) for each hour of the next day ($t \in \{1, \ldots, 24\}$). Subsequently, the ISO solves an optimal power flow (OPF) problem concerning submitted bids such that social welfare is maximized at each hour. In this manuscript, the DCOPF problem is adopted as it is employed extensively in power systems operation and is a linear programming (LP) model. The optimal solution of the DCOPF problem at hour t determines the electricity price (λ_i^t) and voltage angle (θ_i^t) at each node, and GenCos' production level (P_i^t). The DCOPF problem formulation is given as follows:

$$\text{minimize}_{P_i^t, \theta_i^t} \qquad z^t = \sum_i b_i^t P_i^t \tag{1}$$

subject to

$$P_i^t - D_i^t = \sum_{ij \in BR} y_{ij} \left(\theta_i^t - \theta_j^t \right) \qquad \forall i \tag{2}$$

$$P_i^L \leq P_i^t \leq P_i^H \qquad \forall i \tag{3}$$

$$-\pi \leq \theta_i^t \leq \pi \qquad \forall i \tag{4}$$

$$|y_{ij}(\theta_i^t - \theta_j^t)| \leq F_{ij}^H \qquad \forall ij \in BR \tag{5}$$

Here, D_i^t is the demand at node i and hour t, BR is the set of available transmission lines, y_{ij} represents the admittance of the connecting line between a pair of nodes (i.e., i and j), P_i^L and P_i^H determine the minimum and the maximum generation capacity of GenCo-i, respectively, and F_{ij}^H specifies the maximum permissible flow in the transmission line connecting node-i to node-j. For the sake of simplicity, we assume P_i^H, P_i^L, and D_i^t to be constant through time in the remainder of this paper.

The objective function in Eq.(1) is to minimize the electricity procurement cost. Eq.(2) balances the flow of electricity by transmitting the extra power of each node into connected nodes. Eq.(3) confines the maximum and minimum permissible capacity of each GenCo. P_i^L can be set to a positive value when the power is already purchased or GenCo-i is selling according to a support mechanism such as feed-in tariffs. Eq. (4) limits the voltage angle within a finite range. Additionally, the value of θ^t at the reference node is set to zero. Finally, Eq.(5) controls the maximum flow through transmission lines. At the optimal solution, the dual variable corresponding to Eq. (2) sets the unit electricity price at each node (i.e., λ_i^t).

After clearing the market by the ISO, GenCos can calculate their payoffs at each specific hour as $r_i^t = P_i^t(\lambda_i^t - c_i)$, where the electricity generation cost of GenCo-i is captured by c_i. It is quite realistic to assume GenCos conceal their payoffs from rivals as it may reveal confidential information regarding their business [17].

Collusive strategy

As mentioned earlier, a feature of any oligopolistic market is the likelihood of participants engaging in collusion. Table 1 displays a simplified case, in which GenCos' payoffs (r_1^t, r_2^t) are displayed with respect to various bid values. According to Table 1, $(b_1^t = 20, b_2^t = 30)$ is the Nash equilibrium of the single-stage game (i.e., $t \in \{1\}$), as no GenCos can get a better payoff by deviating from the Nash strategy given the other player keeps bidding the same price; nonetheless, there is another strategy $(b_1^t = 30, b_2^t = 40)$, the so-called collusive strategy, in which both GenCos can obtain higher payoffs. Collusive strategies can lead to exorbitant electric bills for end-users by damaging consumer surplus in favor of producer surplus. The high electricity price is one of the influential factors causing energy poverty in societies [35].

Although the collusive strategy serves both GenCos, it is considered unstable in a single-stage game since GenCo-2 can benefit far more by deviating the collusive strategy, i.e., $(b_1^t = 30, b_2^t = 20)$. What prevents GenCo-2 from doing so is the response of GenCo-1 in the forthcoming hours, which can move the game to the Nash equilibrium and damage the long-term profit of both GenCos in infinitely repeated games (i.e., $t \in \{1, \dots, \infty\}$): GenCo-2's deviation from SCE (by offering €20) forces GenCo-1 to play €20 per hour instead of €30. In the next hour, GenCo-2 has to offer €30. Therefore, both GenCos fail by playing less a profitable strategy for the reminder of the time horizon.

Based on the UK's competition market authority [43], algorithmic pricing may help to improve the stability of collusion by allowing cartel members to identify deviations from the negotiated bid prices more rapidly.

Table 1 Payoff profile of GenCo-1 and GenCo-2. The arrows display the transformation of offers in subsequent hours when GenCo-2 deviates from the collusive strategy.

$B_1 \backslash B_2$	**20**	**30**	**40**	**50**
20	(857,0)	(3428, 785)	(6000, 0)	(6000, 0)
30	(416, 2500)	(3428, 785)	**(6000, 1571)**	(6000, 0)
40	(0, 6000)	(0, 6000)	(0, 6000)	(5000, 0)
50	(0, 6000)	(0, 6000)	(0, 6000)	(0, 7500)

In this manuscript, we adopt terminology and definitions similar to that which is available in [1] for the strong collusive equilibrium (SCE) and the most collusive equilibrium (SCE*).

Given the payoff table, discovering collusion using heuristics has been studied in [14]; however, as mentioned earlier, GenCos have imperfect knowledge about rivals' payoffs in the real world [17].

Methodology

In this manuscript, two learning mechanisms are discussed in detail. The first section is devoted to a simple QL method with time-dependent parameters, which has been employed in [2]. GenCos that benefit from this QL model exploit their past experiences alone. The next section discusses the proposed DQN method. Although GenCos have no information regarding the dispatched power and the generation cost of rivals, the submitted bids to the ISO are assumed to be common knowledge in the proposed DQN model. The outcomes of the mentioned learning methods will be contrasted.

Q-learning with decay

For each hour, agents submit their bid prices to the ISO in order to satisfy the demand. The ISO determines the winning bids and LMPs, taking the transmission network structure into account. For this algorithm, GenCos calculate the profit corresponding to the submitted bid prices, assuming that they have no information regarding the submitted bids by rivals. Consequently, the optimal action of GenCos can vary based on rivals' responses.

To capture the dynamism of such markets, players should associate uneven significance to the information, based on accumulated knowledge. Thereby, the following time-dependent parameters are introduced:

- Recency rate (α_i^t) determines the importance of the recent outcomes for ith GenCo at iteration t. The value of α_i^t is expected to decline as GenCo-i collects information.
- Exploration parameter (ϵ_i^t) adjusts the exploration rate versus exploitation. As GenCo-i becomes mature, it tends to rely more on collected information than searching for undiscovered solutions.

GenCo-i chooses a bid price randomly with the probability ϵ_i^t, whereas the best-known bid, $b_i^* = \underset{b_{ij} \in B_i}{arg\,max}\{Q_{ij}^t\}$, with the probability $1 - \epsilon_i^t$. In the literature, this mechanism is called the ϵ-greedy action selection rule [42]. Contrary to generic RL algorithms, ϵ_i^t decreases linearly over time to a value near zero, i.e., $\epsilon_i^t = max\{0.001, \frac{8t(\epsilon_i^0 - 1)}{max_t} + \epsilon_i^0\}$, as GenCo-$i$ explores the state-action space sufficiently.

Furthermore, at each iteration, $t \in \{1, \ldots, max_t\}$, GenCo-$i$ updates the Q-value (Q_{ij}^t) corresponding to each bid price ($b_{ij} \in B_i$) based on modified α_i^t and the realized payoff (r_{ij}) as described in Eq. (6).

$$\begin{aligned} \alpha_i^t &= \alpha_i^0 - (0.9t/max_t)\alpha_i^0 \\ Q_{ij}^t &= (1 - \alpha_i^t)Q_{ij}^{t-1} + \alpha_i^t r_{ij} \end{aligned} \tag{6}$$

Deep Q-Networks approach

In this section, the detail of the proposed DQN model is described, by which GenCos enhance their understandings of the environment and optimize their actions accordingly. The critical elements of the proposed model are as follows:

- **Environment:** The platform whereby ISO clears the market and determines agents' rewards.
- **Agents:** Myopic GenCos that desire to increase their long-term rewards through learning.
- **State:** vector s_i^t encapsulates the state of the system for GenCo-i at time t. In our setting, s_i^t consists of the submitted bid prices by all GenCos at time t in addition to private information related to GenCo-i, such as c_i and P_i^t.
- **Action:** The response of GenCo-i to improve its reward, based on observed state (i.e., $b_i^t \in B_i$).
- **Reward:** the obtained payoff of GenCo-i, r_i^t, based on assigned power and cleared price after submitting its bid price.

The overall workflow of the proposed DQN model is depicted in Figure 1. Agents submit random bids at the beginning of the time horizon and store results until the number of records in their replay memory ($\mathfrak{B}_i$) exceeds a minimum level. Then, GenCo-i chooses a batch of experiences from memory using the last-in, first-out (LIFO) scheme[1]. The LIFO scheme is used to prioritize and capture recent interactions among players. The selected experience $\{s_i^t, b_i^t, r_i^t, s_i^{t+1}\}$ are normalized and fed into a feed-forward multi-layer neural network to predict the expected reward for the submitted bid price b_i^t using Eq.(7).

$$Q_i^{t+1}(s_i^t, b_i^t|\,\vec{w_i}) = (1-\alpha_i^t)\,Q_i^t(s_i^t, b_i^t|\,\vec{w_i}) + \alpha_i^t\,(r_i^t + \gamma\,\mathbb{E}[max_{b_i^{t+1}}\{Q_i^t(s_i^{t+1}, b_i^{t+1}|\,\vec{w_i})\}] \tag{7}$$

$$\alpha_i^t = \alpha_i^0\, e^{-0.1\left(|s_i \in \mathfrak{B}_i : s_i = s_i^t| - 1\right)} \tag{8}$$

In Eq.(7), the discount factor ($\gamma \in (0,1)$) presents GenCos' perceived significance of future rewards compared to immediate payoff. According to Eq.(7), the expected future reward, $\mathbb{E}[max_{b_i^{t+1}}\{Q_i^t(s_i^{t+1}, b_i^{t+1}|\vec{w_i})\}]$, is calculated since s_i^{t+1} is established based on the collective actions of all GenCos (b_i^t, $\forall i \in I$), and not a GenCo solely. As is evident, the Markov property does not hold, considering the action space of each GenCo at the beginning of the simulation, i.e., $p(s_i^{t+1}|s_i^t, b_i^t) \neq p(s_i^{t+1}|s_i^1, b_i^1, s_i^2, b_i^2, \dots, s_i^t, b_i^t)$; however, this property may hold if all GenCos act optimally and choose the best bid, b_i^*, corresponding to a given state at time t. Thus, the Markov property asymptotically holds true if the learning process converges.

Eq.(8) reduces α_i^t value from an initial level of α_i^0 based on the number of recorded identical s_i^t entries in $\mathfrak{B}_i$. Hence, when state s_i^t appears more frequently, $Q_i^t(s_i^t, b_i^t|\vec{w_i})$ converges to a fixed function, i.e., $Q_i^*(s_i^t, b_i^t|\vec{w_i})$, as the solution space is being explored sufficiently [45]. To improve the network stability during the learning process, the weight vector ($\vec{w_i}$) of the target network is synchronized periodically (i.e., every 2500 iterations). In theory, this technique assists smoother convergence by preventing instantaneous oscillations while accelerating the process by not training the target network separately [31].

The rectified linear unit (ReLU) function is adopted as the activation function of hidden layers in both target and prediction networks. In contrast, a regression

[1] Our approach is different from [48], which uses the First-in, First-Out scheme

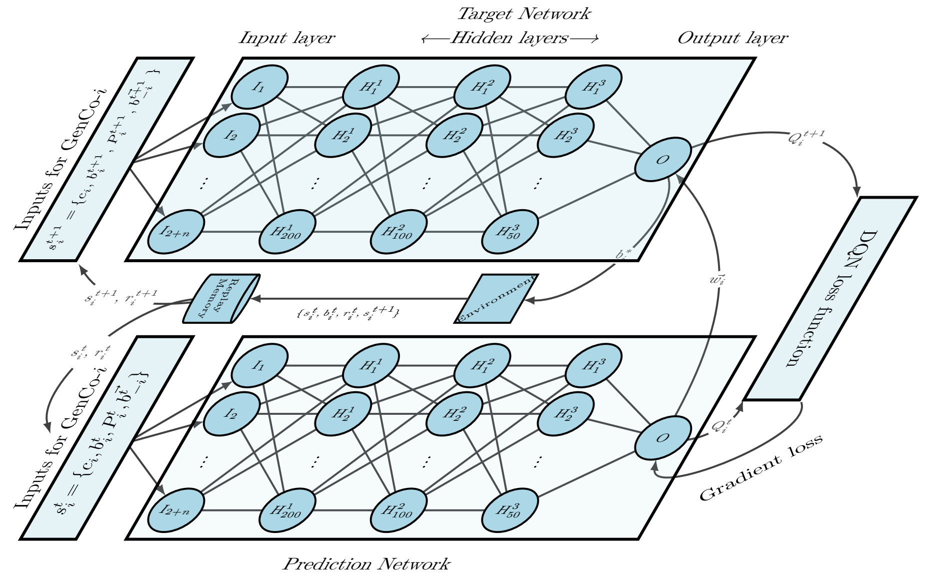

Figure 1 Schematic diagram of DQN model for GenCo-i.

layer is added to the output layer. In order to train the network, the loss function is minimized using the widely-used Adam [23] algorithm with the following values for hyper-parameters: the batch size of 32, the learning rate of 0.01, $\beta_1 = 0.99$, $\beta_2 = 0.999$, and weight decay of L2 regularization of 0.015.

Algorithm 1 displays the method by which GenCo-i evaluates bids before submitting. At first, GenCo-i determines whether to offer a random bid price with the probability of ϵ_i^t or to exploit collected knowledge to submit the best-known bid price otherwise. ϵ_i^t is computed similar to the Q-learning with decay algorithm.

Algorithm 1 Submitting a bid at time t by GenCo-i

```
 1: $r \leftarrow \mathcal{U}(0,1)$
 2: if $r \leq \epsilon_i^t$ then
 3:     if $b_i^{t-k} \neq \ldots \neq b_i^{t-1}$ then
 4:         $b_i^t \leftarrow$ choose a bid randomly from $B_i$
 5:     else
 6:         $k \leftarrow k + 1$
 7:         $b_i^t \leftarrow b_i^{t-1}$
 8:     end if
 9: else
10:     $b_i^* \leftarrow \underset{b_i^t \in B_i}{arg\,max}\{Q_i^t(s_i^t, b_i^t|\vec{w}_i) + \mu_i^t\}$
11: end if
```

To improve the stability, lines 3-8 in Algorithm 1 do not allow GenCo-i to exercise its right for choosing a random bid if k previous bids are unchanged for some reason.

When GenCo-i decides to submit the best-known bid, it feeds the current state, s_i^t, into the prediction network and chooses the bid that maximizes the reward according to the equation in line 10. In line 10, μ_i^t incentives not altering GenCo's best-known bid if the Q-values of other options are just slightly better, i.e., $b_i^* \approx b_i^{t-1}$. The μ_i^t parameter also penalizes smaller bid prices than b_i^{t-1} to prevent a price war between GenCos.

To implement the proposed DQN model, the ConvNetSharp library [22] has been utilized in EMSimulator [15]. Moreover, EMSimulator employs the Microsoft Solver Foundation [30] library to clear the wholesale electricity market at each hour, through which the simulation process is accelerated by generating the DCOPF model on the fly. We assume that ISO uses a lookup table for the optimal solutions of previously solved problems. Doing so helps speed up the simulation even further at the expense of eliminating possible alternate optimal solutions.

Results

Figure 2 illustrates a case study with seven nodes and four active agents, in which strong collusive strategies exist in 13 states. The presented case study is the modified version of the real Pennsylvania-New-Jersey-Maryland (PJM) five node power system, which is widely used in economic papers [3, 25, 26, 32] due to its simplicity. We developed a script to adjust structure-related parameters such that SCE is available, given the set of bid prices.

The maximum generation capacity of GenCos (P_i^H) and load at demand centers (D_i) are written within the boundary of each node. Also, the maximum permissible flow between the source and destination nodes (F_{ij}) are mentioned next to the transmission lines. The dedicated set of bid prices for each GenCo, B_i, and the unit cost of generating electricity, c_i are shown at the top-right corner. We devise bid prices such that no two GenCos have the same offer. Doing so will decrease the possibility of alternative optimal solutions per se.

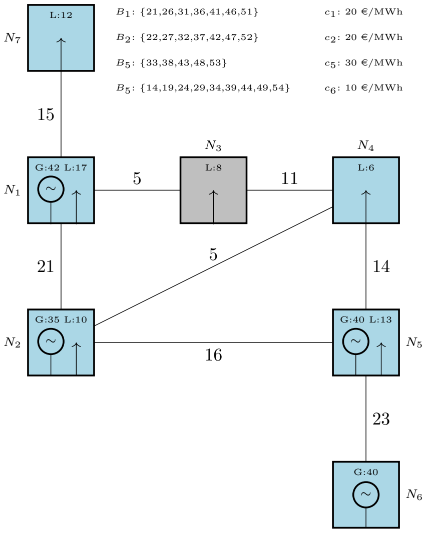

Figure 2 The case study with seven nodes and collusive strategies. The third node is considered as the base node.

Among 2205 states, the most collusive strategy is at ($b_1^{SCE} = 51, b_2^{SCE} = 47, b_5^{SCE} = 43, b_6^{SCE} = 39$) with the payoff tuple of ($r_1^{SCE} = 279, r_2^{SCE} =$

$513, r_5^{SCE} = 195, r_6^{SCE} = 667$). The two Nash equilibria strategies are at ($b_1^N = 31, b_2^N = 27, b_5^N = \{33, 38\}, b_6^N = 29$) with the same payoff tuple ($r_1^N = 225.5, r_2^N = 157.5, r_5^N = 0, r_6^N = 437$). It is clear that $r_i^{SCE} > r_i^N, \forall i \in I$.

The simulation is conducted ten times over 100,000 iterations in a computer with 16 GB memory and an Intel Core i7-10510U processor. The program dedicates a thread with its exclusive memory space to each GenCo; hence, four logical cores out of eight are utilized thoroughly in this case study.

The initial recency (α_i^0) and initial exploration rates (ϵ_i^0) of all GenCos are set to 0.1 and 0.9, respectively. The prediction network is trained using Adam algorithm as mentioned earlier. Figure 3 shows the total loss ($\sum_i \mathcal{L}_i(\vec{w}_i)$) of the action-value function for 20,000 iterations and five replications. The pale lines represent different replications, and their average is drawn with a darker line.

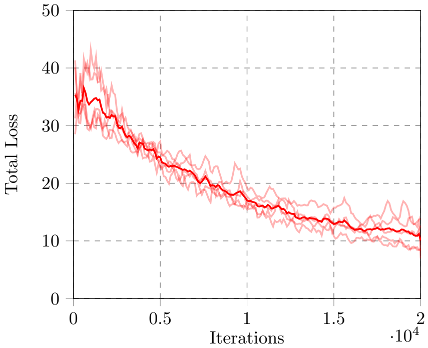

Figure 3 Loss reduction through learning. The light-toned lines represent various replications, and the dark line depicts their average. The total loss decreases as agents minimize the error using Adam algorithm.

Figure 4 demonstrates payoff values of all GenCos as an instance when GenCos converge to an SCE. GenCo-2 and GenCo-6 gradually increase their payoffs while GenCo-1 and GenCo-5 struggle to hold their position in the market. Finally, all parties settle on an SCE strategy, ($b_1^{SCE} = 51, b_2^{SCE} = 47, b_5^{SCE} = 43, b_6^{SCE} = 34$), at around 87K.

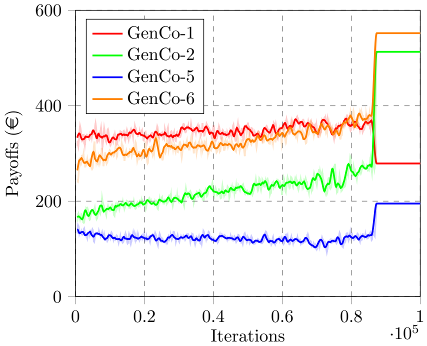

Figure 4 GenCos' payoff transformation through iterations. Pale lines represent payoffs with a resolution of 100 iterations. To observe directions clearly, payoffs smoothed out in listed trends.

As depicted in Table 2, the converged tuple of bids using the proposed DQN outperforms the Nash equilibria and Q-learning with decay, in terms of payoffs.

The bold rows mean convergence to an SCE as defined in [1]. However, GenCo-i's average payoff ($\mathbb{E}[r_i^{DQN}]$) is not greater than the corresponding payoff in the SCE*. The proposed DQN algorithm was able to find an SCE in 70% of replications and the SCE* in 30% of occurrences. On average, DQN equipped GenCos could earn €1466 per hour versus €1018 in QL with decay. If all GenCos agree to act according to the SCE*, they can acquire €1654 per hour. This means that DQN-equipped GenCos can obtain 77.5% of acquirable profit if they diverge from the Nash equilibrium to the SCE*.

Table 2 Converged bid and payoff for each GenCo under two learning mechanisms

	DQN				Q-Learning with decay			
#	b_1^*/r_1^*	b_2^*/r_2^*	b_5^*/r_5^*	b_6^*/r_6^*	b_1^*/r_1^*	b_2^*/r_2^*	b_5^*/r_5^*	b_6^*/r_6^*
1	**51/279**	**47/513**	**43/195**	**34/552**	46/182	32/324	33/96	34/0
2	**51/279**	**47/513**	**43/195**	**39/667**	41/147	32/324	33/27	29/437
3	51/217	42/594	53/207	49/897	41/430	37/328	48/0	34/552
4	**46/234**	**42/418**	**38/120**	**29/437**	36/328	32/270	38/0	29/437
5	**51/279**	**47/513**	**43/195**	**39/667**	36/387	37/0	33/160	34/43.2
6	**51/279**	**47/513**	**43/195**	**29/437**	41/430	37/328	43/0	34/552
7	**46/234**	**42/418**	**38/120**	**29/437**	36/112	32/324	33/96	34/552
8	**51/279**	**47/513**	**43/195**	**39/667**	36/112	32/324	33/96	34/552
9	31/367	32/116	33/0	29/437	36/112	32/324	33/96	34/552
10	51/217	42/594	43/117	34/552	41/147	32/324	43/117	34/552
$\mathbb{E}[r_i^*]$	266	471	154	575	239	287	69	423

The readers should note that the convergence to SCEs is not guaranteed since a simulation method is employed. However, we have a few significant observations:

1 The average payoff using DQN is higher than the conventional Q-learning methods.

2 Participants often receive payoffs larger than Nash equilibria of the single-stage game, which might be caused by a price war and competition among players.

3 The proposed setting unveils the possibility of players unintentionally engaging in collusion in an oligopoly market.

Discussion

It is well-known in the literature that transparent markets facilitate maintaining tacit collusion via coordination of GenCos' actions [34]. However, we designed a DQN model in this manuscript, which has no information regarding the rivals' utilized technology, LMPs, and dispatched powers. The developed model discovers and sustains collusive strategies only by knowing the rivals' offered prices even though GenCos' objective is to improve the long-term payoff.

While the proposed DQN model proves the possibility of tacit collusion among players in deregulated electricity markets, it should be noted that submitted bids are assumed as common knowledge. Although GenCos may learn rivals' actions under pay-as-bid pricing, this information usually stays hidden behind the curtain of market-clearing prices under uniform and DCOPF pricing. Hence, for agents to collude using the proposed algorithm, the bidding curve should be available immediately, which is not the case in many countries [28]. According to [46, 47], information disclosure varies extensively among countries: some countries release bidding curves almost immediately while others experience a delay of multiple weeks or months.

There are also supporting discussions regarding the immediate release of bidding curves by ISO [10, 28, 29]. All in all, the general trend around the globe confirms that markets are moving toward full transparency, notably with data concerning historical bidding behaviors [17]. Hence, market designers and policymakers should consider the joint impact of autonomous pricing and information disclosure on GenCos' behavior prior to crafting market regulations.

Conclusions

Liberalized electricity markets should overcome empirical challenges to materialize predicted objectives completely. One major challenge is to achieve a fully competitive market by eliminating collusion of any type. While revealing the exercise of market power by participants is a tough row to hoe, autonomous pricing algorithms add extra complexity to the problem. In this paper, we aim to investigate the impact of emerging DQN models on the behavior of players. The outcomes suggest that GenCos may be able to collude unintentionally while trying to ameliorate long-term profits. Therefore, policymakers and market designers should be vigilant regarding the combined effect of information disclosure and autonomous pricing, as new models exploit information more effectively.

Although the proposed DQN model does not need the solution of the DCOPF problem for rivals, it still requires knowing other GenCos' bidding curves. Consequently. one future research direction might be to design an algorithm that only relies on publicly available information such as LMPs with monthly delay.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
All generated or analyzed data in this study are included in this manuscript.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
DEA Conceptualization, Methodology, Visualization, Perform formal analysis, Investigation, Data curation, Coding, Writing - original draft. **KC** Original draft

Acknowledgments
The authors would like to thank Prof. Tiago Pinto and Prof. Zita Vale for their invitation to the 19th EPIA Conference on Artificial Intelligence, where the preliminary results of this manuscript were presented.

References

1. Aliabadi DE, Kaya M, Şahin G (2016) Determining collusion opportunities in deregulated electricity markets. Electric Power Systems Research 141:432–441
2. Aliabadi DE, Kaya M, Şahin G (2017) An agent-based simulation of power generation company behavior in electricity markets under different market-clearing mechanisms. Energy Policy 100:191–205
3. Aliabadi DE, Kaya M, Sahin G (2017) Competition, risk and learning in electricity markets: An agent-based simulation study. Applied Energy 195:1000–1011
4. Avşar B, Aliabadi DE (2015) Parallelized neural network system for solving euclidean traveling salesman problem. Applied Soft Computing 34:862–873

5. Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems 13(1-2):41–77
6. Bernhardt L, Dewenter R (2020) Collusion by code or algorithmic collusion? when pricing algorithms take over. European Competition Journal 16(2-3):312–342
7. Calvano E, Calzolari G, Denicolo V, Pastorello S (2020) Artificial intelligence, algorithmic pricing, and collusion. American Economic Review 110(10):3267–97
8. Çelebi E, Şahin G, Aliabadi DE (2019) Reformulations of a bilevel model for detection of tacit collusion in deregulated electricity markets. In: 2019 16th International Conference on the European Energy Market (EEM), IEEE, pp 1–6
9. Chen L, Mislove A, Wilson C (2016) An empirical analysis of algorithmic pricing on Amazon marketplace. In: Proceedings of the 25th International Conference on World Wide Web, pp 1339–1349
10. Darudi A, Moghadam AZ, Bayaz HJD (2015) Effects of bidding data disclosure on unilateral exercise of market power. In: 2015 International Congress on Technology, Communication and Knowledge (ICTCK), IEEE, pp 17–24
11. David AK, Wen F (2001) Market power in electricity supply. IEEE Transactions on energy conversion 16(4):352–360
12. Dubois U, Meier H (2016) Energy affordability and energy inequality in europe: Implications for policymaking. Energy Research & Social Science 18:21–35
13. Erev I, Roth AE (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review pp 848–881
14. Esen B (2019) Utilizing genetic algorithm to detect collusive opportunities in deregulated energy markets. Master's thesis, Sabanci University
15. Esmaeili Aliabadi D (2016) Analysis of collusion and competition in electricity markets using an agent-based approach. PhD thesis, Sabanci University
16. Esmaeili Aliabadi D, Çelebi E, Elhüseyni M, Şahin G (2021) Modeling, simulation, and decision support. In: Pinto T, Vale Z, Widergren S (eds) Local Electricity Markets, Academic Press, pp 177–197,
17. Guo H, Chen Q, Xia Q, Kang C (2021) Deep inverse reinforcement learning for reward function identification in bidding models. IEEE Transactions on Power Systems
18. Harrington JE (2018) Developing competition law for collusion by autonomous artificial agents. Journal of Competition Law & Economics 14(3):331–363
19. Heim S, Götz G (2021) Do pay-as-bid auctions favor collusion? evidence from Germany's market for reserve power. Energy Policy 155:112,308
20. Hemmati M, Nili M, Sadati N (2010) Reinforcement learning of heterogeneous private agents in a macroeconomic policy game. In: Progress in Artificial Economics, Springer, pp 215–226
21. Huisman R, Huurman C, Mahieu R (2007) Hourly electricity prices in day-ahead markets. Energy Economics 29(2):240–248
22. Karpathy A (2016) Convnetsharp
23. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
24. Klein T (2021) Autonomous algorithmic collusion: Q-learning under sequential pricing. The RAND Journal of Economics pp 1–21,
25. Krause T, Andersson G (2006) Evaluating congestion management schemes in liberalized electricity markets using an agent-based simulator. In: 2006 IEEE Power Engineering Society General Meeting, IEEE, pp 8–pp
26. Krause T, Andersson G, Ernst D, Vdovina-Beck E, Cherkaoui R, Germond A (2004) Nash equilibria and reinforcement learning for active decision maker modelling in power markets. In: Proceedings of the 6th IAEE European conference: modelling in energy economics and policy
27. Kuhn H, Tucker A (1951) Nonlinear programming. In: Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, pp 481–492
28. Lazarczyk E, le Coq C (2017) Information disclosure in electricity markets. In: Heading Towards Sustainable Energy Systems: Evolution or Revolution?, 15th IAEE European Conference, Sept 3-6, 2017, International Association for Energy Economics
29. Markard J, Holt E (2003) Disclosure of electricity products—lessons from consumer research as guidance for energy policy. Energy Policy 31(14):1459–1474
30. Microsoft (2017) Microsoft solver foundation. Available from https://www.nuget.org/packages/Microsoft.Solver.Foundation
31. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
32. Mohammad N, Mishra Y (2018) The role of demand response aggregators and the effect of gencos strategic bidding on the flexibility of demand. Energies 11(12):3296
33. Naghibi-Sistani MB, Akbarzadeh-Tootoonchi M, Bayaz MJD, Rajabi-Mashhadi H (2006) Application of Q-learning with temperature variation for bidding strategies in market based power systems. Energy Conversion and Management 47(11-12):1529–1538
34. Overgaard PB, Møllgaard HP (2008) Information exchange, market transparency and dynamic oligopoly. University of Aarhus Economics Working Paper (2007-3)
35. Papada L, Kaliampakos D (2016) Measuring energy poverty in greece. Energy Policy 94:157–165
36. Pinto T, Falcão-Reis F (2019) Strategic participation in competitive electricity markets: Internal versus sectorial data analysis. International Journal of Electrical Power & Energy Systems 108:432–444
37. Poplavskaya K, Lago J, De Vries L (2020) Effect of market design on strategic bidding behavior: Model-based analysis of european electricity balancing markets. Applied Energy 270:115,130
38. Poplavskaya K, Lago J, Strömer S, de Vries L (2021) Making the most of short-term flexibility in the balancing market: Opportunities and challenges of voluntary bids in the new balancing market design. Energy Policy 158:112,522

39. Razmi P, Buygi MO, Esmalifalak M (2020) A machine learning approach for collusion detection in electricity markets based on nash equilibrium theory. Journal of Modern Power Systems and Clean Energy
40. Ritchie H, Roser M (2020) Access to energy. Our World in Data https://ourworldindata.org/energy-access
41. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
42. Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press
43. UK Competition and Markets Authority (2018) Pricing algorithms: Economic working paper on the use of algorithms to facilitate collusion and personalised pricing. Retrieved Sep. 13, 2021 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/746353/Algorithms_econ_report.pdf.
44. Vázquez-Canteli JR, Nagy Z (2019) Reinforcement learning for demand response: A review of algorithms and modeling techniques. Applied Energy 235:1072–1089
45. Watkins CJ, Dayan P (1992) Q-learning. Machine Learning 8(3-4):279–292
46. Wolak FA (2014) 4. regulating competition in wholesale electricity supply. In: Economic regulation and Its reform, University of Chicago Press, pp 195–290
47. Yang Y, Bao M, Ding Y, Song Y, Lin Z, Shao C (2018) Review of information disclosure in different electricity markets. Energies 11(12),
48. Ye Y, Qiu D, Papadaskalopoulos D, Strbac G (2019) A deep Q network approach for optimizing offering strategies in electricity markets. In: 2019 International Conference on Smart Energy Systems and Technologies (SEST), IEEE, pp 1–6
49. Ye Y, Qiu D, Sun M, Papadaskalopoulos D, Strbac G (2019) Deep reinforcement learning for strategic bidding in electricity markets. IEEE Transactions on Smart Grid 11(2):1343–1355