ORIGINAL RESEARCH **Open access**

Towards Autonomous Virtual Power Plants: Deep Reinforcement Learning for Distributed Energy Management

Nathan Carter[1*], Emily Parker[1], Oliver Evans[2], Ahmad Fauzi[3]

Abstract

This research applies machine learning, particularly reinforcement learning (RL), to enable the optimized, real-time management of numerous decentralized, flexible assets at the residential level within the electricity sector. We illustrate the capabilities and existing challenges of RL, offering a framework for practitioners to approach comparable problems without requiring extensive expertise in programming neural networks. Within the energy sector, we show that advanced RL algorithms can be trained to manage what could amount to millions of small residential assets. Specifically, the implemented RL algorithm surpasses standard heuristic methods and performs only marginally below linear optimization outcomes, while requiring less than one-thousandth of the computation time. Consequently, RL creates opportunities for flexible energy asset aggregators to maximize profitability across various smart grid applications, thereby delivering critical grid services and promoting the sustainable management of private energy resources.

Keywords Reinforcement learning, Virtual power plant, Aggregation of energy, Value stacking, Flexibility of decentral energy assets

*Correspondence:
Nathan Carter
nathan.carter@auckland.ac.nz
[1] Department of Artificial Intelligence for Grid Resilience, Faculty of Engineering, University of Auckland, Auckland, New Zealand
[2] Department of Smart Energy Infrastructure, Faculty of Engineering, University of Canterbury, Christchurch, New Zealand
[3] Department of AI for Sustainable Smart Grids, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

Introduction

Recently, deep learning has evolved from a specialized technique into a robust methodology across diverse domains, yielding significant advancements in areas like visual and speech recognition [1], robotics [2], manufacturing processes [3], and multi-agent frameworks [4]. This progress is largely attributed to the capacity to train deep learning models on massive datasets, fostering rapid and optimized decision-making capabilities. This is particularly valuable when traditional dynamic programming and econometric methods encounter the “curse of dimensionality” [5, 6].

Notwithstanding its notable benefits in other fields, several recent prominent studies suggest that the practical application of deep learning remains somewhat limited within business analytics and operations research [7, 8]. This limitation is similarly evident in industry practice: although 85% of business leaders anticipate that artificial intelligence (AI) will yield a competitive edge, a mere 5% have deeply integrated AI into their operational workflows [9]. It appears many industry professionals either bypass AI due to apprehensions about needing profound expertise in neural network coding, or they grow disheartened by the multitude of hurdles often encountered when attempting to deploy a functional system.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

The electrical grid represents a sector where AI adoption could serve as a pivotal catalyst, especially as the system undergoes two major disruptive transformations. First, the surge in digitalization enables the linking of potentially millions of decentralized energy units—encompassing generation, storage, and consumption—into a comprehensive “smart grid” [10, 11]. Furthermore, the shift toward sustainable, low-carbon energy solutions necessitates structural transformations across the electricity, mobility, and heating sectors, thereby fostering sector coupling. Devices such as electric vehicles, heat pumps, and residential battery storage introduce substantial risks to grid stability, as their heavy power requirements threaten to overload current electrical infrastructures [12, 13]. Such shifts in consumption patterns could necessitate immense investments in grid expansion, particularly if grid operators must plan for extreme scenarios (e.g., numerous consumers simultaneously extracting massive volumes of power). Nevertheless, alongside these vulnerabilities, the operational flexibility of these devices presents an opportunity to unlock diverse applications advantageous to consumers, utility providers, and grid operators **(Figure 1)**. For instance, flexible resources can be managed to (1) capitalize on fluctuating power prices, (2) alleviate grid strain by flattening peak demand, (3) supply reserve capacity to maintain electrical frequency, (4) minimize the degradation of battery storage,1 or (5) optimize the on-site usage of local (solar) power generation [14-17].

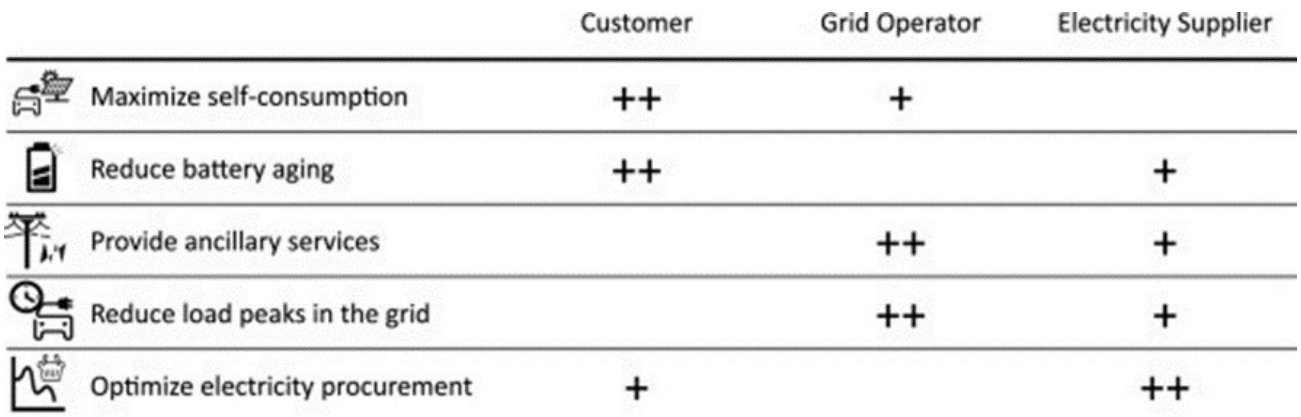


Figure 1. Flexible resources within residential properties, including battery storage or electric vehicles, can be managed to optimize benefits across various value streams for diverse stakeholders (+ … small advantage; ++ … significant advantage).

Nonetheless, end-users like residential consumers often dislike constant modifications to the operational schedules of their equipment. Consequently, managing the flexible capacity of these assets has increasingly been assumed by “aggregators.” These organizations virtually consolidate decentralized units and commercialize their capabilities across the aforementioned applications, thereby creating new streams of income for both their clients and themselves [18, 19].

In an earlier related investigation, we showed that deploying a linear optimization framework to manage flexible resources could yield up to 150 €/a for an average German residence [20]. Such figures illustrate the absolute theoretical maximum, assuming flawless prediction. Practically, these optimal figures are unattainable since local power generation, user consumption, and market prices cannot be predicted with absolute certainty. Furthermore, nonlinear factors, like the degradation trajectories of battery systems, create optimization dilemmas that require minutes or even days to compute (when calculating thousands of future intervals). This extensive processing time is highly impractical for scenarios involving millions of residences that require power schedule adjustments every few minutes (for instance, due to shifts in spot market prices, fluctuations in reserve power needs, the activation of high-demand appliances, or variations in the renewable energy supply mix).

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Recently, deep learning techniques have been implemented in this environment, demonstrating capabilities that exceed traditional decision-making models [8, 21, 22]. Reinforcement learning (RL), a distinct subset of machine learning, emerges as a particularly viable strategy here. In RL scenarios, massive datasets are beneficial rather than problematic, as they facilitate the rapid training of an algorithmic model, known as an agent. Following training, these agents can respond to fluctuating real-world conditions with minimal processing time and computational strain. Additionally, they operate without the need for predictive forecasts or human oversight to analyze or clarify events. Rather, RL systems can learn autonomously from incoming environmental observations. This autonomous learning is a critical asset, given the implausibility of assigning human analysts to monitor the behavioral trends of millions of consumers.

Building upon the preceding research by Kraus *et al*. [8], this paper seeks to fill the gap in deep learning literature within practical operations research, while also providing necessary methodologies for the optimal management of residential smart grid units across various revenue channels. To achieve this, we analyze whether deep learning systems can surpass linear optimization techniques in maximizing profitability through the flexible management of electricity across concurrent value streams, drawing on the concepts outlined by Madlener [11].

This manuscript offers a twofold primary contribution. First, the outlined methodology for engineering a reinforcement learning algorithm for specific applications can be executed without requiring specialized proficiency in neural network coding. This allows professionals and scholars across diverse disciplines to utilize the framework, thereby enhancing the precision or computational efficiency of their optimization tasks. Second, we demonstrate that deep reinforcement learning models can serve as critical facilitators in integrating distributed flexible energy resources to support the transition toward sustainable energy. This is validated through a comprehensive quantitative economic analysis and a performance comparison against traditional methods relying on mixed-integer linear programming or basic heuristics. Furthermore, the use cases examined, and particularly the reward function formulated herein, are likely to be highly beneficial for subsequent research endeavors in this sector.

Milestones in the Development of Reinforcement Learning

In the following text, we first present an overview of deep learning, with a specific focus on reinforcement learning, followed by a brief survey of deep learning research within the energy sector. This progression helps to distinctly articulate the novel contributions our investigation brings to the prevailing body of literature.

Background: machine learning

The literal notion of “machine learning” is frequently linked to the 16th-century formulation of Bayes’ Theorem and encompasses various milestones, including the 1950s creation of artificial neural networks. Similarly, while the foundation of “deep learning” originated in the 1940s, the phrase only became widely recognized around 2006 [23]. It generally narrows the overarching field of machine learning to multi-layered neural networks trained on massive datasets. Following 2006, academic efforts concentrated on deploying deep learning models for image and audio categorization. Typically, these models were trained via “supervised learning” on specific datasets (like apparel items or handwritten digits). Such a method necessitates “labeled data” to provide the model with feedback on the accuracy of its predictions ([24]). By approximately 2011, the most advanced agents began to exceed human capabilities in various image classification benchmarks.

Subsequently, deep learning frameworks acquired the ability to master intricate, sequential games. A notable example is “AlphaGo,” an algorithm designed for the game of Go, which triumphed over top human competitors after being trained on extensive datasets of human matches2 [25]. Notwithstanding these triumphs, supervised learning encountered substantial constraints. The necessity for colossal training datasets confined deep learning applications to a narrow range of use cases—like Go—where such data was abundant. Moreover, training a model exclusively on pre-existing data inherently dictates that its performance cannot exceed the expertise of the individuals who originally produced that data.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Reinforcement learning

Within the framework of RL, an agent (or policy) is trained to take actions that maximize the output of a value function. Precisely, the agent participates in a Markov decision process: it gathers environmental context as an "observation," executes an "action," and subsequently receives a "reward," such as a signal that a specific objective has been achieved (**Figure 2**, Sutton and Barto [26]). The fundamental distinction between pure supervised learning and RL lies in RL's capacity to engage dynamically with its surroundings. This core characteristic permits the agent to independently probe an environment without relying on an expert dataset, subsequently "exploiting" this gathered experience to optimize its rewards [26].

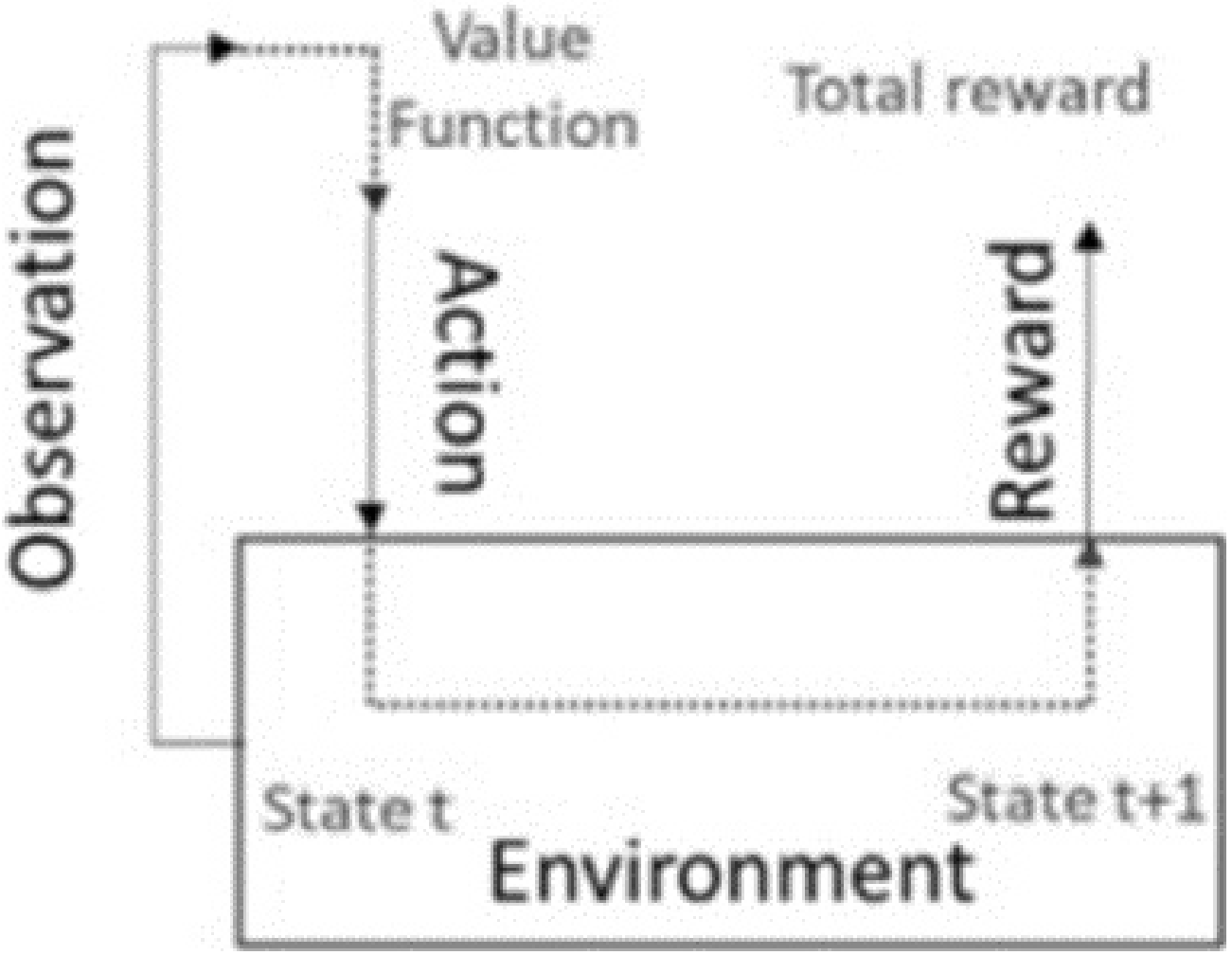


Figure 2. Visualization of one step in a Markov decision process.

The practical value of RL is effectively demonstrated through a prominent example. Barely two years following AlphaGO's supervised-learning-driven victory over human experts, DeepMind introduced its successor, AlphaZero [27]. This updated software relied entirely on RL, learning by playing matches against itself rather than using human-generated data. AlphaZero rapidly outperformed AlphaGO and demonstrated the capacity to master additional games, including shogi and chess. Armed solely with the fundamental rules of each game, AlphaZero eclipsed both top-tier human competitors and existing computational engines across multiple domains. During these matches, specialists noted that while AlphaZero independently reconstructed numerous traditional human tactics, it also formulated highly effective novel strategies previously unrecorded by professional human players [28].

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Initial RL models were restricted to processing discrete action and observation spaces (such as specific board coordinates and chess moves). The deep deterministic policy gradient (DDPG), introduced by Google DeepMind in 2015, marked the inaugural methodology capable of processing continuous action and observation spaces [29]. Nevertheless, DDPG is prone to overestimation, leading to instability. Consequently, it has been refined over time through the integration of critic networks that more accurately assess policy network actions, alongside other stabilizing mechanisms (most prominently, “Double Q-learning”, cf. e.g. [30]). Comprehensive evaluations have highlighted two particularly effective methodologies: the soft actor-critic (SAC [31]) and the twin-delayed deep deterministic policy gradient (TD3 [32]). Both are categorized as “off-policy” frameworks, meaning they populate a “buffer” database during their operation, utilizing this historical experience to determine subsequent actions. Conversely, “on-policy” algorithms continuously modify their policy—or “behavior”—in real time, discarding experience to progress solely with the newly refined policy. While on-policy frameworks generally achieve faster, more stable learning and require less hyperparameter tuning than their off-policy counterparts, this usually comes at the expense of sample efficiency. Proximal policy optimization (PPO, [33]) stands out as one of the most proficient on-policy algorithms available.

Ultimately, dominating board games is not RL’s primary objective; rather, these games act as preliminary testing grounds for real-world implementation. The constrained adoption of RL frameworks in practical settings likely stems from professional skepticism regarding the steep learning curve associated with this intricate subject, coupled with apprehension about navigating multiple implementation hurdles. To address this, our research enriches current literature by offering a methodology for deploying advanced algorithms that circumvents the need for deep neural network programming expertise. Furthermore, we provide an explicit, step-by-step illustration of how to identify and navigate significant structural challenges within a sophisticated use case.

Deep learning in the energy domain

According to Kraus *et al*. [8] and Huck [7], deep learning implementations have yet to reach their full potential, particularly within the energy sector. These authors highlight several areas requiring further investigation, including benchmarking requirements, feature selection methodologies, and the complexities of determining optimal control strategies to leverage energy flexibility within multi-energy vector infrastructures. Expanding upon this premise, Kathirgamanathan *et al*. [34] deliver an extensive evaluation of data-driven predictive control as a mechanism for unlocking the flexibility inherent in building heat management systems.

Lissa *et al*. [35] explored how Deep Reinforcement Learning (DRL) could be leveraged to maximize electrical self-consumption for domestic hot water and indoor heating. Their approach yielded energy savings between 8%–16% while preserving user comfort. For subsequent research, they advocate incorporating “other factors, such as dynamic prices or neighbourhood energy peak reduction,” as well as enabling the algorithm to facilitate electricity sales back to the main grid. In a comparable study, Kathirgamanathan *et al*. [36] utilized a Soft Actor Critic DRL model to lower electricity usage by harnessing heating flexibility within commercial office spaces. Additionally, Nakabi and Toivanen [37] conceptualized a micro-grid incorporating battery storage, wind energy generation, and flexible heating loads. While they evaluated the efficacy of various reinforcement learning models, they underscored the necessity for continued research into algorithms that prioritize consumer comfort parameters.

Zhang *et al*. [38] utilized reinforcement learning to develop algorithms capable of coordinating energy storage, consumption, and generation trading within localized energy markets. Their findings indicate a specific need for further research concerning the structural integration of battery storage. Furthermore, the local markets examined represent merely an intermediate model bridging centralized platforms and peer-to-peer trading, alongside the alternative strategy of managing flexible resources through a virtual aggregator.

The current study endeavors to expand upon prevailing literature by addressing several of these acknowledged academic voids. To achieve this, we incorporate battery mobility as a prominent, emerging vector for electrical consumption flexibility, complementing residential battery storage, which serves as a bidirectional flexibility source for both power

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

consumption and supply. The resulting flexibility is leveraged to optimize utility across five distinct operational scenarios, encompassing fluctuating grid pricing for both energy procurement and feed-in, thus operationalizing the innovative paradigm of a virtual aggregator.

Setting Up an RL Model for the Exemplary Application

Description of the exemplary application

The foundational premise of this deep RL investigation focuses on an aggregator managing distributed energy resources to optimize financial returns (calculated as total revenues less costs) on behalf of private residential clients, structured around the subsequent optimization problem.

$$\max\left\{\sum_{t\in T} P_{Feed,t} * p_{Sale} - P_{Draw,t} * \left(p_{EPEX,t} + p_{Taxes} - p_{Grid_relief}\right) - C_{Aging} + \left(R_{FCR} - C_{FCR}\right)\right\}. \quad (1)$$

Opportunities for supplementary financial gains stem from residential energy systems capable of flexible scheduling (specifically, domestic battery storage and on-site battery electric vehicle (BEV) charging infrastructure in this study). Out of the multitude of viable use cases theorized for generating secondary economic value, we isolated five distinct scenarios to represent a broad spectrum of utility adequately. Each selected scenario targets a unique variable within the profit maximization formula detailed in **Eq. 1** (this operational environment was configured to parallel the linear optimization framework established by Specht and Madlener [20] to ensure comparability). These five scenarios are elaborated upon below:

(i) The cost of electricity for a residential consumer can be conceptualized as a blend of an invariable base rate, p_{Taxes} (26 €-ct/kWh, encompassing grid charges, taxes, and prospective deductions detailed in (ii)), alongside a variable element, $p_{EPEX,t}$, which reflects market volatility derived from actual German electricity exchange rates. Although current end-users generally adhere to static tariff structures, energy providers must fulfill this consumption demand against fluctuating wholesale prices. Consequently, providers could extract significant value by reallocating flexible energy loads to periods of reduced market pricing, thereby lowering aggregate procurement expenses, $P_{Draw,t} * p_{EPEX,t}$.

(ii) As outlined previously, the proliferation of transitional energy technologies, notably heat pumps and electric vehicles, threatens to generate acute demand spikes, compelling grid managers to undertake costly infrastructural upgrades. As an alternative, grid administrators could incentivize consumers (or their representative aggregators) to curtail energy usage during critical peak intervals. An early legislative effort to mitigate such demand spikes is formalized in Article 14a of the German Energy Economy Act (Energiewirtschaftsgesetz, EnWG). In accordance with this legislation,3 we incorporated a functional mechanism into our model that linearly diminishes the static cost variable, p_Taxes in **Eq. 1**, by a maximum of 5 €-ct/kWh, concurrent with a restriction of the permissible peak load ceiling for grid electricity exchange from 22 kWpeak down to 5 kWpeak.

(iii) Daily, the managing aggregator has the discretion to allocate a segment of the available battery capacity for grid stabilization purposes, achieved by absorbing surplus power or injecting necessary energy during deficit periods. We operate under the assumption that the aggregator receives compensation from the grid operator, denoted as R_{FCR}, for maintaining this frequency containment reserve, estimated at roughly 0.15 €/kW per daily contract.4 However, dedicating this capacity precludes its utilization for alternative functions, thereby generating opportunity costs. Simultaneously, it incurs supplementary degradation expenses, R_{FCR}, because the perpetual micro-adjustments required to balance grid frequency necessitate the continuous cycling of minor electrical loads.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

(iv) Modern lithium-ion battery architectures generally experience accelerated degradation when pushed to extreme high or low states of charge (SOCs), translating into financial depreciation,C_{Aging} . The mathematical curve of this degradation profile is fundamentally U-shaped. Comprehensive specifics regarding the applied aging function are available in Specht and Madlener [20], while Ecker *et al*. [39] offer an in-depth technical analysis of these dynamics. Strategic, optimized scheduling can mitigate exposure to these detrimental boundary states (e.g., by preventing the total depletion of the storage unit during winter days, maintaining a baseline capacity buffer instead).

(v) The financial compensation, p_sale, for exporting locally generated PV energy to the broader grid stands at 10 €-ct/kWh. This rate is substantially lower than the equivalent cost of importing electricity from the grid ($p_{EPEX,t}$ + p_{Taxes}, approximately 30 €-ct/kWh in the German market before the recent pricing anomalies triggered by the conflict in Ukraine). Consequently, the AI agent is incentivized to optimize self-consumption—effectively curtailing P_{Feed} during generation periods to subsequently offset $P_{Draw,t}$. This is achieved by buffering excess energy in storage systems for use during non-generation windows, and by preferentially routing surplus PV power to charge a BEV whenever operationally viable.

Modeling the exemplary environment

We developed a simulated environment utilizing a combination of empirical data (such as residential PV generation and electricity consumption profiles) and artificially generated data (including usage patterns for electric vehicles). These datasets, alongside the operational scenarios, were integrated into a framework encompassing both economic and technical dimensions **(Figure 3)**. The primary objective was the minimization of annual electricity costs—or the maximization of net profit, should revenues surpass expenses—across a one-year timeframe (divided into 35,040 intervals of 15 minutes each). Pertinent operational limits were incorporated, such as charging capacity ceilings and the requirement for non-negative states of charge (SOC) for both vehicles and battery systems.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Environment:

Technical dimension:

1. Actions suggested by the agent are clipped to technically possible values

2. The total electricity balance around the household is calculated as:
   - \+ PV production
   - \+ Battery discharge (adjusted for efficiency losses)
   - – Regular electricity consumption
   - – Battery charging (adjusted for efficiency losses)
   - – BEV charging (charging losses were already accounted for in the WLTP data used)
   - – Efficiency losses caused by the battery operation for the provision of reserve power

   Σ Total exchange with the grid

Economic dimension:

1. Monetary streams are calculated as:
   - \+ Feed-in remuneration for PV production
   - \+ Remuneration for the provision of frequency containment reserve[1]
   - – Cost of electricity purchase from the grid as the product of the amount withdrawn times the sum of:
     - the volatile electricity exchange price[2]
     - a fixed price component including grid fees, taxes, levies, etc.[3]
     - a discount (of the grid fee) of up to 5 ct/kWh if committed to intense peak shaving (or a linear increasing penalty if the exchange with the grid exceeds the agreed limit)
   - – Battery depreciation due to aging (for storage and due to reserve service)

   Σ Total monetary profit/cost

[1] An increasing commitment leads to higher revenues for this service (we assumed a usual market price of €1000/MW per week, equivalent to about €0.3 per kWh blocked for one day)
[2] Real historic price data from the EEX electricity exchange was used
[3] For Germany, fixed price components of 26 EUR-ct/kWh were assumed
[4] In Germany, flexible loads can get a reduction in grid fees of up to 5 EUR-ct/kWh. This discount on the grid fees was distributed linearly between a maximum load of 22 kW and the maximal commitment to load reduction to a minimum of 5 kW.
[5] Typically, lithium-ion batteries age faster when operated at very high or low states of charge, resulting in a "U"-shaped aging function. Integrating this U-shaped function results in a function similar to the one depicted to the right. Based on this integrated function, costs for a (dis-) charging action can be calculated as the vertical difference between two SOCs.



2. Virtual rewards/penalties:
   - \+ Final disbursement: At the end of the game the value of the electricity stored in both the stationary battery and BEV is paid out to the agent (to prevent a deliberate draining of all storage units towards the end of the game)
   - – Penalty for attempts to violate technical restrictions of assets
   - – An exponentially increasing penalty for missing power in the BEV (resulting in an increasing pressure to charge the BEV the lower the battery is)
   - – A slight penalty for electricity consumption from the grid to motivate the maximization of self-consumption

   Σ Total virtual rewards/penalties

Reward function:

The sum of the monetary as well as the virtual profits and costs is considered for the rewards of one simulation step that will later on be passed back to train the agents.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 3. Comprehensive overview of the simulated environment, illustrating the technical variables driving the energy balance, alongside the financial and virtual penalties/rewards that collectively determine the agent's total remuneration.

The action space dictated four distinct responsibilities for the trained agent: (a) establishing a weekly maximum threshold for grid power consumption (cf. use case (ii)); (b) designating a daily power reserve capacity for grid stabilization (cf. use case (iii)); (c) regulating the 15-minute intervals of power charged to or discharged from the domestic battery; and (d) dictating the 15-minute charging volume allocated to the BEV.

To formulate these decisions, the agent receives environmental state observations. This defined observation space comprises eight distinct metrics: (1) the prevailing electricity rate; (2) net electricity (on-site generation minus usage); (3&4) individual state of charge metrics for the battery storage unit and the electric vehicle; (5&6) indicators detailing the active long-term operational scenarios; and (7&8) temporal markers indicating the specific time and date (**Figure 4**). The continuous variables within both the observation and action spaces were normalized to ranges of [0,1] and [−1,1], respectively.

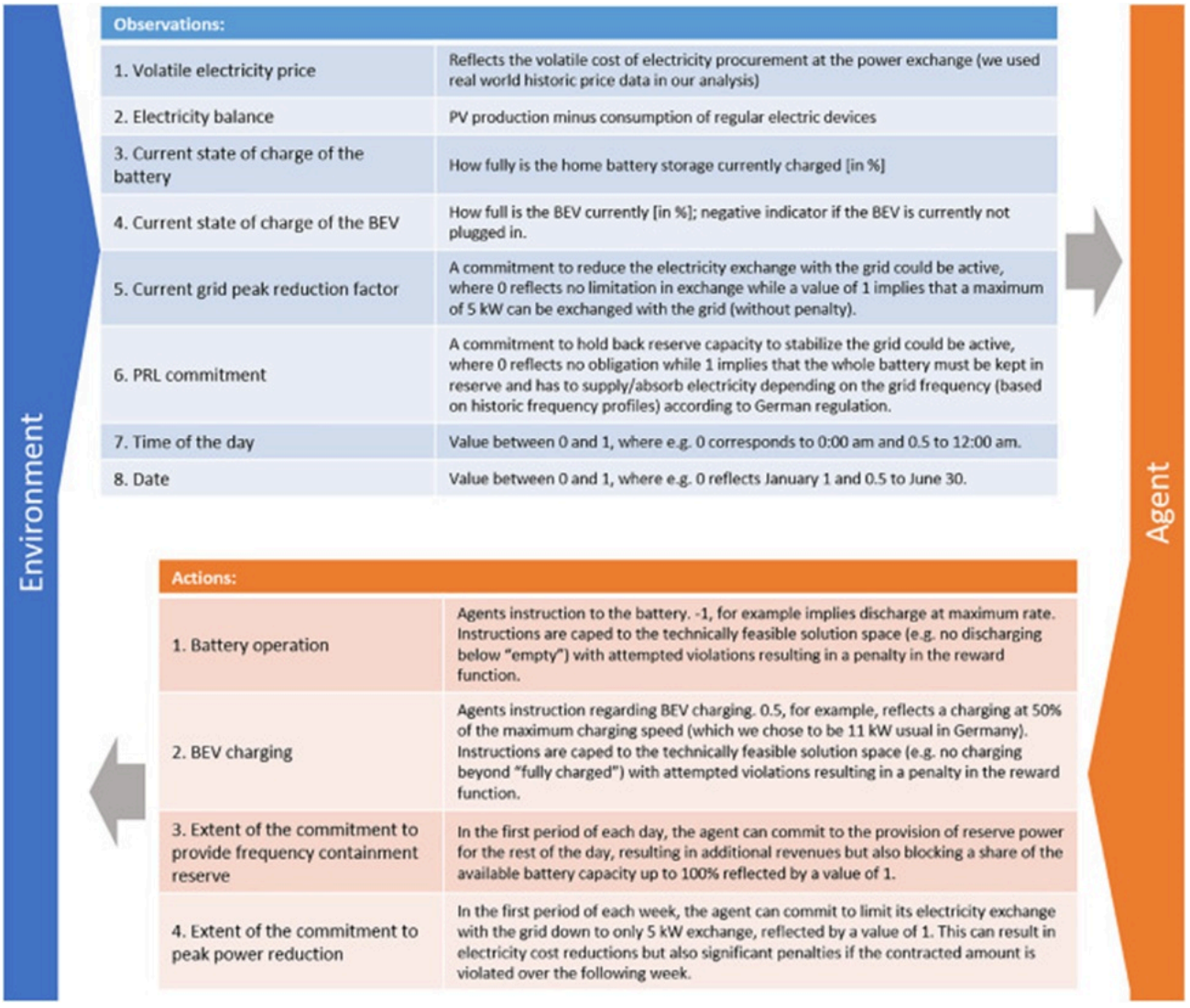


Figure 4. In-depth outline of the eight environmental state observations processed by the agent, alongside the four corresponding actions generated as output.

The reward architecture assimilated various practical requirements inherent to real-world deployments. Initially, standard financial revenues and expenses were integrated. Grid electricity procurement incurred costs calculated by multiplying the

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

volume consumed during a 15-minute block by the fluctuating market rate, [5] whereas exporting surplus PV power yielded a static financial return. Subsequently, asset depreciation was evaluated: we integrated a battery degradation function modeled after Ecker *et al*. [39], which specifically penalized operation at extreme (both low and high) states of charge. Furthermore, the ensuing supplementary revenues and expenses can be achieved based on the specific use cases evaluated:

(i) Daily, the algorithm could dedicate a portion of battery storage to supply frequency containment reserves. Although this generated supplementary revenue per committed kWh, it simultaneously necessitated constraints on standard battery operation, enforcing financial penalties if the agent violated these predetermined limitations. Furthermore, the continuous micro-cycling of electricity required for frequency stabilization accelerates battery degradation, incurring associated aging costs.

(ii) Weekly, the agent possessed the option to cap its maximum grid power exchange to mitigate peak demand. While securing a lump-sum financial reward, any breach of this self-established threshold within a 15-minute window triggered a proportional financial penalty per excess kW.

Finally, “virtual” financial metrics (profits and costs) were integrated. Primarily, these functioned as penalties for constraint violations (such as trying to draw grid power beyond established ceilings or attempting to discharge an already depleted battery). Moreover, to mirror the end-user’s desire for an adequately charged vehicle, an escalating penalty was applied for low BEV SOCs. The aggregation of all these financial and virtual components constituted the ultimate reward signal transmitted to the agent.

To conclude, we constructed a standard heuristic algorithm, depicted via pseudocode in **Figure 5**. This heuristic was isolated from the other advanced use cases, establishing a baseline to benchmark the AI’s optimized, multi-use-case operation against contemporary, conventional management strategies.

```
Pseudocode of the heuristic used:
For each separate period:
    Subtract regular electricity demand from PV-production
    If excess electricity is left over:
        Try to charge the BEV
        If the BEV is "not at home", "full" or the max. charging speed is reached:
                Try to store remaining power in the home battery storage
                If home storage is "full" or at max. charging speed:
                        Feed excess electricity to the grid
    If electricity demand is higher than local production:
        Try to cover difference with the battery
        If the battery is "empty" or at max. discharge rate:
                Use electricity from the grid
```

Figure 5. Pseudocode detailing the functionality of the baseline heuristic algorithm, mirroring the current operational standard for most residential systems.

General challenges and solutions

During the development of our framework, we navigated nine primary hurdles frequently documented in existing research ([37, 40, 41]). The following subsections outline these impediments for prospective practitioners and detail the resolutions we implemented.

The need for a training model

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

An initial, apparent hurdle is the industry consensus that training RL models within live physical systems is excessively risky, costly, or unacceptable. This is particularly relevant when engineering an electricity flexibility aggregator, as erratic explorative actions could severely strain the power grid, inflict substantial financial damages on the aggregator, and disrupt consumer service. Consequently, we chose to formulate the scenario within a simulated environment, allowing for algorithmic pretraining before any physical deployment.

The definition of a reward function

RL agents require continuous evaluation through reward signals to refine their decision-making. Since live environmental feedback was unfeasible for the aforementioned safety reasons, we had to engineer a comprehensive reward function capable of scoring every potential agent action. While simply maximizing fiscal profit suffices in many scenarios, our specific framework necessitated the monetization of intangible variables, such as a user's preference for an immediately available, fully charged EV. Consequently, the total reward calculus encompasses tangible financial outcomes—like grid service revenues and energy trading costs—alongside subjective factors, such as vehicle unavailability, which were mathematically translated into simulated cost/benefit metrics.

Complex features of the environment

Fundamentally, an agent's problem-solving efficacy correlates directly with its situational awareness. Regrettably, perfectly detailing real-world dynamics is impossible within our context—a limitation common to most practical applications. Specifically, our setup represents a partially observable environment, as it is impossible to articulate every variable driving a specific event (e.g., the myriad factors culminating in a specific market electricity price). Additionally, the system is inherently stochastic; state transitions are not solely dictated by the AI's choices (for instance, the agent cannot control a user's spontaneous decision to drive their vehicle, removing it from grid availability). Nevertheless, providing eight discrete data points per 15-minute state proved adequate for modern algorithms to grasp the fundamental environmental mechanics. Correspondingly, a constrained set of only four actions per state (such as regulating vehicle and battery charging) was sufficient for effective environmental interaction.

Unlimited options in continuous observation and action spaces

Mirroring many practical deployments, our system predominantly utilizes continuous, rather than discrete, observation and action variables, theoretically presenting an infinite array of choices. To illustrate, BEV charging could occur at any continuous variable between absolute zero and the 22 kW maximum ceiling. The capability for RL algorithms to optimize decisions across infinite possibilities within continuous spaces only materialized recently with the advent of advanced actor-critic methodologies, which inherently restricted the pool of viable algorithms for our research.

Long games with delayed rewards

The sheer volume of sequential steps per episode presents a formidable obstacle in long-term frameworks. Whereas Go and chess conclude in roughly 150 and 80 moves respectively [41], our simulated year comprised over 35,000 discrete 15-minute intervals. Condensing this timeline is counterproductive due to conflicting objectives: the model must possess the high-resolution granularity required to capture fluctuating market prices and energy generation, while simultaneously encompassing a broad enough timeframe to learn annual seasonal variations. To mitigate this, we avoided relying exclusively on delayed, end-of-episode rewards and instead supplied a continuous cadence of feedback every 15 minutes.

The need to define boundaries and constraints

Because reinforcement learning relies heavily on trial and error, the system must permit the agent to test constraints and receive negative reinforcement. At the same time, the underlying simulation prevents actual breaches of regulatory or physical laws. For instance, storage units cannot be charged beyond maximum capacity or depleted past zero. To manage this, we enforced strict environmental boundaries on the action space, immediately issuing algorithmic penalties whenever the agent attempted an impossible action. Handling this required bifurcating the feedback loop: one stream

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

managed the “real” economic calculations detailing monetary profits, while the other handled virtual rewards and penalties to guide the agent’s behavior.

The difficulty of defining a benchmark

Evaluating the proficiency of an AI’s strategy is complex; a seemingly poor financial outcome may stem from a procedurally generated, highly challenging scenario (such as a residence exhibiting both heavy EV usage and massive baseline energy consumption). To navigate this, we adopted the heuristic model previously outlined by Specht and Madlener [20] to serve as a comparative baseline. This benchmark executes foundational logic rules, closely resembling the standard control mechanisms presently governing residential energy systems.

Differences in the scale of values

The magnitude of the agent’s actions and observational inputs can fluctuate drastically. This creates training friction, as minor shifts in one action could disproportionately impact low-scale variables, inducing algorithmic instability. In alignment with standard machine learning practices, we standardized all inputs and outputs to normalized ranges of [−1,1] or [0,1] . To demonstrate, the SOC observation is scaled from 0 to 1, rendering a 50% charged battery as 0.5. Conversely, battery operational commands span from −1 (‘discharge at maximum rate’) to 1 (‘charge at maximum speed’).

Need for training data

A widely acknowledged barrier in current research is the scarcity of robust training data. Although virtual simulations permit infinite iterations, repeatedly exposing the agent to identical scenarios triggers “overfitting,” where the algorithm memorizes specific responses but fails to generalize to novel situations. To counteract this, we synthesized diverse data matrices during the training phase. Our framework incorporated 85 distinct residential consumption profiles (derived primarily from [42]), five normalized PV production profiles that were multiplied by different system sizes to get the actual electricity production, 10 standardized BEV usage patterns that were scaled by eight different mileage options (based on data from [43]), and 35,040 timesteps per profile per year, yielding billions of unique combinations. Notably, as detailed later, the efficiency of our tuned agents generally plateaued near the 500,000-step mark, after which further iterations induced behavioral instability and a decline in overall performance.

Practical Experiences: Training the Agent in the Exemplary Environment

Setting up the model environment and selecting an adequate algorithm

Preparing the training process involves two main steps: first, the environment must be implemented so that a reinforcement learning (RL) agent can interact with it. Second, an algorithm suitable for the defined constraints must be chosen to train the RL agent.

Our implementation uses a “gym” environment. Gym, developed by OpenAI, is a collection of test problems (usually small games) that share a standardized interface [13]. Gym allows new RL algorithms to be trained on the same tasks, simplifying result comparisons. Conversely, one can also design a new “environment” according to the standardized interface, enabling multiple RL algorithms to be evaluated and compared without major changes to the environment.

Manually implementing a complex algorithm typically requires weeks to study the method in the literature and then implement it using specialized libraries such as PyTorch or Keras. Fortunately, several projects provide pre-built, customizable RL algorithm frameworks. We chose “Stable Baselines” (https://stable-baselines.readthedocs.io), a fork of OpenAI Baselines, which offers about 14 state-of-the-art algorithms with numerous customization options.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

By combining a standardized environment with RL algorithm templates, it is possible to define a machine learning task and train an agent using basic programming skills (e.g., Python) without requiring in-depth AI programming knowledge. Once the task and suitable training algorithms are chosen, a specific RL algorithm must be selected.

Selecting an adequate algorithm for the specific task

Next, a specific algorithm was selected to train agents in the exemplary environment, aiming to minimize operation costs for randomly generated households. A large number of RL algorithms exist in recent literature, many with numerous modifications. Systematically listing the task requirements, helps narrow down the suitable options. For example, feature-limit considerations (iii) allowed us to exclude “model-based” RL algorithms, since they rely heavily on a complete and accurate ground-truth model, which is often considered “fundamentally hard” and unstable [44], even in well-defined technical contexts.

Moreover, the continuous action and observation spaces in our environment further restricted the choice to a few recent algorithms. **Figure 6** compares three promising candidates—SAC, TD3, and PPO2—in their “vanilla” form (i.e., before hyperparameter tuning or modifications).

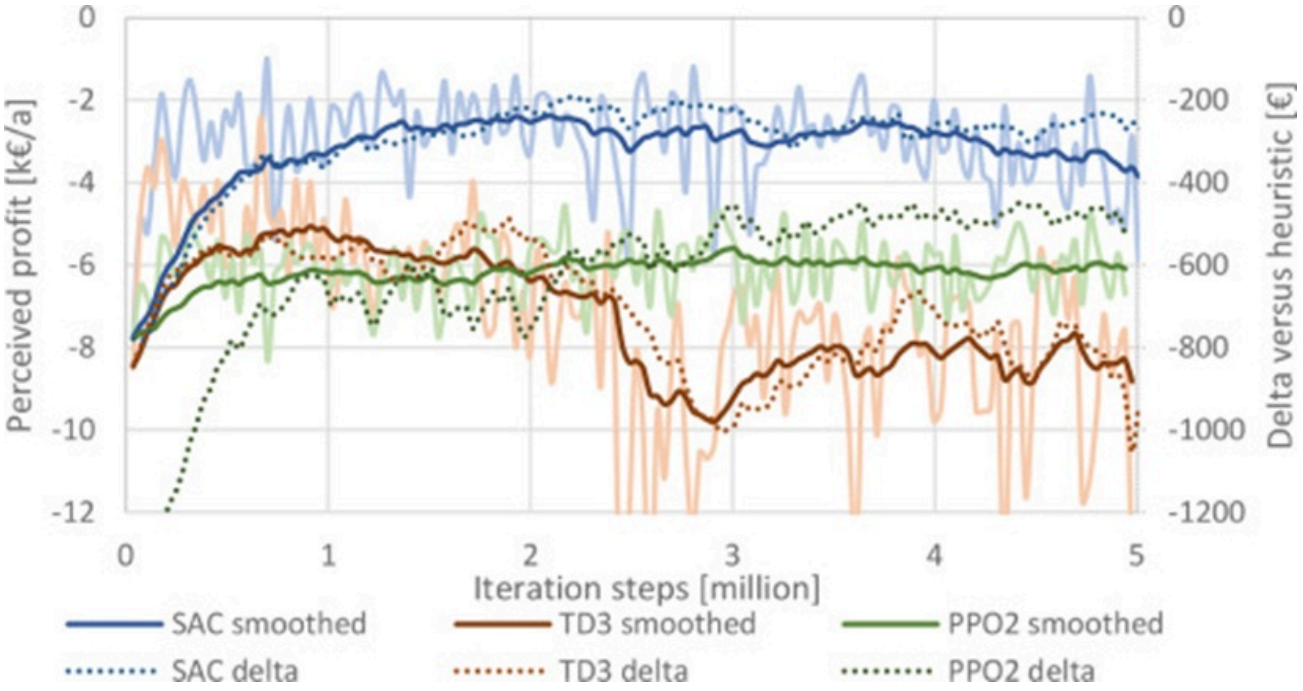


Figure 6. Comparison of selected RL algorithms’ performance in terms of perceived profits (**Figure 3**, ‘Economic dimension’). Continuous lines depict SAC, TD3, and PPO2 over up to 5 million training iterations in the exemplary environment , with darker lines showing smoothed versions. Dotted lines indicate each algorithm’s loss relative to the heuristic on a secondary axis. Vanilla SAC performed best and was most stable among the RL algorithms.

Light-colored lines show smoothed results for specific settings, while dotted lines display differences from the heuristic, demonstrating that all vanilla algorithms underperform relative to the heuristic. Comparing the RL algorithms shows SAC as the most promising. **Figure 6** reveals that both vanilla SAC and PPO2 reach peak performance after approximately two million timesteps (equivalent to simulating 57 households at 15-min resolution for a full year) before plateauing. The perceived profit per year reaches only €−2000 for SAC. Perceived profits account for expected “real” revenues minus household electricity costs and expected “virtual” penalties for restriction violations. Vanilla TD3 peaks at roughly 5000 €/a after 1 million steps, but maintaining this is unstable. PPO2 results in the highest annual costs of around €6000.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

PPO2 also exhibited the fastest computation time, completing 5 million training steps in 1.5 h, whereas SAC and TD3 required 4.75 h and 3.75 h, respectively, on a desktop PC (CPU: i5–7500, 12 GB RAM).

These initial results suggest SAC is promising for further analysis. However, comparison with the heuristic shows consistent underperformance for all three vanilla algorithms, with losses up to €200 per year, indicating a strong need for performance improvements through hyperparameter tuning.

Over several months, we attempted to improve the model setup. Automated hyperparameter tuning software (Optuna) provided reasonable setups, but suggested parameters were often difficult to interpret and worked only in specific environments. For instance, one top automated architecture featured [16,8,4,1,4] neurons for both actor and critic networks. Although it performed well on training samples despite its simplicity, the single-neuron bottleneck led to significant information loss, preventing further learning. Consequently, we manually optimized the architecture through trial and error. A network of [128, 64, 32, 16, 4] using an LnMlp-Policy yielded promising performance and sufficient complexity to develop advanced strategies. Initial training involved 1,000,000 iterations, a learning rate of 0.0003, gamma of 0.999, buffer size of 200,000, and 40,000–150,000 random steps to fill the buffer before actual learning. Even with this setup, most instances performed poorly, but a few approximated the heuristic's performance. These promising instances were saved and further trained until general performance plateaued. Subsequent improvements required specialized training on a specific household instead of random samples, allowing the agent to familiarize itself with an individual customer, which improved performance further.

Results and Discussion

Observations during the training process

Consistent with other studies [45], RL agents exploited any weaknesses in the environment. While fewer issues were expected, given prior use of the environment with linear optimization, the RL agent identified loopholes. For example, although the battery SOC was set to a minimum of 0 at each round start, the agent learned to discharge below 0 virtually, sell the nonexistent energy, accept the penalty, and reset to 0 in the next round.

When trained with SAC and tuned hyperparameters, a single decision for the four tasks (a–d) took approximately 0.0003 s (or 3 s for 10,000 decisions) on a personal computer (CPU: i5–7500, 12 GB RAM). This indicates RL agents can control many households in near real-time, satisfying optimization time requirements.

Economic findings in the chosen application

To evaluate optimization quality, we ran sequences of 1000 simulation steps for households with different BEV usage profiles. With all use cases activated, a pretrained general RL agent increased average annual profit by €126 compared to the conventional heuristic, as shown in **Figure 7**. Certain settings yielded additional profits up to €315 per year. At the same time, some households experienced higher costs relative to the heuristic.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

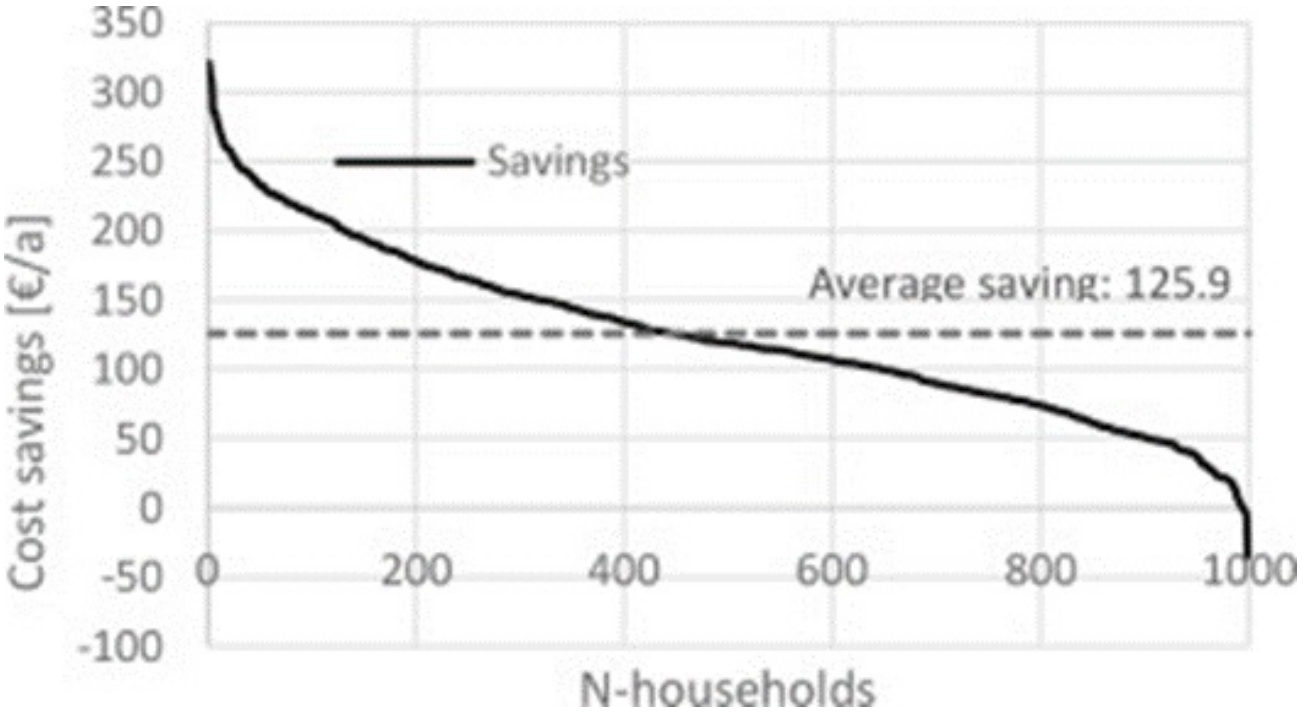


Figure 7. Annual cost savings for 1000 households in the exemplary environment operated by the trained SAC-based agent. Households are sorted by achieved savings. Most random households realized up to €315 in annual savings, with an average of €125.9 per year.

One household, selected based on a VDI norm [46] as the "most typical," achieved annual savings of approximately €150. Pretraining the general agent on this household resulted in additional savings of €28–178 annually, demonstrating the potential benefit of customer-specific training. Notably, these savings compare to €235 achieved via linear optimization in Specht and Madlener [20] in a nearly identical setting, despite RL having only current knowledge and limited experience, while linear optimization used perfect foresight. Additionally, RL completed 35,040 steps in about 10 s, whereas linear optimization required roughly 10 h.

Comparison of the behavior of the RL agent and the heuristic

Comparing the actions of the RL agent with those of the heuristic provides insights into their characteristic operational behavior. **Figure 8** shows the sum of local generation and regular household consumption (lighting, cooking, etc.), the net exchange with the grid (both in kW), and the SOC of the battery and BEV (in kWh) for the same day around the equinox for both a generally trained RL agent and the heuristic. The total of local generation and ordinary consumption (yellow) is identical in both cases, as neither algorithm can alter PV output or household consumption. The PV system begins generating electricity around 7 a.m., rapidly exceeding consumption and resulting in positive net power. The RL agent left the BEV (green) at approximately 35 kWh SOC from the previous day and primarily used the morning surplus to charge the vehicle until its departure at 10 a.m., after which it was unavailable for charging. Simultaneously, the battery (blue) charges slowly until the BEV departs, then continues charging steadily to reach 6.2 kWh at 5 p.m. At this time, PV generation no longer meets demand, prompting the agent to discharge the battery gradually. When the BEV returns at 7:30 p.m., the RL agent observes that its SOC is about 6 kWh lower than at departure. The agent replenishes the BEV by charging slowly and waits for PV generation the following day.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

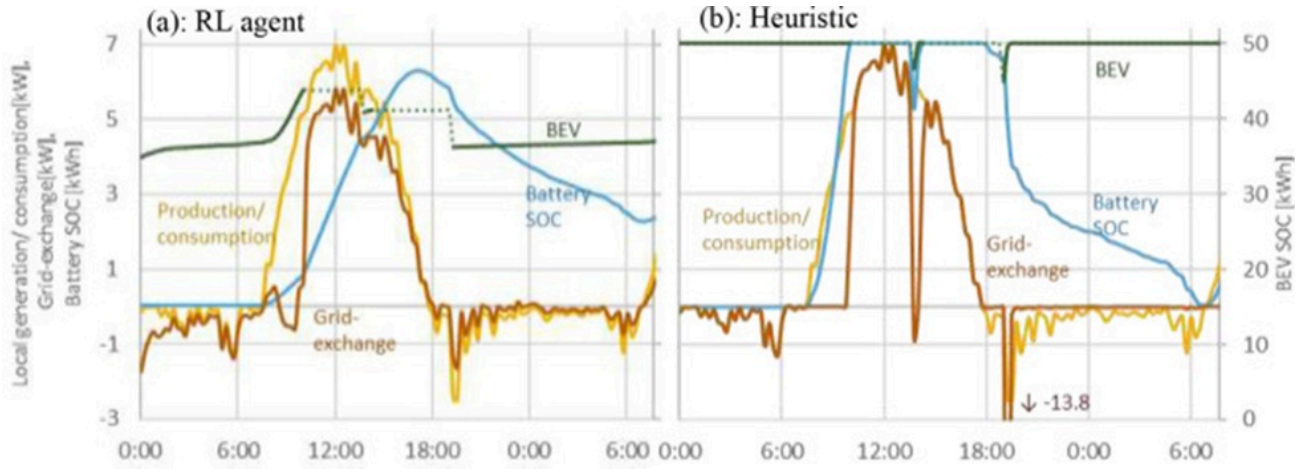


Figure 8. Operation of BEV and battery by a generally trained agent (a) and the heuristic (b) for a day in March.

The heuristic immediately charges the vehicle upon return if not full, missing the opportunity to exploit the early-morning PV surplus. Instead, all excess electricity is fed into the battery, resulting in rapid charging that fully charges the battery by 10:15 a.m., with remaining surplus sent to the grid. Unlike the gradual charging strategy of the RL agent, this behavior produces distinct load peaks. During undersupply, the battery compensates for the deficit. When the BEV returns, both battery and vehicle recharge at maximum speed, drawing significant power from the grid, with a peak of −13.8 kW.

Such peaks could severely stress the grid if implemented at scale. **Figure 9a** illustrates annual positive and negative electricity exchanges with the grid, showing that the heuristic creates demand peaks up to −22 kW (lower blue graph), potentially requiring substantial grid upgrades. In contrast, incentives for reducing load peaks encourage the RL agent to avoid such behavior (lower orange curve). **Figure 9b** demonstrates that the agent gradually limits grid exchanges throughout the year, committing to approximately 5 kW from September onwards. Comparing **Figures 9a and 9b** reveals that the agent occasionally exceeds this commitment slightly, likely because penalties for violations increase exponentially, rendering small infractions “acceptable.”

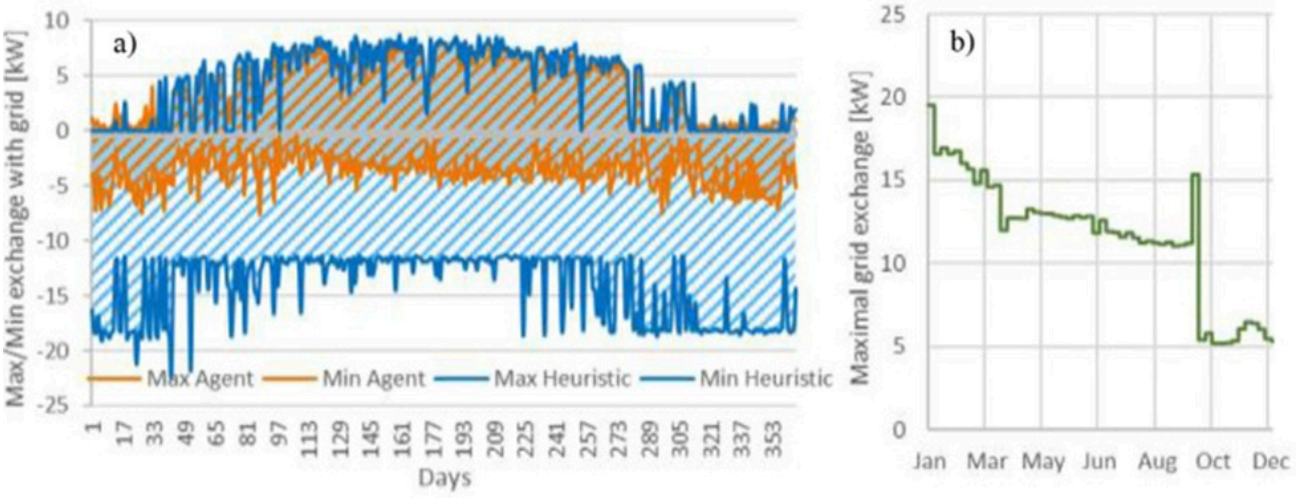


Figure 9. (a) The heuristic (blue line) shows slightly higher production but much larger consumption peaks than the RL agent (orange line). (b) The RL agent limits weekly electricity drawn from the grid (green line) to maximize revenues.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Pros and cons of the dispatch of an RL agent in the given application

Detailed analysis of the RL agent's operations shows that most savings stem from two use cases: peak consumption reduction and reserve power provision. These cases generally require optimization and were not implemented in the heuristic. To ensure a fair comparison, profits from these use cases were set to zero for the heuristic, in which case it slightly outperformed the RL agent in most settings. This advantage arises because the heuristic precisely stores excess electricity in the battery. At the same time, the agent occasionally overcharges (incurring high-cost grid electricity) or undercharges (requiring replacement from the grid at higher cost).

Conversely, the RL agent outperformed the heuristic in scenarios with high PV generation and significant BEV mileage, where careful charging improved self-consumption **(Figure 8)**. Moreover, only the RL agent effectively reduced load peaks, alleviating grid stress and providing reserve power. Overall, performance is roughly balanced for incumbent flexibility use cases; however, upcoming services, such as ancillary grid support, cannot be addressed adequately by conventional heuristics, highlighting the need for innovative solutions like RL.

Discussion of the introduced application for energy management

Although the trained RL agent is not yet perfect, it demonstrates substantial proof-of-concept value. The precise impact of individual use cases requires further evaluation and depends on national regulations. Nonetheless, RL agents appear capable of adapting to variations in regulation. They can robustly respond to changes in (i) household energy demand, (ii) consumption patterns (e.g., individual vehicle usage), and (iii) specific assets with high peak loads. Agent performance improves if it is briefly pretrained on historical data to familiarize itself with customer-specific characteristics. Even on a standard desktop PC, the decision-making speed is sufficient for thousands of schedule updates in real time. Additionally, libraries such as Stable Baselines simplify the training process for non-experts [47].

However, creating a customized setup still demands significant effort. While initial proof-of-life experiments are rapid, achieving reliable performance in complex setups requires extensive model shaping, loophole resolution, hyperparameter tuning, and training. Even under identical settings and after extensive training, only a small fraction of agents (approximately 1 in 25) develop a good understanding of patterns, and very few (about 1 in 250) excel at their task; many agents struggle with local optima, implying that luck plays a role in training. Deep learning agents function as black boxes, so their reasoning is not fully interpretable. Nevertheless, general patterns can be observed in results, providing a rough understanding of agent behavior.

Overall, this study confirms the considerable potential of deep learning, specifically RL. Findings suggest RL is a promising approach for complex optimization problems, particularly when rapid computation outweighs the need for perfect solutions.

Conclusion

This paper explored the potential of deep reinforcement learning (RL) to optimize the operation of numerous distributed flexible energy assets in private households. Additionally, it provides guidance for practitioners on approaching complex control tasks using state-of-the-art deep learning algorithms, common challenges, and possible solutions.

The main operational research insights are threefold: first, deep neural networks trained with RL can optimize schedules for large numbers of assets in near real time. Second, recent RL advancements and freely available AI tools benefit non-experts, allowing users with limited deep learning knowledge to employ customizable algorithms with basic programming skills (e.g., Python). Third, common obstacles were identified, and a procedure for selecting feasible algorithms for

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

specific challenges was proposed. Extensive hyperparameter tuning and model adjustments remain essential to outperform conventional algorithms, confirming concerns noted by practitioners [48-51].

The selected application revealed four key insights relevant to sustainable energy transition research. First, AI agents are promising for controlling numerous flexible assets across multiple value streams, handling nonlinearities, and transforming the “curse of dimensionality” into a “blessing of training options.” Second, a general agent trained on non-specific data can adapt to unfamiliar households, but pretraining on household-specific data significantly enhances performance. Hence, customer energy consumption data is increasingly valuable for energy companies. Third, prior studies indicated that profitable operation of flexible home assets requires perfect foresight and extensive linear optimization. This study demonstrates that RL can capture a substantial share of this potential in realistic settings without perfect foresight, while remaining computationally efficient. Fourth, additional annual gains of approximately €150 compared to conventional algorithms show promising market potential and address energy transition challenges, including heat and mobility electrification. Detailed agent behavior analysis indicates room for technical improvements and higher revenue potential, particularly in scenarios with increased price volatility.

Overall, this study demonstrates the significant potential of deep learning in operations research, especially in the energy sector, which is likely to increase as algorithms and computational power improve. Future research may reveal further application areas where deep learning and RL can supplant traditional methods. Developing customized training environments remains challenging, and additional literature providing practical guidance for practitioners is needed to facilitate access to this technology across disciplines.

Acknowledgements

None

Conflict of interest

None

Financial support

The authors gratefully acknowledge funding by the German Federal Ministry of Education and Research (BMBF) obtained for the Kopernikus Project “ENSURE” (funding nos. 03SFK1HO and 03SFK1C0-2) as well as helpful comments received from two anonymous reviewers.

Ethics statement

None

Received: 12 Dec 2023 Revised: 17 Feb 2024 Accepted: 05 Mar 2024
Published online: 15 November 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
2. Pierson HA, Gashler MS. Deep learning in robotics: a review of recent research. Adv Robot. 2017;31(16):821-35.
3. Wang J, Ma Y, Zhang L, Gao RX, Wu D. Deep learning for smart manufacturing: methods and applications. J Manuf Syst. 2018;48:144-56.
4. Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artif Intell Rev. 2022;55(2):895-943.
5. Gosavi A. Reinforcement learning for long-run average cost. Eur J Oper Res. 2004;155(3):654-74.
6. Ning C, You F. Optimization under uncertainty in the era of big data and deep learning: when machine learning meets mathematical programming. Comput Chem Eng. 2019;125:434-48.
7. Huck N. Large data sets and machine learning: applications to statistical arbitrage. Eur J Oper Res. 2019;278(1):330-42.
8. Kraus M, Feuerriegel S, Oztekin A. Deep learning in business analytics and operations research: models, applications and managerial implications. Eur J Oper Res. 2020;281(3):628-41.
9. Ransbotham S, Kiron D, Gerbert P, Reeves M. Reshaping business with artificial intelligence: closing the gap between ambition and action. MIT Sloan Manag Rev. 2017;59(1).
10. Henry R, Ernst D. Gym-ANM: reinforcement learning environments for active network management tasks in electricity distribution systems. Energy AI. 2021;5:100092.
11. Madlener R. Smart grid economics. In: Dinther C, Flath CM, Madlener R, editors. Smart grid economics and management. 1st ed. Berlin, New York, Heidelberg: Springer; 2022. p. 21-59.
12. Cruz MRM, Fitiwi DZ, Santos SF, Catalão JPS. A comprehensive survey of flexibility options for supporting the low-carbon energy future. Renew Sustain Energy Rev. 2018;97:338-53.
13. Deilami S, Masoum AS, Moses PS, Masoum MAS. Real-time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. IEEE Trans Smart Grid. 2011;2(3):456-67.
14. Greenwood DM, Lim KY, Patsios C, Lyons PF, Lim YS, Taylor PC. Frequency response services designed for energy storage. Appl Energy. 2017;203:115-27.
15. Hao H, Wu D, Lian J, Yang T. Optimal coordination of building loads and energy storage for power grid and end user services. IEEE Trans Smart Grid. 2018;9(5):4335-45.
16. Nykamp S, Molderink A, Bakker V, Toersche HA, Hurink JL, Smit GJM. Integration of heat pumps in distribution grids: economic motivation for grid control. In: Proceedings of the IEEE Power and Energy Society (PES) Innovative Smart Grid Technologies (ISGT) Europe Conference; 2012 Oct 14-17; Berlin. p. 1-8.
17. Ruester S, Pérez-Arriaga I, Schwenen S, Batlle C, Glachant JM. From distribution networks to smart distribution systems: rethinking the regulation of European electricity DSOs: final report. 2013.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

18. Bell K, Gill S. Delivering a highly distributed electricity system: technical, regulatory and policy challenges. Energy Policy. 2018;113:765-77.

19. Specht JM, Madlener R. Energy supplier 2.0: a conceptual business model for energy suppliers aggregating flexible distributed assets and policy issues raised. Energy Policy. 2019;135:110911.

20. Specht M, Madlener R. Quantifying value pools for distributed flexible energy assets. Energy. 2022;263(Pt B):125626.

21. Diamantoulakis PD, Kapinas VM, Karagiannidis GK. Big data analytics for dynamic energy management in smart grids. Big Data Res. 2015;2(3):94-101.

22. Tu C, He X, Shuai Z, Jiang F. Big data issues in smart grid – a review. Renew Sustain Energy Rev. 2017;79:1099-107.

23. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge (MA), London: MIT Press; 2016. 785 p.

24. Li Y. Deep reinforcement learning [Internet]. 2018 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1810.06339v1

25. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484-9.

26. Sutton RS, Barto A. Reinforcement learning: an introduction. Cambridge (MA), London: The MIT Press; 2018. 526 p.

27. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm [Internet]. 2017 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1712.01815v1

28. Sadler M, Regan N. Game changer. New in Chess. 2019.

29. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning [Internet]. 2019 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1509.02971v6

30. Dankwa S, Zheng W. Twin-delayed DDPG. In: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing (ICVISP 2019); 2019 Aug 26-28; Vancouver, BC, Canada. New York (NY): ACM; 2019. p. 1-5.

31. T. Haarnoja, A. Zhou, P. Abbeel, S. Levine. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. http://arxiv.org/pdf/1801.01290v2 .Google Scholar

32. Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods [Internet]. 2018 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1802.09477v3

33. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms [Internet]. 2017 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1707.06347v2

34. Kathirgamanathan A, De Rosa M, Mangina E, Finn DP. Data-driven predictive control for unlocking building energy flexibility: a review. Renew Sustain Energy Rev. 2021;135:110120.

35. Lissa P, Deane C, Schukat M, Seri F, Keane M, Barrett E. Deep reinforcement learning for home energy management system control. Energy AI. 2021;3:100043.

36. Kathirgamanathan A, Mangina E, Finn DP. Development of a soft actor critic deep reinforcement learning approach for harnessing energy flexibility in a large office building. Energy AI. 2021;5:100101.

37. Nakabi TA, Toivanen P. Deep reinforcement learning for energy management in a microgrid with flexible demand. Sustain Energy Grids Netw. 2021;25:100413.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

38. Zhang S, May D, Gül M, Musilek P. Reinforcement learning-driven local transactive energy market for distributed energy resources. Energy AI. 2022;8:100150.

39. Ecker M, Nieto N, Käbitz S, Schmalstieg J, Blanke H, Warnecke A, et al. Calendar and cycle life study of Li(NiMnCo)O2-based 18650 lithium-ion batteries. J Power Sources. 2014;248:839-51.

40. Dulac-Arnold G, Mankowitz D, Hester T. Challenges of real-world reinforcement learning [Internet]. 2019 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1904.12901v1

41. OpenAI, Berner C, Brockman G, Chan B, Cheung V, Dębiak P, et al. Dota 2 with large scale deep reinforcement learning [Internet]. 2019 [cited 2023 Jan 15]. Available from: http://arxiv.org/pdf/1912.06680v1

42. Tjaden T, Weniger J, Quaschning V. Repräsentative elektrische Lastprofile für Wohngebäude in Deutschland auf 1-sekündiger Datenbasis. Berlin: HTW Berlin; 2015 [cited 2022 Oct 2]. Available from: https://solar.htw-berlin.de/wp-content/uploads/HTW-Repraesentative-elektrische-Lastprofile-fuer-Wohngebaeude.pdf

43. Ecke L, Chlond B, Magdolen M, Eisenmann C, et al. Deutsches Mobilitätspanel (MOP) - Wissenschaftliche Begleitforschung und Auswertung, Bericht 2017/2018: Alltagsmobilität und Fahrleistung. Karlsruhe: Karlsruher Institut für Technologie; 2019 [cited 2022 Oct 10]. Available from: https://www.bmvi.de/SharedDocs/DE/Anlage/G/mop-jahresbericht-2017-2018.pdf?__blob=publicationFile

44. Achaim J. Spinning up documentation [Internet]. 2020 [cited 2020 Jun 10]. Available from: https://www.amazon.de/Prime-Video/b/ref=nav_shopall_aiv_piv?ie=UTF8&node=3279204031

45. Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, et al. Emergent tool use from multi-agent autocurricula. arXiv [Preprint]. 2019 [cited 2023 Jan 15]. arXiv:1909.07528. Available from: https://arxiv.org/abs/1909.07528

46. VDI - Verein Deutscher Ingenieure. Reference load profiles of single-family and multi-family houses for the use of CHP systems. 27th ed. 2008.

47. Simonini T. On choosing a deep reinforcement learning library [Internet]. 2019 [cited 2020 Nov 4]. Available from: https://medium.com/data-from-the-trenches/choosing-a-deep-reinforcement-learning-library-890fb0307092

48. Hubbs C. Deep reinforcement learning and hyperparameter tuning: using Ray's tune to optimize your models [Internet]. 2020 [cited 2020 Nov 6]. Available from: https://towardsdatascience.com/deep-reinforcement-learning-and-hyperparameter-tuning-df9bf48e4bd2

49. Liessner R, Schmitt J, Dietermann A, Bäker B. Hyperparameter optimization for deep reinforcement learning in vehicle energy management. In: Proceedings of the 11th International Conference on Agents and Artificial Intelligence; 2019 Feb 19-21; Prague, Czech Republic. Setúbal: SCITEPRESS - Science and Technology Publications; 2019. p. 134-44.

50. Consentec. Description of the balancing process and the balancing markets in Germany. Aachen (2020). https://www.regelleistung.net/ext/download/marktbeschreibung_cons. retrieved October 19, 2020. Google Scholar

51. Mantovani RG, Rossi ALD, Alcobaça E, Vanschoren J, Carvalho ACPLF. A meta-learning recommender system for hyperparameter tuning: predicting when tuning improves SVM classifiers. Inf Sci (Ny). 2019;501:193-221.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press