The Challenges of Using Reinforcement Learning for Controlling Industrial Energy Systems

Tobias Lademann[1], **Théo Vincent**[2,3], **Jan Peters**[2,3,4], **Matthias Weigold**[1]
t.lademann@ptw.tu-darmstadt.de

[1]**Institute for Production Management, Technology and Machine Tools (PTW), Technical University of Darmstadt**
[2]**DFKI GmbH, SAIROL**
[3]**Department of Computer Science, Technical University of Darmstadt**
[4]**Hessian.ai, Technical University of Darmstadt**

Abstract

Reinforcement learning has shown promising results for optimizing the control of industrial energy systems, yet most existing studies remain limited to the application in simulation environments. We investigate the challenges of deploying reinforcement learning in a real-world industrial energy system, considering a thermal heating network as a use case. We formulate the task as a Markov Decision Process and systematically analyze the associated challenges along the structure of the formal description, including partial observability, action space design, reward design, and the simulation-to-reality gap. The challenges are grounded in an existing real-world deployment, where reinforcement learning achieves operational stability but shows a significant performance gap compared to simulation.

1 Introduction

Industrial energy systems convert final energy such as electricity or natural gas into useful energy for production processes and building operation (Thiede, 2012). Given their significant share of industrial energy demand (Rohde & Arnold-Keifer, 2023), improving the efficiency and flexibility of these systems while reducing operational costs is a key objective in industry (Thiede, 2012; Thiel & Stark, 2021).

Controlling industrial energy systems is a complex task due to interlinked thermal networks, multiple energy converters, the integration of energy storage systems, complex energy pricing models and strict requirements regarding security of supply (Kohne et al., 2020; Thiede, 2012). Conventional control strategies are typically designed based on expert knowledge and implemented as static rule-based approaches, combining hysteresis controllers for discrete decisions with PID controllers for continuous control tasks (Frank et al., 2024). Although these approaches are generally robust, they are limited in their ability to define optimal control actions given a complex system behavior and conflicting optimization targets (Stavrev & Ginchev, 2024).

To address these limitations, reinforcement learning (RL) has gained increasing attention in the energy system domain (Perera & Kamalaruban, 2021), as it promises several advantages compared to conventional controllers, such as the ability to handle complex, uncertain and dynamic environments (Stavrev & Ginchev, 2024). Existing studies demonstrate the potential of RL to improve the control of industrial energy systems. However, the application of RL in real-world environments remains limited: Most existing studies evaluate RL-based control strategies exclusively in simulation (Ranzau, 2025), neglecting the real-world challenges in an industrial setting (Dulac-Arnold et al., 2021).

This paper addresses this gap by investigating the challenges associated with the real-world deployment of RL for controlling an industrial energy system. It can be seen as an open call for RL methods to tackle application-relevant challenges, thereby guiding RL researchers in framing their research problems.

Contributions. (1) We model the control task of a representative industrial use case as an RL problem, tailored to the characteristics and constraints of real-world industrial applications. (2) We formulate challenges within the RL problem formulation. (3) We concretize the identified challenges based on a real-world deployment of an RL control strategy in the considered use case.

2 Use case

The *ETA Research Factory* serves as a real-world testbed for the deployment and evaluation of optimized control strategies under conditions representative of industrial energy systems. The energy system of the factory consists of three thermal networks, operating at different temperature levels (Frank et al., 2024). For this work, we consider the *Heating Network High Temperature*, as it captures key characteristics of industrial energy systems, including multiple energy converters, thermal storage, and coupled energy carriers (Frank et al., 2024). At the same time, it is the only thermal network that enables reproducible real-world experiments, as ambient influences are negligible (Lademann et al., 2026a). The use case offers realistic industrial complexity and safety constraints while reducing experimental costs and risks compared to productive industrial environments.

The considered system consists of multiple energy converters and thermal storage components that jointly supply thermal power to production processes and building operation as shown in Figure 1. In particular, it includes **two combined heat and power units** and **one condensing boiler**, which convert gas into thermal energy. The combined heat and power units additionally generate electricity, which is either used within the factory or fed into the grid, depending on the factory demand. Both reduce operating costs, depending on current electricity market prices and feed-in tariffs.

Thermal energy can be stored in an **active storage** system, allowing the temporal decoupling of energy production and consumption. Depending on the operating mode, the storage acts either as a thermal producer (discharge) or consumer (charge). This flexibility enables the operation of the combined heat and power units in periods of favorable electricity prices. The active storage is equipped with vacuum insulation and a layering system to reduce thermal and exergy losses. In addition, a passive **buffer storage**, which cannot be directly controlled, hydraulically decouples systems for thermal energy production from thermal energy consumption.

On the demand side, the system supplies both production processes and building processes within the factory. All relevant energy flows, including thermal, electrical, and gas power, are measured for each energy converter and consumer. In addition, each storage is equipped with three temperature sensors located at the lower, middle, and upper parts of the storage.

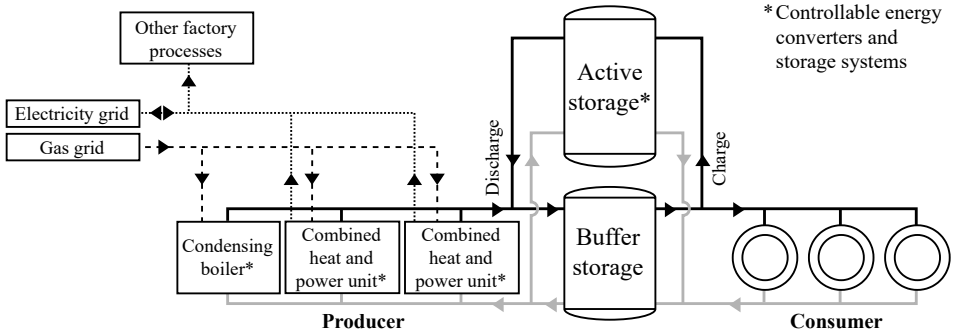

Figure 1: Simplified scheme of the use case system based on Frank et al. (2025).

The main objective of the system is to reliably supply thermal power to consumer systems at the required temperature level. The target temperature depends on the overall operating mode, with 70 °C in production mode and 55 °C in building mode. To ensure safe operation, the middle buffer storage

temperature must remain within predefined limits, ranging from $45\,^{\circ}\mathrm{C}$ to $85\,^{\circ}\mathrm{C}$ in production mode and from $40\,^{\circ}\mathrm{C}$ to $85\,^{\circ}\mathrm{C}$ in building mode.

RL has been applied to industrial energy systems in simulation and real-world settings, but deployment challenges remain largely unaddressed, as discussed in Appendix A.

3 Problem formulation

This section provides a general formulation of the control of the considered industrial energy system as an RL problem. The control task is described with a Markov Decision Process (MDP) (Sutton & Barto, 2018), where an agent interacts with the real-world environment. Given a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$, the environment transitions to a successor state $s' \in \mathcal{S}$ according to the transition probability $p(s' \mid s, a)$. A deterministic reward function of the state, action and next state is received after each transition. The agent learns a policy $\pi(a \mid s)$ that maps system states to action distributions.

State Space: A reasonable state space should include all observable system states and exogenous inputs that influence the system behavior and optimal control decisions. It should also include forecasts of relevant exogenous inputs, such as thermal demand of consumer systems, electrical demand of other factory processes, and energy prices, to enable optimal control decisions.

The state space is constructed from available sensor measurements acquired and provided by the building automation system (e.g. storage temperatures or heat flow rates) and exogenous inputs provided through external data interfaces (e.g. market prices). The state space can include both continuous (e.g. temperature) and discrete (e.g. energy converter state on/off) values.

Action Space: The RL agent controls the energy converters and active storage (see Figure 1). Each component is controlled by a boolean activation variable and a continuous setpoint, whose interpretation and value range depend on the component (Fuhrländer-Völker et al., 2022). The setpoint only affects the component when it is active.

The RL agent operates at a supervisory control level and interacts with the real-world system through a control interface on the programmable logic controller (PLC). The RL actions overwrite the default rule-based control implemented on the PLC. A fallback mechanism ensures safe operation by restoring the rule-based controller if the buffer storage temperature limits defined in section 2 are violated. As only the supervisory control is overwritten, the local control implemented on the PLC remains active and continues to handle actuator sequencing (e.g. pumps and valves) and low-level control tasks such as temperature control (Fuhrländer-Völker et al., 2022).

Transition Dynamics: The system state transitions according to the thermal, hydraulic and electrical dynamics of the system. State transitions of the physical system are influenced by the control actions, thermal demand, and the local control logic implemented in the building automation system. In addition, exogenous variables such as energy prices and production modes evolve independently of the agent actions according to external influences.

Reward: The performance of a control strategy can be evaluated using several evaluation metrics addressing different objectives. These objectives are partially conflicting, requiring the control strategy to balance trade-offs between economic, technical, and environmental performance (Posch, 2011). The following evaluation metrics are relevant for the presented use case, as defined by Lademann et al. (2026a):

A. **Operating costs** comprise electricity costs, gas costs, and maintenance costs. Electricity-related costs include expenses for external electricity purchase under a tariff based on the day-ahead market, as well as revenues from self-consumption and feed-in of locally generated electricity by the combined heat and power units.

B. **Security of supply** is essential to guarantee that thermal demands of production and building systems are reliably met at all times. The security of supply is evaluated based on the mode-

dependent temperature target $55\,^{\circ}\mathrm{C}$ or $70\,^{\circ}\mathrm{C}$ and the actual upper buffer storage temperature. Only deviations below the target temperature are problematic.

C. **Energy efficiency** is considered to minimize overall energy consumption and improve the effective use of available energy resources. It is calculated as the ratio of total useful energy output, including thermal and electrical energy, to the total energy input from gas and electricity.

D. **System wear** is considered to prevent excessive component degradation caused by frequent switching. For the combined heat and power units, the manufacturer specifies a minimum mean runtime of $3\,\mathrm{h}$. In general, longer runtimes are beneficial, as they reduce start-stop cycles, wear, and associated early component degradation. Additionally, the combined heat and power units are subject to a maximum allowable return temperature of $65\,^{\circ}\mathrm{C}$.

E. **CO_2 emissions** are considered to reduce the environmental impact of system operation and support decarbonization goals.

Terminal States: There are no terminal states. The building automation system enforces admissible temperature ranges and temporarily overrides the RL control actions with rule-based fallback control if these limits are violated. However, this does not terminate the deployment, as the RL control resumes once the system state returns to the admissible operating range.

4 Challenges

This section discusses challenges related to the problem formulation and real-world deployment of RL in the considered system. For clarity, the challenges are presented following the structure of the previously introduced MDP formulation.

System State: The system state is only partially observable, as it is often the case for real-world systems (Dulac-Arnold et al., 2021; Xiang & Foo, 2021). In particular, the state of charge of the storage systems, especially the active storage, cannot be accurately determined due to strong temperature stratification and the limited number of temperature sensors. While the presented use case provides extensive sensor information for all components, this level of observability is typically not available in industrial systems (Thiede, 2012).

Furthermore, relevant future state information, such as forecasts of thermal or electrical demand depend heavily on production-related-factors and may be unavailable or uncertain (Walther, 2022). Additional uncertainty arises from sensor inaccuracies or malfunctions.

Action Space: The design of the action space is critical for real-world deployment, as industrial energy systems offer multiple possible ways to formulate control actions. While the action space must be compatible with the existing building automation interface, its formulation can still include several design choices, for example removing redundant actions, combining multiple control signals into one action, or discretizing continuous control variables (Kanervisto et al., 2020).

A priori, it is not clear which formulation facilitates learning most effectively and leads to the best performance. Integrating domain knowledge into the action space formulation can speed up the learning process. However, it also restricts the agent's degrees of freedom and may exclude the true optimal solution (Kanervisto et al., 2020), thereby limiting the achievable performance gains compared to a rule-based controller that already incorporates substantial inductive bias.

Transition Dynamics: All works discussed in Appendix A rely on a simulation model to obtain transition dynamics during training. However, the implementation and validation effort for high-fidelity simulation models is substantial. Moreover, detailed information on the local controller logic and component behavior is often not available from manufacturers.

This issue is further intensified by the fact that industrial energy systems are subject to continuous changes, ranging from gradual efficiency degradation of energy converters to major system modifications such as additional components or changed energy pricing models. Such changes may require repeated adaptation of the simulation model and retraining of the RL agent, which generally includes new hyperparameter searches.

Furthermore, detailed simulation models can result in stiff differential-algebraic equations and long simulation times due to the coupling of dynamics with different time constants, such as slow thermal dynamics and discrete switching operations (Blum et al., 2021). As a result, RL training and especially hyperparameter optimization become computationally expensive.

Reward: In industrial energy systems, relevant objectives are often conflicting (e.g. operational costs and security of supply) and must be translated into a reward function that reflects the priorities of the system operator.

A common approach is to formulate the reward as a weighted sum. As proper weighting of multiple objectives is essential (Schäfer et al., 2025), the reward design should reflect realistic operational priorities from the perspective of a system operator. However, many operational objectives, such as security of supply, energy efficiency, or system wear, cannot be directly quantified in monetary terms and therefore require relative weighting within the reward design.

Additionally, rewards are often delayed, particularly for operational costs when storage systems are used to shift energy production and consumption over time to exploit periods of favorable electricity prices.

Agent's Objective: The objective of the RL agent can be formulated in different ways, including finite-horizon, average-reward, or discounted-return objectives (Sutton & Barto, 2018). A finite-horizon setting is not suited as the system is operated continuously. An average-reward setting appears favorable, as it mitigates the short-sighted effect of the discounted-return setting. However, this setting is still largely unexplored. Finally, the discounted setting requires specifying a discount factor, which should be tuned to balance the trade-off between capturing delayed rewards and prioritizing immediate rewards due to future uncertainty.

Since operational costs are strongly influenced by day-ahead market prices, the agent's objective should capture at least daily electricity price cycles. In addition, weekly factory demand patterns should be considered, as thermal and electrical demand typically differ between weekdays and weekends. However, increasing uncertainty in future states, thermal and electrical demand forecasts, and market conditions favors a stronger emphasis on immediate rewards.

Practical RL training and evaluation are usually performed in an episodic setting, requiring the definition of terminal states, suitable time horizons, and realistic and diverse initial state distributions (Schäfer et al., 2025). These design choices can strongly affect learning and performance (Schäfer et al., 2025).

Control Frequency: The control frequency must be high enough to capture the relevant system dynamics. However, thermal systems typically exhibit slow dynamics compared to other control tasks such as power grids (Schlegel et al., 2025). High control frequencies lead to longer planning horizons, which can make the control problem harder.

Furthermore, if a simulator is used, the control frequency can be constrained by the computational effort of the simulation model, as smaller communication step sizes may force a variable-step solver to decrease its step size, which significantly increases runtime.

Explainability: Industrial deployment requires not only good performance but also system operator acceptance. Explainable RL can support the interpretation of learned policies (Bekkemoen, 2024). However, the challenge is broader than explaining individual control actions. Even a well-performing RL agent may show operating patterns that differ substantially from a well-known rule-based baseline, which can reduce operator acceptance.

5 A Study of Real-World Application

This section analyzes the real-world deployment of RL in the considered system by Lademann et al. (2026a) and links the formulation choices to the challenges we identified. Again, we present the application following the MDP structure.

State Space: The selected state space comprises all available storage temperatures (three per storage), the aggregated thermal demand of the consumer systems, the current operational state of the controlled systems (on/off;charge/discharge), the current runtime of the combined heat and power units, the production and heating mode of the factory, the electrical demand of other factory processes as well as time-dependent energy prices and remunerations. Although these signals would be available in the considered system, the authors do not include forecasts, thermal power of producer systems, or electrical and gas energy consumption in the state space, reflecting the partial observability typically encountered in industrial systems with limited sensor availability.

Action Space: The authors address the challenge of redundant actions by deriving a discrete action space from the rule-based control strategy. For each energy converter i, activation and temperature setpoints are combined into an action $a_i \in \mathcal{A}_i = \{0, 1, 2\}$, denoting off and operation at two predefined setpoints. For the active storage, the same action set represents discharge, idle, and charge at a fixed minimum setpoint. The overall action space is $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_4$. Thereby, the continuous setpoints are discretized to improve operational stability in real-world deployment, but this also limits the agent's degrees of freedom and, therefore, its optimization potential.

In addition, the authors use an action mask for the combined heat and power units to enforce a minimum runtime of $3\,\mathrm{h}$, further supporting operational stability in deployment.

Transition Dynamics: The system is modeled in *Modelica* (Mattsson & Elmqvist, 1997) using the *Thermal Systems Control Library* (Borst et al., 2023), a library dedicated for developing optimized control strategies. The model includes physical component models as well as supervisory and local control logic, thereby representing the real-world system behavior and control interface.

Additionally, the simulation model is calibrated using real-world data to reduce the simulation-to-reality (sim-to-real) gap. The authors report an absolute relative error between simulated and measured accumulated energy of $6\,\%$ for electrical energy and $5\,\%$ for gas energy over a two-day validation horizon. The calibration results further indicate unresolved model mismatch or measurement errors, as several calibrated parameters converge to the boundaries of their predefined ranges. This highlights the modeling and validation effort required when RL training relies on simulation-based transition dynamics.

Reward: The authors address the reward-design challenge by formulating the reward as a weighted sum of monetary and non-monetary cost terms. Monetary operating costs include electricity, gas, and maintenance costs, while non-monetary reward terms include temperature deviations, switch-on operations, minimum runtime violations, and episode termination.

The resulting weighting challenge is addressed by weighting only the non-monetary cost terms relative to the monetary terms. The rule-based strategy is used as a reference to obtain realistic and justifiable weights.

Agent's Objective: The authors chose a discounted return. To cope with the challenge of defining a discount factor, the authors include it in the hyperparameter search space and chose the undiscounted return as an evaluation metric.

The training horizon is set to $7\,\mathrm{d}$ to capture daily and weekly variations in electricity prices as well as thermal and electrical demand. The evaluation horizon depends on the experiment setup.

To obtain a diverse initial state distribution, the authors use a pseudo-random generator to sample from the available training data, while storage temperatures are initialized from predefined temperature ranges. In real-world evaluation, the initial state cannot be controlled and is determined by the actual operating conditions at the experiment start.

Control Frequency: The control frequency is set to $180\,\mathrm{s}$, favoring fast simulation times and a simplified control problem. However, whether this frequency is sufficiently high to fully capture the relevant system dynamics is not further analyzed.

The authors implement the experiments within the open-source Python framework *ETA Factory Thermal System Operation* (Lademann et al., 2026b) using proximal policy optimization with invalid action masking (Huang & Ontañón, 2022). Table 1 compares the rule-based and RL controllers in simulation and real-world deployment. Overall, the RL controller demonstrates operational stability in both simulation and real-world deployment.

Both controllers are evaluated under identical scenarios and initial states in simulation. In the real-world deployment, the scenario conditions (production/building mode, thermal demand of consumer systems and electrical power of other factory processes) are controlled as well. However, the initial states differ between real-world experiments, highlighting the challenge of comparability in real-world deployment using the undiscounted return.

To address this, the authors propose metrics A–E, whose normalization enables comparison despite differing initial states and time horizons. Performance is evaluated without termination to reflect realistic operation, since RL control resumes once the temperature returns to the admissible range after a fallback intervention. Temperature violations are therefore captured by the security of supply metric.

Table 1: Performance evaluation of rule-based and RL control strategy over horizon T. The arrows indicate whether the value should be maximized (↑) or minimized (↓).

		Simulation ($T = 90\,\mathrm{d}$)		**Real-world** ($T = 3\,\mathrm{d}$)	
Metric	**Unit**	**Rule-Based**	**RL**	**Rule-Based**	**RL**
A. Operating costs (↓)	$\mathrm{ct/kWh}$	8.8	**7.2**	**8.3**	8.6
B. Security of supply (↓)	K	1.1	**0.9**	**1.1**	3.5
C. Energy efficiency (↑)	%	69.7	**72.1**	**74.7**	72.4
D. System wear (↑)	h	**15.6**	4.4	**11.2**	10.8
E. CO_2 emissions (↓)	$\mathrm{gCO_2/kWh}$	266.2	**242.3**	279.6	**275.8**
Sum of reward (↑)	–	-1839	**-1345**	**-95**	-297

In both simulation and real-world deployment, the agent permanently activates the active storage and frequently switches between charging and discharging, while the second combined heat and power unit remains completely inactive. This behavior supports operational stability by increasing the effective thermal inertia of the system, but does not exploit volatile energy market prices. We attribute this primarily to three coupled issues: first, the absence of forecasts in the state space limits the agent's ability to anticipate favorable market conditions; second, the reward weighting and delayed returns may not sufficiently support long-term cost optimization; and third, the discount factor underweights future rewards. Moreover, this behavior differs fundamentally from the rule-based strategy, which may reduce operator acceptance.

In real-world deployment, the performance of RL is less promising than in simulation with respect to the metrics A–D. The RL controller shows a significantly larger deviation in security of supply, resulting in a substantially lower sum of rewards compared to the rule-based controller. Although a calibrated simulation model based on a dedicated simulation library is used, the RL performance achieved in simulation does not fully transfer to the real-world system. This highlights the challenge of simulation-based transition dynamics.

6 Limitations, Conclusion & Future Work

This study only discusses the challenges of real-world RL deployment based on a single industrial energy system. Therefore, effects caused by other system topologies, energy converters, storage technologies, or demand structures are not covered. Studying different use cases could reveal more challenges.

A further limitation is the selected use case itself. The ETA Research Factory is a research testbed that enables controlled real-world deployment and representative safety constraints. However, additional technical, organizational, and economic challenges are expected when transferring RL-based control to productive industrial systems.

To conclude, we provide a structured discussion of the challenges associated with formulating and deploying RL for real-world industrial energy systems. The analysis shows that many challenges arise at the level of the RL problem formulation, including partial observability, action space design, reward weighting, delayed rewards, and the dependence on simulation-based transition dynamics. Grounded in an existing real-world deployment, we demonstrate how the identified formulation challenges arise in practice. Thus, our contribution can be seen as a starting point for developing RL to address application-relevant challenges.

In future work, we will investigate offline RL to reduce the dependence on high-fidelity simulation models and make better use of available real-world operational data (Levine et al., 2020). In addition, online learning in the considered use case appears feasible, but requires methods for safe exploration and high sample efficiency (Gu et al., 2024; Dulac-Arnold et al., 2021).

Acknowledgments

The authors thankfully acknowledge the financial support of the project "ENIPRO" (grant no. 03EN4111A) by the Federal Ministry for Economic Affairs and Energy (BMWE) and project supervision by the project management organization Projektträger Jülich (PtJ). This research was supported by "Third Wave of AI", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art, and by the grant "Einrichtung eines Labors des Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) an der Technischen Universität Darmstadt". We gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant no. 01IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003).

References

Yanzhe Bekkemoen. Explainable reinforcement learning (XRL): A systematic literature review and taxonomy. *Machine Learning*, 113(1):355–441, January 2024. ISSN 0885-6125, 1573-0565. DOI: 10.1007/s10994-023-06479-7.

David Blum, Javier Arroyo, Sen Huang, Ján Drgoňa, Filip Jorissen, Harald Taxt Walnum, Yan Chen, Kyle Benne, Draguna Vrabie, Michael Wetter, and Lieve Helsen. Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings. *Journal of Building Performance Simulation*, 14(5):586–610, September 2021. ISSN 1940-1493, 1940-1507. DOI: 10.1080/19401493.2021.1986574.

Fabian Borst, Michael Geord Frank, Lukas Theisinger, and Matthias Weigold. ThermalSystemsControlLibrary: A Modelica Library for Developing Control Strategies of Industrial Energy Systems. *Proceedings of the 15th International Modelica Conference 2023, Aachen, October 9-11*, January 2023. DOI: 10.3384/ecp204.

Yan Du, Fangxing Li, Kuldeep Kurte, Jeffrey Munk, and Helia Zandi. Demonstration of Intelligent HVAC Load Management With Deep Reinforcement Learning: Real-World Experience of Machine Learning in Demand Control. *IEEE Power and Energy Magazine*, 20(3):42–53, May 2022. ISSN 1540-7977, 1558-4216. DOI: 10.1109/MPE.2022.3150825.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, September 2021. ISSN 0885-6125, 1573-0565. DOI: 10.1007/s10994-021-05961-4.

Michael Frank, Fabian Borst, Lukas Theisinger, Tobias Lademann, Daniel Fuhrländer-Völker, and Matthias Weigold. Framework for Implementation of Building Automation Control Programs for Industrial Heating and Cooling Systems. *Energies*, 17(21):5361, October 2024. ISSN 1996-1073. DOI: 10.3390/en17215361.

Michael Frank, Fabian Borst, Lukas Theisinger, Tobias Lademann, and Technische Universität Darmstadt. Hydraulic scheme ETA Factory Thermal Supply Systems. October 2025. DOI: 10.48328/TUDATALIB-1437.2.

Daniel Fuhrländer-Völker, Fabian Borst, Lukas Theisinger, Heiko Ranzau, and Matthias Weigold. Modular data model for energy-flexible cyber-physical production systems. *Procedia CIRP*, 107: 215–220, 2022. ISSN 22128271. DOI: 10.1016/j.procir.2022.04.036.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A Review of Safe Reinforcement Learning: Methods, Theories, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, December 2024. ISSN 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2024.3457538.

Kun He, Qiming Fu, You Lu, Yunzhe Wang, Jun Luo, Hongjie Wu, and Jianping Chen. Predictive control optimization of chiller plants based on deep reinforcement learning. *Journal of Building Engineering*, 76:107158, October 2023. ISSN 23527102. DOI: 10.1016/j.jobe.2023.107158.

Shengyi Huang and Santiago Ontañón. A Closer Look at Invalid Action Masking in Policy Gradient Algorithms. *The International FLAIRS Conference Proceedings*, 35, May 2022. ISSN 2334-0762. DOI: 10.32473/flairs.v35i.130584.

Anssi Kanervisto, Christian Scheller, and Ville Hautamäki. Action Space Shaping in Deep Reinforcement Learning. (arXiv:2004.00980), May 2020. DOI: 10.48550/arXiv.2004.00980.

Thomas Kohne, Heiko Ranzau, Niklas Panten, and Matthias Weigold. Comparative study of algorithms for optimized control of industrial energy supply systems. *Energy Informatics*, 3(S1):12, October 2020. ISSN 2520-8942. DOI: 10.1186/s42162-020-00115-7.

Tobias Lademann, Andreas Clement, Jan-Niklas Witt, Michael Frank, Jan Zangenberg, Stefan Niessen, and Matthias Weigold. Real-World Benchmarking of Control Strategies in an Industrial Energy System. 2026a. Submitted to Advances in Applied Energy.

Tobias Lademann, Andreas Clement, Jan-Niklas Witt, Michael Frank, Jan Zangenberg, and Matthias Weigold. ETA Factory Thermal System Operation v1.1.0. May 2026b. DOI: 10.5281/ZENODO.20050236.

Tobias Lademann, Arthur Stobert, Heiko Ranzau, Jonas Klingelhöfer, Manuel Scharfe, and Matthias Weigold. Deep Reinforcement Learning for Control Strategy Optimization of an Industrial Cooling Supply System in the Chemical and Pharmaceutical Industry. In Holger Kohl, Günther Seliger, Franz Dietrich, and Giampaolo Campana (eds.), *Safe and Sustainable Value Creation by Design*, pp. 122–130. Springer Nature Switzerland, Cham, 2026c. ISBN 978-3-032-21153-8 978-3-032-21154-5. DOI: 10.1007/978-3-032-21154-5_14.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. (arXiv:2005.01643), November 2020. DOI: 10.48550/arXiv.2005.01643.

Sven Erik Mattsson and Hilding Elmqvist. Modelica - An International Effort to Design the Next Generation Modeling Language. *IFAC Proceedings Volumes*, 30(4):151–155, April 1997. ISSN 14746670. DOI: 10.1016/S1474-6670(17)43628-7.

A.T.D. Perera and Parameswaran Kamalaruban. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137:110618, March 2021. ISSN 13640321. DOI: 10.1016/j.rser.2020.110618.

Wolfgang Posch. *Ganzheitliches Energiemanagement für Industriebetriebe*. Gabler, Wiesbaden, 2011. ISBN 978-3-8349-2585-5 978-3-8349-6645-2. DOI: 10.1007/978-3-8349-6645-2.

Heiko Ranzau. Application of Deep Reinforcement Learning for the Operational Optimization of Industrial Energy Supply Systems. January 2025. DOI: 10.26083/tuprints-00030077.

Clemens Rohde and Sonja Arnold-Keifer. Erstellung von Anwendungsbilanzen für die Jahre 2021 bis 2023 für die Sektoren Industrie und GHD. October 2023. URL https://ag-energiebilanzen.de/wp-content/uploads/2024/01/Anwendungsbilanz_Industrie_2022_vorlaeufig-update_20231030.pdf. Accessed: 2026-05-22.

Georg Schäfer, Tatjana Krau, Jakob Rehrl, Stefan Huber, and Simon Hirlaender. The Crucial Role of Problem Formulation in Real-World Reinforcement Learning. (arXiv:2503.20442), March 2025. DOI: 10.48550/arXiv.2503.20442.

Matthew Kyle Schlegel, Martha White, Mostafa Farrokhabadi, and Matthew E. Taylor. Towards understanding the challenges of applying reinforcement learning to the power grid. In *RLC 2025 Workshop on Practical Insights into Reinforcement Learning for Real Systems*, 2025. URL https://openreview.net/forum?id=VnoIY8IKUU. Accessed: 2026-05-22.

Thomas Schreiber, Sören Eschweiler, Marc Baranski, and Dirk Müller. Application of two promising Reinforcement Learning algorithms for load shifting in a cooling supply system. *Energy and Buildings*, 229:110490, December 2020. ISSN 03787788. DOI: 10.1016/j.enbuild.2020.110490.

Stefan Stavrev and Dimitar Ginchev. Reinforcement Learning Techniques in Optimizing Energy Systems. *Electronics*, 13(8):1459, April 2024. ISSN 2079-9292. DOI: 10.3390/electronics13081459.

Richard S. Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts London, England, second edition edition, 2018. ISBN 978-0-262-03924-6.

Sebastian Thiede. *Energy Efficiency in Manufacturing Systems*. Sustainable Production, Life Cycle Engineering and Management. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-25913-5 978-3-642-25914-2. DOI: 10.1007/978-3-642-25914-2.

Gregory P. Thiel and Addison K. Stark. To decarbonize industry, we must decarbonize heat. *Joule*, 5(3):531–550, March 2021. ISSN 25424351. DOI: 10.1016/j.joule.2020.12.007.

Jessica Walther. Hierarchical Electrical Load Forecasting of Industrial Production Systems in the Manufacturing Industry based on Deep Learning. 2022. DOI: 10.26083/TUPRINTS-00021767.

Xiao Wang, Xuyuan Kang, Jingjing An, Hanran Chen, and Da Yan. Reinforcement learning approach for optimal control of ice-based thermal energy storage (TES) systems in commercial buildings. *Energy and Buildings*, 301:113696, December 2023. ISSN 03787788. DOI: 10.1016/j.enbuild.2023.113696.

Matthias Weigold, Heiko Ranzau, Sarah Schaumann, Thomas Kohne, Niklas Panten, and Eberhard Abele. Method for the application of deep reinforcement learning for optimised control of industrial energy supply systems by the example of a central cooling system. *CIRP Annals*, 70(1): 17–20, 2021. ISSN 00078506. DOI: 10.1016/j.cirp.2021.03.021.

Xuanchen Xiang and Simon Foo. Recent Advances in Deep Reinforcement Learning Applications for Solving Partially Observable Markov Decision Processes (POMDP) Problems: Part 1—Fundamentals and Applications in Games, Robotics and Natural Language Processing. *Machine Learning and Knowledge Extraction*, 3(3):554–581, July 2021. ISSN 2504-4990. DOI: 10.3390/make3030029.

Li Yi, Pascal Langlotz, Marco Hussong, Moritz Glatt, Fábio J.P. Sousa, and Jan C. Aurich. An integrated energy management system using double deep Q-learning and energy storage equipment

to reduce energy cost in manufacturing under real-time pricing condition: A case study of scale-model factory. *CIRP Journal of Manufacturing Science and Technology*, 38:844–860, August 2022. ISSN 17555817. DOI: 10.1016/j.cirpj.2022.07.009.

Dafeng Zhu, Bo Yang, Yuxiang Liu, Zhaojian Wang, Kai Ma, and Xinping Guan. Energy management based on multi-agent deep reinforcement learning for a multi-energy industrial park. *Applied Energy*, 311:118636, April 2022. ISSN 03062619. DOI: 10.1016/j.apenergy.2022.118636.

A Related Work

RL has been applied to the control of industrial energy systems in both simulation and real-world applications comparable to the considered use case. However, existing research primarily focuses on the implementation and benchmarking of RL, while challenges related to real-world deployment remain largely unaddressed.

RL has been applied to various industrial energy systems including cooling supply systems for buildings and process cooling (Schreiber et al., 2020), multi-energy industrial parks (Zhu et al., 2022), cooling systems of an industrial production site (Weigold et al., 2021), or cooling supply systems in the chemical and pharmaceutical industry (Lademann et al., 2026c). RL has also been applied in the building energy sector (Wang et al., 2023; He et al., 2023), which exhibit comparable network topology, although building energy systems are typically less complex regarding control complexity and safety requirements. While the presented studies demonstrate the potential of RL, all implementations are exclusively in simulation environments, with real-world deployment challenges remaining an open problem.

A very limited number of studies address real-world deployment of RL in industrial energy systems (Ranzau, 2025). Ranzau (2025) deploys RL in the real-world system of the ETA Research Factory and develops methods for improving the sim-to-real transfer. Lademann et al. (2026a) consider the use case described in section 2 to develop common evaluation metrics for different control strategies and benchmark the performance of RL and other controllers in real-world application. Real-world RL applications have also been demonstrated in related domains, such as building climate control (Du et al., 2022) and industrial battery control (Yi et al., 2022). Again, these application areas exhibit less complexity compared to industrial energy systems. However, despite considering real-world deployment, the systematic investigation of challenges associated with applying RL in real-world industrial environments is not the primary focus of any of the mentioned studies.

Prior work has also investigated challenges associated with the real-world deployment of RL in other applications. Dulac-Arnold et al. (2021) identify and formalize a series of challenges in the real-world deployment of RL. The study does not consider a specific application area. Schlegel et al. (2025) analyze the real-world challenges presented by Dulac-Arnold et al. (2021) for the application in power grids. Schäfer et al. (2025) demonstrate and discuss challenges associated with RL problem formulation choices and their effect on performance for the control of a helicopter testbed considering simulation-based and real-world training.

This work builds upon the real-world RL deployment presented by Lademann et al. (2026a). We formulate the RL problem and identify the associated challenges. Finally, we analyze these challenges in the real-world application.