# Recommendations to Make Reinforcement Learning Practical in Building Control Applications

**Sourav Dey – University of Colorado Boulder – sourav.dey@colorado.edu**
**Gregor Henze – University of Colorado Boulder – gregor.henze@colorado.edu**

## Abstract

The paper provides an analysis of the application of reinforcement learning (RL) in the domain of building controls, summarizing four years of research. The primary focus is exploring RL's potential to adaptively learn from building data, bypassing the need for individualized extensive building modeling efforts and enabling the transfer and adaption of trained agents to similar building environments. Despite its promising prospects, RL faces challenges such as extended training durations, instability during early exploration phases, and issues in interpreting the actions of trained agents. The research was focused on two core areas. The first area investigates strategies to enhance RL agents' learning efficiency and stability in building control contexts with approaches such as imitation learning, inverse RL, and online learning with guided exploration with surrogate models utilizing rule-based controls, showing significant improvements in the training process. The second area addresses the critical aspects of scalability and interpretability of RL agents. It examines the feasibility of transferring trained agents to various buildings, potentially with new objectives, highlighting RL's adaptability and practical applicability in real-world building control scenarios. In summary, this paper consolidates critical findings from the research and offers actionable insights and recommendations for practical deployment and training RL in building energy management systems without extensive building modeling efforts. It emphasizes the transformative potential of RL in this field and suggests avenues for future exploration and development.

## 1. Introduction

Reinforcement learning (RL) as an advanced control strategy holds significant promise in revolutionizing building controls by offering a dynamic, adaptable approach to optimization without extensive modelling efforts for distinct buildings. (Sutton et al., 2014; Chen et al., 2020). Each action taken by the agent results in a reward, positive or negative, which serves as feedback to guide learning. The agent follows a policy—a strategy for selecting actions based on the current state—that can be either deterministic or stochastic. Critical to RL are the value functions: the state value function (V(s)), which estimates the expected return from a state under a given policy, and the action-value function (Q(s, a)), which estimates the expected return from taking an action in a state and thereafter following a policy. The return itself is a total accumulated future reward, typically discounted by a factor ($\gamma$) to prioritize immediate rewards over distant ones and to ensure the return is finite. RL involves a balance between exploration, where the agent tries new actions to learn about their effects, and exploitation, where the agent chooses actions that have previously resulted in high rewards. This dynamic of exploration and exploitation enables the agent to refine its policy through trial-and-error, applying to diverse applications such as robotics, gaming, finance, and healthcare, where systems are required to make sequential decisions to achieve optimal outcomes (Mnih et al., 2013; Lillicrap et al., 2015; Arora & Doshi, 2018; Watkins & Dayan, 1992).

Unlike traditional methods like rule-based controls (RBC) and model predictive control (MPC) (Richalet et al., 1978), RL learns optimal actions through repeated interactions and can adapt to changing dynamics over time. Despite its potential, RL faces

Part of
Pernigotto, G., Ballarini, I., Patuzzi, F., Prada, A., Corrado, V., & Gasparella, A. (Eds.). 2025. Building simulation applications BSA 2024. bu,press.
https://doi.org/10.13124/9788860462022

CC BY-SA

challenges such as long training times and unstable exploratory behavior in the early stages (Wang & Hong, 2020). However, addressing these challenges could unlock its potential for practical implementation in building control systems, paving the way for more efficient and adaptive management of indoor environments while reducing energy consumption and maintaining occupant comfort.

This paper summarizes insights gathered over a four-year research period on the application of RL in building energy management. It acknowledges certain limitations, notably the reliance on supervisory control actions and a lack of optimization for multiple low-level control points, while also neglecting multi-agent coordination across multiple buildings. The primary focus of the research was to improve the learning efficiency, assessing RL's scalability across diverse building types, and enhancing the interpretability of trained agents without extensive modelling efforts but sometimes assisted with "surrogate" models. A "surrogate" model is an approximate representation of a complex real-world system used to facilitate efficient and safe training of reinforcement learning (RL) agents. In this research practical challenges related to hardware and software implementation are omitted, with emphasis instead placed on theoretical advancements and practical implications. This section below serves as a brief overview of the research approaches investigated.

The approaches investigated in this research over the course of the study consist mainly of four types and are mentioned below:

- **Online learning** without surrogate models
  - Pure direct training
  - Imitation learning
- **Offline learning** without surrogate models
- **Hybrid approaches** with surrogate models
  - Offline learning
  - Inverse RL
  - Online learning with guided exploration
  - Online learning with guided exploration and imitation learning
- **Transfer learning**
  - Transductive learning
  - Inductive learning

The first two approaches, online learning and offline learning methods, examine approaches without the need for any surrogate modelling techniques, while the hybrid approaches utilize surrogate models to assist with the learning process. There are two types of online learning: i) pure direct training and ii) imitation learning. These are mentioned in detail in Section 2. Section 3 covers offline learning on historical data. There are four hybrid approaches explored in Section 4, which are i) offline learning, ii) inverse RL, iii) online learning with a guided exploration, and iv) online learning with guided exploration and imitation learning.

The research on transfer learning in Section 5 builds on the previous sections addressing the unstable behaviour of the learning agent as well as the issue of scalability of a trained agent such that it can be utilized to optimize for similar tasks on different environments.

## 2. Novelty and Contributions

This paper's contributions are multi-faceted, reflecting four years of intensive research into practical reinforcement learning (RL) applications in building control systems. Key contributions include strategies to enhance RL agents' learning efficiency and stability through imitation learning, inverse RL, and online learning with guided exploration using surrogate models, significantly improving training processes. Efforts were also made to improve the interpretability of RL agents' actions, essential for practical deployment. Hybrid learning approaches combined offline learning, inverse RL, and online learning with guided exploration, utilizing surrogate models to enhance learning outcomes and mitigate negative impacts of exploratory actions. The paper consolidates findings, offering actionable insights and recommendations for practical RL deployment in building energy management systems, avoiding extensive modelling efforts. Additionally, transfer learning techniques, such as transductive and inductive learning, were investigated to enhance RL agents' scalability, showing potential for reduced training times and improved initial performance. This research significantly advances practical RL applications in building control systems for building energy management with supervisory controls, addressing key challenges and paving the way for future developments.

# 3. Online Learning Without Surrogate Models

## 3.1 Pure Direct Training

Pure direct training is the traditional direct training approach where the RL agent starts without any prior training or knowledge, the so-called tabula rasa, and learns to optimize only by virtue of the interactions with the building environment without any form of assistance from surrogate models or without any pre-training or other learning approaches utilizing rule-based data. The agents start without any prior knowledge with randomly initialized parameters. The agent explores the environment by trial and error and with environmental feedback.

**Findings and recommendations:**

This type of RL agent should not be applied directly to an actual commercial building without pretraining or some kind of assistance from surrogate models. As evident from the results, depending on the complexity of the problem, it can take months or potentially years of training to learn an optimal operational strategy. Such protracted training periods would cause thermal discomfort for the building occupants, and it is unacceptable to bear thermal discomfort for months during the training phase of the RL agent. Moreover, the exploration phase intrinsic to the training phase can inflict damage on the building systems and components during the exploration stage of the training.

## 3.2 Imitation Learning

Imitation learning has been a key learning approach in the domain of autonomous behavioral systems commonly seen in robotics, computer games, industrial applications, and manufacturing, as well as autonomous driving. Imitation learning aims at mimicking a human behavior or another agent that is considered to perform well in a particular task. This is essentially learning to directly map observations to actions. It aids in reducing the task of teaching an agent, by showing the agent the actions to take to complete a specific task. Here an artificial dataset covering the state-space and its corresponding rule-based actions were developed, and the agent was trained to imitate the actions in a supervisory fashion (Dey et al., 2023).

**Findings and recommendations:**

The quality of the starting agent depends on the quality of the rule-based policy developed. With the increasing complexity of a multi-objective problem, it becomes difficult to design a good rule-based policy. The advantage is that it does not require extensive model development, and the agent adapts its starting imitation policy to be more suited to the environmental problem. The challenge of imitation learning is that since the agent blindly follows the imitated policy without any understanding of the environment, this leads to a ‘performance dip’ where the agent may start to explore in wrong directions before finding a better solution. We have found that even with the ‘performance dip’ the agent still has a better training progress curve than a pure direct training approach. The ‘performance dip’ can be addressed by a hybrid approach as mentioned in Section 4.4. The ‘performance dip’ is a characteristic of the direct imitation learning approach which suggests a superficial level of learning at the beginning, where the agent lacks a deeper comprehension of the consequences and implications of its actions.

# 4. Offline Learning Without Surrogate Models

Buildings are typically managed using rule-based energy management strategies, which consist of conditional rules based on various indoor and outdoor conditions. Offline training relies on learning from regular operational building data that follows these rule-based strategies. Prior to the application of an RL agent, the agent has access to historical building data, which is then pre-processed to create a tuple comprising state (*s*), action (*a*), next state (*s*'), reward (*r*), and a completion flag (*d*). The agent is exclusively trained on this operational data. (Dey et al., 2023)

**Findings and recommendations:**

Offline learning on historical rule-based data has led to poor and unacceptable performance. Rule-based building data tends to be sparse, which is often insufficient for effective RL training because it fails to adequately explore the state space. We have found that data augmentation leading to a sufficiently rich dataset exploring the state-space is necessary for learning a good policy at the onset of interacting with a real building.

# 5. Hybrid Approaches

## 5.1 Offline Learning With Metamodel

Pre-training with a metamodel often proves to be an effective strategy in reinforcement learning when considering deployment in an actual commercial building. This approach not only reduces the training duration but also mitigates the unstable exploratory behavior that can be observed with direct training methods. The RL agent can systematically and comprehensively explore the state space within a simulation environment where there are no repercussions on people and systems as in an actual building. This aids the agent in forming more accurate value estimates regarding the potential outcomes of specific actions in particular states. Furthermore, the metamodel enables extended training periods in the simulation, eliminating the need to wait for months or even years in real time to achieve optimal action learning. While training with a metamodel appears promising, it presents several challenges.

Specifically, the quality of the metamodel depends on the quality of the building data available. Large variability in building data is beneficial as it leads to developing a more accurate metamodel. The RL agent can then explore the state-space using the metamodel without incurring any of the negative consequences of exploration and can also discover regions of high rewards that the agent can exploit in the real building environment. To ensure a fair comparison, we extract the metamodel from the same training rule-based building data which was used for the inverse RL process. (Dey et al., 2023)

**Findings and recommendations:**

The effectiveness of metamodel-based training relies on the accuracy of the metamodel in replicating the real building environment. Any discrepancies or model mismatches between the metamodel and the actual building environment can lead to subpar performance when the RL agent is deployed in the real world. Also, the metamodel extraction process is often challenging and requires engineering expertise. Furthermore, it is difficult to extract an accurate metamodel for complex building environments involving many features. Buildings are unique and thus it becomes difficult to extract an accurate metamodel for each building.

## 5.2 Inverse Reinforcement Learning

Inverse reinforcement learning (IRL) (Ziebart et al., 2008) is an advanced method within the field of machine learning that seeks to infer the underlying reward function that a demonstrator (often an expert or an optimal policy) implicitly follows, based on their observed behavior in a specific environment. Unlike traditional RL, where the reward function is predefined and the agent learns the best actions to maximize this reward, IRL works in reverse by analyzing the actions of an expert to determine what rewards they appear to be maximizing (Dey et al., 2023). Here, three months of rule-based data from an existing building are utilized to extract a reward function, which was used in determining the value-function of the states. The value-function is then learned by the RL agent, which results in the agent having a similar policy to the rule-based policy thus limiting exploration.

**Findings and recommendations:**

IRL has proved to be effective in shortening training time, providing a stable learning experience, and outperforming standard RL agents trained for the same duration. IRL can partially deduce the intent behind rule-based controls without environment interaction. However, in the research while optimizing for energy consumption, thermal discomfort, and demand costs, it struggled with power demand limit penalties due to inaccuracies in the transitional dynamics, which are challenging to determine from limited rule-based data as a sparse dataset result in an incomplete extraction of the reward function’s intent. When a building has operational data for over three months for a specific season, an inverse reinforcement learning (IRL) strategy is recommended, even if metered data or thermal discomfort metrics are absent. IRL tends to maintain existing rule-based policies initially, aiding in reducing early exploration. It serves as an indirect method to mimic and improve upon RBC strategies. If the RBC is poorly designed, however, the RL agent needs to forget the value-function from the extracted reward function and start utilizing the actual reward function. This is done by adopting an importance parameter in the reward function that controls the transition between the extracted reward function to the actual rewards, which requires further research to further stabilize the training progress.

### 5.3 Online Learning W/Guided Exploration

In this approach, the agent starts with no prior training but uses the help of surrogate models to limit exploratory activities in the early learning stages. The surrogate models use regression techniques to learn the system dynamics and the reward function in real time. The models generate artificial trajectories guiding the RL agent away from states that previously led to high penalties by suggesting alternatives until the RL agent develops a more accurate understanding of the value functions. Several types of approaches were conducted where explorations were limited to within bounds around rule-based policy vs. full random explorations, short-term exploration vs. long-term exploration (Dey & Henze, 2024).

**Findings and recommendations:**

We found that training on long trajectories from surrogate models with full random explorations, although successful in reducing the exploration of the agent, had the worst performance on test days. Artificial long trajectories with full random exploratory paths yielded unsatisfactory results as training on paths with imperfect predictions from unseen states in the trajectories led to poor training. Limiting the exploration within a certain bound around a rule-based control strategy by control action paths led to better performance. We regard actions based on established rules as “safe”. Initially prioritizing these actions helped improve the accuracy of the system's estimates and the dynamics of rewards related to actions within the rule-based space, as modeled by the surrogates. Consequently, this approach allowed the RL agent to gradually adjust its policy away from these rule-based actions towards areas where it anticipates higher rewards. This incremental adjustment prevents the agent from abruptly moving into unfamiliar state spaces, where it might encounter inaccurate value estimates. Due to the nature of the application, buildings with large thermal time constants might not be effective with this approach as approaches with longer trajectories were not successful.

### 5.4 Online Learning With Guided Exploration and Imitation Learning

This approach is similar to the previous approach except that the agent starts with an imitated rule-based policy but is still assisted with surrogate models for artificial exploration.

**Findings and recommendations:**

This approach addresses the ‘performance dip’ that is inherent in the imitation learning approach. The surrogate model trajectories prevent the agent from moving further away from the rule-based policy in its early stages.

## 6. Transfer Learning

Transfer learning (TL) involves leveraging knowledge gained from one domain and task to improve the learning performance in a different yet similar domain. TL addresses the challenges of scalability in building domain where it is challenging to develop accurate building models for each commercial building and use that model for pre-training with RL. We investigated two types of TL and present the findings below.

### 6.1 Transductive Learning

In the study on transductive learning, the task and the state-space remain the same but the building domain changes.  We trained an agent on a large commercial building, optimizing for costs of energy, thermal discomfort and demand and applied the trained agent to a mid-sized commercial building in the same climate (Lissa et al., 2020; Zhang et al. 2020).

**Findings and recommendations:**

Although this direct transfer of weights approach is beneficial in reducing the training time as well as unstable early-stage training unstable behaviour in the agent, a key challenge lies in transferring the agent to scenarios involving additional tasks or tasks with different input and output architecture of states and action of the RL agent. Moreover, this approach does not lend itself to offering any insight into how the transferred agent will behave in a new environment, which might not help in gaining trust with the adaptation by building control managers.

## 6.2 Inductive Learning

In inductive learning, the source and the target domain are the same while the source and the target tasks are different. A key challenge arises in transferring knowledge to scenarios involving additional tasks or states, despite application in similar building. In this approach, a conditional rule extraction method from a trained agent was employed utilizing decision trees. The target agent learns this conditional policy with supervisory learning on artificial datasets generated from the states and action based on conditional rule-set extracted.

**Findings and recommendations:**

This strategy reduces the reliance on the specific input-output architecture of the trained RL agent and provides an interpretable starting policy due to its conditional ruleset. This interpretability is often absent in direct or partial weight transfer methods used in ML transfer learning methods, where control managers might have difficulty to predict the agent’s behavior in new settings. Furthermore, these rules can be integrated with or augmented with human-generated rules, facilitating the transfer of domain knowledge when introducing new tasks to the RL agent. The approach allows for flexibility as building control managers can review and adjust the rules to better suit the target building agent’s starting policy. Although this rule extraction method was employed in inductive learning, the method can also be used for transductive transfer learning. A minor setback with the rule extraction method is the agent's tendency to replicate actions without truly understanding the value function of the states and actions. The progress plot shows a slight decrease in performance due to the ‘performance dip’ before it finds a better policy than the starting policy.

# 7. Future Work

For effective adaptation of AI in building controls, the decision-making processes of AI models must be transparent and comprehensible. However, the inherent complexity of deep neural networks, which underpin many advanced AI algorithms, poses a challenge to their interpretability. Despite this, their depth and structure equip them to efficiently tackle intricate tasks and derive near-optimal solutions. Gaining insights into how these models function can assist developers in pinpointing inaccuracies or biases in the established policies. A practical method for this is observing an AI agent's actions in a simplified simulator, especially under extreme conditions not covered in training. Creating artificial datasets for these scenarios can help in understanding the model’s decisions through inductive learning, leading to performance refinement by adjusting or adding rules. This hands-on verification and validation serve as a tangible bridge between AI's abstract computations and real-world expectations.

Human-in-the-loop (HITL) reinforcement learning (RL) is an emerging interest in artificial intelligence that incorporates human input into the training process to enhance the performance of AI agents in complex environments (Nagy et al., 2023). This approach is especially beneficial in sectors such as robotics, autonomous vehicles, and healthcare, where exploration errors could lead to significant real-world consequences. In building controls, HITL RL integrates established operational strategies and occupant needs, thus mitigating the risks associated with trial-and-error learning. This is achieved by considering their expressed preferences on indoor environmental conditions such as thermal comfort, indoor air quality, visual comfort, or acoustic conditions while optimizing for factors such as energy consumption, demand management, and other user needs (e.g., EV charge scheduling). Examples of feedback received could be occupant preferences collected through edge devices and information about impending electric grid stress events by the system operator. Recent discussions among RL researchers also underscore the importance of HITL in building management, highlighting its potential to improve occupant productivity, reduce energy costs, and aiding in building decarbonization goals.

# 8. General Recommendations

In the absence of operational data, the combination of online learning with guided exploration and imitation learning is recommended if extensive modelling efforts are to be avoided. Buildings having slower thermal response will require a grey-box model or a physics-informed neural network (PINN) as surrogate models to assist with the training.

Normalizing input features or states is beneficial for speeding up the convergence of learning algorithms. By ensuring that all input features are scaled similarly, normalization facilitates faster convergence in gradient-based optimization algorithms. This process results in models that are more generalizable and less prone to overfitting specific scales or magnitudes in the training data. Additionally, many RL algorithms are sensitive to reward scales. Normalizing rewards not only stabilizes the learning process but also makes hyperparameter tuning more uniform across various environments and tasks. This standardization of rewards keeps value function estimates, such as Q-values in Q-learning (Watkins & Dayan, 1992), within a consistent range, which is advantageous during the transfer learning process when weights are transferred directly from one agent to another.

Building control systems present distinct challenges due to their real-world operational context, where data collection can be both time-intensive and expensive. In terms of sample efficiency, algorithms like Soft Actor-Critic (SAC) (Haarnoja et al., 2018) and Proximal Policy Optimization (PPO) (Schulman et al., 2017) outperform Deep Q-Network (DQN) (Mnih et al., 2013), making them more suitable for situations where data gathering is resource-heavy. SAC and PPO are known for their consistent stability during training, whereas DQN struggles with stability issues, particularly in environments with high variability. The specific nature of the problem often dictates the algorithm of choice. For tasks involving discrete action spaces, such as binary decisions or selecting from a limited set of options, DQN excels due to its effectiveness in simpler contexts where computational resources are a concern. However, for tasks that involve complex dynamics, PPO or SAC are preferred for their nuanced approach and potential to achieve near-optimal performance.

It is essential to have a fail-safe or override system in place for the control actions of RL agents to prevent unexpected consequences. Initially, it is crucial to identify and define the proper constraints of the system. For instance, an RL controller setting the temperature to 15°C during occupied hours in winter should automatically trigger an override, as this decision is clearly inappropriate. Such incorrect actions should incur significant penalties to teach the agent about these constraints. A building domain expert should oversee the setup of these overrides, recognizing potential constraints ahead of time. In practical scenarios, a fail-safe mechanism is necessary for cases where there may be gaps in real-time sensor data, preventing unexpected control behaviour. Additionally, it would be advantageous to implement a system that monitors whether the input states deviate significantly from the expected range, allowing the controller to switch to a baseline control until a building control engineer can verify the appropriateness of the untested control actions.

## Acknowledgement

The study was partially funded by the Building Energy Smart Technologies (BEST) Center, an industry research cooperative research center established under NSF Grants No. 213874 and No. 2113907, which the authors gratefully acknowledge. This work was supported in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding is provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

## Nomenclature

### Symbols

| | |
|---|---|
| AI | Artificial intelligence |
| BMS | Building management systems |
| DNN | Deep neural network |
| DQN | Deep Q-Network |
| MPC | Model predictive control |

| | |
|---|---|
| PPO | Proximal Policy Optimization |
| PINN | Physics-informed neural network |
| RL | Reinforcement learning |
| SAC | Soft actor-critic |
| IRL | Inverse Reinforcement learning |
| TL | Transfer learning |
| $\gamma$ | Discount Factor |
| $s$ | Current state |
| $a$ | Action |
| $s'$ | Next state |
| $d$ | Done flag |

## References

Arora, S., and P. Doshi. 2018. "A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress," Accessed June. http://arxiv.org/abs/1806.06877

Bingqing, C., Z. Cai, and M. Bergés. 2020. "Gnu-RL: A Practical and Scalable Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy." *Frontiers in Built Environment* 6. Accessed November. https://doi.org/10.3389/fbuil.2020.562239

Dey, S., and G. P. Henze. 2024. "Reinforcement Learning Building Control: An Online Approach With Guided Exploration Using Surrogate Models." *ASME Journal of Engineering for Sustainable Buildings and Cities* 5(1). https://doi.org/10.1115/1.4064842

Dey, S., T. Marzullo, and G. Henze. 2023. "Inverse Reinforcement Learning Control for Building Energy Management." *Energy and Buildings* 286: 112941. Accessed May. https://doi.org/10.1016/j.enbuild.2023.112941

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor." *35th International Conference on Machine Learning, ICML 2018* 5: 2976–89.

Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. 2015. "Continuous Control with Deep Reinforcement Learning." *ArXiv Preprint ArXiv:* 1509.02971. http://arxiv.org/abs/1509.02971

Lissa, P., M. Schukat, and E. Barrett. 2020. "Transfer Learning Applied to Reinforcement Learning-Based HVAC Control." *SN Computer Science* 1 (3): 1–12. https://doi.org/10.1007/s42979-020-00146-7

Mnih, V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. 2013. "Playing Atari with Deep Reinforcement Learning." *ArXiv Preprint ArXiv:1312.5602*, 1–9. http://arxiv.org/abs/1312.5602

Nagy, Z., G. Henze, S. Dey, J. Arroyo, L. Helsen, X. Zhang, B. Chen, et al. 2023. "Ten Questions Concerning Reinforcement Learning for Building Energy Management." *Building and Environment* 241. Accessed August. https://doi.org/10.1016/j.buildenv.2023.110435

Richalet, J., A. Rault, J. L. Testud, and J. Papon. 1978. "Model Predictive Heuristic Control. Applications to Industrial Processes." *Automatica* 14(5): 413–28. https://doi.org/10.1016/0005-1098(78)90001-8

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. "Proximal Policy Optimization Algorithms." *ArXiv*: 1–12.

Sutton, R. S., A. G. Barto, 2014. *Reinforcement Learning: An Introduction*. The MIT Press. https://doi.org/10.4018/978-1-60960-165-2.ch004

Wang, Z., and T. Hong. 2020. "Reinforcement Learning for Building Controls: The Opportunities and Challenges." *Applied Energy* 269: 115036. Accessed March. https://doi.org/10.1016/j.apenergy.2020.115036

Watkins, C. J. C. H., and P. Dayan. 1992. "Q-Learning." *Machine Learning* 8: 279–292.

Zhang, X., X. Jin, C. Tripp, D. J. Biagioni, P. Graf, and H. Jiang. 2020. "Transferable Reinforcement Learning for Smart Homes." *RLEM 2020 - Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings and Cities*: 43–47. https://doi.org/10.1145/3427773.3427865

Ziebart, B. D., A. Maas, J. A. Bagnell, and A. K. Dey. 2008. "Maximum Entropy Inverse Reinforcement Learning." *Aaai* 8: 1433–38. Chicago. www.aaai.org