*38th International Electric Vehicle Symposium and Exhibition (EVS38) Göteborg, Sweden, June 15-18, 2025*

# Scheduling Electric Vehicle Charging for Participation in the Belgian Imbalance Market Using Model-Free Reinforcement Learning

Saeed Naghdizadegan Jahromi[1], Gilles Van Kriekinge[1], Cedric De Cauwer[1], Thierry Coosemans[1]

[1]*Saeed Naghdizadegan Jahromi (corresponding author) EVERGi Research Group, MOBI Research Centre & ETEC Department, Vrije Universiteit Brussel (VUB), Pleinlaan 2, Belgium, saeed.naghdizadeganjahromi@vub.be*

**Executive Summary**

This study explores the potential of using a model-free Reinforcement Learning (RL) approach to optimize Electric Vehicle (EV) charging scheduling for participation in the Belgian imbalance market, on the use case of an office parking. Motivated by changing regulations enabling smaller assets like EVs to participate in ancillary markets, the study aims to develop a smart charging strategy that minimizes charging costs by leveraging imbalance price volatility. The proposed problem was formulated as a Markov Decision Process (MDP), and an RL agent was subsequently trained to optimize the charging schedule. The proposed approach balances economic gains with considering the target SOC, offering promising benefits for grid flexibility and revenue generation. Results demonstrate that the proposed RL strategy outperforms uncoordinated and smart charging in the day-ahead market with respect to charging cost, achieving negative average charging costs by leveraging price fluctuations.

*Keywords: Smart charging, Electric Vehicles, Modelling & simulation, AI – Artificial intelligence for EVs, Charging business models*

## 1 Introduction

With the rapid global adoption of EVs, the increasing number of vehicles poses potential challenges to the grid demand and stability if not properly managed. However, these challenges also present opportunities, particularly through the use of Electric Vehicles' (EVs) battery capacities for ancillary services [1]. European regulations are changing to allow smaller assets with fast responses times, like EVs and Battery Energy Storage Systems (BESSs), to participate in ancillary markets [2]. In response, Elia, Belgium's transmission system operator (TSO), is adjusting its framework to make this possible [3]. In the context of various ancillary markets, promising options for EVs to actively participate as new players include Frequency Containment Reserve (FCR), automatic Frequency Restoration Reserve (aFRR), and the imbalance market. The imbalance market is particularly promising due to its lower regulatory barriers and the potential for direct contracts with Balance Responsible Parties (BRPs). Additionally, it is an attractive market because there is no minimum capacity requirement, making it accessible for more participants. A BRP is a privately owned legal entity responsible for managing and balancing one or more access points on the transmission grid [4]. With the liberalization of the European electricity market, BRPs have taken on increased responsibility for their system balancing, although TSOs continue to play a role in this process [4]. The TSO ensures overall grid reliability by monitoring, coordinating, and taking corrective actions to maintain a balanced and stable power system

[4]. An imbalance for BRPs stands for the discrepancy between the planned and actual energy consumption or generation during a specified time frame. Smart charging strategies offer a solution for BRPs to improve grid stability and flexibility inside their portfolios [5]. This research aims to evaluate the economic feasibility of EV participation in the Belgian imbalance market and to assess the potential of model-free RL in learning optimal policy for smart charging to adapt efficiently to the dynamic of this market.

# 2 Literature Review

Integrating EVs and BESSs to provide ancillary services has gained significant attention in research [5], with authors formulating stochastic and optimization models for BESS [6] [7]. In [6], authors developed a service stacking strategy for residential BESS to enhance renewable energy sources integration by leveraging flexibility in ancillary services markets. Paper [7] introduces a stochastic Model Predictive Control (MPC) methodology for BESS that enables implicit balancing in European balancing markets by optimizing out-of-balance decisions in real-time. However, it is argued in [8] that model-based methods are not suitable for this problem due to their nonlinear and nonconvex nature, so they proposed model-free solutions. Additionally, the uncertainty in integrating BESS and EVs into ancillary markets is important because the accuracy of these model-based methods relies on how precise their approximations are. This means that sudden changes in market dynamics can impact the accuracy of these predictions and, as a result, affect the outcomes.

Moreover, on top of the imbalance price volatility, EVs, as non-stationary assets, have additional uncertainties such as arrival time, departure time, initial state of charge (SOC), required energy, and the number of available EVs in the parking lot, which need to be addressed.. While paper [9] addresses uncertainty in future energy demand within a day by using scenario modeling for EV charging schedules, it also acknowledges a key limitation, the reliance on past requests to predict future demands. This dependence can lead to scheduling inefficiencies, particularly when unforeseen changes in user behavior arise. In this study, the model does not predict future market prices; instead, it relies solely on current market data and EVs information.

# 3 Methodology

## 3.1 Model-Free Reinforcement Learning

Model-free RL methods do not require prior knowledge or an explicit model of the environment. Instead, by interacting directly with the environment, the agent learns to adapt itself to the system dynamics and accommodate for inherent uncertainties through experience to maximize cumulative reward [10]. This method simplicity is especially useful when the environment is complex, stochastic, or partly observable since it does not require an understanding of how actions affect the environment. Although it comes at the cost of reduced sample efficiency, meaning the agent may need more interactions to learn an effective policy. Model-free RL is depicted in Figure 1. In contrast, model-based RL simulates future states and rewards using a model of the environment's dynamics and reward function, allowing the agent to plan and make decisions without solely relying on direct interaction experience.

Some popular model-free RL methods are Q-learning, Deep Q-Networks, policy gradient methods, and Actor-Critic methods. Proximal Policy Optimization (PPO), as a member of policy gradient category, is chosen among these due to its balance of stability, sample efficiency, and ease of implementation, making it suitable for a wide range of applications [11]. The policy gradient method's purpose is to directly optimize the policy, which is the strategy the agent uses to decide what action to take in each state by adjusting the parameters to maximize the anticipated cumulative reward. The set of interactions of an agent with its environment (actions, states, rewards) from a starting state to an ending state is called an episode. PPO is developed to address the instability and high variance inherent in earlier policy gradient methods, which frequently resulted in unstable learning processes and suboptimal performance. By incorporating a clipping mechanism within the objective function of the new policy, PPO ensures that policy updates are significant enough to yield meaningful improvements while remaining sufficiently constrained to prevent instability. In PPO, deep neural networks are utilized as function approximators to represent both the policy and the value function, allowing the algorithm to efficiently navigate complex environments and large state spaces. A

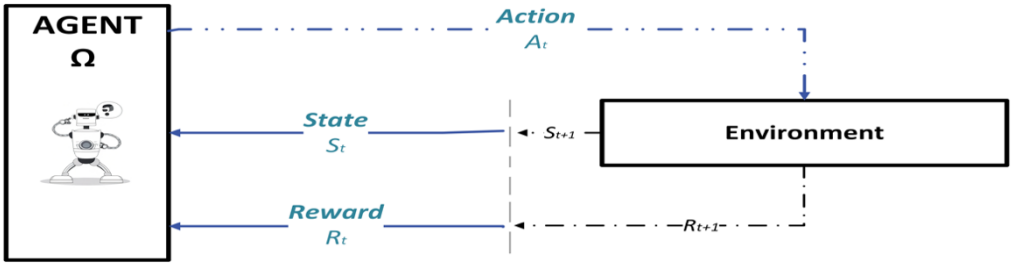


Figure 1. Model-Free reinforcement learning diagram

detailed explanation of the objective function employed for policy updates can be found in [7].

## 3.2 Markov Decision Process Formulation for EVs Smart Charging

The EV smart charging can be formulated as Markov Decision Process (MDP) which provides a mathematical structure for stochastic sequential decision-making problems. The MDP is formally defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. Here $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the discrete action space, $\mathcal{R}: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ indicates the immediate reward function, $\mathcal{P}: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$ signifies the unknown state transition probability distribution. The discount factor $\gamma \in (0,1]$ indicates the importance placed on future rewards relative to immediate ones [12]. At each time step $t$, the agent observes the current environment state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ in response. As a result of this action, the agent receives a reward value $\mathcal{R}(s_t, a_t)$ and transitions to a new state $s_{t+1} \in \mathcal{S}$ according to the state transition probability distribution $\mathcal{P}(s_{t+1} \mid s_t, a_t)$. The MDP formulation for EV smart charging within the imbalance market can be structured as follows:

### 3.2.1 State:

In this study, these selected episode for the RL is a randomly selected specific day and charger. Once training is completed, the developed policy is implemented across all chargers, enabling efficient management of the EV charging process based on the learned strategies. At each timestep, the observation space, referring to the set of possible inputs or states the agent can perceive from its environment, is defined as follows:

$$S_t = \left(T_t, T^{dep}, SOC_t, \hat{\pi}_t^{imb}\right) \quad (1)$$

Where $T_t$ is the timestep on that day, $T^{dep}$ represents the specific timestep during the day when the EV, is scheduled to depart. $SOC_t$ is the SOC of the EV at the timestep *t*, and the $\hat{\pi}_t^{imb}$ is the forecasted imbalance price of the system. A forecast of the imbalance price is used because TSO calculates it at the end of each quarter-hour based on the activated balancing resources. It is assumed that EV users provide their $SOC^{target}$ and departure time ($T^{dep}$) when they connect their car to the charger, and that this information is known in advance.

### 3.2.2 Actions

It requires defining appropriate actions that the agent can select during its interaction with the environment. For this study, a discrete action space is considered, containing three possible actions, as detailed below:

$$a_t \in A, \quad A = \{0, \frac{1}{2} \times P_{max}^{ch}, P_{max}^{ch}\} \quad (2)$$

Where $P_{max}^{ch}$ is the maximum rate of the EV charger in kW. Thus, $a_t$ represents the action chosen by the agent, which can be idle, charge at half rate, or charge at the maximum rate. While the two actions (0 and $P_{max}^{ch}$) are straightforward—indicating whether the EV is charging with maximum rate or not—introducing a third action ($\frac{1}{2} \times P_{max}^{ch}$,) allows for more nuanced control over the charging process. Selecting the half-rate charging option allow the agent to charge during periods of slightly elevated electricity prices without completely forgoing the charging process.

### 3.2.3 **Reward function**

The reward function is depicted in equation (3). It consists of a cost-based term and a term to capture the degree to which the charge is successfully completed. The complexity of the reward function (Equations 3-10) is a direct consequence of the multi-objective nature of the EV charging problem in the context of imbalance market participation, where the primary objective is to maximize charging revenues (minimize charging cost), while trying to fulfill the charging needs of the driver, meaning the EV reaches its target SOC by the departure time set by the driver. Simple reward functions, such as those based solely on imbalance price cannot adequately capture this essential trade-off. Therefore, a well-designed reward system with a clear reward-penalty structure is implemented to promote desirable behaviors, such as achieving the target SOC and minimizing expenses, while also discouraging adverse outcomes like excessive costs or insufficient energy delivered to EVs. This is particularly important, given the complexities of EV smart charging in the imbalance market, which include stochastic imbalance prices, varying EV arrival and departure times, and differing initial SOCs.

The reward at each timestep (3), comprises two components: the stepwise cost and energy delivery-related term, reflecting the agent's need to align with the EV driver's charging preferences. The weighted factors α and β are employed to adjust the influence of each respective term.

$$r_t = \alpha \times Cost_t + \beta \times Energy\_delivery\_factor_t \quad (3)$$

$$Cost_t = -\frac{a_t}{P_{max}^{ch}} \times \frac{\hat{\pi}_t^{imb}}{\hat{\pi}_{max}^{imb}} \times b_t \qquad \in [-1, 1] \quad (4)$$

The cost formula in equation (4) is the chosen action $a_t$ by the agent at time $t$ , normalized by the maximum charging power $P_{max}^{ch}$, multiplied with the imbalance price $\hat{\pi}_t^{imb}$ , normalized by the maximum imbalance price ($\hat{\pi}_{max}^{\text{imb}}$) observed in the dataset, and $b_t$, which is a binary variable that takes the value of 1 when the EV is connected to the charger and 0 when it is not connected. is Normalization is done to ensure $Cost_t$ ranges from [-1, 1]. The imbalance price can be either negative or positive: a negative price means money is received from the TSO, while a positive price indicates a payment to the TSO is required. To respect this convention, a negative sign is added to the cost formulation in equation (4).

The $Energy\_delivery\_factor$ component, as outlined in Equation (5), comprises two essential terms: charging progress (6), which measures the change in the EV SOC, and the penalty component (10), which measures the deviation of the SOC ($SOC_t$) to the target SOC ($SOC^{\text{target}}$), both at each time step $t$ and at the end of the charging session ($t = T^{\text{dep}}$ ). The charging progress term component quantifies the increase in the SOC of the EV by calculating the change in SOC over time, scaled by an urgency factor. The $Time\_Sensitive\_Urgency\_Factor$ in (7) is a dynamic metric that quantifies the urgency based on the remaining time until departure. It acts as a crucial gauge of urgency, facilitating decision-making in time-sensitive situations. Without considering the remaining time, the RL agent might prioritize minimizing cost too heavily in the early stages, leaving insufficient time to reach the target SOC as the departure time approaches. In the equation (7), $\rho$ represents a baseline level of urgency, while $\mu$ indicates the sensitivity to time, defining how urgency increases as the departure time approaches. Equation (7) is normalized in (8) by dividing by $\rho + \mu$, which represents maximum possible value of (7). These values were calculated based on trial and error in this study. Also $\Delta SOC_t$ divided by $\Delta SOC_{max}$ equation (9) to normalize the SOC change. Where $C^{ev}$ is the maximum capacity of the EV battery and $\eta^{ch}$ is the efficiency of the charger in equation (9).

$$Energy\_delivery\_factor_t = \text{Charging\_Progress}_t^{norm} + Penalty_t^{norm} \quad (5)$$

$$Charging\_Progress_t = \Delta SOC_t \times Time\_Senssitive\_Urgency\_Factor \quad (6)$$

$$Time\_Senssitive\_Urgency\_Factor = \rho + \frac{\mu}{max(1,\ T^{dep} - t)} \quad (7)$$

$$Charging\_Progress_t^{norm} = \frac{\Delta SOC_t}{\Delta SOC_{max}} \times \frac{\rho + \frac{\mu}{max(1, I^{dep} - t)}}{\rho + \mu} \qquad \in [0,1] \quad (8)$$

$$\Delta SOC_{max} = \frac{P_{max}^{ch} \times \eta^{ch} \times \Delta t}{C^{ev}} \quad (9)$$

$$Penalty_t = \begin{cases} -\dfrac{\left(SOC^{target} - SOC_t\right) \times b_t}{\beta \times SOC^{target}} & \text{if } t < T^{\text{dep}} \\ -\dfrac{\left(SOC^{\text{target}} - SOC_{T^{dep}}\right)}{SOC^{target}} & \text{if } t = T^{\text{dep}} \text{ and } SOC_{T^{dep}} < SOC^{\text{target}} \\ 1 & \text{if } t = T^{\text{dep}} \text{ and } SOC_{T^{dep}} = SOC^{\text{target}} \end{cases} \quad (10)$$

The penalty component designed as equation (10), when the time t has not yet reached the departure timestep ($T^{\text{dep}}$), the penalty is calculated as the normalized deviation from the requested $SOC^{target}$. This time dependent signal allows the RL model to receive feedback every action step, in contrast to the other component which is a single penalty at the departure time which would not offer sufficient opportunity for the model to adjust its strategy. To avoid this component having disproportionate influence due to being calculated at every connected timestep before departure, the first condition is divided by β, the scaling factor of the full $Energy\_delivery\_factor$ . Upon reaching timestep $T^{\text{dep}}$, the penalty factor assumes a negative value if the departure SOC is less than the level requested by the EV driver, and a positive value of 1 if the SOC satisfies the specified threshold. The penalty factor falls within the range [-1, 1].

### 3.2.4 State Transition

Within the MDP framework, system dynamics are characterized by a state transition probability function, denoted as $P$. In the context of the EV smart charging problem, this probability function is unknown due to the uncertainties related to each EV presence and imbalance price. Consequently, the agent seeks to estimate the state probability distribution through its interactions with the environment. Nonetheless, the state transition for the $SOC_t$ is governed by $a_t$ and can be explicitly represented as follows:

$$SOC_{t+1} = SOC_t + \frac{a_t \times \eta^{ch} \times \Delta t}{C^{ev}} \quad (11)$$

## 3.3 Benchmark charging methods

### 3.3.1 Uncoordinated charging method

Then uncoordinated charging method is the current charging method where EVs charge exclusively on the driver's preferences and schedule, disregarding any broader effects on the power grid or charging infrastructure. Typically, EVs begin charging at a maximum rate as soon as they connect to the charger and continue to consume energy until fully charged [13].

### 3.3.2 Model Predictive Control charging method

The model predictive control charging method effectively addresses the challenges posed by the uncoordinated charging of EVs. Unlike uncoordinated charging, MPC utilizes a proactive approach by considering energy demands and optimizing charging schedules to minimize costs while reaching the desired SOC for users. At each timestep, MPC repeatedly solves an optimization problem that considers current conditions, including grid status, energy prices, and EV charging requirements future energy demand and consumption. We use the model developed in [13] applied to our use case. This smart charging control not only meets the energy demands of EV drivers but also achieves global minimization of charging costs.

# 4 Case Study

The proposed framework is depicted in Figure 2. The simulation consists of an office equipped with 10 EV chargers, operating under exposure to imbalance market prices. The historical data required for the imbalance and day-ahead markets was collected from the TSO, while the EV data consists of a historical dataset of charging sessions at a commercial office building in Belgium. After preprocessing, this data is divided into training and validation datasets to train and fine-tune the RL model. Finally, a test dataset is used to evaluate the final model's capabilities.

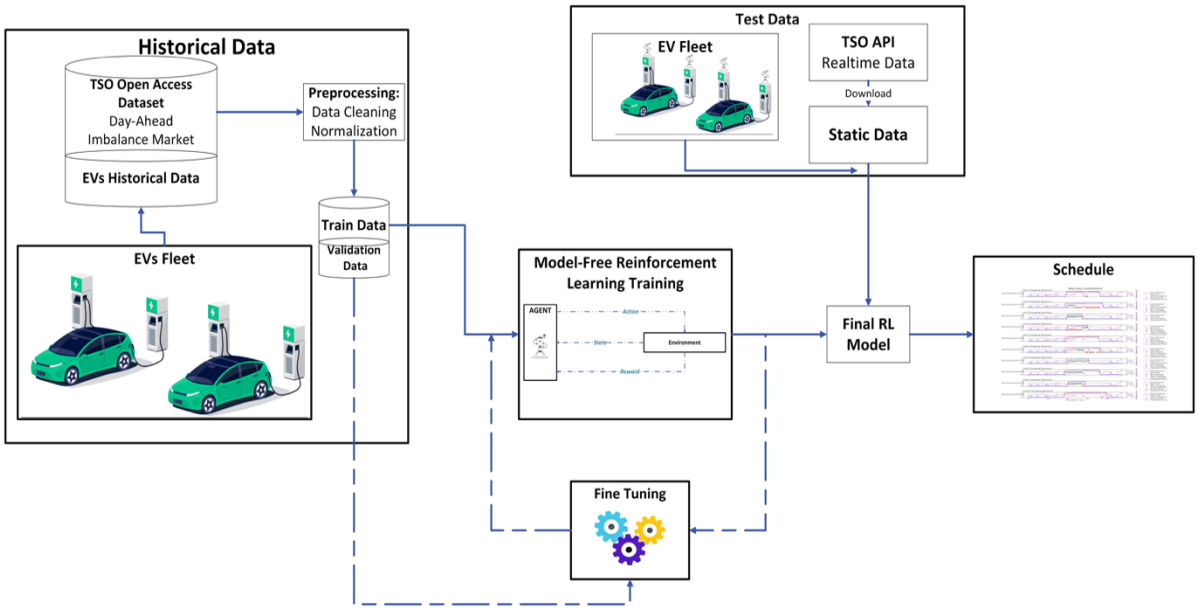


Figure 2. Proposed framework for EV smart charging in the imbalance market

## 4.1 Dataset electricity market

In this study, the electricity data comprises day-ahead and imbalance prices sourced from the Elia open data website, which can be accessed at [14] (the historical data section in Figure 2). The year 2023 was chosen to represent current market conditions due to its recent and relevant data. Elia publishes two types of imbalance prices: the 15-minute-based price and the 1-minute-based price. The 15-minute-based price serves as the reference for imbalance settlements for BRPs. It represents the real imbalance price calculated at the end of each quarter-hour period and is subject to a validation process by TSO to ensure accuracy. In contrast, the 1-minute-based prices are derived from non-validated data. These prices reflect the instantaneous system imbalance and the cumulative activated regulation volumes on a minute-by-minute basis. They are intended to provide BRPs with additional insights into the grid's real-time status and the adjustments being made [15].

BRPs manage imbalances across their entire portfolios, and EV parking lots—whether through aggregators or direct contracts with BRPs—can actively participate in the imbalance market by offering flexibility services. There is no minimum size requirement, as it is dependent on the contractual agreement with the BRP. As depicted in Figure 3, the frequency of negative imbalance prices is particularly high during peak office parking hours (8:00–18:00), with the highest occurrence rate of 35% around noon. Most imbalance prices, in 2023, are concentrated in the pricing bin ranging from -69.95 €/MWh to 344.83 €/MWh, a range that spans approximately 414.8 €/MWh. With an appropriate strategy targeting this pricing window, there is significant potential to decrease the cost of charging. The maximum observed imbalance price during the simulation period was 1450 €/MWh, which was used to normalize the imbalance price data.

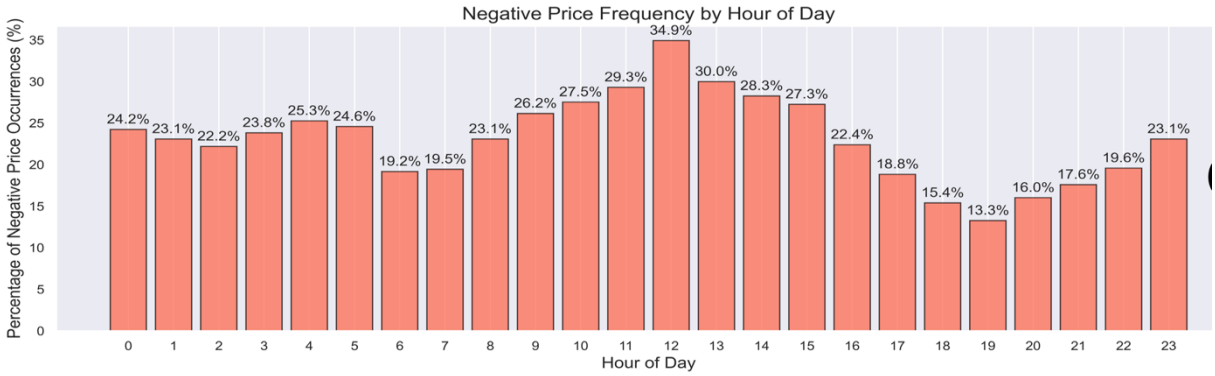


Figure 3. Belgium's market of imbalance price for 2023, Negative price frequency percentage in each hour

In electricity markets, a baseline refers to the scheduled level of energy consumption or generation, serving as a reference point to compare against actual levels and identify imbalances. In the use case, we define a zero-baseline consumption to be cleared day-ahead. By defining a zero-baseline scenario, every power of the EV parking facilities is considered a deviation from the baseline and will be exposed to the imbalance prices, and do not engage the day-ahead market. The setup allows for a focused assessment of the implications of relying exclusively on the imbalance market, giving insight into both its potential cost advantages and risks. Since the EV parking lot participates with a zero baseline and combining individual EVs does not yield additional advantages in this context, a unified smart charging strategy has been developed and applied to all chargers. The approach is scalable as it standardizes decision-making across the entire charging infrastructure, eliminating the need for individualized control logic. This uniformity enables integration and expansion, allowing the system to easily adapt to additional chargers and growing EV fleets.

Table 1. PPO Hyperparameters Optimized by Optuna

| **Hyperparameter** | Explanation | Range | Optimal Value |
|---|---|---|---|
| **learning_rate** | The learning rate | [$1e^{-5}$ - $1e^{-2}$] | $17e^{-5}$ |
| **n_steps** | The number of steps to run for each environment per update | [192 - 672] | 576 |
| **batch_size** | Minibatch size | [192 - 672] | 480 |
| **clip_range** | Clipping parameter | [0.1 – 0.5] | 0.34 |
| **gamma** | Discount factor | [0.9 – 0.999] | 0.985 |
| **ent_coef** | controls the weight of the entropy term in the loss function, promoting exploration by encouraging a more diverse action distribution | [0.1 - 0.8] | 0.34 |
| **vf** | Value Network hidden layer size | [32 - 256] | [32, 64] |
| **pi** | Policy Network hidden layer size | [32 - 256] | [128, 128] |

## 4.2 Dataset charging sessions

The case study focuses on an office parking lot equipped with 10 charging points. The charging sessions, defined by their arrival times, parking durations, and energy requirements, are derived from a historical dataset of charging sessions at a commercial office building in Belgium, as detailed in [16]. The simulation used data from a 90-day period between July 1st and September 30th, 2023. It was trained on the first 70 days, validated on the following 10 days, and tested on the final 10 days. The chargers are considered to have maximum power of 11 kW ($P_{max}^{ch}$). The charger's efficiency ($\eta^{ch}$) is considered at 90%. For simplicity, it is assumed that all EVs have a similar 70 kWh battery capacity ($C^{ev}$). It is also assumed that each charger can accommodate one charging session per day. The arrival and departure times, as well as the SOC, can assume any values in range of (0,1]. The $SOC^{target}$ is set to 1, indicating that EV drivers aim to fully charge their vehicles. In this paper, after several trials and adjustments, the parameters $\alpha$ and $\beta$ were set to 1 and 32, respectively. In equation (8) the $\rho$ and $\mu$ are 0.25 and 2, respectively. These values adjust the magnitude of rewards and penalties at each timestep, thereby considering that the impact of each component is balanced and proportionate within the optimization framework.

## 4.3 Reinforcement Learning hyperparameter

Reinforcement learning hyperparameters play an important role in training agents because they control how quickly and effectively an agent learns optimal policies. A fine-tuning of hyperparameters can enhance performance and convergence rates, whereas a poor choice may result in slow learning or suboptimal solutions. In this study, the Optuna library [17] is utilized to identify the optimal hyperparameters for this problem. Optuna is a robust framework for hyperparameter optimization that updates the search for the best parameters using advanced optimization techniques. The result of the hyperparameters is reported in Table 1. The simulation is executed in a Python environment, utilizing Stable-baselines3 [18] for reinforcement learning. This framework provides a collection of reliable reinforcement learning algorithms in Python, offering a user-friendly and efficient interface for training and assessing RL agents. Given the complexity of

the problem, the PPO model was trained for 20000 episodes, with each episode representing a full day. For each episode, the RL chooses a random day and a random EV on that day, then trains and gathers the required information to use as experience.

### 4.4 Key performance indicator

This subsection outlines the criteria used to evaluate the performance of charging strategies in both the day-ahead and imbalance markets. To evaluate the technical and economic performance of the configurations, as well as the primary reward equation presented in equation (3), the following key performance indicators (KPIs) are utilized:

- *Charging cost:* the total cost related to pay or receive from the market (€/kWh).
- *Energy delivery:* Defined as the percentage of the total energy supplied to EVs during charging relative to the total energy required to achieve a fully charged state, expressed as a percentage (%).

Table 2. Comparative results of charging scenarios

| **Charging Simulation** | Market | kWh Charged | Total Cost (€) | Cost per kWh (€/kWh) | Percentage of Energy delivery |
|---|---|---|---|---|---|
| **Uncoordinated charging** | Day-Ahead | 2282.26 | 216.76 | 0.094 | 100 |
| **MPC method** | Day-Ahead | 2282.26 | 170.43 | 0.074 | 100 |
| **Uncoordinated Charging in Imbalance price** | Imbalance | 2282.26 | 205.38 | 0.09 | 100 |
| **Knowing 15-min Imbalance price** | Imbalance | 2206.93 | -45.84 | -0.027 | 96.7 |
| **1-min imbalance price** | Imbalance | 2079.14 | -29.47 | -0.01 | 91.1 |

## 5 Results

#### 5.1.1 Day-ahead market

To facilitate a comprehensive comparison, the first scenario is in the day-ahead market using two methods: uncoordinated and MPC charging. Table 2 presents the findings for the period from 21 to 30 September 2023. A total of 83 charging sessions took place during this period. The total cost for uncoordinated charging amounts to €216.60 for charging 2282.23 kWh. The MPC method achieved a price reduction to €170.4 while successfully charging all EVs, resulting in 100% energy delivery.

#### 5.1.2 Uncoordinated charging in the imbalance market

This scenario is designed to assess the impact of exposing charging operations to the imbalance market price in the absence of smart charging algorithms. To ensure a fair comparison across all the scenarios, a simulation was also conducted over the same 10-day test period and reported in Table 2. As shown in Table 2, this scenario resulted in a slight reduction of 5.2% in total cost compared to uncoordinated charging in the day-ahead market. Nonetheless, it is observed that it still could not outperform the MPC method in the day-ahead market.

#### 5.1.3 Imbalance market with knowledge of 15-minute prices

As mentioned in Section 4.1, the actual imbalance price is revealed at the end of each 15-minute interval. RL method was tested under the assumption that only the next 15-minute imbalance prices were perfectly known in advance. This idealized scenario provided a simplified environment to demonstrate the RL model's theoretical capabilities. The RL agent must develop an optimal policy without knowledge of future prices

while observing the state represented by $S_t = (T_t, T^{dep}, SOC_t, \hat{\pi}_t^{imb})$.

In this scenario, each day consists of 96 timesteps (24 × 4). The simulation run on a MacBook Pro (with M3 MAX CPU and 48 GB RAM) took 190 minutes. The total cost is €-45.84, indicating that the BRP receives money from the TSO. The total energy delivery is 2206.93 kWh, resulting in an energy delivery rate of 96.7%. Among the 83 charging sessions, two EVs departed with SOC levels below 90%, registering values of 85% and 88%, respectively.

In Figure 4, the policy developed by the RL agent using PPO is displayed. The heatmaps illustrate a dynamic charging policy in function of the imbalance price and SOC. To reduce the four-dimensional state space of Equation (1) to a 2-D plot based on SOC and imbalance price, the other two variables are fixed. For this plot, the arrival time is set at 8:00 AM and the departure at around 6:00 PM, reflecting typical office hours in Belgium. The probability distribution generated by the PPO algorithm (shown in Figure 4) reflects how likely the agent is to choose each possible action in a given state. This framework allows the agent to balance exploration and exploitation by giving higher probabilities to actions that are more beneficial, while still occasionally selecting less likely actions. The probability distribution refines over time as the agent accumulates experience, providing a more optimal policy through iterative learning processes. When the training phase is complete, the learned policy is used deterministically. At each time step, based on the SOC and the normalized imbalance price while also considering the known departure time, the agent consults the policy and selects the action with the highest probability. As it is seen from Figure 4, when the SOC is low (0% to ~30%) the model shows a preference for charging at the maximum rate (100%) even when faced with slightly positive prices (up to ~0.25). This aggressive charging behavior rapidly increases SOC to prevent energy deficits, possibly disregarding cost considerations. As the SOC increases, the model starts to adjust its charging strategy. It tends to reduce the charging rate in response to elevated prices. The width of the no-charge charging also expands slightly during this stage, reflecting a more cautious approach. When the SOC exceeds 70%, the model shows a notable tendency to defer charging, even amid low positive or negative prices. This suggests a strategic wait for further price reductions before engaging in charging at the 100% charging rate. The model shifts from an aggressive, cost-neutral approach at low SOC to a balanced strategy at intermediate SOC, and finally to a conservative, price-conscious position at high SOC.

Figure 5 illustrates the charging progress for a representative day, where the RL agent applies the policy

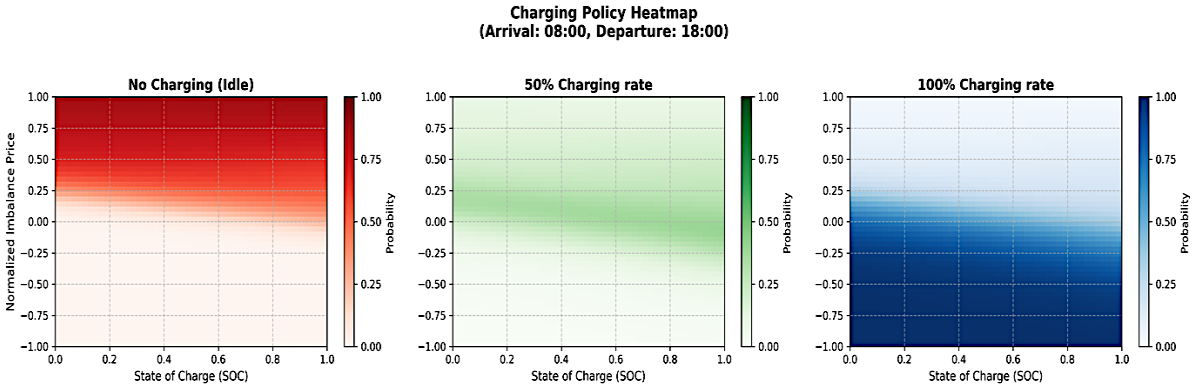


Figure 4. Policy developed with PPO by knowing 15-minute imbalance price

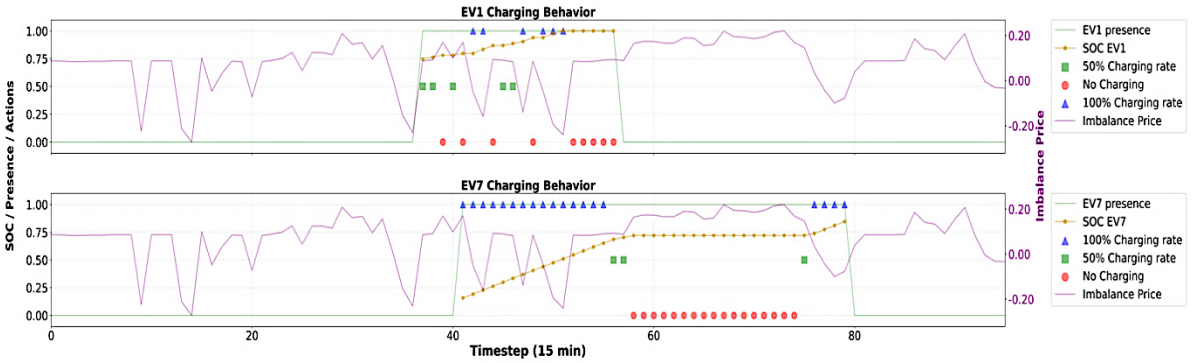


Figure 5. Charging progress of sample day (23/September/2023)

outlined in Figure 4 to make state-dependent decisions for a sample state. The left vertical axis of this plot shows SOC values in yellow and the EV presence with a green line. The right vertical axis displays the normalized imbalance price, which is depicted in purple. For instance, EV1, which connects to the charger with a moderately high SOC (75%), employs a cautious charging strategy by initially selecting a 50% charging rate. This approach enables EV1 to mitigate exposure to periods of marginally elevated imbalance prices, strategically deferring charging until more advantageous pricing conditions arise. Conversely, for EV7, which connects with a low SOC (21%), the agent adopts a full charging rate despite a moderately high positive imbalance price to promptly elevate the SOC to an intermediate level; when more favorable pricing conditions arise later, additional charging is implemented, resulting in a final SOC of 95%, which is deemed acceptable. The RL agent does not guarantee a globally optimal result; rather, it iteratively approximates the best possible pricing outcome by exploiting available observations and experiences.

### 5.1.4 Imbalance market with knowledge of 1-minute prices

Given that having prior knowledge of the exact next 15-minute imbalance price is unlikely, this section assumes that the 1-minute data discussed in Section 4.1 can be used as a prediction to participate in the imbalance market. These 1-minute prices offer an approximation of market conditions, allowing the model to account for real-world conditions. In this section, the RL agent makes decisions at one-minute intervals, consistent with the one-minute granularity of the forecasted imbalance prices. However, the real total cost is calculated using the actual imbalance price, which is validated at the end of 15-minute intervals with the TSO. Due to the increased number of timesteps, this scenario training took 280 minutes.

Here, the final policy is similar to the 15-minute policy. It is depicted in Figure 6. At low SOC levels (0%–30%), the agent predominantly selects a 100% charging rate, displaying insensitivity to moderately positive normalized imbalance prices (up to ~0.25) to quickly accumulate energy. In the intermediate SOC range (30%–70%), the likelihood of using the full charging rate declines under higher prices, with a shift toward either a 50% rate or no charging, reflecting growing cost awareness. At high SOC levels (above 70%), the agent mainly opts for no charging, even when prices are neutral or slightly negative, indicating a strategy to take advantage of potentially lower future prices.

Figure 7 illustrates that for most of the time, the minute-based predictions (shown in black) closely track the actual imbalance prices (shown in purple) measured at the end of each 15-minute interval. However,

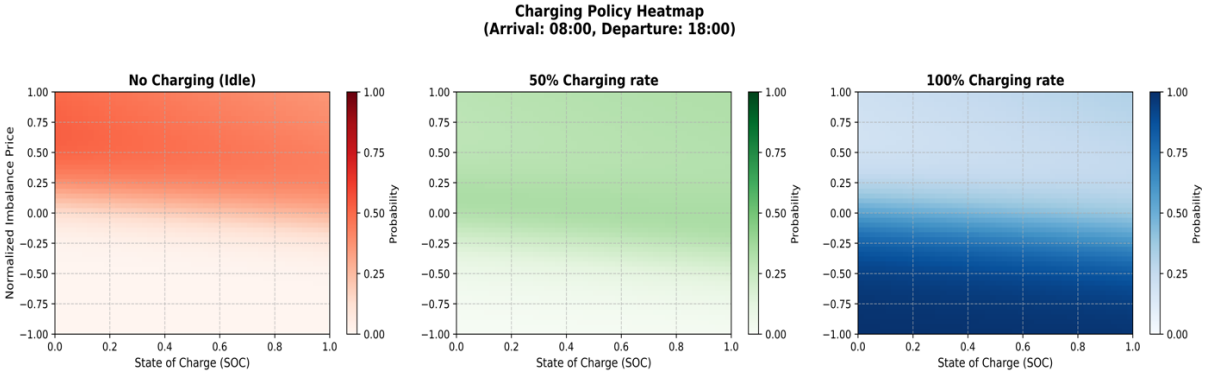


Figure 6. Policy developed with PPO for a 1-minute imbalance price

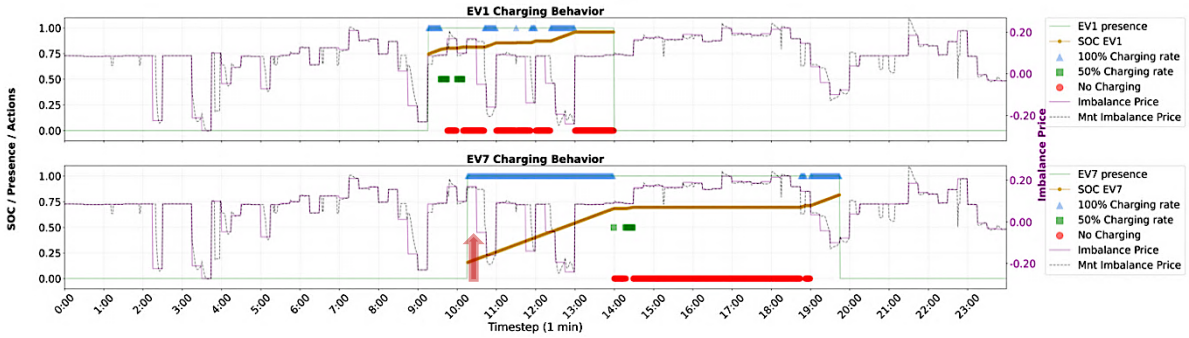


Figure 7. Charging progress of sample day (23/September/2023) in minute granularity

discrepancies do occur. For instance, between 10:00 and 11:00 AM, the predicted imbalance price was approximately 0.2 (normalized), while the actual imbalance price was negative. In this scenario, the agent refrained from charging to avoid high costs, thereby forfeiting an appropriate charging opportunity. As reported in Table 2, compared to the previous scenario, this approach resulted in an approximate 35% increase in cost. Nonetheless, it continues to exhibit clear advantages relative to uncoordinated charging or exclusive participation in the day-ahead market. Among the 83 EV charging sessions, nine vehicles departed with a state of charge (SOC) below 90%, ranging from 77% to 89%. To achieve a full charge, there's a trade-off between penalty and cost optimization—a higher penalty can boost energy delivery but also increase overall costs. For example, when $\beta$ was raised from 32 to 50 in a simulation, the total cost increased by 28.3%, while energy delivery increased by 2%.

# 6 Discussion

The results demonstrate the advantages of smart charging strategies in the day-ahead and imbalance markets. The MPC method in the day-ahead market reduced costs by about 22% compared to uncoordinated charging, highlighting its effectiveness in optimizing expenses without compromising energy delivery. In the imbalance market, uncoordinated charging yielded a total cost of €205.38, a decrease of 5.2% compared to the day-ahead market, but it was still outperformed by the MPC method. Furthermore, the scheduling charging with the RL using the 15-minute imbalance price data resulted in enhanced financial performance with a cost of €-45.84 and corresponding to an energy delivery rate of 96.7%. The RL based on using 1-minute predictions of the imbalance price resulted in €-29.47 and the delivery rate of 91.1. Compared to the 15-minute imbalance price scenario, this is an increase in costs of approximately 35%, and a reduction of energy delivery by 5.8%. These findings indicate that while costs increased in the scenario utilizing 1-minute predictions, this approach, which reflects more closely the real market conditions, still yields good results and thus provides a practical, usable framework. Despite the increased costs, this method still resulted in lower expenses compared to the uncoordinated charging and day-ahead market approaches.

# 7 Conclusion

This study explored the economic feasibility and technical viability of utilizing model-free RL to schedule EV charging to facilitate participation in the Belgian imbalance market. This research demonstrates that an RL-based smart charging strategy effectively employs EV battery flexibility to minimize charging costs and generate revenue by using imbalance price volatility. The approach enabled the RL agent to reduce costs and meet most charging needs by utilizing price fluctuations, especially negative imbalance prices during peak office hours. The proposed framework was tested using 2023 imbalance price data and simulated charging sessions at an office parking lot with 10 charging points under different scenarios, to the research compared the performance of uncoordinated charging and MPC-based smart charging in the day-ahead market, uncoordinated charging in the imbalance market, and two RL-based scenarios in the imbalance market, one with perfect 15-minute price knowledge and another using 1-minute forecasted prices on the specified use case. In the RL scenarios, negative average charging costs (-0.027 €/kWh and -0.01 €/kWh) indicate that RL agents can learn policies that make use of imbalance price volatility, transforming EV charging from a cost center to a revenue possibility. Compared to the higher costs of 0.074 €/kWh for the MPC method, 0.094 €/kWh for uncoordinated charging in the day-ahead market, and 0.09 €/kWh for uncoordinated charging in the imbalance market, this approach offered greater cost-effectiveness.

Despite the highest revenue generated by the scenario assuming perfect knowledge of 15-minute prices, the practical simulation based on 1-minute price data was still able to generate revenue, showing that the RL approach is capable despite inherent market uncertainty and forecast imperfections. However, this economic benefit came at the cost of slightly reduced energy delivery (~91-97%) compared to methods guaranteeing full charges (MPC and uncoordinated methods). This highlights an important trade-off between cost optimization and meeting user charging expectations, suggesting the necessity for further sensitivity analysis on reward parameters and structures, along with the potential for additional safety mechanisms. Multi-year analyses and simulations help to fully evaluate the performance of the developed RL method in the dynamically evolving Belgian imbalance market. With the rapid adoption of EVs and the increasing importance of grid flexibility, such smart charging and market-aware charging strategies are crucial to structure a resilient energy system.

## Acknowledgments

This work has been supported by the ECOFLEX project funded by the FPS economy, S.M.E.s, Self-employed, and Energy, Belgium.

## References

[1] A. Al-Obaidi, H. Khani, H.E. Farag, M. *Mohamed. Bidirectional smart charging of electric vehicles considering user preferences, peer to peer energy trade, and provision of grid ancillary services.* International Journal of Electrical Power & Energy Systems, 124 (2021); 106353.

[2] G. Rancilio, A. Rossi, D. Falabretti, A. Galliani, M. Merlo. *Ancillary services markets in europe: Evolution and regulatory trade-offs.* Renewable and Sustainable Energy Reviews,154 (2022); 111850.

[3] *Innovation Strategy, 2024-2027, Transforming challenges into opportunities,* https://www.elia.be/-/media/ /public-consultations/2023/20230327-innovation-strategy-public-consultation.pdf, accessed on 2025-04-24.

[4] *Terms and Conditions for balance responsible parties (BRPs) ("T&C BRP")*, https://www.elia.be/-tc-brp-into-force-as-of-the-local-go-live-of-the-mfrr-serviceenv12024.pdf, accessed on 2025-04-24.

[5] O. Štogl, M. Miltner, C. Zanocco, M. Traverso, O. Starý. *Electric vehicles as facilitators of grid stability and flexibility: A multidisciplinary overview.* WIREs Energy and Environment, 13 (2024) e536.

[6] G. Rancilio, A. Dimovski, F. Bovera, M. Moncecchi, D. Falabretti, M. Merlo. *Service stacking on residential BESS: RES integration by flexibility provision on ancillary services markets*. Sustainable Energy, Grids and Networks, 35 (2023); 101097.

[7] R. Smets, K. Bruninx, J. Bottieau, J. -F. Toubeau, E. Delarue. *Strategic Implicit Balancing With Energy Storage Systems via Stochastic Model Predictive Control.* IEEE Transactions on Energy Markets, Policy and Regulation, 1 (2023); 373–385.

[8] S.S.K. Madahi, B. Claessens, C. Develder. *Distributional Reinforcement Learning-based Energy Arbitrage Strategies in Imbalance Settlement Mechanism.* arXiv preprint arXiv*, 2401.00015*(2023).

[9] O. Fallah-Mehrjardi, M. H. Yaghmaee, A. Leon-Garcia. *Charge Scheduling of Electric Vehicles in Smart Parking-Lot Under Future Demands Uncertainty.* IEEE Transactions on Smart Grid, 11 (2020); 4949–4959.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov. *Proximal policy optimization algorithms.* arXiv preprint arXiv, *1707.06347*(2017).

[11] Z. Zhu, Z. Hu, K.W. Chan, S. Bu, B. Zhou, S. Xia. *Reinforcement learning in deregulated energy market: A comprehensive review.* Applied Energy, 329 (2023); 120212.

[12] R.S. Sutton. *Reinforcement learning: An introduction.* A Bradford Book (2018).

[13] G. Van Kriekinge, C. De Cauwer, N. Sapountzoglou, T. Coosemans, M. Messagie. *Peak shaving and cost minimization using model predictive control for uni-and bi-directional charging of electric vehicles.* Energy reports, 7 (2021); 8760–8771.

[14] *Elia. Open Data.* https://opendata.elia.be/pages/home/, accessed on 2025-04-24.

[15] *End User Documentation "1-minute publications."*, https://www.elia.be/-/media/project/elia/elia-site/grid-data/balancing/20190827_end-user-documentation-elia1-minute-publications.pdf, accessed on 2025-04-24

[16] G. Van Kriekinge, C. De Cauwer, N. Sapountzoglou, T. Coosemans, M. Messagie. Electric vehicle charging sessions generator based on clustered driver behaviors. *World Electric Vehicle Journal*14 (2023) 37.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama. *Optuna: A next-generation hyperparameter optimization framework.* ACM, 2623–2631.

[18] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, N. Dormann. *Stable-baselines3: Reliable reinforcement learning implementations.* Journal of machine learning research, 22 (2021); 1–8.

## Presenter Biography

Saeed Naghdizadegan Jahromi obtained his Master's Degree in Energy Management in 2016, with a specialization in electric vehicles integration in the power system. He worked in the Power industry for more than six years, and in 2024, he started his PhD at Vrije Universiteit Brussel. His current research focuses on integrating EVs into various electricity markets.