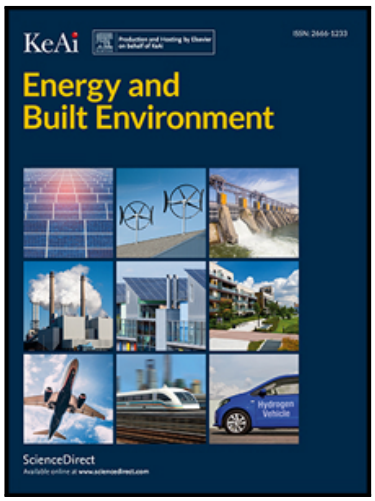

The reinforcement learning method for occupant behavior in building control: a review

Mengjie Han , Jing Zhao , Xingxing Zhang , Jingchun Shen , Yu Li

PII: S2666-1233(20)30087-8
DOI: https://doi.org/10.1016/j.enbenv.2020.08.005
Reference: ENBENV 72

To appear in: *Energy and Built Environment*

Received date: 28 April 2020
Revised date: 17 August 2020
Accepted date: 18 August 2020

Please cite this article as: Mengjie Han , Jing Zhao , Xingxing Zhang , Jingchun Shen , Yu Li , The reinforcement learning method for occupant behavior in building control: a review, *Energy and Built Environment* (2020), doi: https://doi.org/10.1016/j.enbenv.2020.08.005

Highlights

- Literatures of reinforcement learning regarding occupant behavior were presented;
- We provided a comprehensive understanding of how the method works;
- We discussed an overview of implementation requirements and challenges;
- Future research directions surrounding building control were proposed.

The reinforcement learning method for occupant behavior in building control: a review

Mengjie Han[a], Jing Zhao[b], Xingxing Zhang[a], Jingchun Shen[a], Yu Li[c]

[a] School of Technology and Business Studies, Dalarna University, Falun, 79188, Sweden

[b] Leisure Management College, Xi'an Eurasia University, Yanta District, Xi'an, China

[c] Luxembourg Institute of Science and Technology LIST, 5, Avenue des Hauts-Fourneaux, L-4362, Esch-sur-Alzette, Luxembourg

Abstract:

Occupant behavior in buildings has been considered the major source of uncertainty for assessing energy consumption and building performance. Modeling frameworks are usually built to accomplish a certain task, but the stochasticity of the occupant makes it difficult to apply that experience to a similar but distinct environment. For complex and dynamic environments, the development of smart devices and computing power makes intelligent control methods for occupant behaviors more viable. It is expected that they will make a substantial contribution to reducing global energy consumption. Among these control techniques, the reinforcement learning (RL) method seems distinctive and applicable. The success of the reinforcement learning method in many artificial intelligence applications has given an explicit indication of how this method might be used to model and adjust occupant behavior in building control. Fruitful algorithms complement each other and guarantee the quality of the optimization. However, the examination of occupant behavior based on reinforcement learning methodologies is not well established. The way that occupant interacts with the RL agent is still unclear. This study briefly reviews the empirical applications using reinforcement learning, how they have contributed to shaping the modeling paradigms and how they might suggest a future research direction.

Keywords:

Reinforcement learning; occupant behavior; energy efficiency; building control; smart building

1 Introduction

Building energy consumption amounts to approximately 30%-40% of all energy consumed in developed countries [1], [2]. The trend of power demand is still increasing. Not only does this increase the operating cost of energy consumption, it also contributes to the increasing emission of greenhouse gases. Since buildings are also responsible for one-third of global energy-related greenhouse gas emissions [3], developing efficient strategies for reducing the consumption of building energy are urgently required in the future.

Maintaining occupant comfort and use of appliances by occupant generates 80% of building energy consumptions [4]. As is well known, occupant behavior is stochastic and complex. Even when an advanced modeling method is built to include occupant behavior, it is challenging to quickly apply that experience to a similar but distinct environment. There is no general scientific standard outlining appropriate model validation techniques especially when multiple behaviors are modeled [5]. As an extreme case, in a simulation study of different models, occupant behavior with the feature of 'random walk' results in a very large performance gap [6]. It has also been recognized that a building could fail to achieve the desired standards and building designers could miss out on the opportunity of optimizing building design and control for occupancy [7]. Modeling occupant behavior may help to understand and reduce the gap between design and actual building energy performance [8], [9]. However, occupant models are usually context dependent [10]. Simply predicting or simulating occupant behavior in one setting has its intrinsic challenge in transferring the knowledge to a more complex scenario.

Studies of occupant behavior have been grouped into three streams: rule-based models, stochastic models, and data-driven methods [11]. It has been discussed that occupant behavior models do not represent deterministic events, but move into a field where behaviors are described by stochastic laws [12]. Stochastic models consider the occupant behavior to be stochastic because behavior varies between occupants and may evolve over time [13]. Data-driven methods, however, are conducted without an explicit aim to understand occupant behavior [11]. A building's physical environment is dynamic and complex. Occupants can respond quickly to a change of the environment in a process that is often non-stationary. Attempts to model all possible features for building operation systems can be intractable and systems accommodating more features often have significant lag times. Data-driven methods do not always set up physical models and often use historical data to characterize features, including occupant behavior.

Rather than on the understanding of occupant behavior, intelligent control methods used to optimize future reward in building systems seem to be an alternative approach. These create an agent that learns from historical behaviors and is trained to adjust the control actions by utilizing occupant behavior. The occupant interacts with the building control system via presence, actual activity and providing comfort feedback through linked building systems, e.g. HVAC, lighting and windows. Thus, an optimal control method integrating building performance and occupant impact offers a novel way of modeling. In a control problem, generally, an agent is built to complete decision-making tasks in a system to achieve preset goal. Building control system, which is a compound of multiple engineering fields, refers to centralized and integrated hardware and software networks [14] and considers the improvement of energy utilization efficiency, energy cost reduction, and renewable energy technology utilization in order to serve local energy loads while keeping indoor comfort [15]. Control targets usually include shading systems, window, lighting systems and heating/cooling systems.

A recently realized Markov decision process based machine learning method, known as reinforcement learning (RL), can work in both model-based and model-free environments

[16]. Nevertheless, it is the classic model-free learning algorithms, such as Q-learning and $TD(\lambda)$, that makes RL much more attractive and efficient in artificial intelligence applications [17]–[20]. The effort to solve deep RL problems, for example [21], [22], opens up the possibility of working on large continuous datasets. The distinctive feature of RL is that the agent, via trial-and-error search, can make optimal actions without having a supervisor, which fits the goal of a control problem.

These building control systems are able to make decisions based on data-driven modeling outcomes. The RL method is able to work in a stochastic environment and to adapt existing data to extract underlying logic for decision-making, that is, a data-driven method. The agent of RL treats occupant behavior as an unknown factor and learns to adapt itself form what has been observed of human interactions. The RL method has been in existence for over seventy years, but it was not until the past decade that researchers started to commit themselves to expanding its applications. Neither systematic approaches to applying RL on occupant behavior nor relevant literature reviews have been analyzed from the methodological point of view. The indication for future RL application is still unclear. Therefore, the aim of this study is to review the empirical articles on how RL methods have been implemented for adjusting occupant behavior in buildings, and provide instructive directions for future research.

Thus, contributions of this study are threefold. Firstly, we present the results of our literature search and identify the key points emerging from this research topic in recent years. Secondly, we provide a comprehensive understanding of how RL works for building control and an overview of its implementation requirements. Finally, we identify the current research gap surrounding building control and propose future research ideas for modeling occupant behavior.

In the second section of this study, we present the literature searching scope and the outcomes. In Section 3 we briefly introduce the philosophy of RL and its corresponding algorithms. Section 4 then analyzes the empirical articles. A discussion is presented in Section 5 and Section 6 concludes with some findings and possible new research directions**.**

2 Methods and search outcomes

2.1 Methods

We conducted our literature search using the search engine *Scopus.* The first reason is that it provides us with multiple document features that we can adjust such as funding details and conference information. The second reason is that an interface to the R package *bibliometrix,* an open-source tool for executing science mapping analysis, can be created for conducting analytical bibliometrics where three steps are considered for the workflow [23]. In step 1, data is loaded and converted to the R data frame. In step 2, the descriptive analysis and citation networks are produced; the visualization is made available in step 3.

Our searching keywords and operations are

(("reinforcement learning" OR "Q-learning" OR "policy gradient" OR "A3C" OR "actor-critic" OR "SARSA*") AND "occupant*"),

where some prevalent algorithms for RL, for example, Q-learning and policy gradient, are also included to guarantee adequate coverage. Adding the wildcard to *occupant** ensures hits using both singular and plural forms are returned. The same was done for *SARSA** because there are a number of variants of the SARSA algorithm that can be used for some algorithm-specific articles. We exclude the words behavior* or behaviour* because the RL agent does not only take action based on particular behaviors, but also adjusts its policy by

collecting occupant feedback for the control system. We do not limit the search by article type or publication year.

2.2 Search outcomes

The original search returns a total number of forty articles. One of the selection criteria was that articles where either the occupant behavior or occupancy was explicitly considered as an element in a Markov decision process (see Section 3.1) or had an impact on the transition of environmental states were included. In other words, an agent that tried to learn the optimal control strategy only to satisfy occupant comfort and did not include dynamic interactions with the environment was excluded from this analysis. See a relevant review work [24] that examined the RL control for occupant comfort for more articles that we exclude here. Careful reading of each of the forty articles resulted in thirty-two articles that are considered for this analysis. Even though it is not exhaustive, the outcome of this search, we believe, can form a representative sample of current understandings within the field.

2.2.1 Publication sources

The thirty-two documents were published in twenty-three difference sources including international journals, conference proceedings and book chapter. A summary of the top five publication sources from the search is shown in Figure 1. Most of the articles were published in the Elsevier journal *Building and Environment*, followed by a second Elsevier journal *Energy and Buildings* and the Buildingsys 2019[1] conference. Each of remaining eighteen sources has published one article. Even though full-text articles of some publications are not included in the Scopus search engine, the long-tailed Poisson-like distribution for publication sources covers a range of topics including energy, building, computer science, optimal control, sustainability and engineering. The variety of publication sources establishes a multidisciplinary collaborative framework for future studies. We also anticipate that the emergence of new publication sources may attract studies of RL for occupant behavior and increase public awareness of the topic.

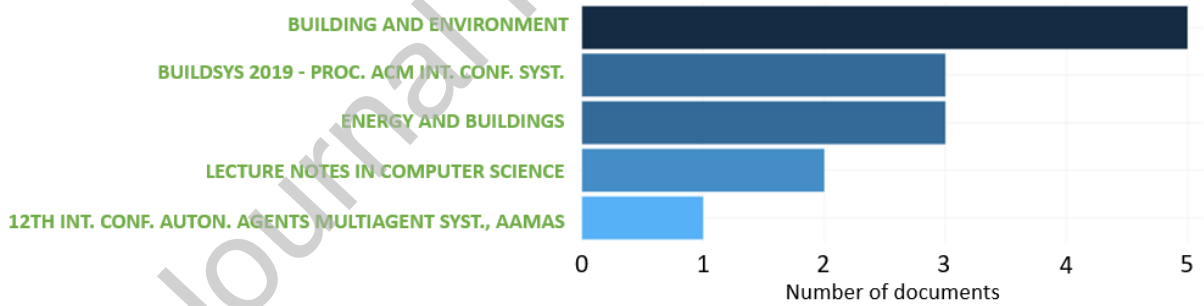

Fig. 1 Top five publication sources

2.2.2 Publication types, years and citations

Of the total articles in this search, the earliest was published in 2007. After that, no article was published until 2013 (Figure 2). This strongly suggests that difficulties in the implementation of complex problems has hindered the development of RL applications. The success of many deep learning paradigms in the early 2010s, however, seems to have promoted a revival of the use of RL applications, including those in building control. It has generated the publication of a number of articles by fusing deep RL for solving complex problems. Nevertheless, overall citations are still low. More attention could be paid to this RL literature when intelligent control systems for occupants are developed.

[1] Full name of the conference: BUILDSYS 2019 - PROCEEDINGS OF THE 6TH ACM INTERNATIONAL CONFERENCE ON SYSTEMS FOR ENERGY-EFFICIENT BUILDINGS, CITIES, AND TRANSPORTATION

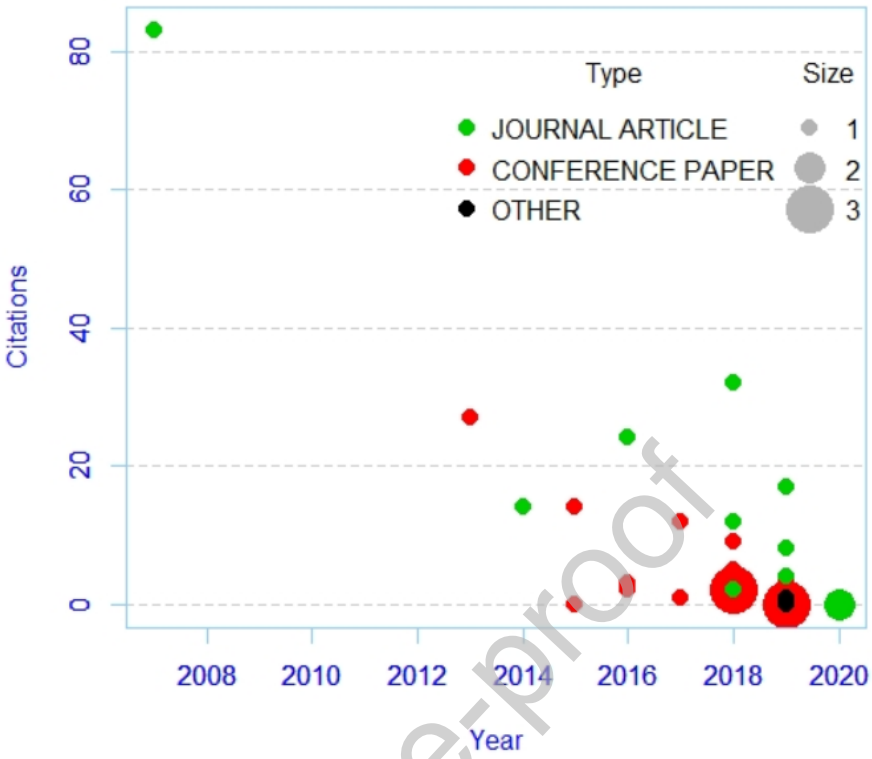

Fig. 2 Type and year of publication and number of citations

2.2.3 Country collaboration

Collaboration between countries allows researchers to share knowledge, data and research infrastructures. The development of RL control for occupant behavior has just started to be noticed and needs worldwide collaboration for fast growth. Most historical collaborations have been carried out between researchers in the United States and some countries in Europe, as well as in China (Figure 3). These three regions/countries will likely take the lead in future contributions to the topic. In the meantime their pioneer activity is setting the stage for comprehensive impacts from other regions and countries.

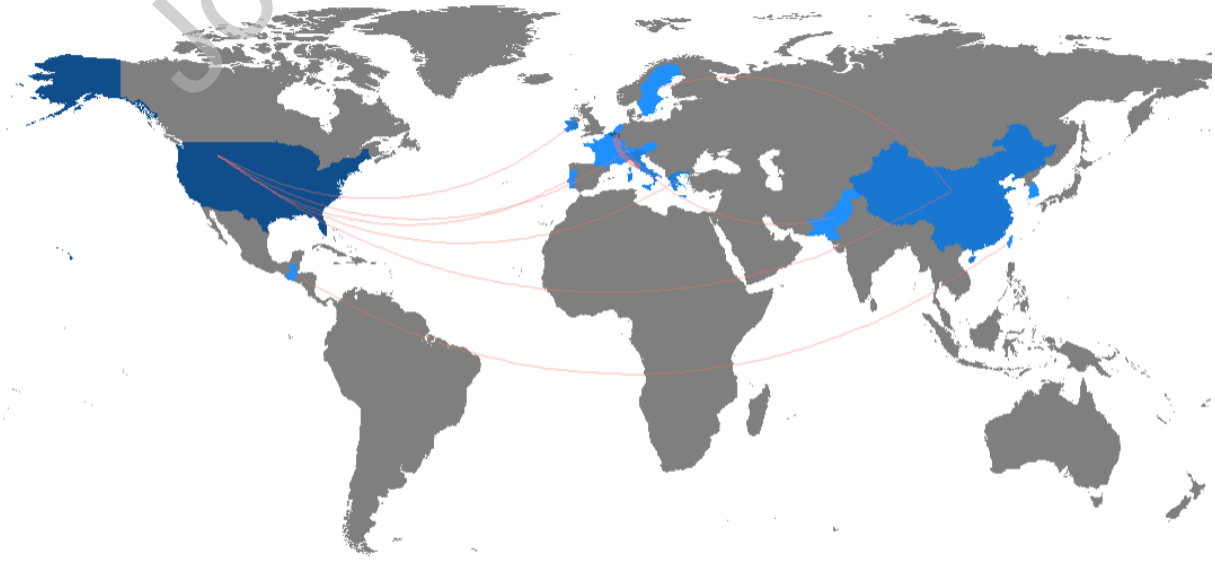

Fig. 3 Country collaboration map

3 The reinforcement learning method

Various studies have reviewed the classification of different control methods in buildings. For example, Shaikh et al. [14] reviewed the intelligent control system for building energy and occupant's comfort, whereas Dounis and Caraiscos [25] focused on the agent-based control system. Aste et al. [26] summarized the model-based strategies for building simulation, control and data analytics. The previous surveys provide a framework of how the different methods relate to each other and the pros and cons of each. A generic challenge of conventional methods (e.g. PID, on-off, model predictive control, etc.) lies in the difficulty of including all unknown environmental factors in the models. Even there is much room to increase model performance, complex model specifications usually bring heavy computations [27].

Compared to the conventional methods the RL technique is still not well developed for buildings. It has not drawn much attention and the performance of RL algorithms has thus not been evaluated yet. Even though Royapoor et al. [28] realized that RL methods are notable, a framework of implementations and explorations on efficient RL methods needs to be systematically investigated and discussed.

The shortage of scientific research publications prevents building users, building managers, device controllers, energy agencies and other related parties from being aware of the neglected technique. An integration with explicit occupant behavior has not been comprehensively examined. The curse of dimensionality, the fact that the number of representative environment states grows exponentially with complex problems, is an inherent problem. Approximate solution methods provide the possibility to overcome this. Deficient consideration of it hinders the development of solutions. Thus, the necessity for investigating current studies and indicating future studies first requires an overview.

The idea of RL derives from the concept of "optimal control", which emerged in the 1950s as a way of formulating problems by designing a controller to minimize a measure of the behavior of a system over time [29]. Bellman [30] came up with the concept of Markov decision processes (MDPs) or finite MDPs, a fundamental theory of RL, to formulate optimal control problems. Unlike conventional control methods, RL does not require a model. A benefit of a model-free approach is that it simplifies the problem when the system is complex. Different from independent and identically distributed (i.i.d.) data that some conventional models require, the RL agent receives subsequent reward signals from its actions. Another benefit is that the trade-off between exploration and exploitation can be balanced via experiment design. Furthermore, a rich class of learning algorithms fused with deep neural networks [20] provide a potential for accurate estimation of value functions.

3.1 Markov decision processes

In a dynamic sequential decision-making process, the state $S_t \in \mathcal{S}$ of a RL agent refers to a specific condition of the environment at discrete time steps $t = 0,1,\ldots$. By realizing and responding to the environment, the agent chooses a deterministic or stochastic action $A_t \in \mathcal{A}$ that tries to maximize future returns and receives an instant reward $R_{t+1} \in \mathcal{R}$ as the agent transfers to the new state S_{t+1}. A sequence of state, action and reward is generated to form an MDP (Figure 4 [24], [29]).

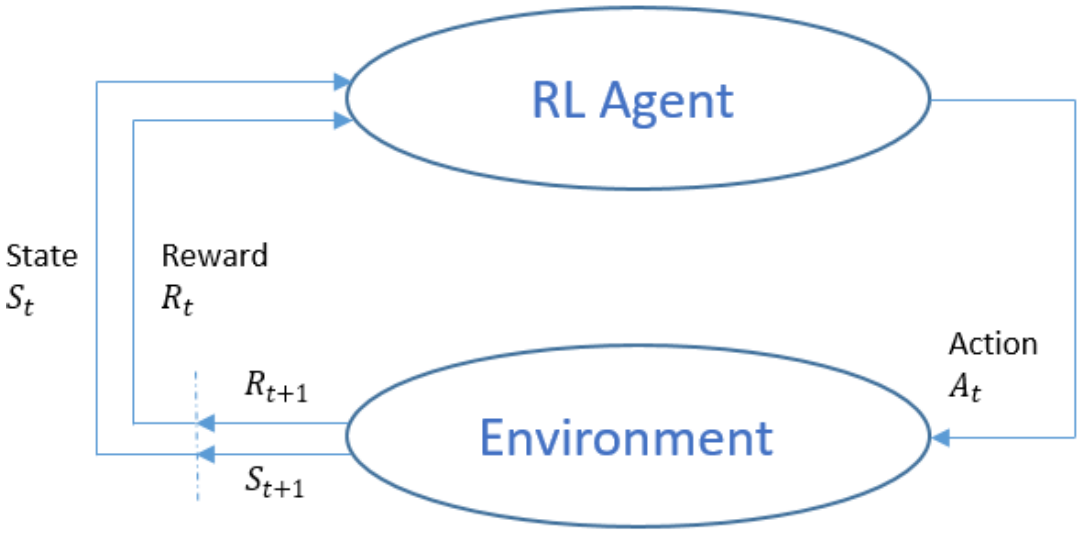

Fig. 4 The interaction between agent and environment in an MDP

The Markov property highlights that the future is independent of the past and depends only on the present. In Figure 4, S_t and R_t are the outcomes after taking an action and are considered as random variables. Thus, the joint probability density function for S_t and R_t is defined by:

$$p(s', r|s, a) = \mathbb{P}[S_t = s', R_t = r|\, S_{t-1} = s, A_{t-1} = a], \tag{1}$$

where $s, s' \in \mathcal{S}$, $r \in \mathcal{R}$ and $a \in \mathcal{A}$. It can be seen from Eq. (1) that the distribution of state and reward at time t depends only on the state and action one step before. From Eq. (1), it is straightforward to obtain the transition probabilities $p(s'|s, a)$ and the expected reward $r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a]$ that are used for formulating the Bellman optimality equation in Section 3.3.

3.2 Policies and value functions

A *policy* π is a distribution over actions given states and can be considered as a function of actions. It fully defines the behavior of an agent by telling the agent how to act when it is in different states. An arbitrary policy targets on evaluating the expected future return when making an action a from time t: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$ under a given state s, where $0 \leq \gamma \leq 1$ is the discount parameter, namely:

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right], for\ all\ s \in \mathcal{S}\ and\ a \in \mathcal{A}. \end{aligned} \tag{2}$$

The task of finding the optimal policy in Eq. (2), π_*, is thus achieved by evaluating the optimal action-value function $q_\pi(s, a)$:

$$q_*(s, a) = \max_\pi q_\pi(s, a). \tag{3}$$

3.3 Value-based algorithms

Strategies to solve Eq. (3) are usually achieved by updating the Bellman optimality equation [31]:

$$q_*(s, a) = r(s, a) + \gamma \sum\nolimits_{s' \in \mathcal{S}} p(s'|s, a) \max_{a'} q_*(s', a'). \tag{4}$$

The recursive relationship assists in splitting the current action-value function into the immediate reward and the value of the next action. Eq. (4) directly provides us with the formulation of value-based algorithms within temporal-difference method[2], where either tabular methods or approximation methods can be adopted for obtaining $q(s,a)$. There is always an explicit state exploration of state-action space for value-based algorithms.

For problems with small and discrete state or state-action sets, it is preferable to formulate the estimations using look-up tables with one entry for each state or state-action value. The tabular method is easy to implement and guarantees convergence [29]. The tabular Q-learning algorithm [32] is the most common RL algorithm used in building control [24]. Easy implementation and accurate solutions make it robust in different building control problems. Other tabular algorithms include tabular SARSA, i.e. the so-called state–action–reward–state–action, value-iteration, and policy-iteration.

For large MDP problems, we do not always want to see separate the trajectory of each entry in the look-up table. The parameterized value function approximation $\hat{q}(s,a;\mathbf{w}) \approx q_\pi(s,a)$ gives a mapping from the state-action to a function value, for which there are many mapping functions available, for example, linear combinations, neural networks, and so on. It generates the state-actions that we may not directly observe. A common way of updating the weight vector, $\mathbf{w}$, is the gradient descent, which yields deep Q-learning. Algorithms like $SARSA(\lambda)$ and fitted Q-iteration can also be found in the earlier studies. More recently developed value-based algorithms [33] have also provided a great number of opportunities for training the agent in a more flexible way.

3.4 Policy-based and actor-critic algorithms

Another way to solve large MDP or continuous state RL problems is to apply the policy-based method [34], where the policy is explicitly represented by its own function approximator, independent of the value function, and is updated according to the gradient of expected reward,

$$J(\theta) = \mathbb{E}_{\pi\sim p_\theta(\tau)}[r(\tau)], \tag{5}$$

with respect to the policy parameters θ. $r(\tau)$ is the total reward for a given trajectory τ, representing the interactions between the agent and the environment in an episode. $p_\theta(\tau)$ depicts the probability of getting a specific τ from a stochastic environment under fixed θ. The approach to finding optimal J can be converted to solve the maximization problem using gradient ascent with regard to a set of parameters $\boldsymbol{\theta}$, for example, the weights and biases in a neural network. The policy-based method has an innate exploration strategy and the variance of the gradient is large for episodes with long time steps. Some recent algorithms such as Proximal Policy Optimization [35] and Trust Region Policy Optimization [36] have been developed for complex problems. Subtracting a baseline b from $r(\tau)$ may reduce the variance while keeping the gradient still unbiased. One option is to apply the state-value $v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$ to the policy gradient methods, known as an actor-critic algorithm. These algorithms work with parameterized policies by relying exclusively on value function approximation [37]. In practice, the actor-critic algorithms use deep neural networks to estimate the value function [38], [39].

3.5 RL for building control

[2] The Monte Carlo method and dynamic programing method are also value-based. See [29] for more details.

It has been challenging to apply the trained RL agent to buildings irrespective of the type occupant behavior due to rigorous training requirement, control security and robustness, and the ability of method generalization [40]. However, real implementations may validate and improve the method by observing reliable state transitions and reward signals. Appropriate specifications of state, action and reward in MDP have significant impacts on learning outcomes and practical settings.

The states partly determine the complexity of RL control problems. In building applications, states are mostly defined by the variables that are associated to physical environment and weather condition for a building, for example, outdoor temperature, airflow rate, indoor CO_2 level and so on. Sufficient changes of state variables will alter indoor comfort level and energy use, which also updates building environment for RL agent to take action. Accurate representation of states will lead to efficient training process and avoid curse of dimensionality. For continuous state or state with large number of levels, building environment becomes too complex to get fully explored. Dimension reduction is an alternative way for resolving the problem [41]. However, it is a collaborative work between building management expert and data scientist to figure out applicable state representation.

The action of an agent is taken based on observed state and the action levels can also affect the problem complexity. For a building system, controlling HVAC (heating, ventilation, and air conditioning) is the most complicated due to various components and control levels [40]. Actions like setting constant temperature set point or airflow rate will cause high energy use, because room occupancy change, outdoor environment and pre-heating/cooling strategy may also generate effects to HVAC performance and energy use. Actions of an RL agent do not only try to immediately improve current reward, but also aim to maximize future return. For simpler control problems, for example, window opening/closing [42], action can also be generalized to a continuous domain, which requires more efforts on making acceptable simplifications.

Two types of rewards have been examined in most of the studies: comfort level and energy saving. It seems that occupant comfort gets more priorities when optimization is considered for these two contradictory factors in developed areas. Nevertheless, reward is more related to contextual, psychological, physiological, and social background of an occupant. Using same comfort criteria to different individuals will bring bias to learning process. It is also reasonable to take $\gamma = 1$ indicating that time factor will not give any discount to future comfort.

4 Empirical articles of RL control for occupant behavior

In this section, we will scrutinize the RL applications in two categories: those where occupant behavior or occupancy is explicitly characterized as a state, action or reward in the MPD; and those which not use occupant behavior to directly train an agent, but interact with the environment by adjusting the state transition, estimating the disturbance of reward, providing feedback and changing occupancy schedules.

4.1 Occupant behavior in MDP

Nine representative articles were selected to illustrate the first category of applications. Their workflows are summarized in Table 1 where occupant behavior or occupancy interacting with RL agent will be examined in detail. We also present a breakdown of the specific state, action, reward and algorithms each application uses.

There is always some doubt when selecting state variables. Selecting too many will increase the learning inefficiency exponentially while selecting too few will not fully depict the Markov property. Thus, evaluating the computation power and model accuracy should be considered for making a selection balance. Looking at the actions made on the building systems, the

main interventions have been taken with the HVAC system, which directly contributes to affecting occupant thermal comfort and indoor air quality. It is not surprising that comfort and energy consumption are the most studies objectives, represented by reward, for different learning tasks. Incorporating learning efficiency to the reward also provides us with innovative method in designing the experiment [43].

Table 1 Occupant behavior in MDP

References	state	action	reward	algorithms
Jia (2019) [44]	*occupancy*, room temperature, weather, time of day, energy consumption	supply air temperature	energy and comfort	policy gradient
Park (2019) [45]	*occupancy*, light switch position, indoor light level, time of a day	switching lights on/off, doing nothing	energy and comfort	value iteration
Valladares (2019) [46]	*number of people,* indoor/ ambient temperature, levels of CO_2, PMV index, etc.	setting temperature and ventilation system	CO_2 levels, PMV index, and power consumption	deep Q-learning and double Q-learning
Marantos (2019) [47]	*occupant's existence, number and activity,* indoor/outdoor temperature, humidity, solar radiation, etc.	temperature set-point	thermal comfort and energy	Neural Fitted Q-iteration
Kazmi (2018) [43]	*environment including occupant behavior*, embodied energy content of vessel, heating mechanism	reheating the storage vessel or not	comfort, energy, exploration bonus	Model-based RL
Lee (2018) [48]	*occupant's feeling of cold, comfort, and hot*	*occupancy, occupant's overriding the set point*	point tracking error and energy	policy gradient
Zhang (2018) [49]	*occupancy,* day of the week, hour of the day, outdoor air temperature, outdoor air relative humidity, etc.	supply water temperature set point	energy demand and indoor thermal comfort	Asynchronous Advantage Actor-Critic (A3C)
Barrett (2015) [50]	*occupancy*, room temperature; outside temperature	turning on/off heating turning on/off cooling	indoor temperature, energy	Q-learning
Fazenda (2014) [51]	time that the system has been in operation, lifetime desired for the	on/off heating/cooling:, temperature set	*user interaction of thermal comfort*, energy	Q-learning with function approximator

	system, heating on/off	points, opening windows

4.1.1 Occupant behavior as a state for HVAC control

Most of the applications focused on controlling HVAC by setting occupancy as the state [44], [46], [47], [50]. This was because the occupant's schedule usually followed a fixed routine or could be predicted with stochastic models. For example, Barrett and Linder [50] developed a HVAC control system by including the prediction of occupancy, where a modified Bayes rule was applied. Initial prior probability and environmental experience were used to obtain the posterior probability. The predicted occupancy followed a multinomial distribution of occupancy for specific times and returned a binary outcome of true and false.

One of the recent studies [44] added expert experience when they considered occupancy as one of the states to control HVAC, where the availability of state-action pairs helped to initialize the neural network and expert policy was used as a baseline for better policies. Valladares et al., [46] believed that occupant has strong influence on CO_2 level and included the number of occupants as one of their states, arguing that CO_2 control requires additional fresh air from the outside environment and increases HVAC loading. Simulations were carried out in their initial study using between 2-10 occupants, a number that was extended to 60 occupants in a subsequent study. A pre-training loop was used for the exploration of state-action pairs to guarantee that the agent was able to observe sufficient information for deep Q-learning. Combined with supervised learning for estimating energy consumption given occupant activity, Marantos et al., [47] developed a Neural Fitted Q-iteration, where the Q function was represented in parametric form by a multi-layer perceptron.

4.1.2 Occupant behavior other than as a state for HVAC control

In addition to setting occupancy as the state, Zhang and Poh [49] also used a smart phone app to collect thermal preferences from the occupants. The RL agent figured out the control policy by using the collected feedback. A Bayesian model calibration was implemented for heating energy demand and average indoor air temperature before training RL agent. The training was carried out in OpenAI Gym with customized design, which provides them with flexible options to build an RL agent.

Besides occupancy, other studies used occupant's feeling of cold, comfort, and hot as a state. One simulation-based work [48] also included occupancy, as represented by uniform distribution, and the occupant's override at a set point, as actions. A sample average method was developed for approximating the gradient, a method that was shown to be applicable for complicated stochastic problems. The occupant's interaction with the thermostat was also set as the reward in one study, where the behavior of the occupant was simulated with "out", "working", and "uncomfortable" [51]. All of these studies, however, are based on the assumption that occupant behavior stays constant. If occupants change their behavior from time to time, the learning outcomes demonstrated here may fail to work.

4.1.3 Control for lighting and vessel

Two of the studies used lighting and vessel control respectively as a way to explore occupant behavior. In a study of lighting control [45], occupant was detected by smart device. Their feedback on the control was collected through a survey. RL agent was able to gather the information and the learning were continuously updated to adapt the control parameters via occupant interactions. It has been discussed that the developed method can also control a dimmable light. For vessel control [43], future occupant behavior was modeled as an uncontrollable environmental factor for hot water consumption. This was because of the

limitations of the prediction model. Nevertheless, the study did show that specific behavior can be learnt from data and that the RL agent was able to adapt the policy.

4.2 Indirect influence of occupant behavior on MDP

In contrast to the studies that directly characterize occupant behavior in MDP, there are various ways for the occupant to influence the building control method. The RL agent in these studies optimizes its policy not by taking occupant behavior as an immediate input to MPD, but by measuring its indirect effect on the system. A summary of the literatures generates three categories for understanding occupant behavior: occupancy, actual behavior and providing feedback to the control system. For MDP, occupant behavior can have an effect on changing the state or state transition. In most of the studies, occupant behavior can be modeled as a stochastic factor to adjust the reward. Only a few studies associated occupant behavior with action. Detailed references for each application are shown in Table 2. For the building systems, HVAC is the mostly examined one, because it makes a substantial contribution to occupant thermal comfort and indoor air quality. RL controls for lighting, window and vessel, for example, are relatively uncommon in the existing literature; this gap should be addressed in future studies.

Table 2 Indirect influence on MDP

Interactions	MDP		
	State/state transition	Reward	Action
Occupancy	-	HVAC ([52]–[56]); HVAC and window ([57]); HVAC, lighting, blind and window ([58])	-
Actual behavior	Vessel ([59]–[61]); PV system ([62]); Lighting ([63])	HVAC ([53], [64]); Vessel ([65]); Space heating ([66]); Lighting ([63])	HVAC ([67])
Feedback	-	HVAC([68], [69])	-

4.2.1 Actual behavior and state

Actual behavior includes any activities that occupants carry out to interact with the building system, for example, using hot water, turning on the light, and opening the window. The stochastic behavior will lead to frequent updates of the state in the Q-table. As some studies show, the inclusion of actual behavior in controlling vessels seems to be a viable approach [59]–[61]. Occupant behavior together with current state and action, contributing to the state transition, can be modeled as a stochastic time series sequence using real world occupant behavior when the RL agent develops its policy [61]. Occupant behavior was considered as a perturbations of the vessel states: energy content inside the storage vessel and temperature [59]. The state transitions were modeled based on this assumption. Higher hot water consumption might require shorter episodes to preserve occupant comfort. A SARIMA model learned occupant behavior, with adjustments for the seasonality of individual occupant demand. Similarly, individual occupant behavior, or consumption profiles, was modelled, which defines vessel state transitions [60]. Occupant models were built to offer additional insight into individual occupant behavior types and were used for clustering households. The SARIMA models also provided reliable predictions for houses with regular consumption patterns. Non-stationary, nonlinear and highly irregular consumption profiles were dealt with using the additional bias term. In these case, different occupant behavior might be the reason for the variance of energy savings.

The RL method has also been applied to photovoltaic systems. In [62], stochastic occupant behavior capturing tap water use was included in a heat pump buffer model. It was counted as energy loss to the environment. The tap water model used historical data to relate occupant behavior to hot water demand. This historical data was used to construct a conditional probability, but it could also be used to generate samples of occupant behavior. Besides the stochastic occupant behavior associated with hot water consumption, other behaviors, such as those associated with the use of cooking appliances, lighting, washing machines entertainment devices and other electrical loads, could also be studied. Occupant behavior is the result of complex decisions that are dependent on unpredictable personal factors. One study used a hidden Markov model (HMM) to demonstrate occupant behavior around light usage, where a RL was applied without the need to consider hidden states [63]. The authors considered the whole building as a set of spaces and for each space the occupant occupied a HMM.

4.2.2 Actual behavior and reward

The studies reviewed here also show that occupant behavior can affect the reward. For example, using hot water and having the lights on at the same time can increase energy consumption. When the RL agent specifies the reward, insufficient consideration of human activities can lead to errors. Because it is very challenging to develop explicit physical models that are both accurate and fast, deep RL (DRL) algorithms are necessary to adapt for occupant activities [64]. A deep deterministic policy gradient was developed for a HVAC system in [53]. Occupant behavior was concluded to affect the reward in two ways. First, the system was set to occupied and unoccupied periods. The unoccupied spaces did not have to maintain thermal comfort. Second, variable-air-volume boxes controlling the volume of conditioned air were installed based on the set points set by the occupants. These provide more accurate air temperature controls. The percentage of discomfort occupants in the experiment experienced was represented by averaging the sensor readings from the boxes. In this study, the authors used a long-short-term-memory (LSTM) method to model historical HVAC operational data in order to build a training environment for the DRL agent to interact with. In the LSTM, the environment took the state and the action chosen by the DRL agent as inputs and returned the new state and reward for action as outputs. The DRL agent was able to learn the optimal control policy for a HVAC system by interacting with the training.

For studies that considered heating systems, the profiles of individual occupant behavior were averaged and then applied to simulate the results [65]. When this was done the SARSA(λ) algorithm was then able to learn the desired behavior – the occupant's domestic hot water use - to enhance the heating cycles. The results, however, showed a large difference in the number of heating cycles between the individual and averaged profiles. This was due to individual occupant behavior. Occupants' clothing insulation and activity level, such as sitting, cooking or sleeping, were used to calculate Predicted Mean Vote (PMV) [66]. The simulations considered the number of occupants and their metabolic rate. Typical behaviors during the week (working or studying during the day, eating dinner at home) and activities during the weekend were also simulated to evaluate energy consumption. Because occupants may feel and act differently and wear different clothes, room temperature has to be adjustable to obtain good thermal comfort.

4.2.3 Occupancy and reward

Occupancy is a more general concept where actual occupant behavior is not formulated. A number of occupancy detection methods have been be developed [70]–[72]. From these techniques, it is now possible to identify if a room is occupied or not and how many occupants it has. Like actual behavior, the level of occupancy is also a stochastic factor to be rewarded. In one study of HVAC systems, the transition function of the MDP was assumed

unknown to the agent [52]. The occupants were assumed to affect the CO_2 concentration and to generate heat emission. When the occupancy level changed, the RL agent had sense this change and adjust the CO_2 levels and temperature accordingly. The reward, including CO_2, thermal and energy, was calculated based on a negative sigmoid function. More simply, the indoor air quality was modeled in proportion to the number of occupants [54], where a 24 hour period was used to form an episode in which the number of occupants in a building could change. In the simulation, two peak periods for the number of occupants and CO_2 concentrations were found, one at approximately 9:00 am and one at 7:00 pm.

Besides air quality, one of the studies examined thermal comfort in a single-family residential home [55]. The authors assumed that the occupants were at home between 6pm and 7am the next day and that the house was unoccupied between 7am and 6pm. Thus, the RL agent tried to keep a desired temperature range whenever the occupants were at home, and remained indifferent to home temperature when the occupants were out. The setting led to a straightforward setback strategy that turned the system off when the occupants were out and turned it back on once the occupants were at home. Occupancy schedules and counts were used as a future disturbance in another recent study [56]. By the end of the experiment, the agent was able to perform well, irrespective of the number of occupants. In this study, occupancy count was not an initial part of the model the authors used for the real test. When examining the results, however, they found that the amount of cooling required varied drastically with the number of occupants and so occupancy count was added to their subsequent calculations. Another approach is to replace default occupancy schedules with actual occupancy schedules collected from real target buildings [58]. This system was installed in a test building and the collection of accurate occupancy pattern data at the zone level was then obtained. The RL control system developed in this case could also accept occupants' feedback allowing it to train the agent where only minor modifications were needed.

4.2.4 Feedback and reward

Providing comfort feedback to the control system makes RL agents react more efficiently. Even though comfort standards, for example thermal comfort [73], can help RL agents to figure out the appropriate comfort level, this can be challenging because of data availability and individual variation.

In one study an adaptive occupant satisfaction simulator was used as a measure of user dissatisfaction that originated from the direct feedback of the building occupants [69]. Every time a signal from the simulator became available, the simulator was updated to incorporate the new information. It should be noted that this study was the earliest publication in our document set. The learning speed was slow and the agent was still making errors after four years of training. For example, it was still turning on the heating in summer and cooling during winter. This may have been because the exploration was not enough. It may also have been because the use of the recursive least-squares algorithm $TD(\lambda)$ requires high computational demands and large amounts of memory. Further training should eliminate these wrong decisions. On the positive side, this study clustered thermal conditions to produce homogeneous environments, where the classification was implemented to predict the level of thermal comfort by using the state space, including clothing insulation, indoor air temperature and relative humidity [68]. A confusion matrix was then created to evaluate its performance. It formed a function mapping the state to the reward, which enabled the occupant's feedback to be collected by the RL agent for HVAC control. This approach was able to reach the optimal policy from any start state after a certain number of episodes. The authors pointed out that when new occupant provides feedback to the agent, the model needs to be calibrated for new training.

4.2.5 Actual behavior and action

There are a limited number studies considering occupant behavior as an indication to action, because optimal action is usually learnt by the agent. One exception is to make recommendations [67]. Occupants' historical location and the shift schedule of their arrival and departure times was used for operational recommendations. The occupants' location preferences, consisting of the distribution of time over the spaces, were extracted.by using historical data. Location data was also extracted for the arrival and departure times of each occupant. The occupants could change location after receiving a move recommendation. The Q-table was maintained for learning both move and shift schedule recommendations.

4.3 Training RL agent with deep neural networks

Curse of dimensionality refers to high number of levels for state variable or continuous state, which hinders efficient exploration of the state space and leads to insufficient learning. In Table 3, three simplification methods have been compared for their pros and cons. For value-based methods with continuous state, variable discretization take a set of single values to represent the whole state space [50], [54], [63]. However, including too many such type of variables may easily lose important information in the data and increasing the size of the data will not help to compensate the loss. On the other hand, dimension reduction aims to utilize all dimensions in the variable space to extract principal features that are in relatively low dimensions [41]. Although larger amount of data can utilize more information and extract more representative features, bridging the extracted features to the original values is not straightforward and thus the policies may be misleading.

Table 3 Comparison of simplification methods

	benefit	weak point
variable discretization	easy to implement, problem can quickly become simple	may lose important information
dimension reduction	able to capture all features	inaccurate description to original data
function approximation	efficient for really complex problem	not easy to find perfect function

Artificial neural networks are widely used for nonlinear function approximation. It is a network of interconnected units that have some of the properties of neurons, the main components of nervous systems. Function approximation avoids to create a look-up table to store action values. Instead, approximate value is represented as a parameterized function. Actions are quickly generated by using a neural network to map the state into a set of action-value pairs [51]. The number of hidden layers in a neural network is associated to the degree of nonlinear transformations. High number of hidden layers indicate more sophisticated mathematical modeling and better mapping ability, which is also called deep neural network (DNN). A direct application is to extend Q-learning to deep Q-learning where the demand of data is high [46], [64]. Insufficient data input to DNN is not able to optimize thousands of parameters in DNN. Thus, high quantity and quality of data guarantees the convergence of the loss function for a DNN. An alternative way to overcome the data insufficiency is to apply transfer learning technique by freezing most layers of a deep neural network that are pre-trained on data from other source. The model can be then re-trained with much less trainable parameters from the target data. The performance of this transfer learning deep neural network model will keep improving over time while more operational data are streaming into the model [74]. For policy-based implementations [53], [56], [75], the parameters in the policy network, θ, connect the DNN layers in Eq.(5). Unlike deep Q-network, policy network maps a state to an action that maximizes the expected reward from sampled trajectories. Training policy DNN requires intensive experiments to generate actual behaviors, which is time-consuming and costly in terms of data collection. In Section 5, we will discuss the details of implementing an alternative off-policy strategy.

4.4 The algorithms

Algorithm selection is problem dependent. For problems with small state-action space, value based algorithms are preferred because the optimization can converge quickly. For problems with large state-action space, creating a table to update learnt action values is not feasible. For building control applications, it is common to adopt continuous variables such as temperature, solar radiation, and occupancy duration for the analysis. Discretization to such variables may mitigate the problem, but can also generate bias. Thus, variants of Q-learning algorithms and policy-based algorithms have emerged as ways to achieve more exploration to the state space. As seen in Figure 5, tabular Q-learning is still the most commonly used algorithm any more, but the relative frequency of this has reduced in recent years compared to earlier work [24]. The variants of Q-learning, for example Q-learning with approximation, and policy-based algorithms now also supply various strategies for dealing with continuous state. The class of actor-critic algorithms seem to be an alternative approach; more applications need to be developed.

4.5 Keywords

The growth of authors' keywords in recent years depicts how the topic in this study has evolved. In Figure 6, we present keyword growth by using the loess smoothed occurrence. Loess is a nonparametric regression strategy for fitting smooth curves to empirical data [76]. The phase "deep reinforcement learning" is a subclass of RL algorithms. "Deep" in this case refers to the number of layers in a neural network. A shallow network has one so-called hidden layer and a deep network has more than one. Training deep neural networks usually requires a large amount of data and extensive computing resources. Thus, a deep RL agent will outperform over the long run [77]. For the control target, "energy" and "thermal comfort" are the most relevant words and are also likely to be important topics for future study.

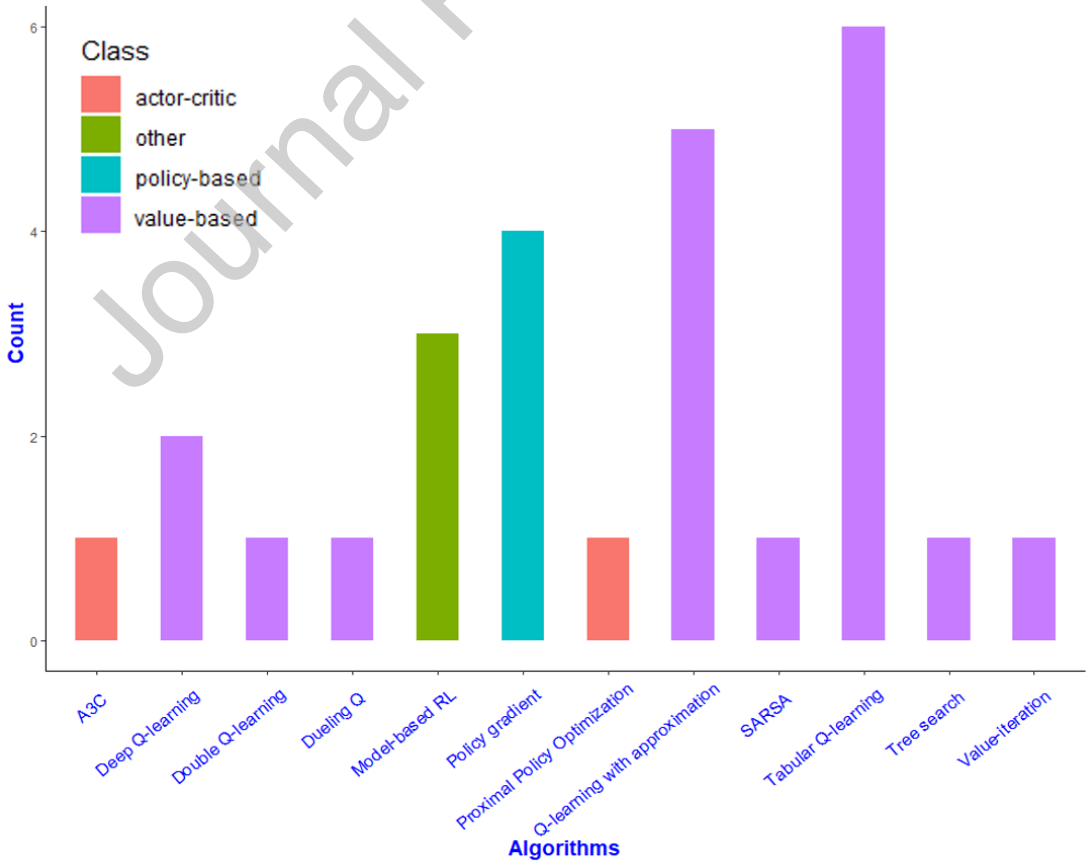

Fig. 5 Algorithms used in the literatures

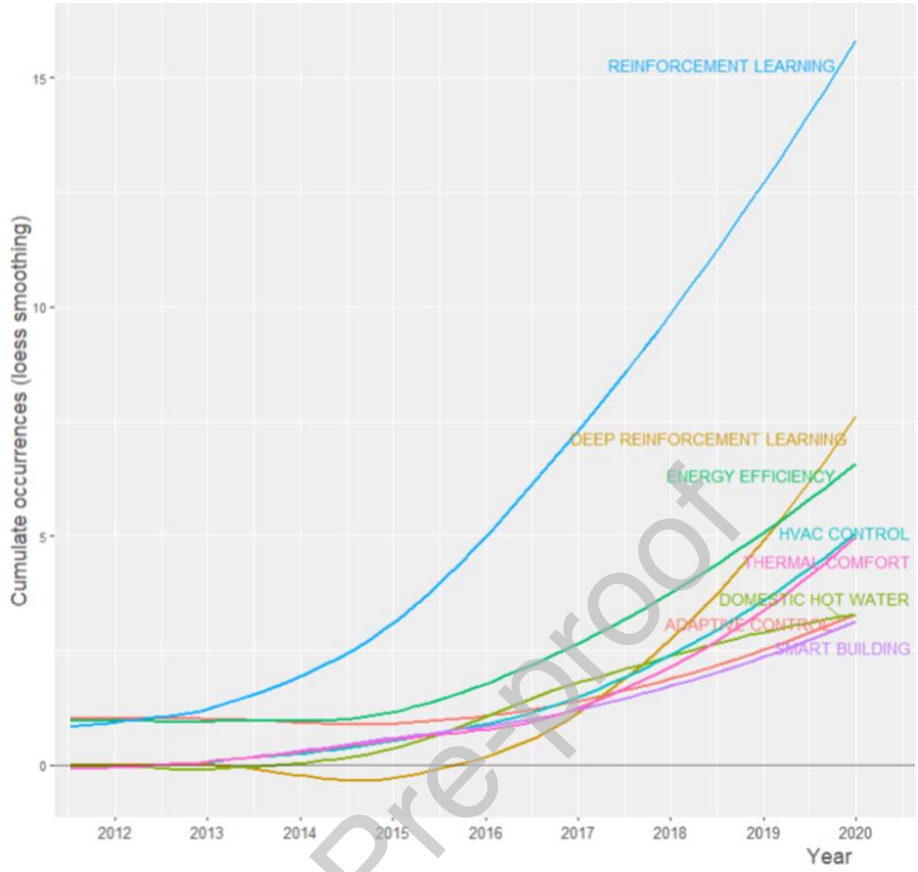

Fig. 6 Keywords growth

5 Discussions

Before training an RL agent, one of two strategies must be selected: on-policy or off-policy. For on-policy training, the agent learning and interacting with the environment is the same. For value-based methods, it estimates the value of the policy being followed. SARSA is on-policy when the agent starts from a state, makes an action, receives a reward, and is transited to next state. Based on the new state, the agent takes an action. The process will be conservative and sensitive to errors, but will be efficient when the exploration penalty is small. On the other hand, agents trained by off-policy are different from those interacting with the environment. Off-policy methods can find the optimal policy even if the agent behaves randomly. Thus, ignoring the interacting agent's policy may lead to a suboptimal policy when most of the rewards are negative. For policy-based methods, there is also a need to consider the gains of applying off-policy learning, because the problems can emerge with large or continuous state-action space and exploration is not feasible. The agent interacting with the environment is usually making policies under the parameter setting θ' that differs from θ for the agent to be trained. Approximations can be made by importance sampling [78] in order to get the gradient. Thus, when an agent is exploring in error-insensitive systems, SARSA may be the preferred option. Agents that do not explore should use Q-learning.

Another issue that needs to be considered is the actual implementation of collecting occupant behavior. On-policy for policy-based methods can only update its gradient when actual actions are made and $J(\theta)$ are observed. Actual deployment of devices in buildings should be able to provide frequent reward and state signals to the agent. Moreover, the repetition of the signals' provision allows the agent to update policy parameter θ. This is still a challenge, not only for devices but also for the occupant to remember to repeatedly react in

the same environment so that more sampled trajectories can be collected. Thus, shifting to off-policy methods makes learning more efficient for complex control tasks.

6 Conclusions

This study has briefly reviewed the reinforcement learning methods for building control that incorporate occupant behavior. Since RL methods assume that the agent interacts with a stochastic environment and works in a data-driven fashion, they are of great importance when forming intelligent building systems where occupant behavior has a significant influence on building performance.

Historical publications on this topic were searched for in Scopus to understand the publication sources, types, years, citations and country collaborations of the existing published literature. It can be seen that, because of the success of deep reinforcement learning in game playing, the number of publications in this field has been growing. The topic covers multiple disciplines including energy, building, computer science, optimal control, sustainability and engineering. Integration of diverse domain knowledge may accelerate the construction of more intelligent systems. However, the current number of citations is not high and international collaborations are still only between a small number of countries. Thus, joint efforts should be made in order to strengthen the research around the topic.

In this study, we first analyzed those studies that examined occupant behavior within the MDP framework. Most of the studies we examined considered occupant behavior as a state for controlling HVAC systems. It is likely that this will remain the focus of new and upcoming work. The rest of the literature can be grouped into three categories regarding the ways of interaction: occupancy, actual behavior and providing feedback where occupant behavior poses an indirect effect on MDP. The reward is the MDP element that is most sensitive to occupant behavior, which makes it essential to design the reward in an efficient way [79], because for occupants with different profiles, their preferences for comfort factors will vary [80], [81].

Over the course of this review we have noticed that the classical tabular Q-learning algorithm has become insufficient for building control with stochastic and complex occupant behavior. Adopting a Q-table to store action values may yield an unreliable policy. As more approximation algorithms have been applied to actual studies, future research should be able to implement, test and verify these in different scenarios. We also compared simplification method and highlighted the function approximation with deep neural network due to the curse of dimensionality. Finally, we discussed some of the issues to be taken into consideration when using off-policy strategy. The implementation of off-policy control requires frequent signal collection from the occupant.

Acknowledgements

The authors are thankful for the financial support from IMMA project of research network, Dalarna University, Sweden.

References

[1] M. P. Fanti, A. M. Mangini, and M. Roccotelli, “A simulation and control model for building energy management,” *Control Engineering Practice*, vol. 72, pp. 192–205, Mar. 2018, doi: 10.1016/j.conengprac.2017.11.010.

[2] P. Xu, E. H.-W. Chan, and Q. K. Qian, "Success factors of energy performance contracting (EPC) for sustainable building energy efficiency retrofit (BEER) of hotel buildings in China," *Energy Policy*, vol. 39, no. 11, pp. 7389–7398, Nov. 2011, doi: 10.1016/j.enpol.2011.09.001.
[3] P. Nejat, F. Jomehzadeh, M. M. Taheri, M. Gohari, and M. Z. Abd. Majid, "A global review of energy consumption, CO 2 emissions and policy in the residential sector (with an overview of the top ten CO 2 emitting countries)," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 843–862, Mar. 2015, doi: 10.1016/j.rser.2014.11.066.
[4] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, vol. 40, no. 3, pp. 394–398, Jan. 2008, doi: 10.1016/j.enbuild.2007.03.007.
[5] T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati, "Advances in research and applications of energy-related occupant behavior in buildings," *Energy and Buildings*, vol. 116, pp. 694–702, Mar. 2016, doi: 10.1016/j.enbuild.2015.11.052.
[6] K.-U. Ahn, D.-W. Kim, C.-S. Park, and P. de Wilde, "Predictability of occupant presence and performance gap in building energy simulation," *Applied Energy*, vol. 208, pp. 1639–1652, Dec. 2017, doi: 10.1016/j.apenergy.2017.04.083.
[7] W. O'Brien, I. Gaetani, S. Gilani, S. Carlucci, P.-J. Hoes, and J. Hensen, "International survey on current occupant modelling approaches in building performance simulation," *Journal of Building Performance Simulation*, vol. 10, no. 5–6, pp. 653–671, Nov. 2017, doi: 10.1080/19401493.2016.1243731.
[8] J. Li, Z. (Jerry) Yu, F. Haghighat, and G. Zhang, "Development and improvement of occupant behavior models towards realistic building performance simulation: A review," *Sustainable Cities and Society*, vol. 50, p. 101685, Oct. 2019, doi: 10.1016/j.scs.2019.101685.
[9] T. Hong, Y. Chen, Z. Belafi, and S. D'Oca, "Occupant behavior models: A critical review of implementation and representation approaches in building performance simulation programs," *Build. Simul.*, vol. 11, no. 1, pp. 1–14, Feb. 2018, doi: 10.1007/s12273-017-0396-6.
[10] A. Mahdavi and F. Tahmasebi, "The deployment-dependence of occupancy-related models in building performance simulation," *Energy and Buildings*, vol. 117, pp. 313–320, Apr. 2016, doi: 10.1016/j.enbuild.2015.09.065.
[11] S. Carlucci *et al.*, "Modeling occupant behavior in buildings," *Building and Environment*, vol. 174, p. 106768, May 2020, doi: 10.1016/j.buildenv.2020.106768.
[12] T. Hong, D. Yan, S. D'Oca, and C. Chen, "Ten questions concerning occupant behavior in buildings: The big picture," *Building and Environment*, vol. 114, pp. 518–530, Mar. 2017, doi: 10.1016/j.buildenv.2016.12.006.
[13] D. Yan *et al.*, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and Buildings*, vol. 107, pp. 264–278, Nov. 2015, doi: 10.1016/j.enbuild.2015.08.032.
[14] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim, "A review on optimized control systems for building energy and comfort management of smart sustainable buildings," *Renewable and Sustainable Energy Reviews*, vol. 34, pp. 409–429, Jun. 2014, doi: 10.1016/j.rser.2014.03.027.
[15] P. Zhao, S. Suryanarayanan, and M. G. Simoes, "An Energy Management System for Building Structures Using a Multi-Agent Decision-Making Control Methodology," 2013, vol. 49(1), pp. 322–330.
[16] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.
[17] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.
[18] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
[19] D. Silver *et al.*, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017, doi: 10.1038/nature24270.

[20] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning," *arXiv:1312.5602 [cs]*, Dec. 2013, Accessed: Jan. 26, 2019. [Online]. Available: http://arxiv.org/abs/1312.5602.
[21] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous Deep Q-Learning with Model-based Acceleration," New York, NY, USA, 2016, vol. 48.
[22] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv:1509.02971 [cs, stat]*, 2016, Accessed: Feb. 02, 2019. [Online]. Available: http://arxiv.org/abs/1509.02971.
[23] M. Aria and C. Cuccurullo, "bibliometrix : An R-tool for comprehensive science mapping analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.
[24] M. Han *et al.*, "A review of reinforcement learning methodologies for controlling occupant comfort in buildings," *Sustainable Cities and Society*, vol. 51, p. 101748, Nov. 2019, doi: 10.1016/j.scs.2019.101748.
[25] A. I. Dounis and C. Caraiscos, "Advanced control systems engineering for energy and comfort management in a building environment—A review," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 6–7, pp. 1246–1261, Aug. 2009, doi: 10.1016/j.rser.2008.09.015.
[26] N. Aste, M. Manfren, and G. Marenzi, "Building Automation and Control Systems and performance optimization: A framework for analysis," *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 313–330, Aug. 2017, doi: 10.1016/j.rser.2016.10.072.
[27] F. Ascione, N. Bianco, C. De Stasio, G. M. Mauro, and G. P. Vanoli, "A new comprehensive approach for cost-optimal building design integrated with the multi-objective model predictive control of HVAC systems," *Sustainable Cities and Society*, vol. 31, pp. 136–150, May 2017, doi: 10.1016/j.scs.2017.02.010.
[28] M. Royapoor, A. Antony, and T. Roskilly, "A review of building climate and plant controls, and a survey of industry perspectives," *Energy and Buildings*, vol. 158, pp. 453–465, Jan. 2018, doi: 10.1016/j.enbuild.2017.10.022.
[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, Second edition. Cambridge, Massachusetts: The MIT Press, 2018.
[30] R. Bellman, "A Markovian Decision Process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.
[31] R. Bellman, *Dynamic Programming*. Princeton University Press, Princeton, 1957.
[32] C. J. C. H. Watkins, "Learning from Delayed Rewards," *Ph.D. thesis, University of Cambridge*, 1989.
[33] M. Hessel and J. Modayil, "Rainbow: Combining Improvements in Deep Reinforcement Learning," pp. 3215–3222.
[34] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," pp. 1057–1063.
[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv:1707.06347 [cs]*, Aug. 2017, Accessed: Apr. 12, 2020. [Online]. Available: http://arxiv.org/abs/1707.06347.
[36] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, "Trust Region Policy Optimization," in *Proceedings of the 31st International Conference on Machine Learning*, France, 2015, vol. 37, pp. 1–9.
[37] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," Denver, Colorado, 2000, vol. 12, pp. 1008–1014.
[38] M. Babaeizadeh, I. Frosio, S. Tyree, J. Clemons, and J. Kautz, "Reinforcement Learning through Asynchronous Advantage Actor-Critic on a GPU," *arXiv:1611.06256 [cs]*, Mar. 2017, Accessed: Apr. 12, 2020. [Online]. Available: http://arxiv.org/abs/1611.06256.
[39] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," *arXiv:1602.01783 [cs]*, Feb. 2016, Accessed: Feb. 03, 2019. [Online]. Available: http://arxiv.org/abs/1602.01783.
[40] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Applied Energy*, vol. 269, p. 115036, Jul. 2020, doi: 10.1016/j.apenergy.2020.115036.
[41] F. Ruelens, S. Iacovella, B. J. Claessens, and R. Belmans, "Learning Agent for a Heat-Pump Thermostat with a Set-Back Strategy Using Model-Free Reinforcement Learning," *energies*, vol. 8, pp. 8300–8318, 2015, doi: doi:10.3390/en8088300.

[42] M. Han *et al.*, "A novel reinforcement learning method for improving occupant comfort via window opening and closing," *Sustainable Cities and Society*, vol. 61, p. 102247, Oct. 2020, doi: 10.1016/j.scs.2020.102247.
[43] H. Kazmi, F. Mehmood, S. Lodeweyckx, and J. Driesen, "Gigawatt-hour scale savings on a budget of zero: Deep reinforcement learning based optimal control of hot water systems," *Energy*, vol. 144, pp. 159–168, Feb. 2018, doi: 10.1016/j.energy.2017.12.019.
[44] R. Jia, M. Jin, K. Sun, T. Hong, and C. Spanos, "Advanced Building Control via Deep Reinforcement Learning," *Energy Procedia*, vol. 158, pp. 6158–6163, Feb. 2019, doi: 10.1016/j.egypro.2019.01.494.
[45] J. Y. Park, T. Dougherty, H. Fritz, and Z. Nagy, "LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning," *Building and Environment*, vol. 147, pp. 397–414, Jan. 2019, doi: 10.1016/j.buildenv.2018.10.028.
[46] W. Valladares *et al.*, "Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm," *Building and Environment*, vol. 155, pp. 105–117, May 2019, doi: 10.1016/j.buildenv.2019.03.038.
[47] C. Marantos, C. Lamprakos, K. Siozios, and D. Soudris, "Towards Plug&Play Smart Thermostats for Building's Heating/Cooling Control," in *IoT for Smart Grids*, K. Siozios, D. Anagnostos, D. Soudris, and E. Kosmatopoulos, Eds. Cham: Springer International Publishing, 2019, pp. 183–207.
[48] D. Lee, S. Lee, P. Karava, and J. Hu, "Simulation-Based Policy Gradient and Its Building Control Application," in *2018 Annual American Control Conference (ACC)*, Milwaukee, WI, Jun. 2018, pp. 5424–5429, doi: 10.23919/ACC.2018.8431592.
[49] Z. Zhang and K. P. Lam, "Practical Implementation and Evaluation of Deep Reinforcement Learning Control for a Radiant Heating System," presented at the The 5th ACM International Conference on Systems for Built Environments, Shenzen, China, 2018, doi: https://doi.org/10.1145/3276774.3276775.
[50] E. Barrett and S. Linder, "Autonomous HVAC Control, A Reinforcement Learning Approach," in *Machine Learning and Knowledge Discovery in Databases*, vol. 9286, A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, and M. Spiliopoulou, Eds. Cham: Springer International Publishing, 2015, pp. 3–19.
[51] P. Fazenda, K. Veeramachaneni, P. Lima, and U.-M. O'Reilly, "Using Reinforcement Learning to Optimize Occupant Comfort and Energy Usage in HVAC Systems," *Journal of Ambient Intelligence and Smart Environments*, vol. 6, no. 6, pp. 675–690, 2014, doi: 10.3233/AIS-140288.
[52] L. Eller, L. C. Siafara, and T. Sauter, "Adaptive control for building energy management using reinforcement learning," in *2018 IEEE International Conference on Industrial Technology (ICIT)*, Lyon, Feb. 2018, pp. 1562–1567, doi: 10.1109/ICIT.2018.8352414.
[53] Z. Zou, X. Yu, and S. Ergan, "Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network," *Building and Environment*, vol. 168, p. 106535, Jan. 2020, doi: 10.1016/j.buildenv.2019.106535.
[54] S. Baghaee and I. Ulusoy, "User comfort and energy efficiency in HVAC systems by Q-learning," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, Izmir, Turkey, May 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404287.
[55] D. Urieli and P. Stone, "A Learning Agent for Heat-Pump Thermostat Control," Saint Paul, Minnesota, USA, 2013.
[56] B. Chen, Z. Cai, and M. Berges, "Gnu-RL: A Precocial Reinforcement Learning Solution for Building HVAC Control Using a Differentiable MPC Policy," New York, NY, USA, 2019, pp. 316–325, doi: https://doi.org/10.1145/3360322.3360849.
[57] Y. Chen, L. K. Norford, H. W. Samuelson, and A. Malkawi, "Optimal control of HVAC and window systems for natural ventilation through reinforcement learning," *Energy and Buildings*, vol. 169, pp. 195–205, Jun. 2018, doi: 10.1016/j.enbuild.2018.03.051.
[58] X. Ding, W. Du, and A. Cerpa, "OCTOPUS: Deep Reinforcement Learning for Holistic Smart Building Control," New York, NY, USA, 2019, pp. 326–335, doi: https://doi.org/10.1145/3360322.3360857.

[59] H. Kazmi and S. D'Oca, "Demonstrating model-based reinforcement learning for energy efficiency and demand response using hot water vessels in net-zero energy buildings," in *2016 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, Ljubljana, Slovenia, Oct. 2016, pp. 1–6, doi: 10.1109/ISGTEurope.2016.7856208.

[60] H. Kazmi, S. D'Oca, C. Delmastro, S. Lodeweyckx, and S. P. Corgnati, "Generalizable occupant-driven optimization model for domestic hot water production in NZEB," *Applied Energy*, vol. 175, pp. 1–15, Aug. 2016, doi: 10.1016/j.apenergy.2016.04.108.

[61] H. Kazmi, J. Suykens, and J. Driesen, "Valuing Knowledge, Information and Agency in Multi-agent Reinforcement Learning: A Case Study in Smart Buildings," in *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, 2018, pp. 585–587.

[62] A. Soares, D. Geysen, F. Spiessens, D. Ectors, O. De Somer, and K. Vanthournout, "Using reinforcement learning for maximizing residential self-consumption – Results from a field test," *Energy and Buildings*, vol. 207, p. 109608, Jan. 2020, doi: 10.1016/j.enbuild.2019.109608.

[63] X. Pan and B. Lee, "An Approach of Reinforcement Learning Based Lighting Control for Demand Response," Nuremberg, Germany, 2016, pp. 558–565.

[64] T. Wei and X. Chen, "Model-based and Data-driven Approaches for Building Automation and Control," San Diego, CA, USA, 2018, pp. 1–8, doi: https://doi.org/10.1145/3240765.3243485.

[65] A. Ali and H. Kazmi, "Minimizing Grid Interaction of Solar Generation and DHW Loads in nZEBs Using Model-Free Reinforcement Learning," in *Data Analytics for Renewable Energy Integration: Informing the Generation and Distribution of Renewable Energy*, Cham, 2017, vol. 10691, pp. 47–58, doi: 10.1007/978-3-319-71643-5_5.

[66] J. Zhu, F. Lauri, A. Koukam, and V. Hilaire, "A Hybrid Intelligent Control System based on PMV Optimization for Thermal Comfort in Smart Buildings," in *Advances in Intelligent Systems and Computing*, 2015, vol. 358, pp. 27–36, doi: https://doi.org/10.1007/978-3-319-17996-4_3.

[67] P. Wei, S. Xia, and X. Jiang, "Energy Saving Recommendations and User Location Modeling in Commercial Buildings," Singapore, 2018, pp. 3–11, doi: https://dl.acm.org/doi/10.1145/3209219.3209244.

[68] S. Lu, W. Wang, C. Lin, and E. C. Hameen, "Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884," *Building and Environment*, vol. 156, pp. 137–146, Jun. 2019, doi: 10.1016/j.buildenv.2019.03.010.

[69] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis, "Reinforcement learning for energy conservation and comfort in buildings," *Building and Environment*, vol. 42, no. 7, pp. 2686–2698, Jul. 2007, doi: 10.1016/j.buildenv.2006.07.010.

[70] K. Sun, Q. Zhao, and J. Zou, "A review of building occupancy measurement systems," *Energy and Buildings*, vol. 216, p. 109965, Jun. 2020, doi: 10.1016/j.enbuild.2020.109965.

[71] W. Kleiminger, C. Beckel, T. Staake, and S. Santini, "Occupancy Detection from Electricity Consumption Data," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings - BuildSys'13*, Roma, Italy, 2013, pp. 1–8, doi: 10.1145/2528282.2528295.

[72] H. Saha, A. R. Florita, G. P. Henze, and S. Sarkar, "Occupancy sensing in buildings: A review of data analytics approaches," *Energy and Buildings*, vol. 188–189, pp. 278–285, Apr. 2019, doi: 10.1016/j.enbuild.2019.02.030.

[73] "ASHRAE Standard 55--Thermal Environmental Conditions for Human Occupancy." ASHRAE Inc, 2017.

[74] Y. Chen, Z. Tong, Y. Zheng, H. Samuelson, and L. Norford, "Transfer learning with deep neural networks for model predictive control of HVAC and natural ventilation in smart buildings," *Journal of Cleaner Production*, vol. 254, p. 119866, May 2020, doi: 10.1016/j.jclepro.2019.119866.

[75] J. R. Vazquez-Canteli, J. Kämpf, G. Henze, and Z. Nagy, "Demo Abstract: CityLearn v1.0 - An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning," presented at the Buildsys, New York, USA, 2019, doi: DOI: 10.1145/3360322.3360998.

[76] W. G. Jacoby, “Loess: a nonparametric, graphical tool for depicting relationships between variables☆,” *Electoral Studies*, vol. 19, pp. 577–613, 2000, doi: DOI: 10.1016/S0261-3794(99)00028-1.
[77] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, “Benchmarking Deep Reinforcement Learning for Continuous Control,” New York, NY, USA, 2016, vol. 48, pp. 1–10.
[78] H. Hachiya, T. Akiyama, M. Sugiayma, and J. Peters, “Adaptive importance sampling for value function approximation in off-policy reinforcement learning,” *Neural Networks*, vol. 22, no. 10, pp. 1399–1410, Dec. 2009, doi: 10.1016/j.neunet.2009.01.002.
[79] Z. Zheng, J. Oh, and S. Singh, “On Learning Intrinsic Rewards for Policy Gradient Methods,” Montréal, Canada, 2018, pp. 1–11.
[80] M. Frontczak and P. Wargocki, “Literature survey on how different factors influence human comfort in indoor environments,” *Building and Environment*, vol. 46, no. 4, pp. 922–937, Apr. 2011, doi: 10.1016/j.buildenv.2010.10.021.
[81] A. Zalejska-Jonsson and M. Wilhelmsson, “Impact of perceived indoor environment quality on overall satisfaction in Swedish dwellings,” *Building and Environment*, vol. 63, pp. 134–144, May 2013, doi: 10.1016/j.buildenv.2013.02.005.

Confilct of Interest

This manuscript has not been published and is not under consideration for publication elsewhere. All authors are employees of non-profit institutes and have no conflicts of interest to disclose. All authors have also read and understood author’s guidelines and ethical policies.

Individual contributions

Mengjie Han: Methodology, Funding acquisition, Software, Visualization, Roles/Writing – original draft

Jing Zhao: Roles/Writing – original draft, Investigation

Xingxing Zhang: Conceptualization, Writing – review & editing, Funding acquisition

Jingchun Shen: Conceptualization, Writing – review & editing

Yu Li: Writing – review & editing