applied sciences

MDPI

Article

Smart Grid for Industry Using Multi-Agent Reinforcement Learning

Martin Roesch *, Christian Linder, Roland Zimmermann, Andreas Rudolf, Andrea Hohmann and Gunther Reinhart

Fraunhofer Research Institution for Casting, Composite and Processing Technology, 86159 Augsburg, Germany; christian.linder@igcv.fraunhofer.de (C.L.); roland.zimmermann@igcv.fraunhofer.de (R.Z.); andreas.rudolf@igcv.fraunhofer.de (A.R.); andrea.hohmann@igcv.fraunhofer.de (A.H.); gunther.reinhart@igcv.fraunhofer.de (G.R.)

* Correspondence: martin.roesch@igcv.fraunhofer.de; Tel.: +49-821-9067-8142

Received: 19 August 2020; Accepted: 29 September 2020; Published: 1 October 2020

Abstract: The growing share of renewable power generation leads to increasingly fluctuating and generally rising electricity prices. This is a challenge for industrial companies. However, electricity expenses can be reduced by adapting the energy demand of production processes to the volatile prices on the markets. This approach depicts the new paradigm of energy flexibility to reduce electricity costs. At the same time, using electricity self-generation further offers possibilities for decreasing energy costs. In addition, energy flexibility can be gradually increased by on-site power storage, e.g., stationary batteries. As a consequence, both the electricity demand of the manufacturing system and the supply side, including battery storage, self-generation, and the energy market, need to be controlled in a holistic manner, thus resulting in a smart grid solution for industrial sites. This coordination represents a complex optimization problem, which additionally is highly stochastic due to unforeseen events like machine breakdowns, changing prices, or changing energy availability. This paper presents an approach to controlling a complex system of production resources, battery storage, electricity self-supply, and short-term market trading using multi-agent reinforcement learning (MARL). The results of a case study demonstrate that the developed system can outperform the rule-based reactive control strategy (RCS) frequently used. Although the metaheuristic benchmark based on simulated annealing performs better, MARL enables faster reactions because of the significantly lower computation costs for its own execution.

Keywords: smart grid; multi-agent reinforcement learning; energy flexibility; production control

1. Introduction

In order to mitigate the effects of anthropological climate change, efforts are being made worldwide to reduce greenhouse gas emissions (GHG) and increase the share of renewable energies. One of the first industrial countries to do so, Germany announced in 2010 the plan to reduce GHG by 80% by 2050 along with plans to generate 80% of total electricity using renewable sources [1]. So far, in 2019, 42% of the total consumed electricity was generated by renewable sources. However, this trend has also entailed some challenging side effects. In particular, the high costs for new renewable-energy power plants has led to rising electricity prices for industrial companies by 170% compared to the price level in 2000 [2]. In addition, electricity prices have been increasingly fluctuating, especially in the short term (see Figure 1a) since power generation using renewable sources, e.g., wind or solar, is subject to sudden changes in weather and is, therefore, neither controllable nor easily predictable.

Industrial companies are pressured to minimize electricity costs, which can be achieved in three ways. First, energy efficiency can be increased in order to reduce total energy consumption. Second,

manufacturers can reduce energy costs by means of energy flexibility, i.e., adapting the energy demand of a factory to volatile energy market prices, thus consuming less energy in times of high prices and vice versa [3]. Third, power self-generation can also contribute to reducing energy costs, especially when using solar as well as combined-heat and power plants (CHP) (see Figure 1b). Moreover, energy flexibility can be further increased by using stationary batteries. The prices of these systems have decreased sharply in recent years (see Figure 1c) [4] and, as a consequence, the capacity of installed large battery storages in Germany increased by 1500% between 2015 and 2018 [5].

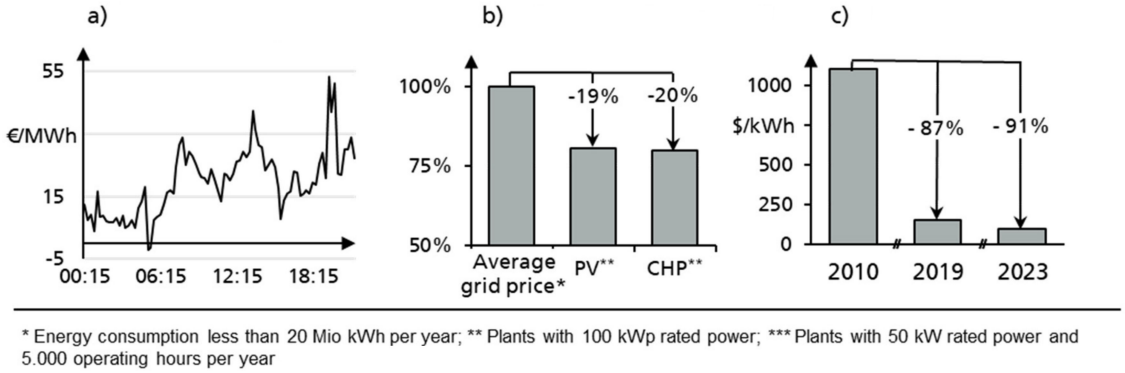

Figure 1. Potentials of energy cost reduction for industrial companies [2,5–8]. (**a**) Exemplary price history of 25 May 2020 at Intraday-Market [7]. (**b**) Exemplary cost of advantage of electricity self-supply [2,8,9]. (**c**) Price trend for battery packs [6].

In order to benefit from the aforementioned aspects and reduce energy costs, a suitable, integrated control approach is required. The approach adopted needs to consider the power consumptions of the manufacturing system as well as the power supply using batteries, self-generation plants, and market trading [9]. This paradigm of locally balancing power supply and demand is widely known as a smart grid [10]. Thereby, an intelligent control strategy for efficient electricity usage is derived, based on online date from smart sensors like energy meters. On the manufacturing side, it is the task of production control to schedule manufacturing jobs in the short term and thus ensure logistic objectives like due time or throughput while also determining the energy demand of the production system [11]. As a result, energy costs can represent an additional objective for production control. On the power supply side, the battery, self-generation, and market trading should be controlled. Since every element regarded underlies stochastic impacts like machine breakdowns, generation forecast errors, or changing prices, the system needs to react quickly to such unforeseen events [12]. Regarding the resulting complex optimization problem, multi-agent reinforcement learning (MARL) provides high potential in this field because of its short reaction time and fair solution quality [13]. MARL has already been applied to both production control (e.g., [14–16]) and smart grid approaches to non-industrial applications (e.g., [17,18]).

This paper presents an energy and manufacturing system which consists of a stationary battery, power plants for self-consumption, short-term electricity trading, and various production resources. The relevant system elements and resulting vision for a control system in the context of smart grid for industry is shown in Figure 2. Given that common production control systems struggle with the resulting complexity, a novel MARL-based approach will be presented here. The fundamentals are then briefly pointed out in the next section, which is followed by a short literature review. The MARL approach will be proposed therein and validated using a case study. The paper will then end with a brief summarizing conclusion.

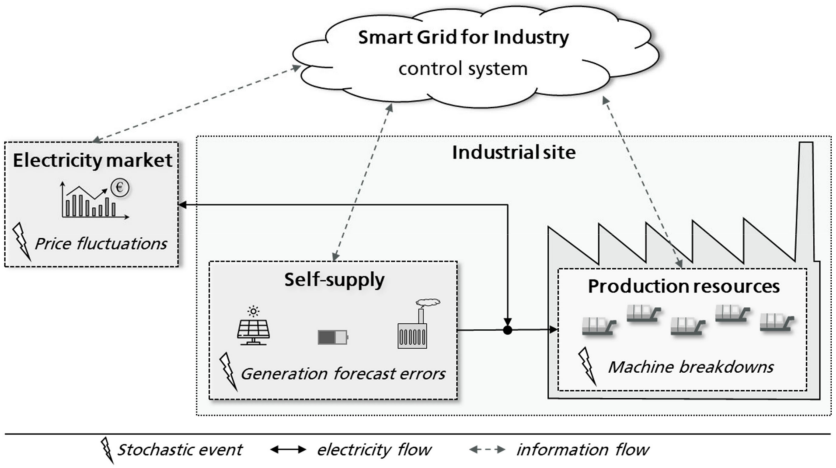

Figure 2. Vision of the smart grid system for industrial companies.

2. Fundamentals

The essential fundamentals will be briefly introduced in the following for further comprehension. The fields of industrial electricity supply as well as approaches to production control and reinforcement learning are outlined for this purpose.

2.1. Industrial Electricity Supply

Based on German regulations, there are several options industrial companies can pursue in order to arrange their power supply. These options will be briefly outlined.

- **Power procurement [19,20]:** In this case, the manufacturing companies buy the required amount of electricity from external sources. The general billing interval of the entire electricity market is 15 min. There are two main options for procurement:
  - **Utility company:** Industrial companies may directly rely on power utility companies, which in general provide constant prices, thus completely bearing the price risks.
  - **Market trading:** Companies can actively take part in the electricity market, either trading directly or relying on a suitable aggregator. There are several electricity markets which can be characterized by their lead time. It is worth noting that there is always some lead time between order purchasing and the actual delivery and consumption of the electricity (see Figure 3). The shortest lead time is offered by the intraday market, in which electricity can be bought at least 5 min before delivery. In case a company has bought electricity within a specified interval, it must ensure that the power is actually consumed. In general, a tolerance bandwidth of ±5 to 10% is granted. Otherwise, there is a risk of receiving high penalties.

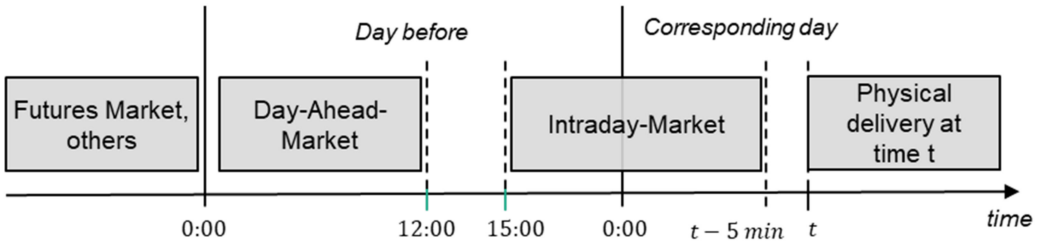

Figure 3. Available short-term energy markets in Germany, depending on the time of delivery t [20].

- **Self-supply [21,22]:** Manufacturing companies may dispose of their own power generation facilities, which enable them to generate their electricity on-site. Depending on the relevant control characteristics, power plants can be divided into two groups:
  - **Variable power sources (VPS):** The power generation of these sources is difficult to control because it is highly dependent on the current weather conditions. The only control option is to cut off their power. However, since VPS do not entail any working expenses, this is not recommended. The most important kinds of VPS are solar and wind power plants.
  - **Controllable power sources (CPS):** The power output of CPS can be controlled and adapted within a specified range. In the manufacturing field, combined heat and power plants (CHP) are widely used and are assigned to this category.
- **Battery systems:** Batteries consist of electrochemical, rechargeable cells and are very suitable for storing electricity for several hours or days [23]. However, the cells suffer from degradation over time. The extent of the degradation strongly depends on the charging cycles the battery is exposed to [24]. As a result, there exist several modelling approaches for considering the degradation process within a battery control strategy [25].

2.2. Approaches for Production Control

The main task of production control is to ensure the manufacturing of the predefined jobs by the production resources, although stochastic events like machine break-downs can occur [11]. Therefore, production control needs to make decisions and react to unforeseen events within a few minutes or seconds. Although the jobs are mostly assigned to resources in advance, production control determines the time and order sequence of the various jobs to be manufactured in every resource. In general, there exist two production control strategies for accomplishing this task, and these are briefly outlined in the following [26,27]:

- **Reactive control strategy:** The jobs are reactively scheduled based on simply dispatching rules like first-in-first-out (FIFO) or earliest-deadline-first (EDF). After finishing a job, the next job is selected, so no fixed production schedule is determined. In doing so, decisions can be made very quickly, whereas the solution quality is limited.
- **Predictive-reactive control strategy:** A deterministic schedule is calculated for a specified production period so as to optimize the given objectives. However, every stochastic event during the production period requires a schedule update, which makes this approach computationally expensive.

Moreover, multi-agent systems (MAS) display a further control strategy and are mainly characterized by their decentralized system architecture, so they cannot be clearly assigned to either one of the aforementioned categories [26]. In most cases of MAS, physical production resources like machines are represented by an individual agent, which pursues given goals and thereby either cooperates or competes with other agents [28]. Although classic MAS approaches are very fault tolerant and robust, they often provide a low solution quality because agents can only provide local optimization and can hardly pursue long-term goals [29]. However, these downsides can be surpassed using multi-agent reinforcement learning (MARL). A basic understanding of MARL is provided in the following.

2.3. Multi-Agent Reinforcement Learning

In reinforcement learning (RL), the focus lies on finding a control policy that is able to achieve a given goal. In this context, RL is formalized as a Markov decision process and builds upon iterative learning. During this process, a RL agent selects an action a_t based on the current policy and state s_t, which is executed in a given environment. Afterwards, the environment transitions into a new state s_{t+1} and the agent receives a reward r_{t+1} (see Figure 4). By interacting with the environment, the policy

is iteratively updated using a RL method to maximize the long-term reward. Various methods are used to compute the optimal policy, value-based and policy-based approaches being the most common ones. Regarding complex problems, neural networks are widely applied in order to predict the optimal policy or value function [30].

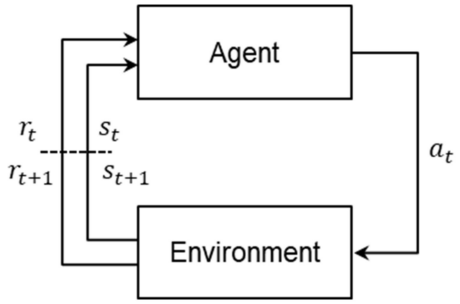

Figure 4. Reinforcement learning paradigm.

For large problems, the RL approach can be extended to several agents, all of which are interacting in the same environment in a collaborative or competitive manner, which then is known as multi-agent reinforcement learning. In the case of collaborative agents, cooperative behavior can be reinforced by giving all agents a common reward. They thus learn to maximize this common reward together. One of the main challenges thereby is the so-called credit assignment problem [31]: In a multi-agent environment, it may be difficult for agents to assess their individual contribution to globally assigned rewards, thus making agents struggle to optimize their own policy.

RL has proven able to solve very large and complex problems, e.g., Alpha Go [32]. Although a noteworthy computational effort is needed for training, the policy can then be executed within seconds while using little computation effort. However, RL—and especially MARL—require a detailed model of the environment. In addition, aspects like the reward assignment, the state space of the agents, and the hyper parameter of the learning algorithm have a strong impact on the solution quality [13].

3. Literature Review

In the context of considering energy aspects within production control, a large number of researchers have focused on increasing the energy efficiency and flexibility in production sites. In the following analysis, the focus will lie on approaches for an energy-oriented production control, which also include the energy supply side. However, only approaches with a focus on energy flexibility aspects will be considered, since this objective is crucial to exploiting fluctuating electricity prices. In this context, production control indicates an ability to rapidly react to events. Approaches for single-machine systems as well as publications in the field of smart grids used for non-industrial buildings will also be excluded from the analysis.

In one of the first approaches, [33] presents a reactive production control method used to adapt the energy consumption of a production system to variable prices. Material buffers and flexibility in personnel and shift planning are used thereby. The proposed system in [34] also takes advantage of material buffers in order to temporarily reduce electricity consumption while also maintaining a given throughput. The measures are implemented and selected based on predefined rules, so this approach can be assigned to the domain of reactive control strategies. The method presented in [35] addresses both energy flexibility and efficiency aspects while applying a mathematical model for predictive-reactive job scheduling. Regarding short-term reactions, there exists an additional rule-based logic used to temporarily adapt the energy consumption, e.g., by shutting down processes. Reactive dispatching rules considering the available amount of energy and the job due time are the central element in [12]. The approach aims at meeting a given energy availability within some tolerance while controlling a production system. In addition, variable spot market prices and self-supply based on renewable

power plants are considered. Moreover, a predictive-reactive, short-term load-management for peripheral energy consumers like heating or ventilation is applied. Reference [36] presents another reactive approach to energy-oriented production control in order to exploit volatile prices. However, no market constraints, e.g., the required due time, are considered.

In order to adapt short-term energy consumption to volatile prices and to changing power availability from renewable sources, ref [37] presents a scheduling approach based on mixed integer programming. In this case, the billing interval was set to one hour, which reduces complexity when compared to the widely used 15 min. A system considering volatile prices, self-supply, and a battery storage is introduced in [38]. A two-stage optimization strategy was developed for this purpose using a limited application for complex systems due to the increasing computational power demand. The focus of [39] lies on the integration of volatile energy prices into the short-term decision making for job scheduling. Here again, a predictive-reactive algorithm is developed, whereas the computational expense increases exponentially with a growing problem size. A MAS used to decrease energy costs based on intelligent scheduling and volatile prices is developed in [40]. The agents are bargaining for jobs, thus minimizing energy costs, production costs, and throughput time.

Figure 5 summarizes the analysis of the contributions considered and unveils the central gaps concerning electricity supply as follows. First, volatile prices from the public grid have been widely integrated into production control approaches. However, the actual organizational and technical boundaries of the short-term electricity market (i.e., the required lead time and market fees) have not yet been considered in the context of production control. As a result, market trading has only been partially considered by the literature in question. Second, the potential electricity supply options—short-term market, self-supply, and battery systems—have not yet been integrated together into one holistic system. Additionally, battery systems have only been regarded in one approach [38], which uses a simple battery model that does not include cycle-dependent battery degradation mechanisms. However, bringing all of the electricity supply options together is a promising way of efficiently reducing electricity costs.

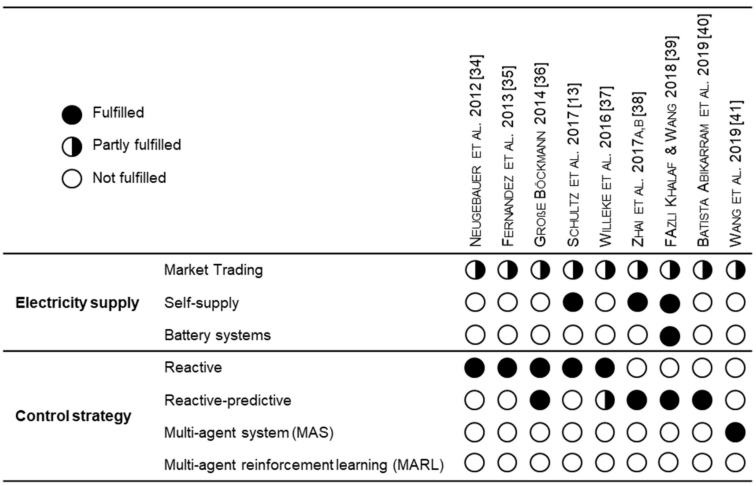

Figure 5. Literature review of energy-oriented production control.

On the control strategy side, mainly reactive and reactive-predictive approaches are applied. It is worth noting that the analyzed reactive-predictive approaches indicate the computation effort and the limited ability for solving large-scale problems. As a result, a joint optimization of both all relevant electricity supply options and production resources is rarely possible when using the applied control approaches because reactive approaches are not suitable for such complex systems, and reactive-predictive strategies require a high level of computational effort, thus contradicting the requirement for short-term reactions. Although a bargaining MAS is developed in [40], RL has not been used in this field of research, but it would be promising given the ability to handle stochastic events and short-term decisions.

Therefore, the main goal of this work is the integration of the three relevant options of electricity supply—a battery system, self-supply, and trading at the short-term market to use volatile prices—into a system for energy-oriented production control. The literature review shows that a system able to combine all these elements and at the same time react in real-time has yet not been developed, and to do so, a new control strategy has to be derived. In doing so, a novel smart grid for an industrial production site is created that optimizes both the energy demand and supply sides and production costs simultaneously. The solution strategy is based on MARL, and it exhibits a promising and novel approach in this given field of research.

4. Proposed MARL Approach

Using MARL as a solution approach has a large impact on the system architecture. In the following, first a system overview is given before the sub-models of the environment are introduced briefly. Using this as a basis, the resulting state and action space of the MARL agents will be described, and the assignment of the complex reward function will be illustrated. Finally, the selected training procedure will be outlined.

4.1. System Overview

The central elements of the developed approach are the RL-enhanced control agents, the complex reward function, and the environment in which the agent can act, gain experience, and learn. Three types of agents exist in this context, and they are characterized by their specific tasks. Whereas the market agent executes the trading of electricity on the short-term market, the battery agent controls the charging and discharging power of the battery system. Every physical production resource with the task to manufacture production jobs, e.g., a machine, is represented by a resource agent. Consequently, there may be several resource agents operating in parallel within one system and defining the manufacturing time and job sequence of the specific production resource they represent.

The agents interact with the environment by selecting actions. In this context, the considered options of energy supply and the production system, including all technical and organizational restrictions, are modelled as the environment. The new state of the environment that results from the agents' actions as well as the impacts the actions entailed can thus be computed. Based on the latter, the reward function assigns the reward to every agent. Due to the system complexity and the different agent types, there are mainly two different types of rewards. The first relates to the energy costs, which are modelled as a global reward that all types of agents can receive. In contrast, the production costs are a reward only assigned to the resource agents, respectively. Figure 6 displays the resulting system structure.

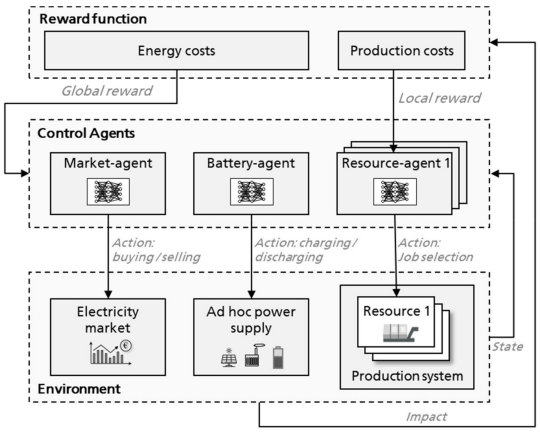

Figure 6. System structure.

4.2. Environment

As shown in Figure 6, the environment consists of three main elements: the electricity market, the ad-hoc power supply, and the production system. These elements will be successively outlined in the following.

4.2.1. Electricity Market

For industrial companies, several markets exist for potential participation (see Figure 2). Since the presented system focuses on the short-term control of a production site, the possible market participation is limited to the intraday market due to its lead time of only five minutes. In the other markets, electricity can be traded only at least one day in advance, so these markets are out of scope for the system in question. However, power which has already been purchased in advance on any of the aforementioned markets (hereinafter referred to as "prior purchased power") is an input into the system and must be consumed.

On the intraday market, the prices change dynamically over time based on the relevant supply and demand. The result is a high level of price uncertainty while trading. In addition, electricity can be purchased during every 15 min time interval, which is more than five minutes in the future. In order to reduce the complexity and to avoid modelling of price uncertainty, market trading is only limited to the next billing interval. This means that the market agent can only buy or sell electricity in the consecutive billing interval and at a defined time of nine minutes before the actual beginning of the corresponding billing interval. There thus exists a time buffer of four minutes in order to ensure that the trading is transacted within the required five minutes of lead time. A proportional order fee is modelled for every trade.

4.2.2. Ad-hoc Power Supply

Power plants used for electric self-supply (VPS and CPS) and batteries are merged into the ad-hoc energy supply. What they all have in common is that their power generation can be adapted on an ad-hoc basis with no actual lead time. An additional factor displays the public grid in which load deviations can be settled. In the following, the modelled sub-elements will be described before the resulting coaction is derived.

Regarding self-supply, VPS as well as CPS are modelled. As mentioned earlier, VPS do not entail any operational costs. Consequently, the resulting costs for a company do not differ whether electricity from a VPS is used or not. Consequently, electricity from this source is modelled as free with respect to price. In contrast, CPS (and especially CHP) rely on fuel and thus directly result in operational costs. Moreover, the efficiency of CHP decreases when the plant is not being operated at nominal capacity [41]. For the sake of simplification, this correlation is assumed to be linear, using an efficiency factor which displays the maximum efficiency loss when the CHP is operating at minimal power. Technically, the maximal power generation of VPS is limited by the current weather conditions and can only be cut off, whereas CPS can be controlled within a specific range.

The charging and discharging power of batteries is limited by the nominal storage capacity and C-rate. The latter indicates the minimal time in hours for completely charging or discharging. In the context of this approach, the battery charging and discharging power is assumed to be constant within a three-minute time period. In addition, there is a specific storage efficiency to be considered. Since the degradation of a battery mainly depends on its cycles, the battery cell life will be affected by the charging strategy. In the relevant context of a stationary battery (due to volatile power availability and consumptions), a flexible battery control strategy resulting in irregular cycles is expected. Therefore, the degradation model presented in [42] is applied, which allows the assessment of the consequences of irregular cycles for the degradation of the battery. Thus, after every work shift, the resulting battery degradation can be calculated, and the working life wear can be transformed into costs on the basis of overall battery purchase costs.

Although various options exist for adapting energy supply and demand, short-term gaps between power availability and demand can still occur. As an ultimate measure, this electricity shortfall can be compensated for via the public grid. However, this means that either more or less energy from the grid is being consumed than has actually been bought on the markets in advance, which can result in penalty costs in case a specific tolerance level is exceeded (see Section 2.1). The amount of the penalty costs depends on the expenses which the grid operators have to face due to this unexpected divergence and is thus determined individually for each event. Therefore, a linear function according to [43] is used to model the penalty costs.

The relevant options for an ad-hoc energy supply—self-supply via VPS and CPS, use of battery storage, and the public grid—need to be coordinated and are thus transferred to an operation scheme. Regarding the modelled control parameters (e.g., charging power) and costs structure (e.g., degradation or electricity production costs), the following key-findings can be extracted:

- The electricity bought from the intraday market and other prior purchased power should always be consumed in order to avoid penalty costs.
- Power from VPS should also be consumed since this does not result in any costs. Only in the case of excessively low power demand from the production system can VPS be temporarily cut off so as to avoid penalty costs for consuming less electricity than previously purchased.
- The remaining power demand which cannot be met by VPS and the electricity purchased from the market has to be provided by the battery and the CPS.
- Compensating for a deficit of power availability and demand with energy from the public grid should be avoided, as this can incur high costs.

Considering these key aspects, the control mechanism of the considered ad-hoc power supply elements works according to the following four guidelines, which are executed in every time step:

1. Calculation of the charging/discharging power of the battery system based on RL (Battery Agent), whereas the total state of all elements in the power supply (battery, VPS, CPS, prior purchased power) and the current demand from production side is considered.
2. The gap between expected power consumption and available power from VPS, battery power as well as purchased electricity from the markets is settled by CPS as much as possible.
3. If there is more energy available than the expected demand, cut-off power generation of VPS to balance demand and supply as much as possible.
4. The remaining deficit is settled by the public grid.

Consequently, the ad-hoc power consumption is based on a single RL agent used to control the battery. All other control operations are based on simple rules. The control system complexity can be significantly reduced in this way. However, in order to minimize total energy costs, the battery agent should learn to adapt the charging and discharging of the battery in a way that the CPS can still optimally be used. Thus, the battery agent indirectly controls the CPS power generation.

4.2.3. Production System

The applied production system model is based on the approach developed in the preliminary work [44]. A defined number of jobs, which are found in the job queue of every resource in the beginning of a manufacturing shift, is assigned to every production resource, e.g., a manufacturing machine. The jobs can be grouped into job types, with jobs of one type having the same machining time and average power demand. However, every job has an individual due date, which lies within the current shift. Furthermore, every job type requires a different set-up state of the resource used for processing, so a type-individual set-up time with a specific energy demand has been modelled.

Regarding production costs, the time-dependent cost model in [45] has been used. In case a job is finished after the given due date, a linear cost-function considering the amount of delay is applied. In contrast, storage costs are implied in cases when the finishing time is earlier than the due date for

the job and outside of a certain time tolerance. In addition, two different kinds of stochastic events, rush jobs and machine breakdowns, are modelled, both of which occur arbitrarily during a shift. The machine breakdowns last for a defined number of time steps, which is not known in advance.

Every production resource is represented and controlled by a single resource agent. It is the task of every resource agent to select the next action after the previous action is finished. For resource agents, this means either choosing one of the jobs currently waiting in the job queue for processing, or going to a stand-by state. The latter entails only a little energy consumption, but no job can be processed during that time. Consequently, it may be beneficial for resources to temporarily go into a stand-by state in order to finish a job on time without incurring storage costs and, at the same time, reducing the energy demand during the stand-by period. If the production resource does not have the required set-up state for processing a selected order, a set-up process is automatically executed prior to the beginning of processing.

4.2.4. System Sequence

On the operational level, the system is discretized into consecutive time steps t with a length of three minutes. In this case, every 15 min billing interval is divided into five partial intervals, so there are several time steps for adjusting the energy consumption within one billing interval. The resulting system sequence is displayed in Figure 6. In the beginning of every time step, all of the resource agents which have completed their action will choose a new action. Based on this, the battery agent and the market agent select their actions simultaneously. Whereas the battery agent comes into action every partial interval, the market agent only acts every 15 min. This is due to the fact that electricity can only be purchased within 15 min intervals (see Section 2.1). In conclusion, it can be noted that, in total, all agents act in an asynchronous manner.

4.3. State and Action Space

The observation space, which represents the current state for every agent, is crucial in MARL. The optimal observation space for the given environment was derived via various simulation experiments based on a step-wise feature reduction. The following shape of the observation space indicates the best solutions of these examinations.

Especially with regard to cooperative behavior, it has proven to be beneficial for the agents to not only receive information on their own state, but also on the state of the cooperative agents and the entire environment. As a result, the agent is able to estimate and predict the behavior of the cooperative agents. However, the information on cooperative agents is not modeled in detail. Hence, the state of cooperative resource agents is abstracted. The total resulting observation spaces of resource, battery, and market agents are summarized in Table 1. For resource agents, their specific set-up state as well as the number and deadline of jobs in the queue in front of the resource (which are grouped by their job type) are important. The cooperative resource agents are, for all three types of agents, represented by the currently selected action and the remaining time steps for its completion, the next three upcoming deadlines of the jobs in the queue in front of the cooperative agent's resource, as well as the total number of waiting jobs and the resource specific energy demand. The state of the battery and market agents are briefly summarized by the features current state of charge (SoC) and the purchased electricity in the current billing interval. In addition, all agents perceive the current state of the environment, the current billing and partial interval, as well as the available VPS power, and the amount of prior purchased electricity. Furthermore, information about the predicted electricity price on the intraday market only proved to be beneficial for the battery and market agents.

Due to the required machining time of jobs, only a few resource agents select an action in a time step because some might still be processing jobs which are not finished yet. Given this fact, part of the required amount of electricity in the next interval can be estimated. Based on this consideration, the features of known power demand and the number of agents with known actions in the next steps are derived. They provide an indication of how much power will be needed in the short term,

and experimental results have proven the ability of this feature to improve the actions of resource agents. Since the battery agent selects an action after the resources do, the power demand can be derived based on the selected actions of all resource agents. Given that the market agent does not operate in the short term, but rather at nine-minute intervals, this feature is not a suitable part of its observation space.

Table 1. Agent-specific observation space.

Observation Feature					Agent		
Category			Name	Explanation	Resource	Battery	Market
Agents	Resource	Specific	Setup state	Current setup state of the resource	x		
			Job deadlines	Three next deadlines of every job type	x		
			Number of jobs	Number of jobs in queue for every job type	x		
		Cooperative	Action	Currently selected action	x	x	x
			Remaining steps	Remaining time steps for currently selected action	x	x	x
			Next deadlines	The three next upcoming job deadlines	x	x	x
			Number of jobs	Total number of jobs in queue	x	x	x
			Energy demand	Total energy demand of all jobs in queue	x	x	x
	Battery		SoC	Current Battery State of Charge	x	x	x
	Market		Purchased electricity	Purchased electricity in the current billing interval	x	x	x
Environment			Billing interval	Count of current billing interval	x	x	x
			Partial interval	Count of current partial interval	x	x	x
			VPS power	VPS power generation forecast for the next 2 h	x	x	x
			Prior purchased electricity	Prior purchased power for the next 2 h	x	x	x
			Electricity price	Electricity price forecast for the next hour		x	x
			Known power demand	Power demand of all resource agents, who have not finished the current action yet	x		
			Number of agents with known actions	Number of all resource agents, who have not finished the current action yet	x		
			Power demand	Total power demand in the current partial interval (only available, when all resource agents have already chosen their actions)		x	

A discrete action space is applied in order to model the actions of the agents. The resource agents can select whether to process a job of every job type or a stand-by step. The action-space of battery and market agent is based on a linear distribution of the maximum and minimum charging rate and allowed trading power, respectively. In case of bad actions, a penalty reward is awarded. This can happen when, e.g., a job type is selected which is actually currently not available in the queue, or when the selected charging rate and duration exceed the battery capacity.

4.4. Reward Calculation and Assignment

The reward function and assignment are crucial for the learning success in RL. Since the given environment consists of various elements and agents, there are several aspects to consider. In order to avoid the credit assignment problem, the reward function is partly divided into global and local rewards (see Figure 7). Whereas local rewards are only assigned to single agents, all agents are receiving global rewards. In addition, some parts of the reward are assigned immediately, and others are assigned at the end of a billing interval or training episode, which means the end of a shift for the case in question.

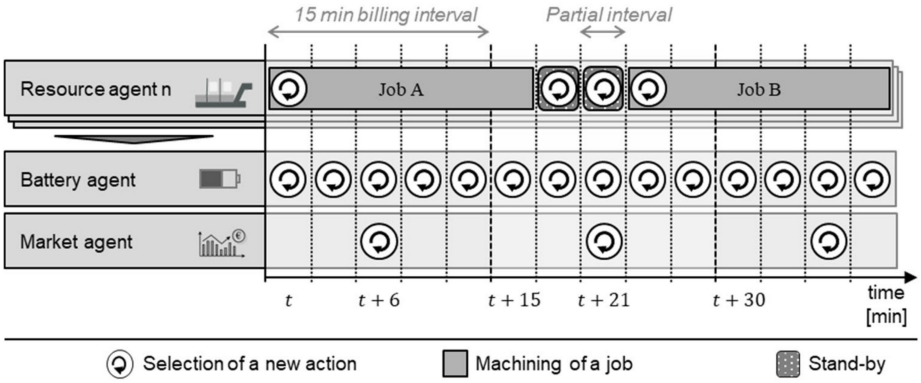

Figure 7. System sequence.

As displayed in Figure 6, there are two components of the reward function: Energy costs and production costs. As it is the overall system objective to minimize total costs, the assigned rewards consist of the negative costs that the agents aim to maximize.

Beginning with the resource agents, the time-dependent production costs of a job can only be influenced by the resource to which the specific job is assigned. As a result, these production rewards $R^{p}_{t,a}$ are granted as an individual and local reward to the resource agents p immediately after an action a is finished and consist based on [45] of time-dependent delay and storage costs $TC_n(t)$, setup costs SC_n, and labor costs LC_n. In case an action a lies outside the currently valid action set $\{A_{RA,t}\}$, a penalty reward r_{BA} is assigned:

$$R^{p}_{t,a} = \begin{cases} -\left(\sum TC_n(t) + \sum SC_n + LC_n\right), \; a \in \{A_{RA,t}\} \\ -\, r_{BA}\;, \; a \notin \{A_{RA,t}\} \end{cases} \tag{1}$$

Since the processing of a job can range over several billing intervals, it might not be possible to assign the specific electricity cost of a job to one time interval and, respectively, to a single action. This is why the resource agents receive the resulting total electricity reward R^{E}_{s} of the entire environment at the end of a training episode of length T. According to Section 4.2, the energy costs consist of CPS generation costs $C_{CPS}(t)$, penalty costs due to deficits settled by the public grid $C_{PC}(t)$, costs for intraday trading $C_I(t)$, and resulting battery degradation $C_{Bat,\,i}(t)$:

$$R^{E}_{s} = -\left(\sum_{t=0}^{T} C_{CPS}(t) + C_{PC}(t) + C_I(t) + C_{Bat,\,i}(t)\right) \tag{2}$$

In contrast, the electricity costs can be assigned to the battery agent directly at the end of every billing interval i. Since it takes actions for every partial interval t, the battery agent receives the reward inspired by Formula (2) after every fifth action. Again, the degradation costs, which are calculated with the battery degradation model and the charging gradient, are integrated into the energy costs.

The market agent buys or sells electricity for the subsequent billing interval nine minutes in advance. Thus, when the market agent chooses a new action, the effects of the previous action cannot be calculated because the respective billing interval is not finished yet. Due to this fact, the market agent receives the resulting electricity reward R^{E}_{s} stated in Formula (2) at the end of the episode, which means the shift respectively. In total, the reward is individually assigned to each agent, as displayed in Figure 8, in order to meet its specific needs and boundary conditions.

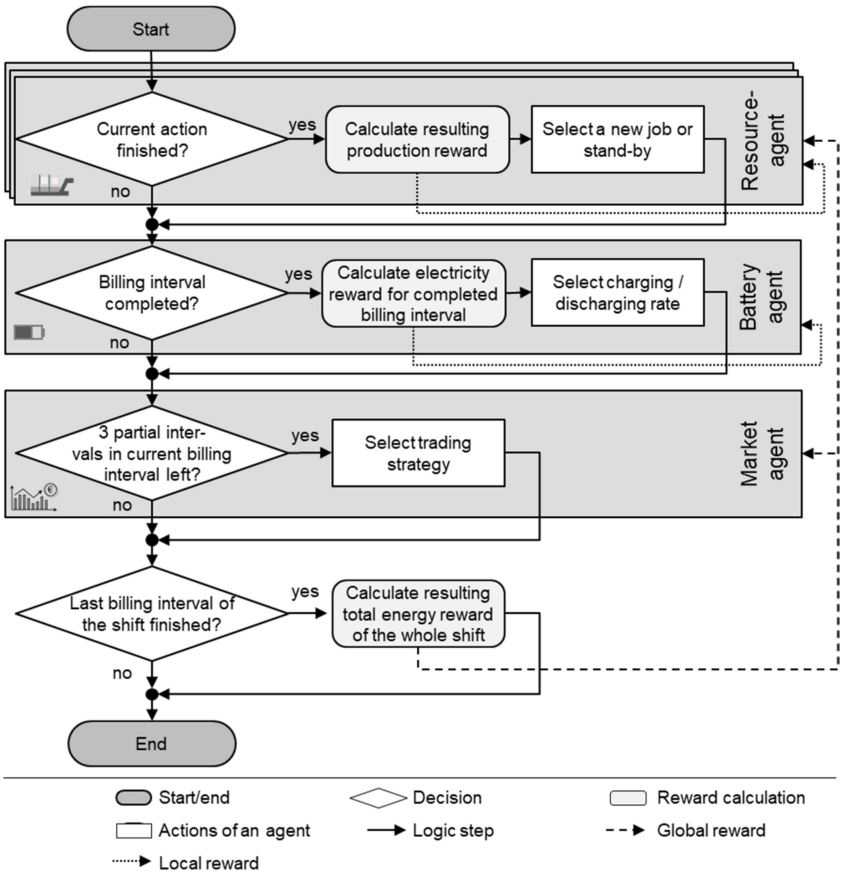

Figure 8. Reward assignment in every time step.

4.5. Training Procedure

By interacting with the environment, training batches consisting of state–action pairs and rewards are collected, which in return are used by the RL method in order to maximize the long-term reward. In this context, the overall costs are reduced because the reward function is derived from the actual costs. Furthermore, a decentralized training approach is applied due to the heterogeneous system, in which different roles have to be learned by the agents. In this regard, an actor-critic RL method using proximal policy optimization [46] and a generalized advantage estimator [47] are used for each agent in order to learn individual policies, as this is a promising approach for discrete control tasks [48]. The value and policy functions are approximated using two separate, fully connected neural networks. Furthermore, a training iteration is completed whenever 40,000 experiences have been gathered. To collect batches of experiences, eight environments are run in parallel. The training is stopped after 3000 iterations, and the resulting policies serve as an input for the following validation scenarios.

5. Case Study

To examine the effect of the proposed energy-oriented production control method, this section will present a case study based on a simulation study with two benchmark algorithms. First, the applied system set-up will be described before the examined validation scenarios are presented. Finally, the results will be depicted and discussed.

5.1. Set-Up

The production system was derived from a contract manufacturing business for surface treatment. The portion of the factory site having the main energy demand consists of five machines in total, where jobs with a fixed lot size are produced. In every resource, between two and four different job types

can be machined, each requiring a specific set-up state. Additional parameters like stochastic events and delays as well as storage costs are considered, like those discussed previously. The modelling data of the production system are briefly summarized in Table 2.

Table 2. Modelling data of the production system.

Category	Parameter	Value
Production resource and jobs	Total number of resources	5
	Number of job types, which can be machined on the same resource	2–4
	Machining duration of a job depending on the job type	21–42 min
	Average power consumption of a job depending on the job type	20–48 kW
	Set-up duration of a job depending on the job type	6–12 min
Stochastic events	Probability for the arrival of rush jobs during a shift (normally distributed)	5% ($\sigma = 2.5\%$)
	Probability for a resource breakdown in every time step	1%
	Average duration of a breakdown (normally distributed)	12 min ($\sigma = 3$ min)
Costs	Delay costs of a job	48–88 €/h
	Storage costs of a job	5.2–11 €/h

On the energy supply side, data from a solar power plant near the production site were used as a VPS. In addition, the company had a local CHP plant with a maximum generation capacity of 32 kW and a lithium battery system with a total capacity of 35 kWh. To settle load deviation from the public grid, a tolerance of ±5% and costs of 200 €/MWh were implied. Considering the electricity market, the energy prices at the German intraday market from the period October 2018 to September 2019 were used. In addition, it was assumed that a price forecast would be available at a standard deviation of 5%. The energy demand of a production shift was determined by the number of jobs to be machined, which is thus not constant. In order to still be able to generate comparable scenarios, the following dimensioning of the elements of energy supply was assumed: there exists a solar power plant which provides roughly 25% of the total energy demand; 55% has already been purchased in advance, thus forming a constant base load, and the rest can be provided by the CPS. Comparable scenarios were calculated in this way, the used data is summarized in Table 3.

In order to generate sufficient training data, a simple training data generator was introduced. At the beginning of every training episode, arbitrary jobs were generated for every production resource. Thereby, the overall net utilization, without considering set-up, was random at between 60 and 80%. On the energy side, a representative day of the solar power and market data was extracted for every episode and changed by a normal distribution. A sufficient amount of realistic training data was able to be provided in this way.

Table 3. Parameters of the energy system.

Category	Parameter	Value
Solar power	Electricity price	0 ct/kWh
	Maximal generation power (in % of total power demand)	25%
CHP	Electricity price	10 ct/kWh
	Nominal generation power	16 kW
	Minimal generation power	6 kW
	Efficiency factor	0.1
Battery system	Nominal storage capacity	35 kWh
	C-rate	1
	Purchase costs	200 €/kWh
	Total charging and discharging efficiency	0.9
Intraday market	Proportional order fee (due to taxes and apportionments)	7.75 ct/kWh
	Prior purchased, constant base load power (in % of total energy demand)	55%
Load compensation from public grid	Additional costs for load deviation	200 €/MWh
	Tolerance band for load deviations	±5%

5.2. Validation Scenarios

In order to assess the system capability, the developed MARL approach was compared with two benchmark scenarios:

- **Reactive control strategy (RCS):** The reactive production control was based on the commonly used dispatching rule EDF. A widespread rule-based heuristic was developed to control the battery and CHP [49]. In this case, the battery power was adjusted, with the goal of maximizing CHP power generation and minimizing the resulting deviation between power demand and supply as much as possible. The generation of VPS was only cut off when there was still excessive power available while the battery was on maximum charging power and the CHP completely turned off. Since market trading is not possible using a rule-based method, there was no electricity trading in this scenario.
- **Predictive-reactive control strategy (PCS):** An optimization approach based on the metaheuristic Simulated Annealing was applied to control the overall system. The production plan was rescheduled both in the beginning of an episode and every time that a stochastic event occurred. The state transitions were thereby selected in an arbitrary manner, either by changing the starting time or the sequence of jobs, trading electricity, or adjusting the battery power. The implementation was based on [50]; for the parameters of starting temperature, minimum temperature, alpha, and iterations, values of 10, 0.01, 0.9999, and 100 were chosen.

The technical implementation of the MARL system was based on RLlib [51] and its corresponding libraries. The learning algorithm Proximal Policy Optimization (PPO) as well as neural networks with two hidden layers, 128 and 1024 neurons respectively, were applied.

5.3. Results and Discussion

The three approaches have been benchmarked on ten selected production shifts based on energy and production data from the production site. Before this, MARL was trained for 2500 episodes, which took about 5 days. The average of resulting costs of the ten shifts is displayed in Figure 9.

It became clear that the presented MARL system outperformed the reactive control approach (PCS) on production and energy costs. In contrast, the predictive, SA-based algorithm (RCS) achieved, in total, roughly 35% better results than the MARL approach. MARL thereby achieved lower average energy costs than SA, although the production costs were up by 40%. This indicated that MARL is stuck in a local optimal, which focused on the energy cost reduction, although there is still great potential on the production cost side. However, regarding the calculation time, an important advantage of MARL became clear. After training, the MARL system was able to compute decisions within seconds, whereas it took up to 2.5 h for the SA algorithm to calculate a new solution when new stochastic events occurred. Due to this weak reactivity, the application of SA was not suitable for the case in question. In a real-life scenario with stochastic events, decisions have to be made within a few seconds. The behavior of both the system and the market and battery agent is displayed in Figure 10 for an exemplary shift. The upper part of the figure shows the initially available electricity, which consisted of the pre-purchased base load and PV power; the resulting availability and total energy consumption are also displayed. It became clear that the available electricity can be widely adapted to the energy consumption. Regarding the ad-hoc power supply, the CHP was mainly run at maximum power, while the flexibility of the battery was being exploited. On the market side, power was bought in the second half of the shift, when the battery was empty and the prices low, then sold at the end of the shift when the energy demand dropped due to the fact that most jobs had already been machined.

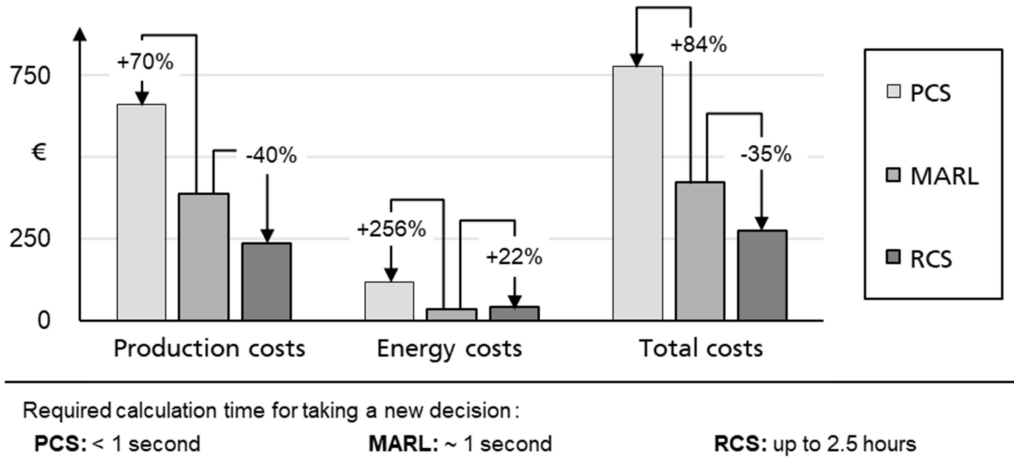

Figure 9. Average results for the ten validation scenarios.

In summary, the results show that the developed approach was able to control the three considered options of energy self-supply, battery, and short-term electricity market as well as a production system. Compared to a simple rule-based control strategy, all of the various cost factors could be minimized jointly while also taking battery degradation costs into account.

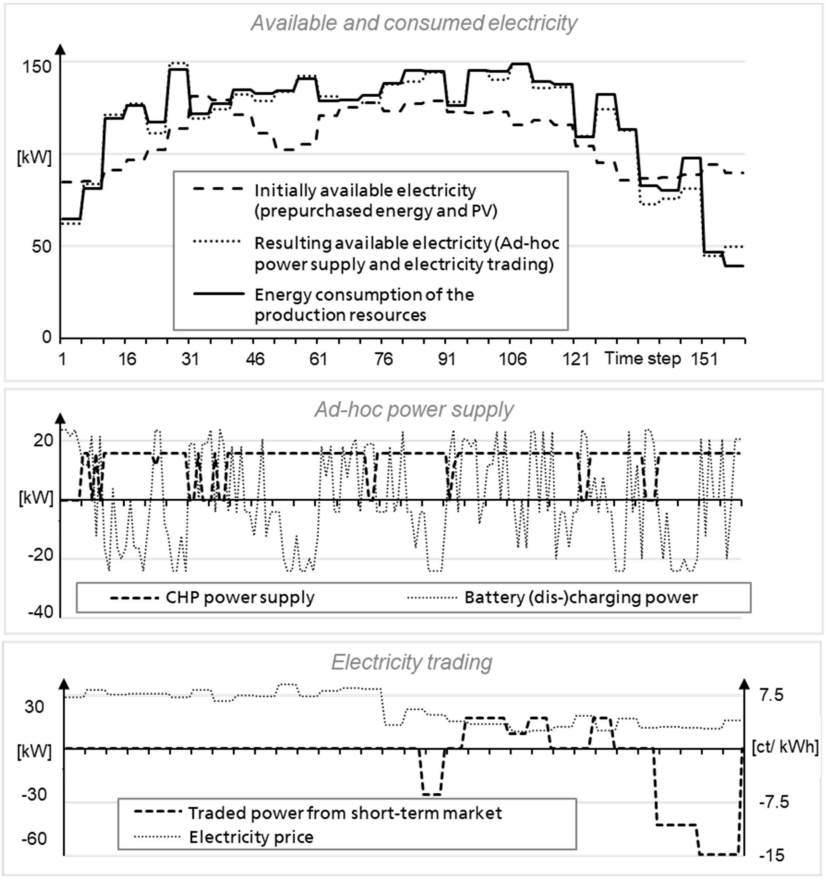

Figure 10. Exemplary system behavior for a given shift.

6. Conclusions

This article presents a new MARL-based approach for energy-oriented production control. Since the existing related studies do not consider all of the available power supply options for industrial companies at the same time, although their combination promises a high potential for cost-reduction,

the focus of this work lies on the integration of self-supply, battery systems, and short-term electricity trading into a production control strategy. Thus, industrial companies are enabled to efficiently and automatically reduce electricity costs, exploiting short-term fluctuations of electricity prices in real-time and, at the same time, considering production costs. Due to the resulting system complexity and the required reactivity of the system, the application of common control strategies is limited in this context, and a new approach based on MARL was developed. For this purpose, the intraday market, a stationary battery with a degradation model, power plants for self-supply, and a production system were modelled in a given environment. Three different agents, a resource agent, a battery agent, and a market agent, were used to minimize the global costs. A distributed and partly global and local reward function based on time-dependent production and energy costs was designed for this purpose. After training the system on a given production system with self-supply and a battery, the agents outperformed a reactive rule-based benchmark-scenario by an average of 84%, although the optimization approach via Simulated Annealing still performed 35% better. In terms of computational expense and reactivity, however, MARL clearly showed its advantages: the computation of a new decision lasted only seconds, whereas up to 2.5 h was needed to solve the Simulated Annealing algorithm. However, compared to the results with SA, the presented MARL seems to be stuck in local optima. This may be caused by a lack of exploration of new state–action trajectories of the environment and may be enhanced by using promising algorithms for this problem, such as [52], in future work.

In the context of the developed system, several simplifications were made with respect to the production and market sides in particular. So far, the method has been limited to a single-stage production system and some market characteristics, e.g., the bidding process and price risks were neglected. In addition, there was a lack of sufficient real training data, so a rudimentary training data generator was introduced. Since no benchmarks were available for the examined scenario, two specific benchmark algorithms were developed within the scope of this work. In future works, the given limitations considering modelling and benchmarks will need to be further reduced. In addition, the aspects of reward function, observation states, and learning will be examined more closely in order to further improve the cooperative behavior of the agents as well as the solution quality in general.

Author Contributions: Conceptualization, M.R.; methodology, M.R., C.L., R.Z.; software and validation, M.R., C.L., A.R., R.Z.; formal analysis, A.H.; investigation, M.R.; writing—original draft preparation, M.R.; writing—review and editing, A.H. and G.R.; supervision, A.H. and G.R.; project administration, M.R.; funding acquisition, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Bavarian Research Foundation (Bayerische Forschungsstiftung).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Prognos, A.G.; Wünsch, M.; Fraunhofer, I.F.A.M.; Eikmeier, B.; Gailfuß, M. *Potenzial-und Kosten-Nutzen-Analyse zu den Einsatzmöglichkeiten von Kraft-Wärme-Kopplung (Umsetzung der EU-Energieeffizienzrichtlinie) sowie Evaluierung des KWKG im Jahr 2014*; Endbericht zum Projekt IC: Berlin, Germany, 2014.
2. BDEW. *Strompreisanalyse Mai 2018: Haushalte und Industrie. Bundesverband der Energie-und Wasserwirtschaft e.V*; BDEW Bundesverband der Energie- und Wasserwirtschaft e.V.: Berlin, Germany, 2018.
3. Reinhart, G.; Reinhardt, S.; Graßl, M. Energieflexible Produktionssysteme. Ein-führungen zur Bewertung der Energieeffizienz von Produktionssystemen. *Werkstattstechnik Online* **2012**, *102*, 622–628.
4. Knupfer, S.M.; Hensley, R.; Hertzke, P.; Schaufuss, P.; Laverty, N.; Kramer, N. *Electrifiying Insights: How Automakers can Drive Electrified Vehicle Sales and Profitability*; McKinsey&Company: Detroit, MI, USA , 2017.
5. Stolle, T.; Hankeln, C.; Blaurock, J. Die Bedeutung der Energiespeicherbranche für das Energiesystem und die Gesamtwirtschaft in Deutschland. *Energiewirtsch. Tagesfragen.* **2018**, *9*, 54–56.
6. EPEX Spot. Market Data. 2020. Available online: https://www.epexspot.com/en/market-data (accessed on 6 May 2020).

7. Kost, C.; Shammugam, S.; Jülich, V.; Nguyen, H.T.; Schlegl, T. Stromgestehungskosten Erneuerbare Energien. Fraunhofer ISE: Freiburg, Germany, 2018.
8. Wünsch, M.; Eikmeier, B.; Eberhard, J.; Gailfuß, M. Potenzial-und Kosten-Nutzen-Analyse zu den Einsatzmöglichkeiten von Kraft-Wärme-Kopplung (Umsetzung der EU-Energieeffizienzrichtlinie) sowie Evaluierung des KWKG im Jahr 2014. 2014. Available online: https://ec.europa.eu/energy/sites/ener/files/documents/151221%20Mitteilung%20an%20KOM%20EED%20KWK%20Anlage%20Analyse.pdf (accessed on 20 May 2020).
9. Beier, J. *Simulation Approach Towards Energy Flexible Manufacturing Systems*; Springer International Publishing: Cham, Switzerland, 2017; p. 22796.
10. Fang, X.; Misra, S.; Xue, G.; Yang, D. Smart grid—The new and improved power grid: A survey. *IEEE Commun. Surv. Tutor.* **2012**, *14*, 944–980. [CrossRef]
11. Lödding, H. *Handbook of Manufacturing Control: Fundamentals, Description, Configuration*; Springer: Berlin/Heidelberg, Germany, 2013.
12. Schultz, C.; Bayer, C.; Roesch, M.; Braunreuther, S.; Reinhart, G. Integration of an automated load management in a manufacturing execution system. In Proceedings of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 10–13 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 494–498.
13. Gabel, T. Multi-Agent Reinforcement Learning Approaches for Distributed Job-Shop Scheduling Problems. Ph.D. Thesis, University of Osnabrück, Osnabrück, Germany, 2009.
14. Kuhnle, A.; Schäfer, L.; Stricker, N.; Lanza, G. Design, implementation and evaluation of reinforcement learning for an adaptive order dispatching in job shop manufacturing systems. *Procedia CIRP* **2019**, *81*, 234–239. [CrossRef]
15. Wang, Y.; Liu, H.; Zheng, W.; Xia, Y.; Li, Y.; Chen, P.; Guo, K.; Xie, H. Multi-objective workflow scheduling with Deep-Q-network-based multi-agent reinforcement learning. *IEEE Access* **2019**, *7*, 39974–39982. [CrossRef]
16. Waschneck, B.; Reichstaller, A.; Belzner, L.; Altenmuller, T.; Bauernhansl, T.; Knapp, A.; Kyek, A. Deep reinforcement learning for semiconductor production scheduling. In Proceedings of the 29th Annual SEMI Advanced Semiconductor Manufacturing Conference 2018 (ASMC 2018), Saratoga Springs, NY, USA, 30 April–3 May 2018; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2018; pp. 301–306.
17. Jiang, B.; Fei, Y. Smart home in smart microgrid: A cost-effective energy ecosystem with intelligent hierarchical agents. *IEEE Trans. Smart Grid* **2015**, *6*, 3–13. [CrossRef]
18. Mbuwir, B.V.; Kaffash, M.; Deconinck, G. Battery scheduling in a residential multi-carrier energy system using reinforcement learning. In Proceedings of the 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Aalborg, Denmark, 29–31 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
19. Berg, M.; Borchert, S. *Strategischer Energieeinkauf: Der Energieeinkauf Zwischen Liberalisierten Märkten und Einer Wechselhaften Energiepolitik in Deutschland*; BME: Frankfurt am Main, Germany, 2014.
20. Bertsch, J.; Fridgen, G.; Sachs, T.; Schöpf, M.; Schweter, H.; Sitzmann, A. *Ausgangsbedingungen für die Vermarktung von Nachfrageflexibilität: Status-Quo-Analyse und Metastudie*; Bayreuther Arbeitspapiere zur Wirtschaftsinformatik 66: Bayreuth, Germany, 2017.
21. Günther, M. *Energieeffizienz durch erneuerbare Energien: Möglichkeiten, Potenziale, Systeme*; Springer: Vieweg, Wiesbaden, 2015; p. 193.
22. Keller, F.; Braunreuther, S.; Reinhart, G. Integration of on-site energy generation into production planning systems. *Procedia CIRP* **2016**, *48*, 254–258. [CrossRef]
23. Köhler, A.; Baron, Y.; Bulach, W.; Heinemann, C.; Vogel, M.; Behrendt, S.; Degel, M.; Krauß, N.; Buchert, M. *Studie: Ökologische und ökonomische Bewertung des Ressourcenaufwands—Stationäre Energiespeichersysteme in der Industriellen Produktion*; VDI ZRE: Berlin, Germany, 2018.
24. Vetter, J.; Novak, P.; Wagner, M.; Veit, C.; Möller, K.C.; Besenhard, J.; Winter, M.; Wohlfahrt-Mehrens, M.; Vogler, C.; Hammouche, A. Ageing mechanisms in lithium-ion batteries. *J. Power Sources* **2005**, *147*, 269–281. [CrossRef]
25. Safari, M.; Morcrette, M.; Teyssot, A.; Delacourt, C. Multimodal physics-based aging model for life prediction of li-ion batteries. *J. Electrochem. Soc.* **2009**, *156*, A145. [CrossRef]
26. Ouelhadj, D.; Petrovic, S. A survey of dynamic scheduling in manufacturing systems. *J. Sched.* **2008**, *12*, 417–431. [CrossRef]

27. Vieira, G.E.; Herrmann, J.W.; Lin, E. Rescheduling manufacturing systems: A framework of strategies, policies, and methods. *J. Sched.* **2003**, *6*, 39–62. [CrossRef]
28. Shen, W.; Hao, Q.; Yoon, H.J.; Norrie, D.H. Applications of agent-based systems in intelligent manufacturing: An updated review. *Adv. Eng. Inform.* **2006**, *20*, 415–431. [CrossRef]
29. Shen, W.; Norrie, D.H.; Barthès, J.P. *Multi-Agent Systems for Concurrent Intelligent Design and Manufacturing*; Taylor & Francis: London, UK, 2001; p. 386.
30. Sutton, R.S.; Barto, A. *Reinforcement Learning: An Introduction*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2018; p. 526.
31. Ngyen, D.T.; Kumar, A.; Lau, H.C. Credit assignment for collective multiagent RL with global rewards. In Proceedings of the 32nd International Conference on Neural Information Processing System NIPS'18, Montréal, QC, Canada, 2–8 December 2018; pp. 8113–8124.
32. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Driessche, G.V.D.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
33. Neugebauer, R.; Putz, M.; Schlegel, A.; Langer, T.; Franz, E.; Lorenz, S. Energy-sensitive production control in mixed model manufacturing processes. In *Leveraging Technology for A Sustainable World, Proceedings of the 19th CIRP Conference on Life Cycle Engineering, University of California at Berkeley, Berkeley, CA, USA, 23–25 May 2012*; LCE 2012; Dornfeld, D.A., Linke, B.S., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 399–404.
34. Fernandez, M.; Li, L.; Sun, Z. "Just-for-Peak" buffer inventory for peak electricity demand reduction of manufacturing systems. *Int. J. Prod. Econ.* **2013**, *146*, 178–184. [CrossRef]
35. Böckmann, M.G. *Senkung der Produktionskosten durch Gestaltung eines Energieregelkreis-Konzeptes*; Apprimus-Verlag: Aachen, Germany, 2014; p. 166.
36. Willeke, S.; Ullmann, G.; Nyhuis, P. Method for an energy-cost-oriented manufacturing control to reduce energy costs: Energy cost reduction by using a new sequencing method. In Proceedings of the ICIMSA 2016 International Conference on Industrial Engineering, Management Science and Applications, Jeju Island, Korea, 23–26 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5.
37. Zhai, Y.; Biel, K.; Zhao, F.; Sutherland, J. Dynamic scheduling of a flow shop with on-site wind generation for energy cost reduction under real time electricity pricing. *CIRP Ann.* **2017**, *66*, 41–44. [CrossRef]
38. Khalaf, A.F.; Wang, Y. Energy-cost-aware flow shop scheduling considering intermittent renewables, energy storage, and real-time electricity pricing. *Int. J. Energy Res.* **2018**, *42*, 3928–3942. [CrossRef]
39. Abikarram, J.B.; McConky, K.; Proano, R.A. Energy cost minimization for unrelated parallel machine scheduling under real time and demand charge pricing. *J. Clean. Prod.* **2019**, *208*, 232–242. [CrossRef]
40. Wang, J.; Zhang, Y.; Liu, Y.; Wu, N. Multiagent and bargaining-game-based real-time scheduling for internet of things-enabled flexible job shop. *IEEE Internet Things J.* **2019**, *6*, 2518–2531. [CrossRef]
41. Schaumann, G.; Schmitz, K.W. *Kraft-Wärme-Kopplung, Vollständig Bearbeitete und Erweiterte*, 4th ed.; Springer: Berlin/Heidelberg, Germany, 2010.
42. Xu, B.; Oudalov, A.; Ulbig, A.; Andersson, G.; Kirschen, D.S. Modeling of lithium-ion battery degradation for cell life assessment. *IEEE Trans. Smart Grid* **2016**, *9*, 1131–1140. [CrossRef]
43. Schultz, C.; Braun, S.; Braunreuther, S.; Reinhart, G. Integration of load management into an energy-oriented production control. *Procedia Manuf.* **2017**, *8*, 144–151. [CrossRef]
44. Roesch, M.; Linder, C.; Bruckdorfer, C.; Hohmann, A.; Reinhart, G. Industrial Load Management using Multi-Agent Reinforcement Learning for Rescheduling. In Proceedings of the 2019 Second International Conference on Artificial Intelligence for Industries (AI4I) (IEEE), Laguna Hills, CA, USA, 25–27 September 2019; pp. 99–102.
45. Roesch, M.; Berger, C.; Braunreuther, S.; Reinhart, G. *Cost-Model for Energy-Oriented Production Control*; IEEE: Piscataway, NJ, USA, 2018; pp. 158–162.
46. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
47. Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv* **2015**, arXiv:1506.02438.
48. Loon, K.W.; Graesser, L.; Cvitkovic, M. SLM Lab: A Comprehensive benchmark and modular software framework for reproducible deep reinforcement learning. *arXiv* **2019**, arXiv:1912.12482.

49. Weniger, J.; Bergner, J.; Tjaden, T.; Quaschning, V. Bedeutung von prognosebasierten Betriebsstrategien für die Netzintegration von PV-Speichersystemen. In Proceedings of the 29th Symposium Photovoltaische Solarenergie, Bad Staffelstein, Germany, 12–14 March 2014.
50. Van Laarhoven, P.J.M.; Aarts, E.H.L. *Simulated Annealing: Theory and Applications*; Kluwer: Dordrecht, The Netherlands, 1992; p. 187.
51. Liang, E.; Liaw, R.; Moritz, P.; Nishihara, R.; Fox, R.; Goldberg, K.; Gonzalez, J.E.; Jordan, M.I.; Stoica, I. RLlib: Abstractions for distributed reinforcement learning. *arXiv* **2017**, arXiv:1712.09381.
52. Badia, A.P.; Sprechmann, P.; Vitvitskyi, P.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; et al. Never Give Up: Learning Directed Exploration Strategies, International Conference on Learning Representations. *arXiv* **2020**, arXiv:2002.06038.