Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

# Reinforcement learning for energy-efficient control of multi-stage production lines with parallel machine workstations

Alberto Loffredo[1 *], Marvin Carl May[2] and Andrea Matta[1]

[1] Politecnico di Milano, Via G. la Masa 1, Milan, 20156, Italy

[2] Karlsruhe Institute of Technology, Kaiserstraße 12, Karlsruhe, 76131, Germany

**Keywords:** Artificial Intelligence, Sustainability, Manufacturing Systems

**Abstract.** An effective approach to enhancing the sustainability of production systems is to use energy-efficient control (EEC) policies for optimal balancing of production rate and energy demand. Reinforcement learning (RL) algorithms can be employed to successfully control production systems, even when there is a lack of prior knowledge about system parameters. Furthermore, recent research demonstrated that RL can be also applied for the optimal EEC of a single manufacturing workstation with parallel machines. The purpose of this study is to apply an RL for EEC approach to more workstations belonging to the same industrial production system from the automotive sector, without relying on full knowledge of system dynamics. This work aims to show how the RL for EEC of more workstations affects the overall production system in terms of throughput and energy consumption. Numerical results demonstrate the benefits of the proposed model.

## 1. Introduction

Energy efficiency has become a crucial focus of research in production systems, alongside productivity and quality improvements. Manufacturing is responsible for nearly 40% of global energy consumption [1], making it a significant area for eco-friendliness and sustainability advancements. Using energy-efficient control (EEC) policies can effectively minimize machines' environmental impact by implementing real-time actions during production to reduce energy consumption during idle periods. EEC aims to switch off machines during idle periods and turn them back on when required for production, achieving the optimal balance between production rate and energy demand. However, the "Always-On" (AOn) policy keeps machines running during idle periods and wastes energy. AOn is still a common practice despite its drawbacks. Existing EEC methods assume complete knowledge of system dynamics and parameters, which can limit the accuracy and generality of EEC models and results.

Recent research shows the potential of Reinforcement Learning (RL) to perform optimal control for complex systems. RL is a Machine Learning approach that enables agents to learn by interacting with their environment, even in the presence of incomplete or uncertain information. RL algorithms are indeed adaptive: they are designed to learn how to deal with the system dynamics and adjust their strategies accordingly. Recent research also demonstrated the effectiveness of RL for the optimal EEC of a manufacturing workstations composed of parallel machines, a widely used layout to obtain a balanced production system in terms of workstations workload. This work is focused on this type of configuration. A literature RL-based model for EEC is applied to more parallel machine workstations that are part of the same production line. The literature model is able to reduce the workstation system energy consumption while maintaining its throughput, even without full knowledge of the system dynamics. The focus is then moved on the effect that these energy-efficient actions have on the overall production system in terms of throughput and energy consumption.

The paper is structured as follows. Section 2 presents a literature review of EEC and RL for production control, and highlights the main contribution of this work. Section 3 includes a

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

description of the industrial case studied in this work. Section 4 presents an overview on the used RL-based algorithm from literature. Section 5 presents the results of the numerical experiments carried out, demonstrating the benefits of the algorithm when applied to more workstations in the same industrial case. Finally, Section 6 concludes the paper and discusses possible further developments.

## 2. Literature Review

### *2.1 Energy-Efficient Control of Manufacturing Systems*

EEC problem for manufacturing equipment is addressed in two ways: (i) controlling systems with a single buffer before a single machine and (ii) controlling systems with a single buffer followed by parallel machines in a workstation. The literature on EEC for manufacturing systems has been growing, and a review can be found in [2].

Recent examples of EEC for the single-buffer-single-machine layout can be found in the following. Mouzon et al. [3] were the first to address the topic, proposing various switch off rules for a non-bottleneck workstation in a production system. Duque et al. [4] contributed to the development of a fuzzy controller that can be used to turn on/off a single workstation. Later, in Frigerio and Matta [5], the authors conducted an analytical study of an EEC policy for a single machine. In Jia et al. [6] an alternative approach was introduced where the authors used Work-In-Process (WIP) data to formulate efficient EEC policies for a multi-stage production line that included single machine workstations. Finally, in a recent work by Cui et al. [7], an optimal EEC technique was suggested for the entire production line using buffer level information.

Research stream for EEC of workstations with parallel machines is less developed. Loffredo et al. [8] introduced a model using a Markov Decision Process (MDP) that generates effective EEC policies for a single workstation with parallel machines. Furthermore, they extended this approach with an MDP-based model for controlling multi-stage production lines with parallel machine workstations [9]. However, these studies were constrained by their reliance on MDP that necessitates complete knowledge of the system dynamics. This approach resulted in a solution that did not utilize any machine learning methods. To fill this gap, in [10] it is possible to find an RL-based model for the EEC of a single parallel machine workstation, but in this model they considered the stand-alone workstation without any interaction with the shop floor.

### *2.2 Reinforcement Learning in Production Systems Control*

RL algorithms consist of two elements: an agent and an environment. The agent continuously interacts with its environment to optimize its behavior and finally achieve a specific goal. This involves recognizing the best action in each state to optimize an objective function, such as the total discounted reward over a given time horizon. A basic overview of the RL framework is explained in the following and is extracted from [11]. Everything begins with the agent observing the environment state $s_t \in \mathbb{S}$, and selecting an action at $a_t \in \mathbb{A}$. The state space, denoted by $\mathbb{S}$, includes all the possible observations an agent can make of the environment, providing information on the current production state to allow the agent to choose actions that optimally solve the control problem. On the other hand, the action space, represented by $\mathbb{A}$, includes all the possible actions the agent can take. After $a_t$ is performed, the environment responds with the resulting state $s_{t+1}$ while the agent is rewarded with a reward $r_{t+1}$ and the next iteration can start (see Fig. 1).

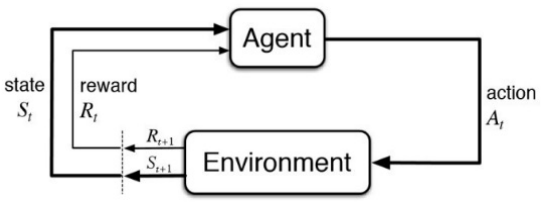


*Fig. 1 Overview of RL-Framework [11].*

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

The primary objective of the agent is to maximize the long term cumulative reward by optimizing its action-selection policy. This involves learning the best control policy, or EEC policy for this work scope, to apply to the environment.

Literature offers plenty of successful RL applications in production systems (complete and recent literature review in [12]). For instance, RL proved to be strongly effective in production scheduling, dispatching and plant-internal logistic of applications (examples in [13, 14, 15]). However, despite RL's proven effectiveness for production planning and control problems, only one work dealing with RL for EEC of manufacturing equipment can be identified [10].

### *2.3 Contribution*

Energy efficiency is a key factor for achieving sustainability in manufacturing. The EEC approach offers solutions to minimize the environmental impact of manufacturing equipment, but it faces limitations due to assumptions of complete knowledge of system dynamics and parameters. Reinforcement Learning can address this challenge by handling uncertain information. Nevertheless, in literature there is only work addressing the RL for EEC approach in manufacturing but, even in this case, the study focuses on a single workstation without considering its interactions with the shop floor. To fill this gap, this work analyzes the impact that using a literature RL for EEC model to more parallel machines workstations has on the overall production system, in terms of throughput and energy consumption. The study analyzes a real-world example in the automotive industry, using various RL agents to identify the optimal one for achieving the best performance. The algorithm parameters are fine-tuned to find the optimal trade-off between throughput and energy consumption.

## 3. System Description

### *3.1 System Layout*

A real industrial system is used as reference case study where the proposed model can be applied and its effects can be analyzed. The production line under investigation is a manufacturing system producing cylinder heads in the automotive sector (see layout in Fig. 2).

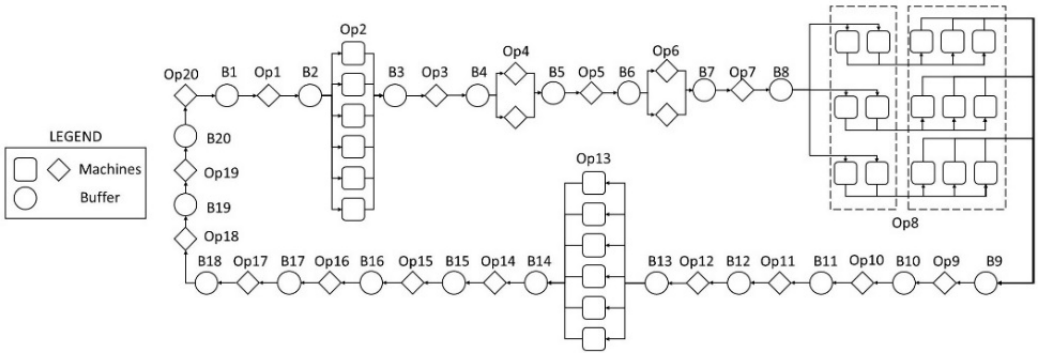


*Fig. 2 Layout of the real industrial system under investigation.*

The production process consists of 20 stages (m = 20). Each stage $S_i | \, i \in \{1, \ldots, m\}$ is characterized by an upstream buffer $B_i | \, i \in \{1, \ldots, m\}$ with a finite capacity $K_i$, followed by $c_i$ parallel machines $M_{ij} | \, i \in \{1, \ldots, m\}, j \in \{1, \ldots, c_i\}$. Machines $M_{ij}$ are considered starved if they have the ability to process parts, but there are none available in $B_i$. Conversely, they are blocked if they can process parts, but $B_{i+1}$ is already full. It is worth noting that machines $M_{mj}$ in $S_m$ cannot be blocked, as there is a downstream infinite capacity buffer that allows processed parts to exit the system immediately after they are completed. All the machines in the line operate on a single type of part, blocking after service rule and first-come-first-served rule are enforced. Also, machines cannot be switched off while they are operating on items: part processing

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

cannot be interrupted by the control. Also five stages are characterized by parallel machines ($c_i > 1$) while the remaining have a single-buffer-single-machine layout ($c_i = 1$). The line is composed by 12 controllable stages that are fully automated while eight are not controllable due to the presence of human operators involved in the operation performed. In detail, the subset of controllable stages is $S_i | \, i = [2, 3, 4, 7, 8, 9, 10, 12, 13, 14, 15, 19]$. The production system parameters are not reported because of a confidentiality agreement with the company owning the system.

The system is characterized by different stochastic processes, including the arrival rate of parts to the first stage $S_1$, machines processing and startup times, and, also time between failures (TBF), and time to repair (TTR) of the machines. All of these are assumed to be independent of each other and stationary. The arrival of parts to $S_1$ follows a stochastic process with an expected value of λ. Each machine $M_{ij}$ has startup and processing times that follow a stochastic process, with expected values equal to $\delta_{ij}$ and $\mu_{ij}$, respectively. Additionally, the machines are unreliable and can be subject to operation-dependent failures. Each machine $M_{ij}$ is characterized by stochastic TBF and TTR, with expected values equal to $\psi_{ij}$ and $\xi_{ij}$, respectively. All the mentioned expected values vary for each stage in the production line. Lastly, all the line machines are consistent with the energetic state model detailed in Section 3.2.

*3.2 Machine States and Associated Power Consumptions*

All the machines are characterized by the following states: working (w), standby (sb), startup (su), and failed (f); the working state is then divided into two sub-states: idle (id) and busy (b) (see Fig. 3 ).

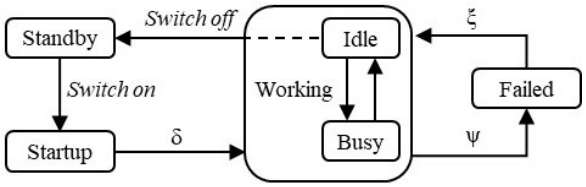


*Fig. 3 Machine state model.*

When the machine is turned off, it is in the standby state. When it is switched on to transition to the working state, the machine goes first through the startup state, which has a stochastic startup time with an expected value of δ. While working, the machine is either busy processing parts with stochastic processing times with an expected value of μ, or it is idle and ready for processing. However, the machine is unreliable and can fail, which causes it to transition to the failed state, determined by the stochastic TBF with an expected value of ψ. If the machine is fully repaired, including being switched off and back on, and tested, it transitions back to the working state, which is ruled by the machine's stochastic TTR with an expected value of ξ. Finally, the machine can be switched off immediately from the idle sub-state sending it back to the standby state, i.e. it is not possible to switch off the machine while it is busy. Each machine state necessitates a power requirement $w_s \mid s = \{w, sb, su, f, id, b\}$, which is non-negative and dependent on the active components in the respective state or sub-state. It is assumed that $w_b > w_{su} > w_{id} > w_{sb} \simeq w_f \simeq 0$. Additionally, it is worth noticing that $w_w$ depends on $w_{id}$ and $w_b$ .

## 4. Literature RL for EEC Model

The elements of the literature RL-based model [10] are summarized in this section as follows: the agent state space in section 4.1, action space in section 4.2, and reward function in section 4.3. The environment is the single workstation to be controlled. The model goal is to achieve the optimum

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

trade-off between system productivity and energy demand for the single workstation without relying on full knowledge of the system dynamics.

*4.1 State Space*

In the RL framework, the state is the representation of the environment at a particular time, which includes all the relevant information necessary for the agent to make decisions about what action to take next. Consequently, the state space $\mathbb{S}$ represents the set of possible agent observations of the environment, i.e. the single workstation $S_i$. In the literature model the state $\boldsymbol{s} \in \mathbb{S}$ is represented with the ordered vector of size $\boldsymbol{s} = [\frac{n_i}{K_i}, x_1, \ldots, x_{c_i}]$. For the generic stage $S_i$, in $\boldsymbol{s}$, $n_i \in \{0, \ldots, K_i\}$ is the number of parts in buffer $B_i$. Furthermore, each binary variable $x_{ij} \in \{0,1\}$ provides information to the agent on the working state of each stage machine: $x_{ij} = 1$ means machine $M_{ij}$ is in working state while if $x_{ij} = 0$ machine $M_{ij}$ is not in working state.

*4.2 Action Space*

In the literature model, the action the agent has to perform is quite straightforward: it has to select, at any time, how many machines should be in working state in the workstation. In this way, starting from the actual number of working machines in the system, some machines might then be switched on or switched off, according to the agent's action. The action space $\mathbb{A}$ is the set of all possible actions that an agent can take. Therefore, the action $\boldsymbol{a} \in \mathbb{A}$ is applied to control the number of machines to be in working state in the controlled stage $S_i$. Therefore, in $S_i$, $\mathbb{A} = \{0, \ldots, c_i\}$, since the control action $\boldsymbol{a}$ can only assume integer values ranging from 0 (all machines in the stage must be switched off and not working) to $c_i$ (all machines in the stage must be switched on and working). Note that system assumptions dictate the allowable actions, such as not interrupting part processing or startup procedures. Therefore, the agent cannot immediately set a machine undergoing startup to a working state.

*4.3 Reward Function*

The reward function is an essential component of RL, as it provides the feedback that guides the agent's learning process. Therefore, it must be designed to reflect the goals of the task being learned. The EEC problem is characterized by a multi-objective nature since there are two goals, or Key Performance Indicators (KPIs), to be considered: reducing the energy consumption and maintaining the system throughput.

The literature model reward function is based on two elements: the *Throughput Component* $R_t$ (see Eq.1) and the *Consumption Component* $R_{cons}$ (see Eq.2):

$$R_t = \frac{\theta}{e} \quad with \quad \theta = e^{\frac{TH_t}{TH_{t,max}}} \tag{1}$$

$$R_{cons} = e^{-z\Delta cons} \quad with \quad \Delta cons = cons_t - cons_{t-1} \tag{2}$$

In Eq. 1, $TH_t$ is the workstation throughput from at the actual time *t*, i.e. number of produced parts until time *t* divided by *t* itself. $TH_{t,max}$ is the maximum throughput the stage $S_i$ can reach in the same time-period by maintaining all the line machines always switched on. Consequently, $0 \leq \frac{TH_t}{TH_{t,max}} \leq 1$, $1 \leq \theta \leq e$, and $\frac{1}{e} \leq R_t \leq 1$, where $R_t$ grows as $TH_t$ approaches $TH_{t,max}$. Through $R_t$, the agent is directed to increase productivity by maintaining the stage machines in working state to produce a larger quantity of parts to increase $R_t$. In Eq. 2, $\Delta cons$ represents the increasing consumption. This is the difference between (i) $cons_t$, i.e. the stage $S_i$ energy consumption in the time-period at actual time *t,* and (ii) $cons_{t-1}$, i.e. the stage $S_i$ energy consumption until time *t*-1. The latter is the time when the previous reward was given to the agent. Considering the use of a

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

scale factor $z$, it is possible to state that $0 \leq R_{cons} = e^{-z\Delta cons} \leq 1$, where $R_{cons}$ grows as $\Delta cons$ approaches zero. This means that, through $R_{cons}$, the agent is directed to decrease productivity and save energy by maintaining the stage machines not in working state to produce fewer parts.

$R_t$ and $R_{cons}$ compose the reward $R$ through the following reward function:

$$R = \phi R_t + (1-\phi)R_{cons} \quad (3)$$

Where $0 \leq \phi \leq 1$ is a key element in the literature model that balances the multi-objective targets of the problem: if $\phi = 0$ then $R = R_{cons}$ and the agent only aims at reducing the consumption, while, if $\phi = 1$ then $R = R_t$ and the agent only aims at increasing the production. $\phi$ determines the weight of energy consumption and throughput in the action-selection process and must be calibrated. However, it is possible to affirm that the optimal $\phi$ will tend to 1, since this leads to a null or almost-null productivity drop. Finally, since $0 \leq \phi \leq 1, 0 \leq R_{cons} \leq 1$, and $0 \leq R_t \leq 1$, then $0 \leq R \leq 1$.

## 5. Numerical Experiments

The objective of the experimental study is to assess the impact that applying the literature RL for EEC model on more parallel machines workstations has on the overall production system in the real-world industrial system (Section 3). In particular, 12 stages are controlled in the system, $S_i | \, i = [2, 3, 4, 7, 8, 9, 10, 12, 13, 14, 15, 19]$. The literature model is applied independently to all the 12 controllable stages in the line: there are 12 single RL agents controlling only the respective workstation observing an environment that is only a part of the overall production system. Different agent-types are compared to identify the most suitable for the use case and then, with the selected agent, ϕ is optimally calibrated for the line under study.

In the experiments, a discrete-event simulator of the system, as described in Section 3, has been implemented using the SimPy library, and the agent, as described in Section 4, has been built using the TensorForce library [16]. Both interact through Python code. Every experiment involves the assessment of two KPIs: throughput loss and energy saving. The former is determined by calculating the difference in system throughput when the AOn policy is implemented versus when the RL-based model is applied to the controlled stages; the latter is determined by measuring the difference in total system energy consumption between the AOn policy and the RL-based model. The KPIs are extracted by comparing the case when the system is managed with AOn policy. 10 replications were carried out for each case, with a simulation length of 30 days. The experiments were characterized by random number generation.

To evaluate the most effective approach for applying the EEC in the real-world case, 4 commonly used types of RL agents are implemented in each stage and compared (results in Fig. 4):

1. The Tensorforce-General agent, an agent included in the Tensorforce library [16].
2. The TRPO agent exploiting the Trust Region Policy Optimization algorithm [17].
3. The PPO agent which exploits the Proximal Policy Optimization method [18].
4. The DQN agent which uses Deep Q-Network [19]

To not jeopardize the throughput, only high values of $\phi$ ($\phi \geq 0.90$) are tested. Among all agents, DQN is best performing since at the same it is characterized by lower throughput loss and higher savings. For all the agents, the default TensorForce library NN hyperparameters are used.

Materials Research Proceedings 35 (2023) 428-436 https://doi.org/10.21741/9781644902714-51

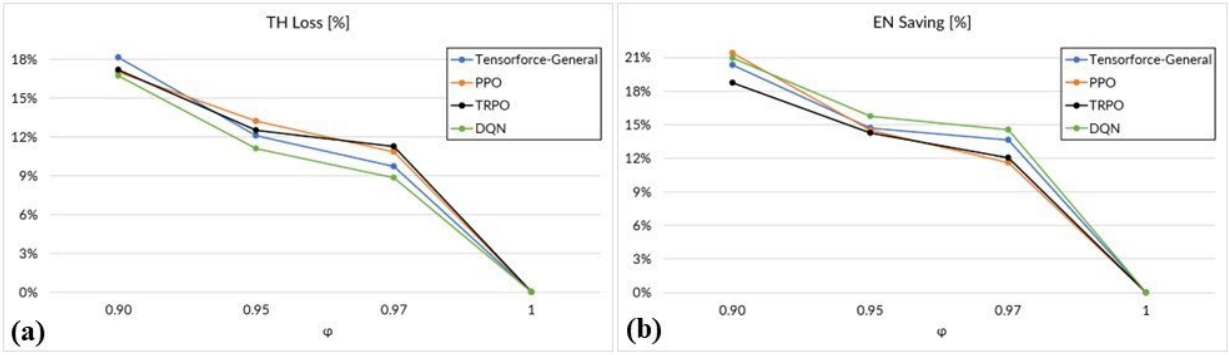


*Fig. 4 Comparison of different agents when applied to the industrial case: throughput loss (a) and energy saving (b) in respect to the AOn Policy. Values are shown considering a confidence level of 90% on the mean value. However, being all the confidence intervals strict, i.e. with a width in all the cases lower than 2%, confidence intervals are not visible with the selected figure scale.*

The subsequent step regarded the optimal calibration of $\phi$ when the DQN agent is applied. $\phi$ has been calibrated only for high values, varying it from 0.95 to 1 with a step of 0.01. Results of this analysis are shown in Fig. 5 .

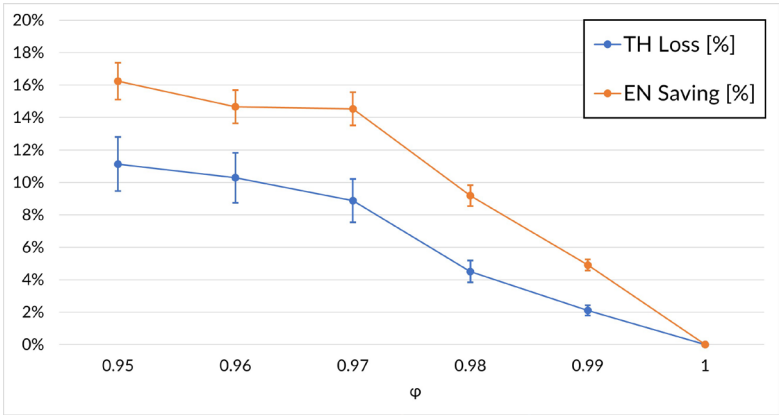


*Fig. 5 Energy saving and throughput loss when $\phi$ varies for the industrial case. A PPO agent is used. Values are shown considering a confidence level of 90% on the mean value.*

The optimal value appears to be $\phi$ = 0.99, leading to a slight throughput loss (2.11 ± 0.07%) and a corresponding significant energy saving equal to 4.91 ± 0.17%. Considering the annual system performance, this would lead, on average, to a throughput decrease of about 8000 units over a total production volume of more than 500,000 parts, while saving more than 105 barrels of oil equivalent. It must be noted that, if the company owning the manufacturing line would prefer to avoid also the slight throughput loss and produce that 2.11 ± 0.07% of lost parts during extra working hours, the system would be subject to additional energy consumption. In this case, the corresponding average savings will be decreased to 3.11 ± 0.14% but with a corresponding null throughput loss. This managerial choice will improve the system sustainability by saving, on average, more than 70 barrels of oil equivalent without any productivity drop. This confirms that the RL-based algorithm significantly enhances the sustainability of the industrial use-case, while maintaining its productivity, even when applied independently to more line stages.

## 6. Conclusions and Further Developments

In this work, an RL for EEC model has been applied to more parallel machines workstation used in a manufacturing system. Afterwards, it has been studied the impact that this energy-efficiency action has on the entire production line. Numerical results are presented, showing the

corresponding benefits when model is implemented. The latter uses Reinforcement Learning strategies to implement energy-efficient control actions efficiently without relying on complete knowledge of system dynamics and parameters. The model adapts and evolves over time during the training process, making it promising for direct application in industry with minimal effort. The potential of the proposed approach, combined with the increasing ease of use and knowledge of reinforcement learning techniques, presents a promising opportunity for direct and successful application in industry. This application can be achieved in a short time and with minimal effort, making it appealing from a managerial standpoint.

However, a challenging topic might be the creation of a novel RL-based model leading to where the control is executed jointly in all the workstations, considering the overall system as the environment to be controlled. Future research will focus on developing this method.

**References**

[1] Bipartisan Policy Center. Annual energy outlook 2022. EIA, Washington, DC, 2020.

[2] Renna, P. and Materi, S. (2021) A literature review of energy efficiency and sustainability in manufacturing systems. Applied Sciences, 11(16), 7366. https://doi.org/10.3390/app11167366

[3] Mouzon, G., Yildirim, M.B. and Twomey, J. (2007) Operational methods for minimization of energy consumption of manufacturing equipment. International Journal of Production Research, 45(18-19), 4247-4271. https://doi.org/10.1080/00207540701450013

[4] Duque, E.T., Fei, Z., Wang, J., Li, S., Li, Y., 2018. Energy consumption control of one machine manufacturing system with stochastic arrivals based on fuzzy logic, in: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE. pp. 1503-1507. https://doi.org/10.1109/IEEM.2018.8607749

[5] Frigerio, N., Cornaggia, C.F. and Matta, A. (2021) An adaptive policy for on-line energy-efficient control of machine tools under throughput constraint. Journal of Cleaner Production, 287, 125367. https://doi.org/10.1016/j.jclepro.2020.125367

[6] Jia, Z., Zhang, L., Arinez, J. and Xiao, G. (2016) Performance analysis for serial production lines with Bernoulli machines and real-time WIP-based machine switch-on/off control. International Journal of Production Research, 54(21), 6285-6301. https://doi.org/10.1080/00207543.2016.1197438

[7] Cui, P.-H., Wang, J.-Q., Li, Y. and Yan, F.-Y. (2021) Energy-efficient control in serial production lines: Modeling, analysis and improvement. Journal of Manufacturing Systems, 60, 11-21. https://doi.org/10.1016/j.jmsy.2021.04.002

[8] Loffredo, A., Frigerio, N., Lanzarone, E., Matta, A., 2021. Energy-efficient control policy for parallel and identical machines with availability constraint. IEEE Robotics and Automation Letters. Vol. 6(3), pp 5713-5719 https://doi.org/10.1109/LRA.2021.3085169

[9] Loffredo, A., Frigerio, N., Lanzarone, E., Matta, A., 2023. Energy-efficient control in multi-stage production lines with parallel machine workstations and production constraints. IISE Transactions. https://doi.org/10.1080/24725854.2023.2168321

[10] Loffredo, A., May, M.C., Schäfer L., Matta, A., & Lanza G. Reinforcement Learning for Energy-Efficient Control of Parallel and Identical Machines. CIRP Journal of Manufacturing Science and Technology. Under Review.

[11] Sutton, R.S., and Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press,

[12] Panzer, M., Bender, B., 2022. Deep reinforcement learning in production systems: a systematic literature review. International Journal of Production Research 60, 4316-4341. https://doi.org/10.1080/00207543.2021.1973138

[13] Baer, S., Turner, D., Mohanty, P., Samsonov, V., Bakakeu, R., Meisen, T., 2020. Multi agent deep q-network approach for online job shop scheduling in flexible manufacturing, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence, pp. 1-9.

[14] Stricker, N., Kuhnle, A., Sturm, R., Friess, S., 2018. Reinforcement learning for adaptive order dispatching in the semiconductor industry. CIRP Annals 67, 511-514. https://doi.org/10.1016/j.cirp.2018.04.041

[15] Malus, A., Kozjek, D., et al., 2020. Real-time order dispatching for a fleet of autonomous mobile robots using multi-agent reinforcement learning. CIRP annals 69, 397-400. https://doi.org/10.1016/j.cirp.2020.04.001

[16] Kuhnle, A., Schaarschmidt, M., Fricke, K., 2017. Tensorforce: a tensorflow library for applied reinforcement learning.

[17] Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P., 2015. Trust region policy optimization, in: International conference on machine learning, PMLR. pp. 1889-1897.

[18] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms.

[19] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning.