FCN Working Paper No. 6/2021

Deep Reinforcement Learning for Optimized Operation of Renewable Energy Assets

Jan Martin Specht and Reinhard Madlener

September 2021
Revised April 2022

Institute for Future Energy Consumer Needs and Behavior (FCN)

School of Business and Economics / E.ON ERC

FCN Working Paper No. 6/2021

Deep Reinforcement Learning for Optimized Operation of Renewable Energy Assets

September 2021
Revised April 2022

Authors' addresses:

Jan Martin Specht, Reinhard Madlener
Institute for Future Energy Consumer Needs and Behavior (FCN)
School of Business and Economics / E.ON Energy Research Center
RWTH Aachen University
Mathieustraße 10
52074 Aachen, Germany
E-Mail: MSpecht@eonerc.rwth-aachen.de, RMadlener@eonerc.rwth-aachen.de

Publisher: Prof. Dr. Reinhard Madlener
Chair of Energy Economics and Management
Director, Institute for Future Energy Consumer Needs and Behavior (FCN)
E.ON Energy Research Center (E.ON ERC)
RWTH Aachen University
Mathieustrasse 10, 52074 Aachen, Germany
Phone: +49 (0) 241-80 49820
Fax: +49 (0) 241-80 49829
Web: www.fcn.eonerc.rwth-aachen.de
E-mail: post_fcn@eonerc.rwth-aachen.de

Deep Reinforcement Learning for Optimized Operation of Renewable Energy Assets

Jan Martin Specht[1,*] and Reinhard Madlener[1, 2]

[1] *Institute for Future Energy Consumer Needs and Behavior (FCN), School of Business and Economics / E.ON Energy Research Center, RWTH Aachen University, Mathieustraße 10, 52074 Aachen, Germany*

[2] *Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology (NTNU), Sentralbygg 1, Gløshaugen, 7491 Trondheim, Norway*

September 2021, last revised April 2022

Abstract

This study utilizes machine learning and, more specifically, reinforcement learning (RL) to allow for an optimized, real-time operation of large numbers of decentral flexible assets in the electricity domain. The potential and current obstacles of RL are demonstrated and a guide for interested practitioners is provided on how to tackle similar tasks without advanced skills in neuronal network programming. For the application in the energy domain it is demonstrated that state-of-the-art RL algorithms can be trained to control potentially millions of small-scale assets in private households. In detail, the applied RL algorithm outperforms commonly used heuristic algorithms and only falls slightly short of the results provided by linear optimization, but at less than a thousandth of the simulation time. Thus, RL paves the way for aggregators of flexible energy assets to optimize profit over multiple use cases in a smart energy grid and thus also provide valuable grid services and a more sustainable operation of private energy assets.

Keywords: reinforcement learning, virtual power plant, aggregation of energy, value stacking

1 Introduction

In recent years, deep learning has developed from a niche into a powerful tool for various applications, with drastic improvements in fields such as speech and visual recognition (LeCun et al., 2015), robotics (Pierson and Gashler, 2017), manufacturing (Wang et al., 2018) and multi-agent systems (Gronauer and Diepold, 2022). This success has been driven by the ability to train deep learning algorithms using vast amounts of data to develop insight into quick, optimized decision making in cases where conventional dynamic programming and econometric approaches suffer from the so-called "curse of dimensionality" (Gosavi, 2004; Ning and You, 2019).

[*] Corresponding author. Tel.: +49 241 80 49 832, MSpecht@eonerc.rwth-aachen.de

Despite promising advantages in other disciplines, several recent, high-profile articles posit that applied research using deep learning in the fields of business analytics and operations research is still relatively scarce (Huck, 2019; Kraus et al., 2020). This issue also extends to practical applications: while 85% of executives believe artificial intelligence (AI) will provide their companies with a competitive advantage, only 5% actually extensively incorporate AI into their business processes (Ransbotham et al., 2017). Many practitioners seem to either avoid AI for fear that this topic requires extensive development of expert knowledge on neural network programming or become discouraged by a seemingly endless number of obstacles that typically appear during implementation of a working setup.

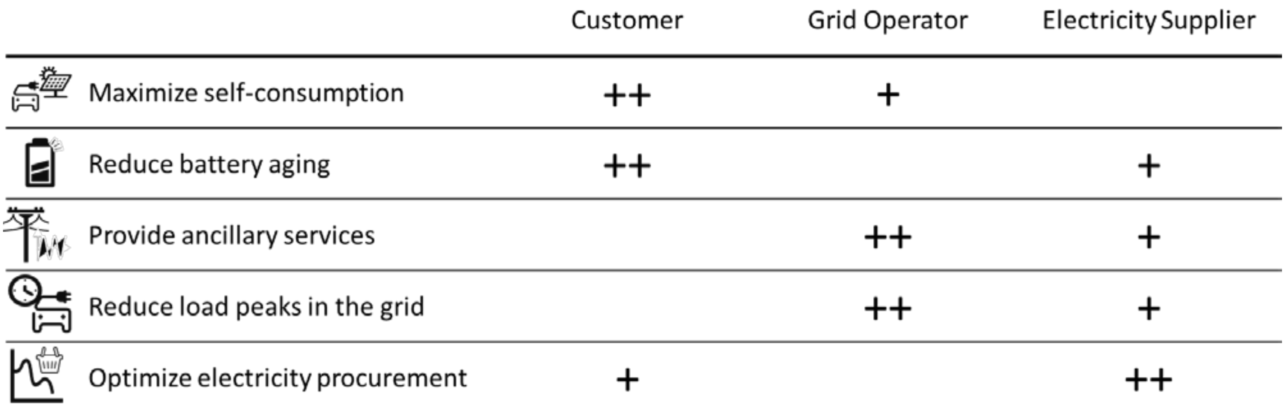
One field where the introduction of AI might become a key enabler is the electricity system, which is currently subject to at least two disruptive changes. First, the rise of digitalization is facilitating the interconnection of potentially millions of decentral energy assets for power generation, storage and consumption within a "smart grid" (Henry and Ernst, 2021; Madlener, 2022). Second, the transition towards low-carbon, sustainable energy technologies involves fundamental changes not only in the electricity sector but also in the mobility and heat sectors. Electric vehicles, heat pumps or home battery storage pose a significant risk to the stability of the electric grid, since their substantial electricity demand could overstrain existing electric grid infrastructure (Cruz et al., 2018; Deilami et al., 2011). This change in demand may lead to massive grid enhancement costs, especially if operators have to design the grid for worst-case scenarios (e.g., multiple customers drawing large amounts of electricity simultaneously). However, despite these challenges, the flexible nature of the operation of these assets also has the potential to provide multiple use cases with benefits to customers, grid operators and electricity suppliers alike (cf. Figure 1). Flexible assets could, for example, be operated to (1) profit from volatile electricity prices, (2) reduce grid stress by leveling out load peaks, (3) provide reserve power to stabilize the electric current frequency, (4) reduce battery storage aging, or (5) maximize the self-consumption of local (solar) electricity generation (Greenwood et al., 2017; Hao et al., 2018; Nykamp et al., 2012; Ruester et al., 2013).

However, energy customers such as private households may not appreciate frequent adjustments in the operation schedule of their assets. Instead, the task of dispatching the flexibility potential of assets has recently fallen to "aggregators," i.e. entities that pool distributed assets virtually and market their potential for some of the use cases mentioned above, allowing them to generate additional revenue streams for themselves and their customers (Bell and Gill, 2018; Specht and Madlener, 2019).

In a previous study, we demonstrated that a linear optimization model could be used to employ flexible assets and generate up to 150 €/a for a typical household in Germany (Specht and Madlener, 2020). These values represent the maximum theoretical potential, i.e. assuming perfect foresight. In reality, these theoretical values cannot be achieved because one cannot perfectly forecast local electricity production and consumption, electricity exchange prices, et cetera. Moreover, nonlinearities such as the aging pattern of battery storage result in optimization problems that may take minutes or even days to solve (if considering thousands of time steps in advance), which is hardly feasible for applications with potentially millions of households that need to update their power consumption schedule every few minutes (e.g. if electricity spot market prices or demand for reserve power changed, if the user activated energy intensive devices, or if the electricity mix changes towards more intermittent renewables.

Deep learning approaches have recently been applied in this context and have shown a performance superior to that of conventional decision algorithms (Diamantoulakis et al., 2015; Kraus et al., 2020; Tu et al., 2017). Reinforcement learning (RL), a subtype of machine learning, is an especially promising approach in this field. For RL applications, large amounts of data are advantageous rather than challenging, since these data can be used to quickly train an instance of the algorithm, called an agent. Once trained, agents can react to changes in their real-world observations with minimal computational effort and time. Furthermore, they do not require forecasts nor human interaction to explain or interpret events. Instead, RL algorithms have the potential to learn independently using their observations. This ability is a very helpful feature, since it is unrealistic to task human employees with analyzing the routines and patterns of millions of customers.

This study extends previous work by Kraus et al. (2020) and aims to address both the dearth of literature on deep learning in the field of applied operations research as well as the lack of tools to optimize the steering of electrical smart grid assets in private households with multiple revenue streams. To this end, we investigate if deep learning can indeed outperform linear optimization methods when maximizing revenue by optimizing flexible electricity operation over multiple value streams, as introduced in Specht and Madlener (2020). The resulting original contributions of this work are twofold. First, the general procedure proposed to develop a reinforcement learning algorithm for a given task can be applied without expert knowledge in the field of neural network programming. It can be applied by researchers and practitioners in various fields of research, allowing these experts to improve the quality or calculation time of optimizations. Second, to the best of our knowledge it is the first study providing quantitative evidence that a significant economic potential and competitive advantage exists when using deep learning techniques for the optimized operation of vast numbers of flexible energy assets.



	Customer	Grid Operator	Electricity Supplier
Maximize self-consumption	++	+	
Reduce battery aging	++		+
Provide ancillary services		++	+
Reduce load peaks in the grid		++	+
Optimize electricity procurement	+		++

Figure 1: Flexible assets in private homes, such as electric vehicles or battery storage, can be operated to maximize value for different stakeholders over multiple value pools.

The remainder of this paper is structured as follows. Section 2 reviews essential developments in deep learning and presents RL as a state-of-the-art approach. Section 3 introduces the challenges in the energy context, matches these challenges with the strength and weaknesses of RL with the aim to maximize the value generation of flexible energy assets in private households while simultaneously considering up to five use cases. Based on the literature and our experiences, we provide a checklist with commonly found obstacles and suggested solutions for readers interested in applying reinforcement learning to applications in their domain. In Section 4, we select an appropriate RL-algorithm based on the challenges identified in Section 3 and briefly outline the training process. Section 5 presents general findings from our application of AI in the energy sector and includes the quantitative results for the selected application. Finally, Section 6 summarizes and concludes.

2 Milestones in the Developments Towards Reinforcement Learning

2.1 Background: Machine Learning

The concept of “machine learning” in its literal meaning is often traced back to the development of Bayes’ Theorem in the 16th century and includes multiple developments, such as the invention of artificial neural networks in the 1950s.

Similarly, the concept of “deep learning” dates back to the 1940s; however, the term gained popularity only around 2006 (Goodfellow et al., 2016) and typically restricts the broad concept of machine learning to concepts involving neural networks of multiple layers that are trained using vast amounts of data. From 2006 onwards, researchers focused on applying deep learning algorithms to audio and image classification. These algorithms were trained on datasets (e.g., handwritten numbers or fashion items), typically using “supervised learning.” This approach requires “labeled data” that informs the algorithm of whether its prediction is correct (cf. Li, 2018). Around 2011, the quality of the image classification of the best agents surpassed human ability in an increasing number of test settings.

Deep learning algorithms next became capable of learning to play complex, sequential games. For instance, the algorithm “AlphaGo,” developed for the board game Go, was able to defeat the best human players after it was trained on large datasets of human gameplay[1] (Silver et al., 2016). Despite of these successes, this approach of supervised learning faced significant limitations. The massive datasets required for training limited the application of deep learning to a small number of appliances, where sufficient training data is available, as in the case of Go. Further, training an algorithm using only available data necessarily implies that the algorithm cannot surpass the capability of whoever generated the training data.

2.2 Reinforcement Learning

In RL, an agent (or policy) learns to act to maximize the returns of a value function. More specifically, the agent conducts a so-called Markov decision process, which includes obtaining an im-

[1] In fact, AlphaGo was actually a transition, also including first elements of Reinforcement Learning but at least with a focus on learning from existing data.

pression from the environment in the form of an "observation," performing an "action," and receiving a "reward," for example, confirmation that a goal was reached (cf. Figure 2, Sutton and Barto, 2018). The main difference between RL and pure supervised learning is the ability of the former to interact with its environment. This central feature allows the agent to actively explore an environment without requiring an expert dataset and later "exploit" its experiences to maximize its reward (Sutton and Barto, 2018).

The utility of RL can be illustrated using an example. Less than two years after AlphaGO surpassed human capabilities primarily as the result of supervised learning, DeepMind released its successor, AlphaZero (Silver et al., 2017). This new program fully exploited RL and was trained by playing against itself, without any expert data. AlphaZero quickly surpassed AlphaGO in the game Go and was also able to learn other games, such as chess or shogi. AlphaZero was only provided with the game rules, and outperformed the best human players as well as computer programs in a variety of games. In these games, experts observed that AlphaZero reinvented many established human strategies but also developed powerful new approaches that were formerly unknown to human professionals (Sadler and Regan, 2019).

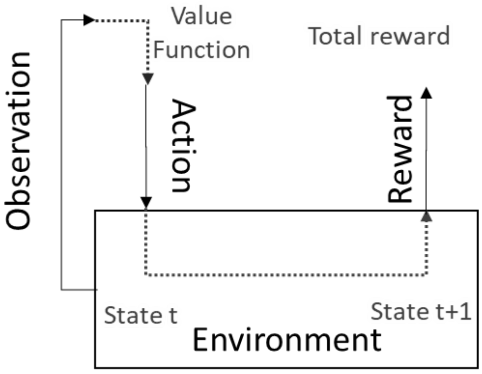


Figure 2: Visualization of one step in a Markov decision process.

Early RL algorithms were only capable of handling discrete observation- and action spaces (e.g., positions and moves in chess). Deep deterministic policy gradient (DDPG) is the first approach to address continuous observation- and action spaces and was presented in 2015 by a Google DeepMind team (Lillicrap et al., 2015). However, this approach was instable and has been modified in recent years by introducing critic networks that better evaluate the actions of the policy network and also reduce instability through different measures (most notably, "Double Q-learning"). Two distinct approaches were very promising in comprehensive tests: the soft actor-critic (SAC, Haarnoja et al., 2018) and the twin-delayed deep deterministic policy gradient (TD3, Fujimoto et al., 2018). Both approaches are considered "off-policy", which means that they fill up a database, called a “buffer”, during operation and then select a new action based on their previous experiences. In contrast to this, on-policy algorithms adjust their policy, that is their “behavior”, on the flight and then discard the previous experience just proceeding with their updated policy. Compared to off-policy algorithms, this typically enables them to learn faster, with more stability and with reduced hyperparameter tuning, but typically at the cost of sample efficiency. One of the most capable on-policy algorithms is proximal policy optimization (PPO, Schulman et al., 2017).

In the long term, optimizing computer games is not the ultimate purpose of RL but instead serves as a test bench for real-world applications. The limited applications of RL algorithms is likely due to the skepticism of practitioners regarding the effort necessary to familiarize oneself with this complex topic as well as concerns of overcoming numerous obstacles. To this end, we contribute to the existing literature by presenting a method to apply state-of-the-art algorithms without extensive background knowledge in programming neural networks. We also present a clearly defined example of how to consider and overcome many of the major existing challenges in a complex application.

3 Setting up an RL model for the Exemplary Application

3.1 Description of the Exemplary Application

The aim of this work is to apply a reinforcement learning algorithm to optimize the operation of flexible energy assets over a variety of different use cases, similar to Nakabi and Toivanen (2021) approach in the context of a micro grid. A real-world application was not possible in this context, so the scenario was simulated by creating a virtual environment for the RL agent to interact with. To facilitate comparison with existing studies, this environment was designed to be similar to the one optimized by linear optimization as introduced in Specht and Madlener (2020).

The underlying motivation is that of an aggregator that operates distributed energy resources. These assets have the potential for flexible operation (in our case, home battery storage systems and the charging of battery electric vehicles (BEVs) in the customers’ premise) that could be used to increase profit. This additional profit could be created either by generating additional revenue or by reducing costs. Of the numerous promising use cases discussed as candidates for additional value, we selected five in an attempt to depict the variety of applications (see Madlener and Specht, 2018 for an in-depth discussion). These five cases are described below:

(i) The electricity price for a household can be regarded as a combination of a static fixed price component (26 €-ct/kWh, comprising taxes, grid fees, etc.) and a flexible price component that captures price volatility and was based on real electricity exchange prices from Germany. While customers today typically pay a fixed electricity rate, the energy supplier has to meet its customers' demand at volatile exchange prices and could gain value from shifting flexible energy consumption to times with low prices.

(ii) As introduced in Section 1, new energy transition technologies such as electric vehicles and heat pumps can cause severe load peaks, which would require grid operators to execute expansive grid enhancement measures. Instead, these grid operators could encourage households (or their aggregators) to avoid consuming power during high load peaks. In Germany, article 14a of the "Energiewirtschaftsgesetz" (EnWG) defines an initial attempt to reduce peak loads. Based on this regulation[2], we implemented a mechanism that reduces the grid fees linearly by up to 5 €-ct/kWh while reducing the allowed peak load from 15 kW_{peak} to 5 kW_{peak}.

(iii) Each day, the aggregator can opt to reserve a fraction of the battery capacity in order to stabilize the grid by either storing excess electricity or providing additional power to the grid in times of undersupply. While we assumed that the grid operator would pay the aggregator 1 €/kW in a contracted week[3], this reserved battery capacity cannot be used for other purposes, thus causing opportunity costs.

(iv) Today's lithium-ion batteries typically age faster when charged or discharged to very high or low states of charge (SOCs). This aging function is approximately U-shaped. A detailed description of the function used can be found in Specht and Madlener (2020); for a detailed technical investigation, we recommend Ecker et al., 2014). An optimized operation could avoid these edge conditions (for example, by not draining battery storage completely each day in winter but instead leaving a residual capacity as a buffer).

(v) The remuneration for electricity from local PV production fed into the grid is 10 €-ct/kWh, which is significantly below the cost of electricity drawn from the grid (about 30 €-ct/kWh in

[2] In Germany, recent regulation allows for reductions in grid fees for controllable load (e.g. heat pumps and BEV); Germanies larges distribution system operator Westnetz, for example, offers a tariff reduction of about 5 €-ct/kWh for devices in the distribution grid that can be switch off if the grid comes to its limit.

[3] Since July 2019, the regulation for frequency containment reserve was subject to three major reforms and corrections, including a shift from a capacity price for a week [€/MW] to an energy pricing on daily basis [€/MWh], cf. (Consentec, 2020). Since a reliable data basis for the new market framework is not yet available, we decided to stay with contracts based on reserved capacity. However, a significant decline in revenues for this services was observed in recent years, which is why we adjusted the typical revenue range of 1500-3000 to only 1000 €/MW per week or 1€/kW/week respectively.

Germany); for this reason, an agent should try to maximize self-consumption both by storing electricity in a battery during times of excess for times without local production and by prioritizing charging a BEV in times with a surplus of PV electricity whenever possible.

3.2 Modeling the Exemplary Environment

We created a virtual environment based both on real data (i.e. for electric consumption profiles, electricity generation profiles for a private PV system) and synthetic data (e.g. for electric vehicle usage profiles). This data and the use cases introduced in Section 3.1 were used to model the environment. Our overall goal was to minimize electricity expenses (or even maximize profit if revenues might surpass costs) over one year (35,040 steps of 15 min duration each). The relevant constraints (for example, maximum charging rates and non-negativity given the SOC of batteries and vehicles) were considered.

The trained agent had four tasks: (a) to commit to a maximal amount of power drawn from the grid for each week, (b) to commit to an amount of power to store as reserve power for the grid on a daily basis, (c) to determine an amount of power to charge into or discharge from the battery storage every 15 min and (d) to decide how much to charge into the BEV every 15 min. The agent's decisions were informed by eight provided values ("observations"), which included the current SOC of the battery and the electric vehicle, the current power generation by the PV system minus the regular power consumption of typical households as well as information on time and date. These agent addresses these tasks by means of for Based on this, the agent's response comprises four actions[4].

To decide on its actions, the agent is provided with eight "observations" [5], needed to understand the given situation. These observations and actions are in a continuous space, so, for example, any arbitrary amount of power (within the given constrains) can be used to charge a vehicle. We implemented some measures to penalize attempts to violate the given constrains (e.g., penalizing attempts to discharge a battery below "empty" or beyond existing constraints as well as penalizing attempts to exceed limitations placed on the maximum load drawn from the grid). Additionally,

[4] The four values in the action space comprised: (1) the amount of electricity to charge/discharge to/from the battery, (2) the energy to charge into the vehicle, (3) once a day, the commitment to provide a certain amount of reserve power (use case (iii)), and, finally, (4) a self-imposed limitation on the maximum amount of power to draw from or feed into the grid (use case ii).

[5] The eight values in the observation space include i.a. (1) the current electricity price, (2) the sum of electricity self-production minus consumption, (3-4) the state of charge of battery storage and electric vehicle, (5-6) information for the agent regarding which long term use cases are currently applied as well as (7-8) information regarding the date and the time.

an increasing penalty for low SOCs in BEVs was implemented to reflect customers' preference for a fully charged vehicle. Finally, a conventional, heuristic algorithm was implemented that simply charges the vehicle at once if it returns from a trip, then uses the battery to store excess power (until it is full) and discharges this power at as soon as electricity demand surpasses local PV generation. No other use cases were subject to this heuristic, allowing us to compare the optimized operation over all use cases to an incumbent operation strategy typically seen today.

3.3 General Challenges and Solutions

When designing our model, we encountered nine major obstacles that reflect typical experiences in the literature (cf. Dulac-Arnold et al., 2019; Nakabi and Toivanen, 2021; OpenAI et al., 2019). This chapter describes these obstacles for interested practitioners and provides our solution to these challenges.

(i) The need for a training model

A first, rather obvious obstacle is that stakeholders typically consider training an RL agent in a real environment to be expensive, dangerous and/or intolerable. This obstacle clearly applies when developing a flexibility aggregator in the electricity domain, since random actions during training could place the grid under significant stress, incur massive costs for the aggregator and inconvenience the customers. For this reason, we opted to depict the problem as a virtual model to pretrain the algorithm before considering a real-world application.

(ii) The definition of a reward function

To evaluate their actions, RL algorithms need constant feedback in the form of a reward. If this reward is not generated in a real setting for the reasons discussed previously, one must design a reward function to evaluate all possible actions of the agent. In many applications, a maximization of the purely monetary profit might be sufficient. However, in the chosen application, aspects such as the customers' preference for the immediate charging of an electric vehicle had to be monetized to be considered by the agent. The return consequently includes both "regular" costs and benefits realized by the algorithm's actions comprising, for example, selling or buying electricity to or from the grid and revenue from providing services for the grid. Aspects such as the unavailability of a fully charged car had to be transformed into "artificial" costs and benefit functions.

(iii) Complex features of the environment

In general, the better an agent understands a given situation, the better its solution. Unfortunately, real-world situations cannot be fully explained in our application, as is likely true for most applications. More specifically, the environment is only partially observable since not all changes can be explained to the agent (as no one actually knows, for example, what resulted in a given electricity price at a given time). Furthermore, the chosen application is stochastic, as the next state is not exclusively determined by the agent’s action (the agent has no influence over whether the customer decides to use their vehicle, which would result in its unavailability). However, we found that eight observations at each state (every 15 min) were enough to allow recent algorithms to gain a sufficient understanding of the underlying causalities in the environment. Moreover, only four actions were required at each state for the agent to react to these observations and interact with the environment (especially to call actions such as “(dis-)charge the battery” and “charge the vehicle”)

(iv) Unlimited options in continuous observation and action spaces

As in many real applications, most of the chosen observations and actions are from a continuous rather than a discrete action space, which theoretically creates an unlimited number of options. In our case, for example, a BEV could be charged by any value between zero and the maximum allowed load (22 kW). Maximizing decisions over an infinite amount of solutions in continuous action spaces became an option only a few years ago, when new (actor-critic) methods made this feature possible for RL algorithms, thus limiting the options for algorithms appropriate for our application.

(v) Long games with delayed rewards

A major challenge in long-term settings is the vast number of steps in each round. While a chess game typically lasts around 80 moves and a game of Go around 150 moves (OpenAI et al., 2019), our chosen environment required over 35,000 steps, representing 15-min intervals over one year. A reduction is problematic here, since two opposing aspects have to be met. It is necessary to depict the volatility in production and electricity prices at a sufficiently detailed resolution, but the agent should also learn to consider seasonal differences that occur throughout a year. A possible solution is to limit long-term rewards (e.g., at the end of a specifically good game) and instead focus on a steady stream of rewards, in our case after each 15-min time step.

(vi) The need to define boundaries and constraints

Reinforcement learning typically involves learning from mistakes. Therefore, the agent must be capable of violating constraints and receiving adequate feedback while ensuring that the virtual environment responds properly, preventing the violation of any physical or regulatory boundaries. For example, batteries must neither be charged below 0 nor above their capacity limit. In these cases, we restricted the environment to the maximum allowed action space and returned a penalty value to the agent if it attempted to overstep this limitation. This penalty required accounting for two reward streams: the virtual rewards, including a penalty given as feedback to the agent, as well the "real" economic balance, which represent actual revenue streams.

(vii) The difficulty of defining a benchmark

Assessing the success of an agent's strategy is non-trivial, since a result well below average might be justified by a randomly selected, extremely difficult setting (e.g., a household with high consumption and high mileage on its electric vehicle). Therefore, we used a heuristic algorithm previously proposed in Specht and Madlener (2020) as a benchmark to compare the quality of the agent's strategy. This benchmark follows a set of simple instructions similar to the algorithms currently used to control the energy assets in households.

(viii) Differences in the scale of values

The agent's observations and actions can vary significantly in their scale. This variation becomes a problem when training the agent since slight variations in the action might be leveraged to significantly change small-scale variables, leading to instability. As frequently suggested by practitioners, we also normalized our observations and actions to a scale either between -1 and 1 or 0 and 1. For example, the observation of the SOC ranges from 0 to 1 so that a half-full battery is normalized to 0.5. The operation the battery is done with actions ranging from -1 (discharge at maximum rate) to 1 (charge at maximum speed).

(ix) Need for training data

A problem reported frequently in literature is a lack of training data. While the simulated environment allows for any number of runs, repetitions on similar data lead to the agent "overfitting" its policy to specific setups, which leads to a loss in performance on new data. To overcome this issue, we mixed different data sets during training. Our environment included 85 different load profiles of different households, five normalized PV production profiles that were multiplied by different system sizes to get the actual electricity production, 10 normalized BEV usage patterns that were scaled by eight different mileage options, and 35,040 timesteps per profile per year, thus allowing for billions of different combinations. In fact, as we will be demonstrated in Section 5 the performance of the tuned agents tended to plateau after approximately 500,000 steps, with more iterations leading to instabilities and a decline in performance.

4 Practical Experiences: Training the Agent in the Exemplary Environment

4.1 Setting up the Model Environment and Selecting an Adequate Algorithm

Once the requirements in Section 3.3 have been defined, two tasks must be considered. First, the identified setting from Section 3.2 has to be programmed as an environment with which a reinforcement learning (RL) agent can interact. Secondly, an algorithm that works under the identified constraints has to be selected to train the RL agent.

We decided to implement our setting as a "gym" environment. "Gym" is a collection of test problems (typically small games) created by OpenAI that all share one interface[6]. Within Gym, new RL algorithms can be trained on the same set of tasks, which facilitates comparison of the results. However, this also works the other way around: One can design a new game (called an environment) in the standardized way, so multiple RL algorithms can be tested and compared without a need for substantial changes to the environment.

Setting up just one of the more complex algorithms manually requires weeks to familiarize oneself with a specific method based on literature and then implement the selected algorithm based on specialized libraries such as pytorch or keras. Fortunately, there are also several projects available that provide precast, customizable RL algorithm setups. We opted to use "Stable Baselines" (https://stable-baselines.readthedocs.io), a fork of OpenAI baselines. Stable Baselines provides around 14 major, state-of-the-art algorithms with numerous options to add or customize specific features.

Using this combination of a standardized environment and RL algorithm blueprints, one can set up a machine learning task and train an agent using selected algorithms with just basic programming skills (optimally in Python) but no in-depth understanding of AI programming. Once a task and training algorithms have been selected, one must select a specific RL algorithm.

4.2 Selecting an Adequate Algorithm for the Specific Task

A significant number of different RL algorithms can be found in recent literature, and each has an even greater number of possible modifications. However, systematically listing the requirements of a given task as shown in Section 3.3 can help restrict the number of eligible algorithms. For example, the considerations regarding feature limits (Section 3.3, iii) allowed us to discard any "model-based" RL algorithms, since those depend heavily on the completeness and accuracy of

[6] Interface in this context means that each came must host a standardized set of specific functionalities such as a "step"-function or a "rest"-function.

the ground-truth model, which is (even in well-defined technical settings) often referred to as "fundamentally hard" and unstable (Achaim, 2020).

Furthermore, the use of continuous action and observation spaces further limits the range of potential algorithms to a small number of very recent approaches, as described in Section 2.2. Figure 3 shows a comparison of the three most promising candidates, SAC, TD3, and PPO2, in their "vanilla" version (i.e., the published version before any hyperparameter tuning or modifications). The graphs in light colors illustrate findings for specific settings and are soothed for easier comparability. The dotted graphs plot the difference to the heuristic algorithm on the secondary axis, indicating that all vanilla versions perform significantly worse than the heuristic algorithm. A comparison of the RL algorithms with each other revealed that SAC was the most promising option.

In detail, Figure 3 shows that both the vanilla SAC and PPO2 reached their peak performance for our setting after around two million timesteps (equivalent to simulating 57 households at 15-min resolution for a year) before the learning progress flattened to a constant level. The performance in terms of "perceived profit" per year, however, reached just -2000 € for SAC. Perceived profits here includs "real" revenues from use cases minus costs to supply the household with electricity as well as "virtual" penalties for intended violations of restrictions). Vanilla TD3 reached peak performance with around 5000 €/a after only 1 million steps but was too instable to maintain this level of performance. PPO2, finally, resulted in the highest overall yearly costs of around €6,000.

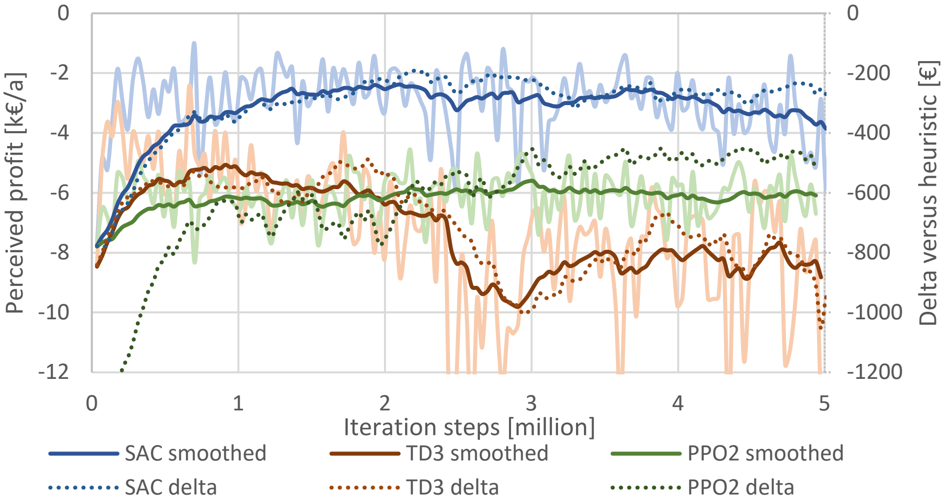


Figure 3: The continuous lines show the performance of SAC, TD3, and PPO2 throughout a training of up to 5 million iterations, with the dark colored graphs being a smoothed version of the lighter pendants. The dotted graphs indicate the loss of the respective RL-algorithm to the simple heuristic on the secondary axis. Amongst the RL algorithms, vanilla SAC performed best and most stable.

However, PPO2 also had the fastest computation time, requiring only 1.5 h to complete 5 million training steps, while SAC and TD3 required 4.75 and 3.75 h, respectively, to complete the same number of simulation steps on a desktop PC (CPU: i5-7500, 12 GB RAM).

These initial trials indicate that SAC is a promising candidate for more detailed analysis. However, comparison with the heuristic algorithm reveals a consistent, significant underperformance of all three algorithms in their vanilla versions of (at best) €200 losses per year compared to incumbent algorithms, indicating an extensive need for performance improvement by means of hyperparameter tuning.

5 Results

5.1 Observations During the Training Process

As with other studies (Baker et al., 2019), we also found that RL agents excel at exploiting any weaknesses in game mechanics. We expected fewer issues from this side since the underlying environment had already been used for previous studies with linear optimization. However, the RL agent was still able to find loopholes in our environment to exploit. For example, the environment was coded to set the battery SOC to a minimum of 0 at the beginning of each round and the agent learned that it was profitable to discharge the battery storage below 0, sell this nonexistent energy, accept the penalty for violating the rules, and allow the battery to be reset to 0 in the next round.

When we applied the agent trained with SAC and with tuned hyperparameters to this task, we learned that a single decision for the four tasks (a–d) takes about 0.0003 seconds (or 3 seconds for 10,000 decisions) on a personal computer (CPU: i5-7500, 12 GB RAM). This finding demonstrates that RL agents can control large numbers of households in nearly real-time, thus meeting the requirements in term of optimization time.

5.2 Economic Findings in the Chosen Application

To assess optimization quality, we conducted sequences of 1000 simulation steps for different households with different BEV usage profiles. With all use cases activated, a pretrained, general RL agent was able to increase the average profit by €126 per year in comparison to the conventional heuristic algorithm, as depicted in Figure 4. More specifically, some settings yielded additional profits (i.e., reduction in electricity costs) of up to €315 per year, while a small number of households experienced negative savings and thus higher costs compared to conventional operation by means of a heuristic algorithm (thus without utilizing any additional use cases).

We selected one household based on a norm developed by the Association of German Engineers (VDI) to resemble the “most typical” household (VDI, May/2008) for which the general algorithm achieved annual savings of around €150 and allowed the general agent to pretrain specifically on this household. This specialization yielded additional annual savings of €28–178, indicating that real-world applications will likely profit from training using individual customers’ data. This result is also remarkable in comparison to the €235 annual savings achieved by linear optimization in Specht and Madlener (2020) in a nearly identical setting[7], not only because this linear optimization utilized perfect foresight (while the RL agent only knows the present situation and has some general experience from the past) but also because the RL agent only took about 10 seconds (instead of 10 h) to optimize all 35,040 steps of a given year.

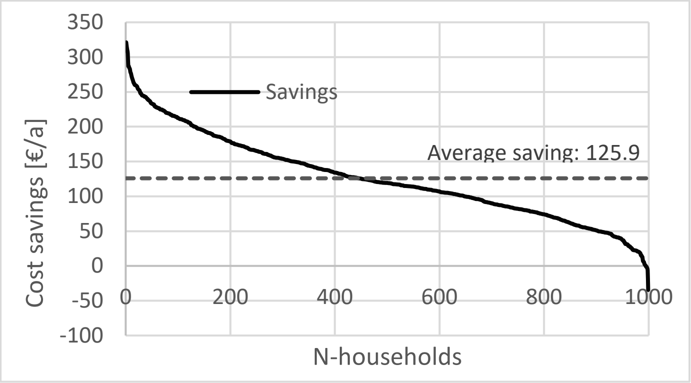


Figure 4: (a)**:** With all use cases activated, most households realized cost savings or profits of up to €315 and an average of €125.9 per annum.

5.3 Comparison of the Behavior of the RL Agent and the Heuristic

A comparison of the agent’s action to the heuristic allows to better understand their characteristic behavior. Figure 5 depicts the sum of local generation and ordinary consumption (lighting, cooking, et cetera); the net exchange with the grid, both in kW; as well as the SOC of the battery storage and the BEV (in kWh) for the same day around equinox for a generally trained RL agent and the heuristic. The sum of local generation and regular consumption (yellow) is the same in both cases as neither of the algorithms can change the PV generation or the ordinary consumption of the household. The PV system starts to generate electricity around 7 a.m., and this generation quickly surpasses the consumption, resulting in positive net power. The RL agent left the BEV (green) with a SOC of approximately 35 kWh the previous day and primarily uses the excess

[7] The only major difference is that this paper introduced a remuneration for the reduction of load peaks, allowing for additional revenue for this use case.

production in the early morning to charge the BEV until it leaves at 10 a.m. and thus becomes unavailable for charging. In parallel, the battery (blue) is charged slowly until the BEV leaves and then is steadily charged to 6.2 kWh at 5 p.m. At that point, PV generation is not sufficient to meet energy demands, so the agent begins to slowly discharge the battery. At 7:30 p.m. the BEV returns home, and the RL agent observes that its SOC is about 6 kWh lower than when it left. The RL agent meets this energy need by slowly charging the BEV and waiting for new PV electricity the next day.

The heuristic algorithm immediately charges the car when it returns less than full. The heuristic is consequently not able to use the early excess electricity in the morning hours to charge the vehicle but instead feeds all excess electricity into battery storage, leading to a rapid charging process that results in a full battery at 10:15 a.m. All the remaining excess electricity is fed into the grid. Unlike the steady charging of the agent, this behavior results in distinct load peaks. In times of undersupply, the battery is used to cover the difference. In particular, when the BEV returns home, the battery is recharged at full speed, but since the BEV allows for significant charging speeds, a massive amount of power is drawn from the grid, with the peak reaching -13.8 kW.

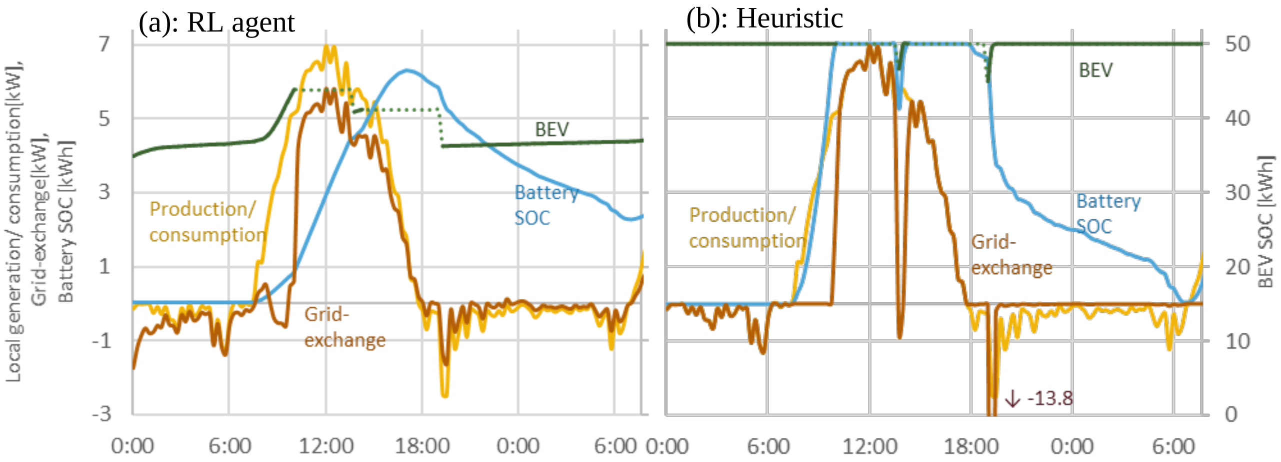


Figure 5: Operation of BEV and battery by a generally trained agent (a) and the heuristic (b) for a day in March

In the future, these massive peaks of the heuristic would significantly stress the grid. Figure 6a depicts the positive and negative electricity exchanges with the electric grid over a year, illustrating that the heuristic causes demand peaks of up to -22 kW (lower blue graph) that would necessitate extensive grid enhancements. In contrast, the implemented revenues for reducing load peaks incentivize the RL agent to avoid this behavior (lower orange graph). Figure 6b shows that the agent is becoming increasingly confident in reducing the electricity exchange with the grid over the year, resulting in commitments to limit this exchange to about 5 kW from September on. The comparison with Figure 6a reveals that the agent frequently surpassed this commitment by a small margin. Apparently, the agent learned that the penalties for violating this restriction were implemented with an exponential increase, so that small violations only resulted in "acceptable" penalties.

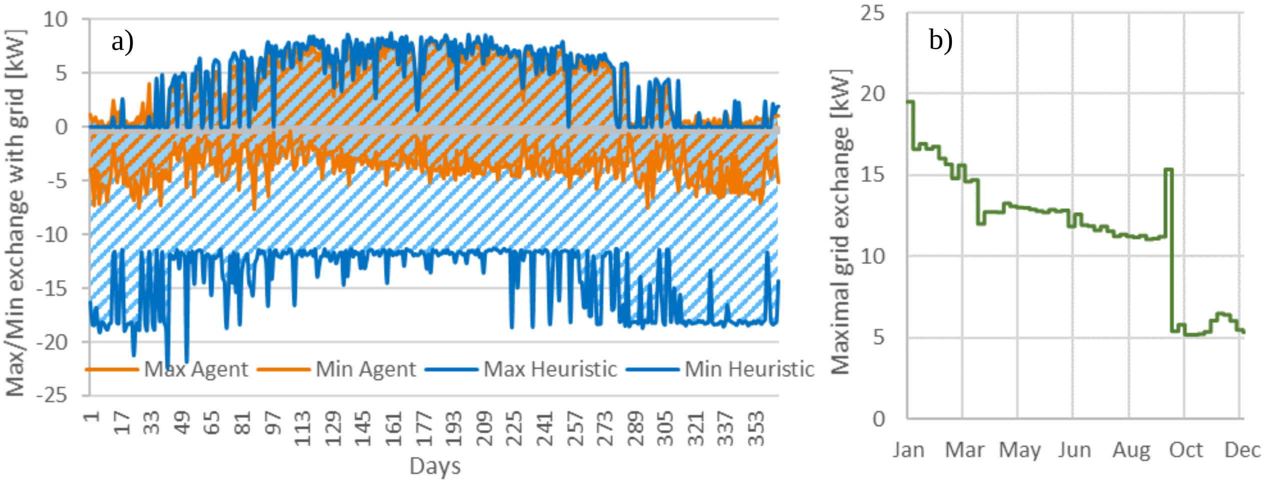


Figure 6: (a) The heuristic (blue) shows slightly higher production but significantly higher consumption peaks compared to the RL agent (orange). (b) The agent limits its own electricity drawn from the grid (green) for each week to gain additional revenues.

5.4 Pros and Cons of the Dispatch of a RL Agent in the Given Application

An in-depth analysis of the data from the RL agent's operation revealed that most of the observed savings were attributed to two use cases: reduction of consumption peaks and reserve power provision. However, these use cases typically require optimization and were therefore not implemented for the heuristic algorithm. For fair comparison, we set the specific profits from the two use cases unavailable for the heuristic to zero. In this case, the heuristic slightly outperformed the RL agent in the majority of the settings. The reason for this is the heuristics ability to store *precisely* the excess electricity in the battery, while the agent sometimes either charged slightly too much (resulting in additional electricity been taken from the grid at high costs) or too little (losing energy that would have been needed in the evening and had to be replaced, again, by expensive grid electricity).

In contrast, the RL agent was able to outperformed the heuristic in settings where high PV production and high BEV mileage enabled significant potential for increases in self-consumption through a more considerate charging strategy (cf. Figure 5). Furthermore, only the RL agent was able to significantly reduce load peaks thus diminishing stress for the electric grid and provide reserve power to stabilize the grid. Summing up, the performance of heuristic and RL roughly balance out, just considering incumbent use cases for flexibility. Upcoming services such as supporting the electric grids, however, cannot be addressed sufficiently by todays heuristics thus demanding for new solutions such as RL.

5.5 Discussion of the Introduced Application for Energy Management

While the trained RL agent is still far from perfect, it nonetheless serves as a valuable proof of concept. The precise value of individual use cases is requires further consideration as well as national regulation, but as long as RL agents can be implemented in an environment, they seem to be capable of adapting to almost any variation in regulation. In fact, general RL agents can robustly adapt to changes in setting caused, for example, by variations in (i) energy demand of different households, (ii) consumption patterns such as customer specific usage of a vehicle or (iii) individual assets with potentially high peak loads. However, RL agent performance can be increased if the agent is allowed to briefly pretrain using previous data (e.g., from the previous year) to familiarize itself with detailed customer characteristics. Even on a regular desktop PC, the decision speed is fast enough to handle thousands of schedule updates in real time. Finally, the availability of blue-prints in the form of libraries, such as from Stable-Baselines, for most aspects of the training process makes this topic accessible to nonexperts without extensive preparation (cf. Simonini 2019 for a comparison of RL libraries).

However, implementation of a customized setup still requires significant time and effort. While a first “proof of life” can be quickly achieved, proper performance in more complex setups requires extensive shaping of the model, fixing of loopholes, hyperparameter tuning and training. Moreover, even under the same settings and after an extensive training cycle, only a small fraction (in our case about 1 in 25) of trained agents develop a good understanding of the underlying patterns and only a very small fraction (about 1 in 250) really excel at its task, while all other agents encounter difficulties in training, e.g. getting stuck in some local optima, so a degree of luck is required in the training process. Finally, deep learning agents are basically a black box, so their reasoning cannot be understood in detail. However, general concepts can still be observed rather clearly in the algorithm results, allowing at least a rough understanding of an agent’s behavior.

Overall, this study confirms the huge potential of deep learning and more specifically RL. Our findings may encourage more practitioners to consider RL as an option for solving complex optimization problems. Complex situations where quick computations times outweigh the need for perfect solutions seem especially promising candidates for the application of deep learning.

6 Conclusion

This paper investigated the potential of deep reinforcement learning (RL) for the optimized control of a large number of distributed flexible energy assets in private households. Further, this work provides practitioners with a guide on how to tackle complex steering tasks with state-of-the-art deep learning algorithms, typical obstacles to expect and potential solutions.

The main insights of this work related to operations research were threefold: First, we confirmed that deep neural networks trained using RL are capable of optimizing the schedule of vast numbers of assets in almost real time. Second, we demonstrated that recent advances in RL algorithms as well as the practice in the AI domain of making these algorithms and tools freely accessible benefit non-experts. Even those with limited knowledge of deep learning have comparatively easy access to potent, customizable algorithms that solve their individual control challenges, and working with these algorithms requires only basic skills in (preferably Python) programming. Third, we developed a list of common obstacles based on a practical example and provided an exemplary procedure to select feasible algorithms suitable for a given challenge. We also confirmed that extensive hyperparameter tuning and model adjustments are required to outperform conventional algorithms, as is often voiced by practitioners (Hubbs, 2016; Liessner et al., 2019; Mantovani et al., 2019).

The application selected as an example provided four interesting insights into research on the sustainable energy transition. First, AI agents are a promising approach to control the large number of flexible energy assets when considering multiple value streams. Advantages of AI over conventional algorithms include their ability to optimize the operation of multiple, distinct use cases; their capacity to handle nonlinearities; and their transformation of the "curse of dimensionality" into a "blessing of training options." Second, we found that a general agent trained on unspecific data could adapt to an unfamiliar setting, for example, a different household with individual energy (and mobility) demand habits. However, allowing the agent to familiarize itself with a given household based on previous data still considerably increased its performance. Thus, data (e.g., on energy consumption habits of customers) will evidently become increasingly important for companies in the energy domain. Third, previous studies have found that working business models of the utilization of flexible home energy assets could theoretically be profitable if one had perfect foresight and hours of computational time for linear optimization. This study finally provided a tool that could realize a significant share of this theoretic potential but in a more realistic setting *without* perfect foresight and at a speed sufficient for the number of assets that would have to be coordinated. Fourth, the additional annual value of around €150 compared to conventional algorithms has promising market potential and offers an economic solution to several issues related to the energy transition as well as the electrification of the heat and the mobility sector. In fact, detailed analysis of the behavior of our trained agent revealed that it is still far from perfect, so technical improvements and increasing revenue potentials (e.g., from increasing price volatility allowing for arbitrage) might allow for even higher gains in future.

Overall, this study demonstrated the enormous potential of deep learning applications in the operations research domain in general and in the energy sector specifically. This potential will be further enhanced by rapid improvements in the underlying algorithms and computational power that are currently underway.

Future research will surely reveal other fields of application where deep learning and RL might replace incumbent algorithms and methods. Furthermore, the task of creating a customized environment to train an agent is still full of obstacles. More literature that provides a practical guide for practitioners is required to provide researchers from multiple disciplines with reasonably easy access to this powerful technology.

Acknowledgments

The authors gratefully acknowledge funding obtained from the German Federal Ministry of Education and Research (BMBF) for the Kopernikus Projects ENSURE 1 and 2 (funding nos. 03SFK1HO and 03SFK1C0-2).

References

Achaim, J., 2020. Spinning Up Documentation. https://www.amazon.de/Prime-Video/b/ref=nav_shopall_aiv_piv?ie=UTF8&node=3279204031, retrieved June 10, 2020.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., Mordatch, I., 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528.* 10.48550/arXiv.1909.07528.

Bell, K., Gill, S., 2018. Delivering a highly distributed electricity system: Technical, regulatory and policy challenges. *Energy Policy* 113, 765–777. 10.1016/j.enpol.2017.11.039.

Consentec, 2020. Description of the Balancing Process and the Balancing Markets in Germany, Aachen. https://www.regelleistung.net/ext/download/marktbeschreibung_cons, retrieved October 19, 2020.

Cruz, M.R.M., Fitiwi, D.Z., Santos, S.F., Catalão, J.P.S., 2018. A comprehensive survey of flexibility options for supporting the low-carbon energy future. *Renewable and Sustainable Energy Reviews* 97, 338–353. 10.1016/j.rser.2018.08.028.

Deilami, S., Masoum, A.S., Moses, P.S., Masoum, M.A.S., 2011. Real-Time Coordination of Plug-In Electric Vehicle Charging in Smart Grids to Minimize Power Losses and Improve Voltage Profile. *IEEE Trans. Smart Grid* 2 (3), 456–467. 10.1109/TSG.2011.2159816.

Diamantoulakis, P.D., Kapinas, V.M., Karagiannidis, G.K., 2015. Big Data Analytics for Dynamic Energy Management in Smart Grids. *Big Data Research* 2 (3), 94–101. 10.1016/j.bdr.2015.03.003.

Dulac-Arnold, G., Mankowitz, D., Hester, T., 2019. Challenges of Real-World Reinforcement Learning. http://arxiv.org/pdf/1904.12901v1.

Ecker, M., Nieto, N., Käbitz, S., Schmalstieg, J., Blanke, H., Warnecke, A., Sauer, D.U., 2014. Calendar and cycle life study of Li(NiMnCo)O2-based 18650 lithium-ion batteries. *Journal of Power Sources* 248, 839–851. 10.1016/j.jpowsour.2013.09.143.

Fujimoto, S., van Hoof, H., Meger, D., 2018. Addressing Function Approximation Error in Actor-Critic Methods. http://arxiv.org/pdf/1802.09477v3.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press, Cambridge, Massachusetts, London, England, 785.

Gosavi, A., 2004. Reinforcement learning for long-run average cost. *European Journal of Operational Research* 155 (3), 654–674. 10.1016/S0377-2217(02)00874-3.

Greenwood, D.M., Lim, K.Y., Patsios, C., Lyons, P.F., Lim, Y.S., Taylor, P.C., 2017. Frequency response services designed for energy storage. *Applied Energy* 203, 115–127. 10.1016/j.apenergy.2017.06.046.

Gronauer, S., Diepold, K., 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review* 55 (2), 895–943. 10.1007/s10462-021-09996-w.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. http://arxiv.org/pdf/1801.01290v2.

Hao, H., Wu, D., Lian, J., Yang, T., 2018. Optimal Coordination of Building Loads and Energy Storage for Power Grid and End User Services. *IEEE Transactions on Smart Grid* 9 (5), 4335–4345. 10.1109/TSG.2017.2655083.

Henry, R., Ernst, D., 2021. Gym-ANM: Reinforcement learning environments for active network management tasks in electricity distribution systems. *Energy and AI* 5, 100092. 10.1016/j.egyai.2021.100092.

Hubbs, C., 2016. Deep Reinforcement Learning and Hyperparameter Tuning: Using Ray’s Tune to Optimize your Models. https://towardsdatascience.com/deep-reinforcement-learning-and-hyperparameter-tuning-df9bf48e4bd2, retrieved November 6, 2020.

Huck, N., 2019. Large data sets and machine learning: Applications to statistical arbitrage. *European Journal of Operational Research* 278 (1), 330–342. 10.1016/j.ejor.2019.04.013.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research* 281 (3), 628–641. 10.1016/j.ejor.2019.09.018.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521 (7553), 436–444. 10.1038/nature14539.

Li, Y., 2018. Deep Reinforcement Learning. http://arxiv.org/pdf/1810.06339v1.

Liessner, R., Schmitt, J., Dietermann, A., Bäker, B., 2019. Hyperparameter Optimization for Deep Reinforcement Learning in Vehicle Energy Management, in: Proceedings of the 11th International Conference on Agents and Artificial Intelligence, Prague, Czech Republic. 2/19/2019 - 2/21/2019. SCITEPRESS - Science and Technology Publications, 134–144.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning, 10. http://arxiv.org/pdf/1509.02971v6.

Madlener, R., 2022. Smart Grid Economics, in: Dinther, C., Flath, C.M., Madlener, R. (Eds.), Smart Grid Economics and Management, 1st Ed. Springer, Berlin, New York, Heidelberg, 21–59.

Madlener, R., Specht, J.M., 2018. Business Opportunities and the Regulatory Framework, in: Hester, R.E., Harrison, R.M. (Eds.), Energy storage options and their environmental impact. Royal Society of Chemistry, Cambridge, 296–326.

Mantovani, R.G., Rossi, A.L.D., Alcobaça, E., Vanschoren, J., Carvalho, A.C.P.L.F. de, 2019. A meta-learning recommender system for hyperparameter tuning: Predicting when tuning improves SVM classifiers. *Information Sciences* 501, 193–221. 10.1016/j.ins.2019.06.005.

Nakabi, T.A., Toivanen, P., 2021. Deep reinforcement learning for energy management in a microgrid with flexible demand. *Sustainable Energy, Grids and Networks* 25. 10.1016/j.segan.2020.100413.

Ning, C., You, F., 2019. Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming. *Computers & Chemical Engineering* 125, 434–448. 10.1016/j.compchemeng.2019.03.034.

Nykamp, S., Molderink, A., Bakker, V., Toersche, H.A., Hurink, J.L., Smit, G.J.M., 2012. 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, Berlin, 1–8.

OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H.P.d.O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., Zhang, S., 2019. Dota 2 with Large Scale Deep Reinforcement Learning. http://arxiv.org/pdf/1912.06680v1.

Pierson, H.A., Gashler, M.S., 2017. Deep learning in robotics: a review of recent research. *Advanced Robotics* 31 (16), 821–835. 10.1080/01691864.2017.1365009.

Ransbotham, S., Kiron, D., Gerbert, P., Reeves, M., 2017. Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review* 59 (1), 1-17.

Ruester, S., Pérez-Arriaga, I., Schwenen, S., Batlle, C., Glachant, J.-M., 2013. From Distribution Networks to Smart Distribution Systems: Rethinking the Regulation of European Electricity DSOs: Final Report.

Sadler, M., Regan, N., 2019. Game Changer. *New in Chess*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. http://arxiv.org/pdf/1707.06347v2.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (7587), 484–489. 10.1038/nature16961.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D., 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. http://arxiv.org/pdf/1712.01815v1.

Simonini, T., 2019. On Choosing a Deep Reinforcement Learning Library. https://medium.com/data-from-the-trenches/choosing-a-deep-reinforcement-learning-library-890fb0307092, retrieved November 4, 2020.

Specht, J.M., Madlener, R., 2019. Energy Supplier 2.0: A conceptual business model for energy suppliers aggregating flexible distributed assets and policy issues raised. *Energy Policy* 135. 10.1016/j.enpol.2019.110911.

Specht, J.M., Madlener, R., 2020. Quantifying Value Pools for Distributed Flexible Energy Assets: A Mixed Integer Linear Optimization Approach. *FCN Working Paper* No. 20/2019, November (revised August 2020).

Sutton, R.S., Barto, A., 2018. Reinforcement learning: An introduction. The MIT Press, Cambridge, MA, London, 526.

Tu, C., He, X., Shuai, Z., Jiang, F., 2017. Big data issues in smart grid – A review. *Renewable and Sustainable Energy Reviews* 79, 1099–1107. 10.1016/j.rser.2017.05.134.

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

Wang, J., Ma, Y., Zhang, L., Gao, R.X., Wu, D., 2018. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems* 48, 144–156. 10.1016/j.jmsy.2018.01.003.

FCN | Future Energy Consumer Needs and Behavior | E.ON Energy Research Center | RWTHAACHEN UNIVERSITY

List of the latest FCN Working Papers

2021

Fabianek P., Glensk B., Madlener R. (2021). A Sequential Real Options Analysis for Renewable Power-to-Hydrogen Plants in Germany and California, FCN Working Paper No. 1/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, January.

Fabianek P., Schrader S., Madlener R. (2021). Techno-Economic Analysis and Optimal Sizing of Hybrid PV-Wind Systems for Hydrogen Production by PEM Electrolysis in California and Germany, FCN Working Paper No. 2/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, February.

Karami M., Madlener R. (2021). Business Models for Peer-to-Peer Energy Trading in Germany Based on Households' Beliefs and Preferences, FCN Working Paper No. 3/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, March.

Rasool M., Khalilpur K., Rafiee A., Karimi I., Madlener R. (2021). Evaluation of Alternative Power-to-Chemical Pathways for Renewable Energy Exports, FCN Working Paper No. 4/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, May.

Madlener R., Sheykhha S., Briglauer W. (2021). The Electricity- and CO_2-Saving Potentials Offered by Regulation of European Video-Streaming Services, FCN Working Paper No. 5/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, May.

Specht M., Madlener R. (2021). Deep Reinforcement Learning for Optimized Operation of Renewable Energy Assets, FCN Working Paper No. 6/2021, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, September (revised April 2021).

2020

Klie L., Madlener R. (2020). Optimal Configuration and Diversification of Wind Turbines: A Hybrid Approach to Improve the Penetration of Wind Power, FCN Working Paper No. 1/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, January.

Klie L., Madlener R. (2020). Concentration Versus Diversification: A Spatial Deployment Approach to Improve the Economics of Wind Power, FCN Working Paper No. 2/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, February (revised May 2021).

Madlener R. (2020). Small is Sometimes Beautiful: Techno-Economic Aspects of Distributed Power Generation, FCN Working Paper No. 3/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, March.

Madlener R. (2020). Demand Response and Smart Grid Technologies. FCN Working Paper No. 4/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, March.

Vartak S., Madlener R. (2020). On the Optimal Level of Microgrid Resilience from an Economic Perspective, FCN Working Paper No. 5/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, April.

Hellwig R., Atasoy A.T., Madlener R. (2020). The Impact of Social Preferences and Information on the Willingness to Pay for Fairtrade Products, FCN Working Paper No. 6/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, May.

Atasoy A.T., Madlener R. (2020). Default vs. Active Choices: An Experiment on Electricity Tariff Switching, FCN Working Paper No. 7/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, May.

Sheykhha S., Madlener R. (2020). The Role of Flexibility in the European Electricity Market: Insights from a System Dynamics Perspective, FCN Working Paper No. 8/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, June.

Wolff S., Madlener R. (2020). Willing to Pay? Spatial Heterogeneity of e-Vehicle Charging Preferences in Germany, FCN Working Paper No. 9/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, June.

Priesmann J., Spiegelburg, Madlener R., Praktiknjo A. (2020). Energy Transition and Social Justice: Allocation of Renewable Energy Support Levies Among Residential Consumers in Germany, FCN Working Paper No. 10/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, July.

Sabadini F., Madlener R. (2020). The Economic Potential of Grid Defection of Energy Prosumer Households in Germany, FCN Working Paper No. 11/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, August.

Schlüter P., Madlener R. (2020). A Global Renewable Energy Investment and Funding Model by Region, Technology, and Investor Type, FCN Working Paper No. 12/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, September.

Ghafuri F., Madlener (2020). A Hybrid Modeling Approach for the Optimal Siting of Mobile Battery-Enhanced Fast-Charging Stations, FCN Working Paper No. 13/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, October.

Ghafuri F., Madlener (2020). A Real Options Analysis of the Investment in Mobile Battery-Enhanced Fast-Charging Stations, FCN Working Paper No. 14/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, October.

Ghafuri F., Madlener (2020). A Virtual Power Plant Based on Mobile Battery-Enhanced Fast-Charging Stations, FCN Working Paper No. 15/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, October.

Saunders H., Roy J., Azevedo I.M.L., et al. (2020). Energy Efficiency: What has it Delivered in the Last 40 Years? Working Paper No. 16/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, November (revised April 2021).

Wimmers A., Madlener R. (2020). The European Market for Guarantees of Origin for Green Electricity: A Scenario-Based Evaluation of Trading under Uncertainty, FCN Working Paper No. 17/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, December.

Walter A., Held M., Pareschi G., Pengg H., Madlener R. (2020). Decarbonizing the European Automobile Fleet: Impacts of 1.5 °C-Compliant Climate Policies in Germany and Norway, FCN Working Paper No. 18/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, December.

Toussaint M., Madlener R. (2020). Mind the Gap! Economic Impacts of Energy Efficiency Policy Targeting Single-Family Residential Buildings in Germany, FCN Working Paper No. 19/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, December.

Toussaint M., Madlener R. (2020). Closing the Gap? A Welfare Analysis of Energy Efficiency Policy Targeting Single-Family Residential Buildings in Germany, FCN Working Paper No. 20/2020, Institute for Future Energy Consumer Needs and Behavior, RWTH Aachen University, December.

FCN Working Papers have been published since 2008 and are free of charge. They can mostly be downloaded in pdf format from the FCN / E.ON ERC Website (www.eonerc.rwth-aachen.de/fcn) and the SSRN Website (www.ssrn.com), respectively. Alternatively, they may also be ordered as hardcopies from Ms Sabine Schill (Phone: +49 (0) 241-80 49820, E-mail: post_fcn@eonerc.rwth-aachen.de), RWTH Aachen University, Institute for Future Energy Consumer Needs and Behavior (FCN), Chair of Energy Economics and Management (Prof. Dr. Reinhard Madlener), Mathieustrasse 10, 52074 Aachen, Germany.