Sustainable Cities and Society 51 (2019) 101748

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

A review of reinforcement learning methodologies for controlling occupant comfort in buildings

Mengjie Han[a], Ross May[a], Xingxing Zhang[a,*], Xinru Wang[a], Song Pan[b], Da Yan[c], Yuan Jin[c], Liguo Xu[d]

[a] School of Technology and Business Studies, Dalarna University, Falun 79188, Sweden
[b] Beijing Key Laboratory of Green Built Environment and Energy Efficient Technology, Beijing University of Technology, Beijing 100124, China
[c] Building Energy Conservation Research Center, Tsinghua University, Beijing 100084, China
[d] School of Management, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an, China

ARTICLE INFO

Keywords:
Reinforcement learning
Control
Building
Indoor comfort
Occupant

ABSTRACT

Classical building control systems are becoming vulnerable with increasing complexities in contemporary built environments and energy systems. Due to this, the reinforcement learning (RL) method is becoming more distinctive and applicable in control networks for buildings. This paper, therefore, conducts a comprehensive review of RL techniques applied in control systems for occupant comfort in indoor built environments. The empirical applications of RL-based control systems are presented, depending on comfort objectives (thermal comfort, indoor air quality, and lighting) along with other objectives which invariably includes energy consumption. The class of RL algorithms and implementation details regarding how the value functions have been represented and how the policies are improved are also illustrated. This paper shows there are limited works for which RL has been explored for controlling occupant comfort, especially in indoor air quality and lighting. Relatively few of the reviewed works incorporate occupancy patterns and/or occupant feedback into the control loop. Moreover, this paper identifies a gap with regard to the performance of implementing cooperative multi-agent RL (MARL). Based on our findings, current challenges and further opportunities are discussed. We expect to clarify the feasible theory and functions of RL for building control systems, which would promote their widerspread application in built environments.

1. Introduction

1.1. Background

Around 90% of people spend most of their time in buildings (Shaikh, Nor, Nallagownden, Elamvazuthi, & Ibrahim, 2013) and they spend 80%–90% of the day indoors, and consequently occupant comfort becomes more and more important. Therefore, maintenance of comfort factors is crucial for improving occupant's feeling of comfort, health, morale, working efficiency as well as productivity (Li, Cui, zhu, Zhang, & Su, 2016). Thermal comfort, visual comfort and indoor air quality (IAQ) seem to be the key parameters that jointly influence the level of comfort of a building occupant (Boodi, Beddiar, Benamour, Amirat, & Benbouzid, 2018; Park & Nagy, 2018; Frontczak & Wargocki, 2011). Thus, reaching a comfortable indoor environment is a multi-objective task and needs comprehensive cooperation between different building components.

Building design and the building management system (BMS) are direct key factors that affect the comfort level of an occupant. The design of buildings relates to the occupancy level, ventilation, use of natural resources, etc., which remains critical for a comfortable indoor climate in future building development (Wang et al., 2018). However, it is a difficult task to find sound design alternatives satisfying different conflicting criteria, such as natural ventilation against heating/cooling loss (Wang, Zmeureanu, & Rivard, 2005). Compared to the design of buildings, the BMS considers both the maintenance and the improvement of the comfort level of the buildings' occupants through a diversity of control methods. A BMS generally refers to the integrated monitoring, transmitting and control of the indoor environment based on various protocols and communication interfaces. Such a characteristic enables the BMS to have a wider application in practice.

In a BMS, the essential function is the building control system (BCS), which is usually designed to maintain indoor comfort at a certain level when responding to dynamic climate and operational conditions. The

* Corresponding author.
E-mail address: xza@du.se (X. Zhang).

https://doi.org/10.1016/j.scs.2019.101748
Received 20 March 2019; Received in revised form 16 June 2019; Accepted 31 July 2019
Available online 07 August 2019

Nomenclature

A2C	advantage Actor-Critic
A3C	asynchronous advantage Actor-Critic
ANN	artificial neural network
BCS	building control system
BMS	building management system
C	continuous
D	discrete
D-DNFQI	double deep neural FQI
eJAL	extended joint action learning
FQI	fitted Q-iteration
GA	genetic algorithm
HRL	hierarchical RL
HVAC	heating, ventilation, and air conditioning
IAQ	indoor air quality
LSTM	long short-term memory
MA	multi-agent
MACS	multi-agent control system
MARL	multi-agent reinforcement learning
MAS	multi-agent system
MDP(s)	Markov decision process(es)
MEC	multi-samples in each cell
MPC	model predictive control
PID	proportional-integral-derivative
RBC	rule-based controls
RBFs	radial basis functions
RL	reinforcement learning
RLCs	reinforcement learning controls
RLS	recursive least-squares
SA	single-agent
SDP	stochastic dynamic programming
TD	temporal difference

Notations

t	discrete time steps
S_t	state at time t stochastically
$\mathscr{S}$	set of states
A_t	action at time t, stochastically
$\mathscr{A}$	set of actions
R_t	reward at time t, stochastically
$\mathscr{R}$	set of rewards
π	policy
γ	discount parameter
s, s'	states
a	action
$\boldsymbol{a}$	joint action between multiple agents
r	reward
$\hat{q}(s, a; \mathbf{w})$	estimate of the Q-function
$\mathbf{w}$	weight vector
$\pi(a\|s)$	prob. of taking action a in state s
$v_\pi(s)$	state-value function
$q_\pi(s, a)$	action-value function
$p(s'\|s, a)$	transition prob.
$r(s, a)$	expected reward
$r(s, a\|a_{others})$	reward for multiple agents
$v_*(s)$	optimal state-value function
$q_*(s, a)$	optimal action-value function
Π_i	policy set for multi-agent
ε	prob. of selecting random actions
$Q(S, A)$	approx. of $q(s, a)$ from data
α, β	step size parameter
$\pi_\theta(a\|s; \theta)$	parametrized policy

advanced control methods are able to, not only take advantage of real-time data - data available as soon as it is created - to produce the desired comfort level, but can also minimise the operational and maintenance cost, and in turn improve the building's energy performance (Marinakis, Karakosta, Doukas, Androulaki, & Psarras, 2013). As a result, there is a high demand for the development of advanced control methods for future smart and economic-friendly building environments.

1.2. Necessity of new methods for the building control system (BCS)

With the development of diverse building systems and the movement towards improving adaptive indoor comfort, buildings are becoming more and more complex to control. In practice, advanced real-time control strategies attempt to make a correct action at a prescribed point in time within defined time tolerances (Gambier, 2004). In a BCS, the controller uses real-time data, which is presented as it is acquired, to make decisions, where its related impact on the indoor habitat is often delayed in such a dynamic setting. The ideal real-time control strategies can deliver the signal so as to avoid the delayed influence on the indoor surroundings. They work effectively based on the building models, the building system models, weather forecast models, and energy tariff forecast models, etc. However, these models are not as accurate in the sense of prediction, thus leading to potential inappropriate control in the future. Therefore, the existing control approaches are facing a serious challenge in real-time adaption/influence to/on occupant comfort and may fail to respond/maintain to/the indoor environment efficiently.

Reinforcement learning (RL), as one of the model-free control techniques, can be an alternative solution to such challenges when it is applied together with real-time control strategies. Model-free control techniques are able to work independently without having a priori knowledge of specific models. For instance, a recently realised Markov-based method, can work in both a model-based and model-free context where the former refers to learning a model and using this to obtain a policy and the latter to learning a policy without learning a model (Kaelbling, Littman, & Moore, 1996). With this approach are the classic learning algorithms, such as Q-learning, $TD(\lambda)$, Dyna, and simulation-based search, that make RL much more attractive and efficient in artificial intelligence applications (2017, Mnih et al., 2015; Silver et al., 2016; Sutton & Barto, 2018). Moreover, the efforts made on solving deep RL problems open up the possibility of working on continuous large datasets (Gu, Lillicrap, Sutskever, & Levine, 2016; Lillicrap et al., 2016). The distinctive property of RL is that the learner or agent, via a trial-and-error paradigm, can make optimal actions without having a supervisor, which essentially fits the goal of a complex control problem.

In BCSs, performances of using RL for occupant comfort have not been analysed from the methodological point of view and the future tasks in this field are still rare. Relevant review works examining RL control methods has been limited (see Table 1 and Fig. 1). Unlike energy demand response (Vázquez-Canteli & Nagy, 2019), this paper considers occupant comfort as the principal optimisation target. Therefore, the aim of this paper is to methodologically review the empirical works on how RL methods have been implemented for comfort control in buildings, and provide instructive directions for future research.

1.3. Contributions and structure

The contributions of this paper are fourfold. Firstly, it summarises the existing relevant review works in different areas of occupant comfort control, including thermal control, indoor air quality (IAQ) control, lighting control, air velocity control, and visual comfort control, etc.

Table 1
Relevant review works.

Ref. & Year	Comfort type	Control methods/algorithms	RL reviewed	Future indications
(Galasiu & Veitch, 2006)	Lighting	Lighting control; shading control	no	Satisfaction and outdoor conditions need to be considered in control system.
(Dounis & Caraiscos, 2009)	Thermal; visual; IAQ	Conventional control; intelligent control; agent-based control	yes	Future trends: balance between thermal comfort and energy usage and random neural networks with RL
(Wenqi & Zhou, 2009)	Thermal	ANN; fuzzy logic	no	More real-time environmental data and human activity level can be collected and applied in the system design.
(Guo, Tiller, Henze, & Waters, 2010)	Lighting	Occupant based control	no	–
(Roetzel, Tsangrassoulis, Dietrich, & Busching, 2010)	Thermal; IAQ (ventilation)	Occupant based control	no	–
(Haq et al., 2014)	Lighting	Occupant based control; daylight-linked control; scheduling control; mixed control	no	Development of control algorithm can be helpful in improving the effectivity of commissioning and lead to better user satisfaction.
(Shaikh et al., 2014)	Thermal; visual; IAQ (CO2, humidity, plug loads)	Conventional control (on/off, PID); intelligent control (learning method, model-based predictive, agent based)	no	Various other artificial intelligent techniques need to be future research objectives.
(Veselý & Zeiler, 2014)	Thermal	Occupant based control	no	–
(Song et al., 2015)	Thermal; IAQ; visual	Conventional control; computational intelligent control	no	Future trends: model-independent control strategies for general purpose use which can reduce the development time for model matching and parameters tuning
(Chenari, Dias Carrilho, & Gameiro da Silva, 2016)	Thermal; IAQ (CO2, contaminant); humidity	Model-based control; rule-based control; GA	no	Study of intelligent window-based hybrid ventilation strategies for maintaining IAQ and reducing energy consumption is missing.
(Merabti et al., 2016)	Thermal; IAQ	PID; fuzzy; fuzzy PID; adaptive fuzzy PD; NN; neuro-fuzzy; GA	no	Intelligent control system need to be upgraded.
(Enescu, 2017)	Thermal	ANN; AR and hybrid AR-ANN; fuzzy, hybrid ANN-fuzzy	no	Refined adaptive comfort models in smart building control systems are needed.
(Wang, Kuckelkorn, & Liu, 2017)	Thermal; humidity; CO2; air velocity	Binary; iterative; PID; MPC; nonlinear; pole-placement; optimal; fuzzy; ANN; adaptive	no	Advanced control strategies combined with HVAC technologies have been currently becoming a new trend in building energy conservation and indoor environment quality research.
(Ye, Zhang, Gao et al., 2017)	IAQ (pollutants)	Occupant-based control	no	–
(Boodi et al., 2018)	Thermal; lighting; IAQ; humidity	MPC; PID; fuzzy logic; RL; ANN predictive; rule-based	yes	Future trends: intelligent building models and adaptive building controller
(Guyot, Sherman, & Walker, 2018)	IAQ (CO2, humidity)	–	no	Smart ventilation is still an emerging technology.
(Kruisselbrink, Dangol, & Rosemann, 2018)	Lighting quality	–	no	–
(Park & Nagy, 2018)	Thermal	Rule-based; optimisation; intelligent control; MPC	no	Only 5.2% and 15.6% of thermal comfort and building control publications cited each other.
(Royapoor et al., 2018)	–	Classic (binary and PID); computational (supervisory, RL, Fuzzy logic, robust, ANN, agent-based)	yes	More advanced computational techniques (ANN or agent-based) have so far largely remained in demonstration stage.

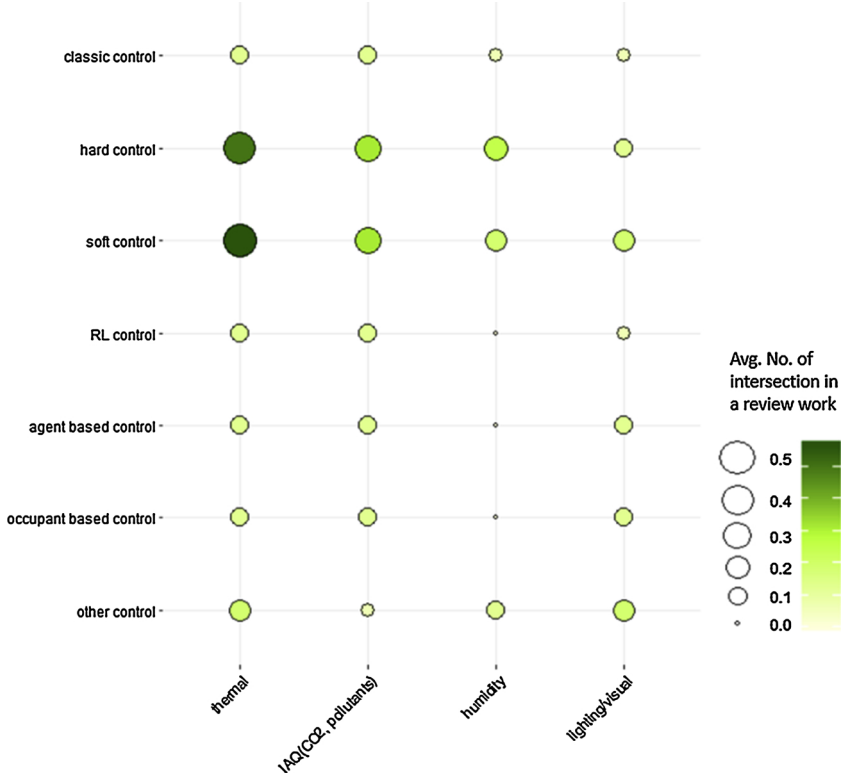

Fig. 1. Average number of intersections of different comfort types and control methods reviewed in each work.

Secondly, under this control setting, it provides a comprehensive review of RL being implemented. Thirdly, it analyses the application of RL for comfort control in multi-agent environments. Fourthly, it highlights the potential of RL as a sustainable forerunner for truly occupant-centric building operation in the evolving smart city. Beyond these, this paper finally identifies the current research gaps and proposes future research from both the application and methodology points of view.

In Section 2 of this paper, we examine relevant review works and their indications. In Section 3, we briefly introduce the philosophy of RL along with some applicable algorithms. Section 4 then analyses the current publications in both single agent and multi-agent systems. In Section 5, we discuss some challenges, and in Section 6 we conclude the findings and give some avenues for future work.

1.4. Review methods

In this paper, we make our search for articles (Table 1 and Table 2) in the search engines, Web of Science, ScienceDirect and Google Scholar. We do not limit the publication time. For the review articles given in Table 1, our searching keywords are, *review* and *control* and *occupant* and *building*; *review* and *control* and "*indoor comfort*" and *building*.

For the core articles given in Table 2, our searching keywords are

$$\left[(building\,(s)) \cap \begin{pmatrix} (reinforcement\ learning) \\ (Markov\ decision\ processes) \\ Q-learning \end{pmatrix} \cap \begin{pmatrix} comfort \\ (thermal\ comfort) \\ (visual\ comfort) \\ (indoor\ air\ quality) \\ occupant \\ (indoor\ environment) \end{pmatrix}\right] \cup \begin{bmatrix} (model\ free\ control) \\ (intelligent\ control) \end{bmatrix}.$$

The intersection notation, ∩, is the relation operation, "and", where the words on both sides are searched for simultaneously. ∪ is equivalent to, "or", where only one word or phrase on either side is taken for searching. For example, (*building* and *"reinforcement learning"* and *comfort)* or *("model free control")* is one of our searching records. We also search *Markov decision processes* (*MDPs) and Q-learning* to guarantee that the underlying theory of RL and the most popular algorithm are covered. We also include "*model free control*" and "*intelligent control*" as alternative keywords because some articles treat RL as a special case of these control methods. We read through every search outcome and excluded irrelevant articles without direct optimisation on comfort. That is, we only included those core articles that have clearly optimised for comfort. Other joint optimisation objectives may have also been considered in the core articles but our main interest was in those articles containing at least one comfort component in the optimisation objectives. Furthermore, only those papers which empirically investigated their proposed approaches, either through the use of synthetic data or real data were included. Doing so, we have identified 33 core articles that are summarised and analysed in Section 4.

2. Summary of relevant review works and their indications

A summary of previous review works that have examined building control methods for occupant comfort factors is shown in Table 1 in chronological order. In addition to examining all mentioned comfort types and control methods in those works, we have also checked if RL has been reviewed and whether there are any future indications in each work. Herein, we focus on those indications that point out some possibilities of implementing RL for comfort control. For instance, Dounis & Caraiscos suggest using random neural networks with RL to control thermal comfort and energy usage (Dounis & Caraiscos, 2009), while Song et al indicate model-independent control strategies for general purpose use which can reduce the development time for model matching and parameter tuning (Song, Wu, & Yan, 2015). Similarly, the development of artificial intelligence control strategies is regarded as an important future task (Merabti, Draoui, & Bounaama, 2016; Shaikh, Nor, Nallagownden, Elamvazuthi, & Ibrahim, 2014). In this sense, the neglected model-free RL technique has been regarded as a promising and attractive method for controlling occupant comfort, and we

Table 2
Empirical applications of RL control on occupant comfort.

Ref. and year	Energy system	Optimisation objectives		Class	Learning control method	Value function representation	Policy improvement	Pre-training	Agent	States	Actions
		Comfort	Others								
(Baghaee & Ulusoy, 2018)	HVAC	IAQ	Energy consumption	Value based	Q-Learning	Tabular	N/A	No	SA	D	D
(Barrett & Linder, 2015)	HVAC	Thermal	Energy cost	Value-based	Q-Learning	Tabular	ε-greedy	No	SA	D	D
(Bielskis et al., 2013)	HVAC and lighting	Thermal and lighting	N/A	Actor-Critic	TD Actor-Critic	Function approximator	Policy gradient	No	SA	C	C
(Bonte, Perles, Lartigue, & Thellier, 2014)	Heating	Thermal	N/A	Value-based	Q-Learning	N/A	N/A	No	SA	N/A	N/A
(Chen et al., 2018)	HVAC and window	Thermal and IAQ	Energy consumption	Value-based	Q-Learning	N/A	ε-greedy	No	SA	N/A	D
(Cheng et al., 2016)	Blinds and lighting	Lighting	Energy consumption	Value-based	Q-Learning	N/A	ε-greedy	No	SA	D	D
(Dalamagkidis et al., 2007)	HVAC	Thermal and IAQ	Energy consumption	Value-based	RLS-TD(λ)	Function approximator	ε-greedy	No	SA	C	D
(Dalamagkidis & Kolokots, 2008)	HVAC	Thermal	Energy consumption	Value-based	RLS-TD(λ)	Function approximator	N/A	No	SA	C	D
(Eller et al., 2018)	Heating and ventilation	Thermal and IAQ	Energy consumption	Value-based	Q-Learning	Tabular	N/A	No	SA	D	D
(Fu et al., 2018)	HVAC	Thermal and IAQ	Energy consumption	Value-based	SARSA	Function approximator	ε-greedy	No	SA	C	D
(Hurtado et al., 2018)	Generic loads	Thermal and IAQ	Demand flexibility	Value-based	eJAL, Q-learning	N/A	N/A	N/A	MA	N/A	N/A
(Jouffe, 1997)	Ventilation	Thermal and IAQ	N/A	Actor-Critic	TD(λ) Actor-Critic	Function approximator	Policy gradient	No	SA	N/A	D
(Khalili, Wu, & Aghajan, 2010)	Lighting	Lighting	Energy consumption	Value-based	MAXQ, N/A	Tabular	HRL Epsilon-Decay, HRL Greedy, Regular Epsilon-Decay, Normal Greedy	No	SA	D	D
(Li et al., 2015)	HVAC	Thermal	N/A	Value-based	MEC, Q-Learning	Function approximator, tabular	PAC, ε-greedy	N/A	SA	D	D
(Li & Xia, 2015)	HVAC	Thermal	Energy consumption	Value-based	Q-Learning	Multi-grid method	Ergodic exploration	No	SA	N/A	N/A
(Lu, Wang, Lin, & Hameen, 2019)	HVAC	Thermal	N/A	Value-based	Q-Learning	Tabular	ε-greedy	No	SA	D	D
(Mozer, 1998)	HVAC, DHW, lighting,	Thermal and lighting	Energy cost	N/A	N/A	N/A	N/A	N/A	SA	N/A	N/A
(Nagy et al., 2018)	Heat pump	Thermal	Energy cost	Value-based	Dyna, D-DNFQI	N/A, function approximator	ε-greedy	No	SA	C	D
(Park et al., 2019)	Smart appliances	Lighting	Energy consumption	Value-based	Value iteration	Tabular	N/A	N/A	MA	D	D
(Pedro et al., 2014)	HVAC	Thermal	Energy cost	Value-based	Q-Learning	Tabular, function approximator	N/A	No	SA	D,C	D,C
(Ruelens et al., 2015)	Heat pump	Thermal	Energy consumption	Value-based	Fitted Q-Iteration	Function approximator	Soft-max	No	SA	N/A	N/A
(Sato et al., 2012)	Air Conditioner	Thermal	Energy consumption	Value-based	Q-Learning	Tabular	N/A	No	SA	D	D
(Schmidt et al., 2017)	HVAC	Thermal	Energy consumption	Value-based	Fitted Q-Iteration	Function approximator	N/A	No	SA	N/A	D
(Sun et al., 2013)	HVAC	ThermaI and IAQ	Energy cost	Value-based	Q-Learning	N/A	N/A	N/A	SA	N/A	N/A
(Sun, Luh et al., 2015)	HVAC	Thermal	Energy cost	Value-based	N/A	N/A	ε-greedy	No	MA	D	D

(*continued on next page*)

Table 2 (*continued*)

Ref. and year	Energy system	Optimisation objectives		Class	Learning control method	Value function representation	Policy improvement	Pre-training	Agent	States	Actions
		Comfort	Others								
(Sun, Somani et al., 2015)	HVAC	Thermal and IAQ	Energy cost	Value-based	Q-Learning	N/A	ϵ-greedy	N/A	SA	N/A	N/A
(Urieli & Stone, 2013)	HVAC	Thermal	Energy consumption	Value based	Simulation based tree search	N/A	N/A	No	SA	N/A	D
(Wang, Velswamy et al., 2017)	HVAC	Thermal	Energy consumption	Actor-Critic	A2C	Function approximator	Policy gradient	Yes	SA	N/A	D
(Wei et al., 2017)	HVAC	Thermal	Energy cost	Value-based	Variant of DQN, Q-learning	Function approximator, N/A	ϵ-greedy, N/A	N/A	MA/ SA	N/A	D
(Yang et al., 2015)	PV/T, thermal storage, heat pump	Thermal	Energy consumption	Value-based	Q-Learning	Tabular, function approximator	ϵ-greedy	No	MA	C,D	D
(Yu & Dexter, 2010)	HVAC	Thermal	Energy cost	Value based	Q(λ)	Tabular	ϵ-greedy	Yes	SA	FD	D
(Zhang et al., 2018)	HVAC	Thermal	Energy consumption	Actor-Critic	A3C	Function approximator	Policy gradient	Yes	SA	N/A	D
(Zhang & Lam, 2018)	HVAC	Thermal	Energy consumption	Actor-Critic	A3C	Function approximator	Policy gradient	No	SA	N/A	D

therefore have a rationale to highlight it.

In all the 19 review works, thermal comfort accounts for the majority of all topics and ANN, model predictive control (MPC) and the fuzzy method seem to be the most often reviewed control methods. RL has appeared three times and only around ten relevant empirical papers have been investigated (Boodi et al., 2018; Dounis & Caraiscos, 2009; Royapoor, Antony, & Roskilly, 2018). To have an overview of the distribution, we divide the control methods into classic control (on/off and PID), hard control (MPC, optimal control, nonlinear control, adaptive control), soft or intelligent control (ANN, fuzzy-based), RL control, agent-based control and occupant-based control. Considering the frequency, we divide comfort factors into the four most common occurring amongst the review articles, namely, thermal comfort, IAQ, humidity and visual comfort. Each intersection of control method and comfort type is counted once if at least one paper was reviewed. We aggregate them and divide them by nineteen to obtain the average examining rate. Specifically, Fig. 1 gives the average number of intersections of different comfort types and control methods reviewed in each work. For thermal comfort, both hard and soft control appear most frequently and reach at least a 50% examining rate. For IAQ and humidity, the percentage ranges from 19% to 31%. For visual comfort, the control methods distribute relatively evenly. Looking at RL, the rates are not more than 13% among all comfort types. This low examining rate makes it appealing to investigate the framework of combining RL with occupant comfort control.

A recent work reviewed the application of reinforcement learning for demand response, which is relevant for integrating renewable energy sources into the smart grid (Vázquez-Canteli & Nagy, 2019). The authors pointed out that human comfort and satisfaction in buildings have been mostly studied in single agent systems. The performance of RL in multi-agent systems needs to be explored, which is one of our goals in our study.

3. The reinforcement learning method

The idea of reinforcement learning originated from the term "optimal control" which emerged in the late 1950s, where a problem was formulated by designing a controller to minimise a measure of the behavior of a system over time (Sutton & Barto, 1998, 2018). Bellman (Bellman, 1957a) came up with the concept of MDPs or finite MDPs, a fundamental theory of RL, to formulate optimal control problems.

The learner or *agent* of RL learns how to map situations to actions to maximise a numerical delayed reward signal. It does not have to have a "teacher" telling it how to take an action but, rather, makes decisions via implementing a trial-and-error search, and recognizing the delayed reward from the environment that the agent interacts with (Sutton & Barto, 2018). RL is neither supervised learning nor unsupervised learning; it is a third category of machine learning. Whereas supervised learning gets signals of correct actions, RL gets signals from the reward of an action without knowing if the action was correct or not. RL, in a sense, is the core of machine learning techniques. In the context of artificial intelligence, RL allows the agent to automatically determine behaviors, which cannot be achieved by supervised learning or unsupervised learning.

3.1. Elements of reinforcement learning and MDPs

3.1.1. Elements

In a dynamic sequential decision-making process, the *state* $S_t \in \mathscr{S}$ refers to a specific condition of the environment at discrete time steps $t = 0,1, \ldots$. By realising and responding to the environment, the agent chooses a deterministic or stochastic *action* $A_t \in \mathscr{A}$ that tries to maximise future returns and receives an instant *reward* $R_{t+1} \in \mathscr{R}$ as the agent transfers to the new state S_{t+1}. The reward is usually represented by a quantitative measurement. Fig. 2 (Sutton & Barto, 1998) shows how a sequence of state, action, and reward is generated to form an MDP.

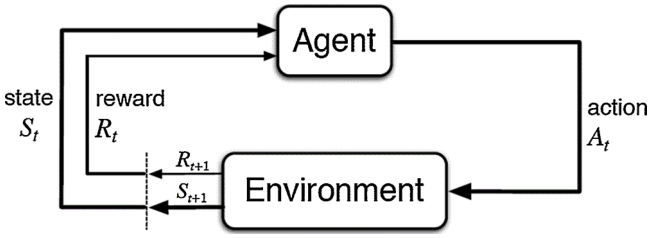

Fig. 2. The interaction between agent and environment in an MDP.

3.1.2. Markov decision processes

The Markov property tells us that the future is independent of the past and depends only on the present. In Fig. 2, 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 $s,\ s' \in \mathscr{S}$, $r \in \mathscr{R}$, and $a \in \mathscr{A}$. 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. Eq. (1) implies the basic rule of how the MDP works and one can easily determine the marginal transition probabilities $p(s'|s, a)$:

$$p(s'|s, a) = \mathbb{P}[S_t = s'|S_{t-1} = s, A_{t-1} = a] = \sum_{r\in\mathscr{R}} p(s', r|s, a). \tag{2}$$

Eq. (3) gives the expected reward by using the marginal distribution of R_t:

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r\in\mathscr{R}} r\sum_{s'\in\mathscr{S}} p(s', r|s, a). \tag{3}$$

Both Eqs. (2) and (3) are used for solving the optimal value functions presented in Section 2.3.

3.2. Policies and value functions

A *policy* π is a distribution over actions given states. It fully defines the behavior of an agent by telling the agent how to act when it is in different states. The policy itself is either deterministic or stochastic (Sutton & Barto, 1998) and the probability of taking an action, a, in state s is:

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]. \tag{4}$$

The policy can be considered as a function of actions. It acts either as a look-up table or in an approximation form (see Section 4 for the discussion). The overall goal of RL is to find the optimal policy given a state.

An optimal policy tries to maximise the expected future return from time t: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...$, where $0 \le \gamma \le 1$ is the discount parameter. The *state-value function,* $v_\pi(s)$, and the *action-value function,* $q_\pi(s, a)$, are two useful measures in RL that can be estimated from the data. The literature defines $v_\pi(s)$, of an MDP, under policy π, as the expectation of the return starting from state s:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s],\ \textit{for all}\ s \in \mathscr{S}. \tag{5}$$

In practical applications, $v_\pi(s)$ is more applicable for model-based problems, whereas the action-value function, $q_\pi(s, a)$, is more useful in the model-free context. When the full environment or the model is unknown, episodic simulations are often used to estimate $q_\pi(s, a)$, that is, under policy π, the expectation of the return starting from state s and taking the action a:

$$\begin{aligned} q_\pi(s,\ a) &= \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a]\ ,\ \textit{for all}\ s \in \mathscr{S}\ \textit{and}\ a \in \mathscr{A}. \end{aligned} \tag{6}$$

The task of finding the optimal policy, π_*, is achieved by evaluating either the optimal state-value function

$$v_*(s) = \max_\pi v_\pi(s), \tag{7}$$

or the optimal action-value function

$$q_*(s, a) = \max_\pi q_\pi(s, a). \tag{8}$$

3.3. Bellman optimality equation

One way of optimising Eqs. (7) or (8) is to make use of the recursive relationships between two states or actions in a sequential order. Since the procedures are similar, we only present the relationship starting from the action-values, i.e. the *Bellman optimality equation* for $q_*(s, a)$ (Bellman, 1957b).

The backup diagrams in Fig. 3 show relationships between the value function and a state or state-action pairs. Fig. 3(a) considers the optimal

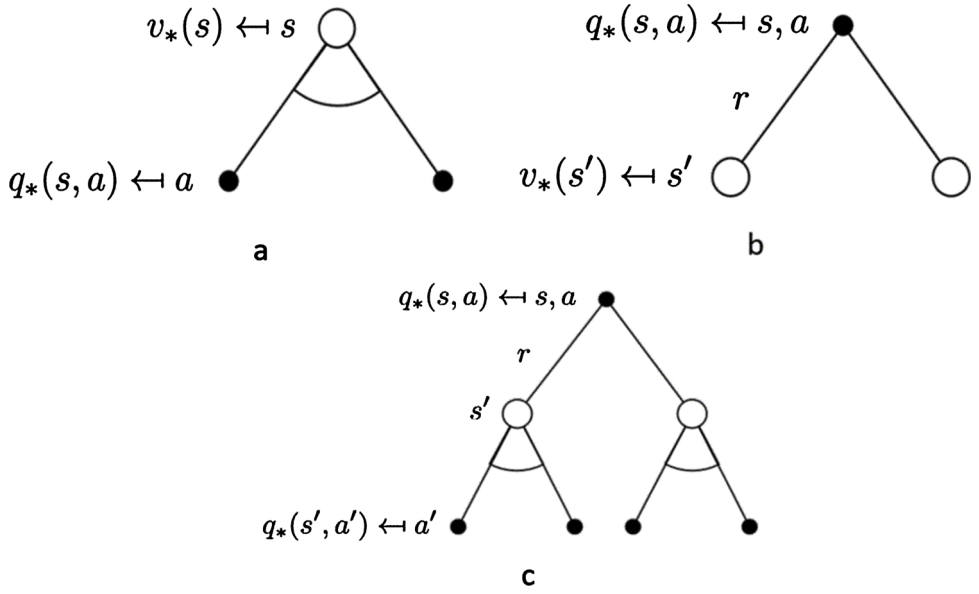

Fig. 3. Backup diagrams for the optimal value functions.

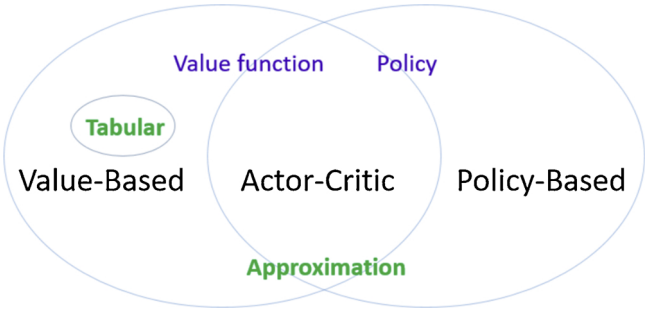

Fig. 4. Classification of RL algorithms.

state-value function when taking an action. The agent looks at each of the possible actions it might take and selects the action with maximum action-value that tells the agent how good the state is. That is,

$$v_*(s) = \max_a q_*(s, a). \tag{9}$$

Similarly, Fig. 3(b) evaluates the dynamic and stochastic environment when an action is taken. Each of the states it ends up in has an optimal value. Thus, the optimal action-value counts the immediate expected reward, $r(s, a)$, from Eq. (3), and a discounted optimal state-value:

$$q_*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) v_*(s'). \tag{10}$$

Thus, as shown in Fig. 3(c), the Bellman optimality equation for $q_*(s, a)$ is obtained by substituting Eq. (9) into Eq. (10):

$$q_*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \max_{a'} q_*(s', a'). \tag{11}$$

In a similar way, we can derive the Bellman optimality equation for $v_*(s)$. Both of them are the fundamental expressions for MDPs. The recursive relationship assists in splitting the current value function into the immediate reward and the value of the next action. Some of the specific learning algorithms, presented in Section 3.4, make use of the Bellman optimality equations to reach optimal policies.

3.4. Categorisation of RL algorithms

In this subsection, we briefly summarise RL algorithms. The purpose is to examine how current research has methodologically explored the algorithms and to propose potential future work. There are many categorisation methods for RL algorithms. As shown in Fig. 4, an iteration-based classification suggests value-based methods, policy-based methods and a fusion of value-based and policy-based methods known as actor-critic methods.

Value-based methods, such as the off-policy Q-learning algorithm (Watkins, 1989), start with a random value function and update to an improved value function in an iterative process until reaching the optimal value function $Q(S, A)$. The optimal policy is made by selecting the action that optimises the value function at a certain state. For some value based methods, e.g. the on-policy *SARSA* and *SARSA(λ)* algorithms (Rummery & Niranjan, 1994), they evaluate policies by constructing their value functions and use these value functions to find improved policies. The distinguishing feature between off-policy and on-policy learning is that in the former the policy being learned is different to the one being followed whereas the latter approach follows the policy being learned. One of the advantages about the former way is that the optimal policy can be learned whilst following a different control strategy, for example, an MPC or rule-based control (RBC) strategy (Sutton & Barto, 2018; Vázquez-Canteli, Ulyanin, Kämpf, & Nagy, 2019).

In systems 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 action-value. The tabular method is straightforward to implement and guarantees convergence (Sutton & Barto, 2018). For large MDP problems, however, we do not always want to separately see the trajectory of each entry of 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, etc. It generates the state-actions that we cannot observe. For the incremental approximation method, $\mathbf{w}$ is updated by gradient descent:

$$\mathbf{w_{t+1}} = \mathbf{w_t} + \beta[q_\pi(S_t, A_t) - \hat{q}(S_t, A_t; \mathbf{w_t})]\nabla\hat{q}(S_t, A_t; \mathbf{w_t}). \tag{12}$$

The learning target $q_\pi(S_t, A_t)$ is iteratively obtained from the Bellman Equation Eq. (11). Whereas the incremental method makes use of the experience once to update the estimate of the value function, and then throwing it away before going to the next step, the batch method is sample efficient and tries to find the best fit to all of the data. (Ernst, Geurts, & Wehenkel, 2005; Xu, He, & Hu, 2002).

Policy-based methods use optimisation techniques to directly search for an optimal policy. Both the tabular and approximation methods work in value-based paradigms where the value functions have to be approximated and the policy is taken by greedy or ε -greedy strategies, whereas the policy-based method directly searches for the parametrised policy:

$$\pi_\theta(a|s; \theta) = \mathbb{P}[A_t = a|S_t = s; \theta_t = \theta]. \tag{13}$$

The policy-based method gives better convergence, especially for the continuous state-action space. In episodic experiments, the expected value of the start state is used as the objective function. The gradient ascent technique iteratively updates θ for the optimisation. The action preference is usually assigned to a probability to avoid the deterministic policy.

Furthermore, a combination of value-based and policy-based methods, e.g. the Actor-Critic algorithm (Grondman, Busoniu, Lopes, & Babuska, 2012; Mnih et al., 2016), is also appealing. The Actor makes an action when it observes a state. The Critic then marks the Actor's performance. The Actor adjusts the policy, e.g. parameters in ANN, according to the score it obtained and the Critic adjusts its marking policy by evaluating the reward. In this way, both the Actor and the Critic improve themselves from random policies to better policies.

3.5. Multi-agent systems

As an agent-based technology, the multi-agent system (MAS) provides promising paradigms in artificial intelligence (Sycara, 1998). The decomposition of complex systems facilitates each agent to share a common environment and work independently on a specific sub-problem. A MAS extends the single agent system by allowing each agent to interact with other agents – not simply by exchanging data, but also by engaging in analogues of social activity: cooperation and negotiation.

3.5.1. Multi-agent control systems

When it comes to a multi-agent control system (MACS) in buildings, each agent implements autonomous actions in order to optimally run building models in a dynamic system. In most of the studies, the hierarchical central-local agent structure is embedded in the building models to balance the energy consumption and the occupants' comfort (Yang & Wang, 2012, 2013). A multi-objective optimisation technique is usually utilized for optimising intelligent management. In Fig. 5 (Yang & Wang, 2013), for instance, a central agent communicates with building managers and zone agents to decide the optimal power distribution for each zone by considering the comfort demand. A zone agent, on the other hand, communicates with occupants and decides power demand. The local agents take care of temperature control, illumination control, and air quality control. An occupant-driven control is studied in MASs including HVAC agents, occupant agents and meeting agents (Klein et al., 2012), where the MDP-based coordination

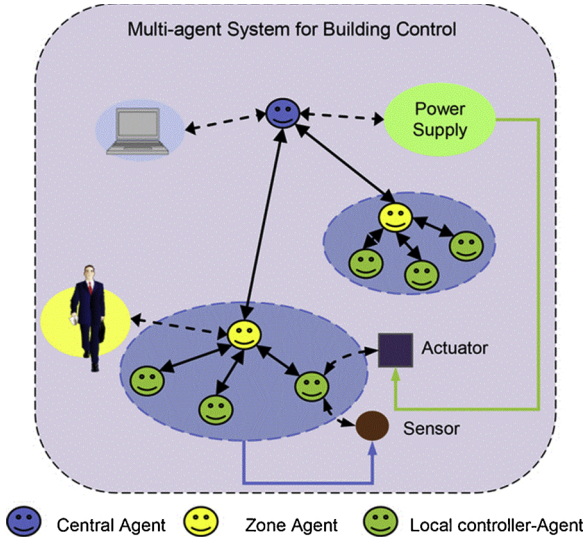

Fig. 5. MAS for building energy control.

tries to find the optimal policy by considering energy consumption, occupant comfort, and scheduling convenience individually.

3.5.2. Multi-agent RL control

The MDP property for multi-agent reinforcement learning (MARL) has been extensively studied in matrix game playing (Littman, 1994) since both cooperative and competitive as well as a mixed environment can be modeled and simulated. The survey work (Buşoniu, Babuška, & De Schutter, 2010) has summarised and explored MARL theory, algorithms and applications. The benefit of MARL comes from experience sharing, information exchange, and skill learning among agents. When one or more agents fail to work in the system, the remaining agents are still able to react optimally by learning from the new environment. Generally, the transition probability for MARL extends Eq. (1) to a multi-action case:

$$p(s^{'}|s,\ \boldsymbol{a}) = P\{S_t = s^{'}|S_{t-1} = s,\ \boldsymbol{a} = (a_1,\ ..., a_n)\} \tag{14}$$

where n is the total number of agents, $s^{'}$, s belong to the set of states S and $\boldsymbol{a} \in A_1 \times \cdots \times A_n$ is the joint action of the agents where A_i, $i = 1, \ldots, n$ is the action set for agent i. In Eq. (14), the stochastic transition is a probability distribution over next states $s^{'}$ given the current state s and joint action $\mathbf{a}$ of the agents. For the policy $\pi_i \in \Pi_i$, the optimal policy π_i^* for agent i fulfills the Nash equilibrium:

$$\begin{aligned} &\sum_{a_1, ..., a_n} q_*(s, \boldsymbol{a})\pi_1^*(a_1|s)\bullet ...\bullet \pi_i^*(a_i|s)\bullet ...\bullet \pi_n^*(a_n|s) \\ &\geq \sum_{a_1, ..., a_n} q_*(s, \boldsymbol{a})\pi_1^*(a_1|s)\bullet ...\bullet \pi_i(a_i|s)\bullet ...\bullet \pi_n^*(a_n|s), \end{aligned} \tag{15}$$

where $q_*(s, \boldsymbol{a})$ is the optimal action-value function for agent i and $\pi_i^*(a_i|s)$ is the individual probability of taking action a_i given the Nash equilibrium policy.

4. Applications of reinforcement learning methods for comfort control in buildings

The following section reviews the applications of RL methods for occupant comfort control in buildings. We consider the algorithms implemented, their exploration vs exploitation strategies, and whether the application was from a single or multi-agent perspective. We close the section by discussing those applications that were implemented in a physical setting.

Table 2 gives a summary of the reviewed literature pertaining to RL methods applied to comfort controls in buildings. We show specific learning algorithms and the classes they belong to for each publication. We also investigate the representation of value functions to highlight optimisation techniques. Pre-training refers to whether or not the agents were implemented with pre-trained policies using existing data or simplified models of the physical system. We further distinguish between discrete (D) and continuous (C) state/action spaces. Unless otherwise stated, any reference to RL methods should be assumed to be model-free methods.

4.1. Comfort factors

In the building literature, there are several well-established comfort factors (Dalamagkidis & Kolokots, 2008). These are thermal comfort, IAQ, lighting, and noise. To ensure the quality of comfort of the buildings' occupants, intelligent controls are seen as an ideal strategy for maintaining the standards of these environmental factors as outlined by, for example, ASHRAE 55 (ASHRAE Standard 55, 2017) and EN 15251 (CEN prEN15251, 2005). We have thus divided the articles according to these factors. In the literature reviewed, we have identified three out of the four factors, outlined above, in which RL control has been applied. These are thermal comfort, IAQ, and lighting. Among these articles were some that also optimised over a combination of these comfort factors. Because of this, we first analyse those articles in which only a single comfort objective was considered. Following this, we then analyse those articles in which at least two comfort components were included in the objectives.

4.1.1. Thermal comfort

Thermal comfort alone has had the most interest compared to the other comfort factors. Dalamagkidis and Kolokotsa implemented an RL control for an HVAC system with the goal of maximising both thermal comfort and energy conservation, where a heavier emphasis on thermal comfort was made (Dalamagkidis & Kolokots, 2008). They compared the performance of the RL control to the performance of a fuzzy-PD and a common on/off control over a 5-year simulated time period. They found that after 4 years of simulation the RL control achieved as good as if not better performance than the other two controls. They also suggested pre-training the control before deploying it in a real environment to mitigate suboptimal performance due to policy exploration. For example, Yu and Dexter used RL to tune a fuzzy rule-based supervisory control for an HVAC system. They found that by pre-training the RL control it was able to improve the performance of a low-energy building system in an acceptable period of time. (Yu & Dexter, 2010). Sato et al, by considering occupants daily action plans, implemented RL to maximise both energy conservation and thermal comfort through controlling the operation of an air conditioning system (Sato, Samejima, Akiyoshi, & Komoda, 2012). They found their proposed method to be effective. Similarly, Pedro et al, using RL to control an HVAC system and based on the tenant's preferences and occupancy patterns, were able to achieve energy efficiency while maintaining the tenant's thermal comfort level (Pedro, Kalyan, Pedro, & Una-May, 2014).

A number of comparisons have been made between RL and other control strategies. Li and Xia applied RL to optimally control an HVAC system with respect to minimizing energy consumption while maintaining thermal comfort inside an acceptable range (Li & Xia, 2015). The objective of the study was biased more towards energy. Compared to a blank control with a constant temperature schedule they found that the RL agent achieved better performance in operating the system with regards to energy conservation. In a similar study, Barret and Linder produced an RL control for an HVAC system and, through a comparative analysis, found their RL control outperformed, in terms of energy cost, two common strategies for controlling HVAC systems, namely, the "Always On" and "Programmable Control" methods. Improved thermal comfort was also demonstrated by offline training. Yang et al compared the performance of RL controls (RLCs) with typical RBCs using numerical simulations on the thermal model of a LowEx full-building

system (Yang, Nagy, Goffin, & Schlueter, 2015). They found that the RLCs outperformed the RBCs in all of the given thermal objectives. Schmidt et al, and Wang, Velswamy, & Huang showed that RL control achieved better thermal comfort results with less energy consumed as compared with traditional methods (Schmidt et al., 2017; Wang, Velswamy, & Huang, 2017). Sun, Somani, & Carroll, in a comparison of a model-free method against model-based bidding strategies, showed that RL gave similar performance to the model-based bidding strategy of smart thermostats (Sun, Somani, & Carroll, 2015). Urieli and Stone applied RL to control an HVAC system in a simulated residential home, with the aim of minimising energy consumption while maintaining an acceptable temperature range (Urieli & Stone, 2013). Compared with a default strategy of thermostat setback, their RL agent learned an effective setback strategy that both reduced energy costs and minimised violations of the temperature constraints. Nagy et al applied a novel RL algorithm to the problem of controlling a heat pump in a building (Nagy, Kazmi, Cheaib, Driesen, & Leuven, 2018). The objective of the controller was to maintain occupant comfort while reducing energy cost. A higher priority was given to occupant comfort. They compared their novel RL method to an RBC, an MPC with perfect information, and a model-based RL method. It was concluded that the model-based controls outperformed the model-free RL method; however, regarding computational complexity their novel RL method was much better. It also outperformed its model-based counterpart regarding changes to environmental dynamics. Furthermore, compared to the RBC, energy and cost savings were observed.

Zhang et al developed a framework that uses a whole building energy model for optimal control of an HVAC system by deep RL (DRL) (Zhang et al., 2018). They found that the DRL control consumed 15% less heating energy with similar thermal comfort as compared to a base case. In a follow up study the same DRL control was deployed in a real-life office building over a period of 3 months (Zhang & Lam, 2018). They included a thermal preference feedback app so that each occupant could state his/her thermal comfort preference. In conclusion, they found that their DRL control saved 16.6% heating energy consumption in the real setting. The thermal preference feedback system, however, had a very low participation rate and thus was not an accurate representation of the thermal comfort level. Wei, Wang, & Zhu implemented DRL for maintaining zonal room temperature within a desired range, while minimising energy cost (Wei, Wang, & Zhu, 2017). Compared with a conventional RL technique (Q-learning) and a rule-based strategy, DRL was found to be the most effective in both keeping the number of temperature violations to a minimum and reducing energy cost.

4.1.2. IAQ

Optimising just IAQ has had the least attention. Indeed, we have identified only one article, namely, Baghaee and Ulusoy's use of RL for operating an HVAC system (Baghaee & Ulusoy, 2018). Here, the objective of the control was to maintain CO_2 concentration at an acceptable range while also minimising energy consumption. In a simulation study they compared their RL control to an on/off and set point control. The RL method was found to be superior to the other two controls regarding energy consumption and CO_2 concentration.

4.1.3. Lighting

The earliest RL control on lighting was established about a decade ago where hierarchical RL (HRL) was developed to enable fast convergence and practcal application (Khalili, Wu, & Aghajan, 2010). The goal was to optimise over both lighting and energy cost. Compared to regular RL, they found HRL to be much faster in adapting to ideal light settings. After a hiatus of six years, lighting control was reconsidered in the work by, Cheng et al, who applied RL with a human feedback mechanism in order to control the blinds and lights in a single-occupant office of a building (Cheng et al., 2016). They jointly maximised over lighting comfort and energy conservation, with a heavier emphasis on comfort. With respect to the test environment, they found their results suggested an improvement in luminosity from both a comfort perspective and an energy saving perspective. With regard to the latter, compared to a manual and a traditional integrated automated control, their RL method reduced energy consumption.

In a more recent article, and the last among those just considering lighting, Park et al developed an RL control, LightLearn, with an occupant feedback mechanism (Park, Dougherty, Fritz, & Nagy, 2019). The aim of the control was to balance energy consumption and occupant comfort. They tested their control framework in five single-occupant offices in a building over a period of five weeks. In comparison with schedule-based and occupancy-based control strategies, LightLearn saved on energy consumption. Furthermore, LightLearn achieved good performance regarding occupant comfort. In conclusion, their RL control outperformed the other two controls in achieving a good balance between energy saving and occupant comfort.

4.1.4. Combinations of factors

As can be seen, studies controlling combined factors are rare in the 1990s. For example, the earliest work in this area was by Jouffe who used RL to tune a ventilation controller for controlling temperature and relative humidity (Jouffe, 1997). The policy obtained from the control was exactly to the experts' specifications. In the seminal work by Mozer, RL was used to control an HVAC and water heating system (Mozer, 1998). The aim of the control framework was to minimise both discomfort (heating and lighting) and energy cost. The RL control was found to outperform alternative control strategies.

There have been more studies in the 21 st century. Dalamagkidis applied RL to control an HVAC system (Dalamagkidis, Kolokotsa, Kalaitzakis, & Stavrakakis, 2007). The objective of the RL control was to minimise energy consumption and maximise user comfort where comfort was made up of two components, namely, thermal comfort and IAQ. In a computer experiment consisting of a simulated period of 4 years, they compared the performance of the RL control with an On/Off control and a Fuzzy-PD control. In terms of occupant comfort, the RL control was superior. This however came at the expense of higher energy consumption. Overall, they found the RL control to have achieved a level of performance close to that of the other control strategies. As in their later work (Dalamagkidis & Kolokots, 2008) they raised the issue of exploration during deployment in a real building setting. Their advice on this was to exhaustively train the controller beforehand, allowing little to no exploration during deployment.

Sun et al implemented a novel RL technique based on events – where an event was defined as a set of states - as opposed to time, to control an HVAC system (Sun, Luh, Jia, & Yan, 2013). The objective of the control problem was to minimise energy cost while satisfying thermal comfort and IAQ constraints. In a numerical simulation, they compared their RL method to two other optimisation techniques, namely, backward stochastic dynamic programming (SDP) and a greedy method. They found the RL approach to be much more efficient in solving the given problem while satisfying the comfort constraints and saving energy costs. Later, Sun et al refined their novel RL method and applied it to the aforementioned problem (Sun, Luh, Jia, & Yan, 2015). In a more detailed numerical simulation, they compared their RL technique to the same algorithms as before, namely, a greedy algorithm and an SDP algorithm. The same conclusions as in their former paper were reached, namely, they found their novel RL approach to be more efficient at solving the given problem as compared to the other algorithms. Similar topics can also be found in recent years (Eller, Siafara, & Sauter, 2018; Fu et al., 2018), especially in MAS (Hurtado, Mocanu, Nguyen, Gibescu, & Kamphuis, 2018).

Apart from IAQ, there have also been articles studying a combination of thermal comfort along with the other factors, namely, lighting and humidity. Bielskis et al applied RL to control an HVAC system and a Red-Green-Blue-Yellow LED lighting system (Bielskis, Guseinoviene, Drungilas, Gricius, & Zulkas, 2012). The goal of the RL controller was to

maximise occupant comfort conditions, which consisted of thermal and lighting conditions. They tested their RL controller in a controlled laboratory experiment and found the performance acceptable. Chen et al applied RL to the control problem of natural ventilation (Chen, Norford, Samuelson, & Malkawi, 2018). In particular, they used RL to control HVAC and window systems. The aim of the control was to minimise both energy consumption and occupant discomfort. The comfort component consisted of two parts, namely, a thermal part and humidity (relative) part. In a numerical simulation over a simulated period of one year, they compared the RL control with a rule-based heuristic control under two different climates. In the two case studies, the RL control exhibited superior performance compared with the other control resulting in less energy consumption and occupant discomfort.

4.2. Algorithm class

As already alluded to in Section 3 we can classify our learning algorithms according to whether they are value-based, policy-based, or exhibit a combination of these two classes, known in the literature as Actor-Critic algorithms. We can see from Fig.6 that value-based algorithms seem to dominate the building literature followed by Actor-Critic algorithms. Furthermore, there does not (at the time of writing) seem to be any applications of policy-based algorithms. A couple of reasons for this may have something to do with the fact that less is known about this class of algorithms as well as the fact that policy evaluation is known to be less efficient with high variance (Silver, 2015; Sutton & Barto, 2018). In Fig. 6, we include the methods that are missed in Sutton and Barto's introduction (Sutton & Barto, 2018) into the group of "Other" for further discussion: RLS-TD(λ), eJAL, MEC, D-DNFQI, MAXQ, simulation based tree search and a variant of DQN.

4.2.1. Value-based

Fig. 6 clearly shows that the most popular value-based algorithm applied to occupant comfort optimisation is Q-learning. This is a temporal-difference (TD) learning method whereby two policies are used, one for generating the behavior and the other being the learned policy, which eventually becomes the optimal policy. We call this process off-policy learning, and it is one of two classes, the other being on-policy learning, used to address the exploration vs exploitation dilemma (Sutton & Barto, 2018). The simplicity of Q-learning leads to a majority of studies in comfort control (Baghaee & Ulusoy, 2018; Barrett & Linder, 2015; Bonte, Perles, Lartigue, & Thellier, 2014; Chen, Norford, Samuelson, & Malkawi, 2018; Cheng et al., 2016; Eller, Siafara, & Sauter, 2018; Khalili, Wu, & Aghajan, 2010; B. Li & Xia, 2015; D. Li, Zhao, Zhu, & Xia, 2015; Lu, Wang, Lin, & Hameen, 2019; Pedro, Kalyan, Pedro, & Una-May, 2014; Sato, Samejima, Akiyoshi, & Komoda, 2012; Sun, Luh, Jia, & Yan, 2013, 2015; Yang, Nagy, Goffin, & Schlueter, 2015). Specifically, two consecutive studies applied Q-learning within a Lagrangian relaxation framework (Sun et al., 2013; Sun, Luh et al., 2015). Moreover, instead of the usual time-based approach to optimisation, they took a novel event-based approach in which an action follows an event (a set of state transitions). Fazenda and Lima implemented Q-learning in two contexts, one in which action and states were discrete and thus tabular Q-learning was used; another in which the action and states were continuous. Here they combined Q-learning with a wire fitted neural network in order to approximate the Q-values (Pedro et al., 2014). Similarly, Yang et al (Yang et al., 2015) implemented tabular Q-learning and batch Q-learning with memory replay, where in the latter various neural network architectures were tested with the best performing being a 5-4 structure. Li and Xia implemented a novel multi-grid Q-learning algorithm. They compared its performance to standard Q-learning and found it to be more efficient (Li & Xia, 2015). Yu and Dexter implemented a simplified version of Q(λ) (Yu & Dexter, 2010), which extends Q-learning by including eligibility traces: these are an efficient technique that allows information to be propagated backwards over multiple time steps (Silver, 2015; Sutton & Barto, 2018). By applying a fuzzy discretization to the state space they were able to implement Q(λ) in tabular form, reducing the learning time. Furthermore, by incorporating a pre-trained policy into their implementation they were able to reduce the learning time even further, which they concluded to be essential for an RL control to be able to improve the operation of a building system online in an acceptable time frame.

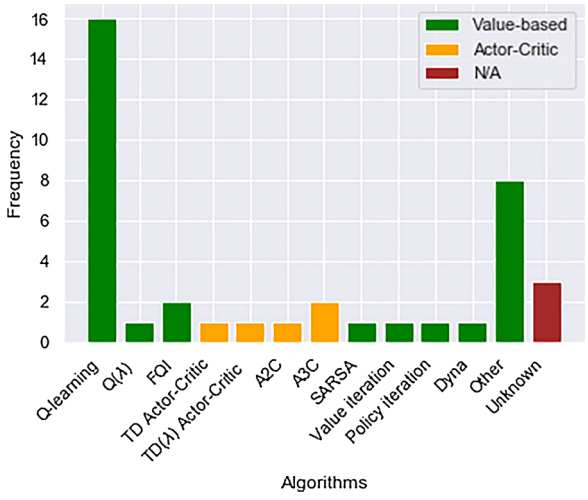

Fig. 6. Distribution of algorithms used for indoor comfort control.

Hurtado et al implemented an extended joint action learning algorithm (eJAL) (Hurtado et al., 2018). The JAL method is an extension of Q-learning to a cooperative multi-agent setting whereby the actions of the other competing agents are included in the optimal policies. Hurtado et al extended this algorithm in two ways, first by defining a conditional action space over all the other agents, and secondly by attaching a preference to each agent in a full cooperative game thereby allowing a joint reward function to be constructed. They also implemented the Q-learning algorithm in a decentralized non-cooperative setting and compared this to their novel eJAL algorithm and a Nash *n*-player game (a game theoretic non-cooperative approach). Overall eJAL achieved the highest average fairness index. Furthermore, they found the Q-learning implementation to be a good non-cooperative method.

As stated in Section 3, approximation methods do not create tables for state-actions. In batch learning the whole dataset is available for learning, which is done in an offline fashion (Sutton & Barto, 2018). For example, Ruelens et al and Schmidt et al implemented the sample efficient batch RL technique, fitted Q-iteration (FQI) (Ruelens, Iacovella, Claessens, & Belmans, 2015; Schmidt et al., 2017). In the paper by Ruelens et al, they used past observations at the end of each episode (a day in their case) to learn a policy that was then used online during the next episode with a Boltzmann exploration strategy. To reduce the state feature space they used an auto-encoder network which mapped features to a smaller dimensional subspace of the original feature space. In combination with this feature extraction technique, they used extremely randomized trees to approximate the Q-values. A novel variant of FQI, double deep neural FQI (D-DNFQI), uses deep neural networks for approximating the Q-values (Nagy et al., 2018). Here, Nagy et al used target Q-networks and prioritised experience replay to address the problems of convergence and learning instability and, furthermore, used double Q-learning to address the upward bias problem of Q-estimates. This novel algorithm exhibited faster compute times with the added benefit of robustness under changes to system dynamics. The on-policy batch learning recursive least-squares TD(λ) (RLS-TD(λ)) is used in adaptive filtering, system identification and adaptive control. Its popularity lies in its fast convergence speed. Along with RLS-TD(λ), radial basis functions (RBFs) can be used to construct the feature vector

used for the approximation of Q-values. (Dalamagkidis & Kolokots, 2008; Dalamagkidis et al., 2007). Similarly, Fu et al used the SARSA algorithm, another on-policy TD learning technique (Fu et al., 2018).

4.2.2. Actor-Critic

Actor-only methods work with parameterized policies and critic-only methods rely exclusively on value function approximation and aim at learning an approximate solution to Eq. (11). Actor-Critic methods aim at combining the strong points of these techniques. (Konda & Tsitsiklis, 2000).

A TD(λ) Actor-Critic technique is used to reduce the variance of the policy gradient method by introducing a baseline function, called the advantage function (Silver, 2015; Sutton & Barto, 2018). Bielskis et al used the backward view of TD(λ), which uses independent eligibility traces for the actor and critic (Bielskis et al., 2013). To compute the action preferences and the state-values, they used a radial basis neural network. In a much earlier work, Jouffe also implemented the backward view of TD(λ) Actor-Critic but in a fuzzy logic network, which he called Fuzzy Actor-Critic Learning (Jouffe, 1997). He defined fuzzy reinforcement functions to smooth the distinction between goal and failure states with the aim to speed up learning. To further increase the speed of learning, a priori knowledge was used.

Wang et al implemented advantage actor-critic (A2C) (Wang, Velswamy et al., 2017). This uses the same advantage function as described before but instead of using TD(λ) for different time scales it uses Monte Carlo (Silver, 2015). They combined this with two long-short-term-memory networks (LSTM), one for the actor and the other for the critic. These were used in order to mitigate against noise and the partial observability of states. Differing from A2C, asynchronous advantage actor-critic (A3C) (Mnih et al., 2016) executes multiple agents asynchronously on multiple instances of the same environment, independent of each other (Zhang & Lam, 2018; Zhang et al., 2018). This approach is attractive from a practical point of view since state of the art results can be achieved on, for example, a standard multi-core CPU as opposed to a GPU, which is the norm for deep learning (Mnih et al., 2016).

4.3. Exploration vs exploitation strategies

From a value-based perspective, the dilemma of exploration vs exploitation has to do with, on the one hand, exploiting current information by choosing the best (greedy) action-value, and on the other hand, exploring non-greedy actions (Sutton & Barto, 2018). By exploring, the agent is able to improve on current estimates of non-greedy action-values and thus, in the long run, can lead to greater returns. Most of the works reviewed applied the naïve ε-greedy approach to addressing this problem but several have taken a more sophisticated approach (Khalili, Wu, & Aghajan, 2010; B. Li & Xia, 2015; D. Li, Zhao, Zhu, & Xia, 2015; Ruelens, Iacovella, Claessens, & Belmans, 2015) with good results. However, none of these works systematically addressed this issue, that is, a parameter study of the given exploration strategies was not investigated. Given the sensitive nature of this topic on optimal control of buildings' energy conservation and, in particular, occupant comfort, further studies should be carried out on this topic.

Policy-gradient methods, on the other hand, approach this dilemma from another perspective. By parameterizing the policy they can learn suitable levels of exploration and asymptotically approach deterministic policies, which for action-value methods is extremely difficult to achieve and would involve placing strong assumptions on the problem (Sutton & Barto, 2018). We have seen examples of policy gradient methods in (Bielskis et al., 2013; Jouffe, 1997; Wang, Velswamy et al., 2017; Zhang & Lam, 2018; Zhang et al., 2018) with good results.

4.4. Agent perspectives

Multiple agents can be treated both cooperatively and non-cooperatively. For non-cooperative agents, RL collapses to independent learnings in multiple zones or scenarios (Park et al., 2019; Wei et al., 2017; Yang et al., 2015). Single-agent algorithms may apply but the performances have seldom been compared with cooperative agents. Nevertheless, Eq. (14) depicts a joint action probability distribution for a Markov game – a combination of MDPs and matrix games. The Nash equilibrium is achieved via communication and interaction among agents. The reward function for multiple agents has become $r_i(s^{'}s, \boldsymbol{a})$, which denotes the reward for agent i in state s' given state s and joint action a (Schwartz, 2014). The corresponding joint action learning method was also extended (Hurtado et al., 2018). Even though an extensive survey of MARL has been developed for many years (Buşoniu et al., 2010), there remains a great number of examples for occupant comfort control that still need to be made.

4.5. Physical implementations

Only six out of the thirty-three articles reviewed involved a case study in an actual real setting rather than a simulated environment. For practitioners we believe it will be useful to have an indication of the state of the art of RL control implementation, given the relatively new adoption of this burgeoning field in building control methods, and the fact that conventional control techniques such as On/Off and PID strategies are still preferred by industry-based experts (Royapoor et al., 2018).

The earliest example of RL being applied to building control was Mozer's Neural Network House (Mozer, 1998). In his seminal work, Mozer implemented a lighting control in a former school residence. It exhibited good performance of occupant comfort and conservation of energy. Much later, Bielskis et al (Bielskis et al., 2013) implemented their ACAR-Controller for HVAC and LED lighting in a laboratory setting. The performance was deemed acceptable with regard to thermal and lighting comfort constraints. Cheng et al applied RL with a human feedback mechanism in order to control the blinds and lights in a single-occupant office of a university building in Beijing (Cheng et al., 2016). It achieved good results, both from an occupant comfort point of view and an energy saving point of view. They pointed out that to achieve a more accurate comfort model, the agent would require more exploration which they saw as infeasible in a working office. Schmidt et al deployed an RL controller on a Spanish school's heating system over two zones, in a building with a low level of thermal insulation (Schmidt et al., 2017). It was successful in improving thermal comfort with reduced energy consumption. Park et al implemented an occupant centric lighting control with an occupant feedback mechanism at the University of Texas (Park et al., 2019). They selected five single-occupant rooms, each with a manually adjustable blind. The control strategy reduced energy consumption with good performance on occupant comfort. They pointed out that further research should be conducted to investigate the adaptability of the control to multi-occupant rooms, as well as the effect of control hyperparameters such as the discount factor and the reward structure. Zhang and Lam deployed an RL control for an HVAC system in the Intelligent Workplace (IW) at the University of Pittsburgh (Zhang & Lam, 2018). The case study consisted of a multi-occupant single office. They included a thermal preference feedback phone app so that each occupant could state his/her thermal comfort preference. The control saved 16.6% heating energy consumption. However, the feedback system had a very low participation rate and thus was not an accurate representation of the thermal comfort level. They suggested investigating the effects of hyperparameter tuning on convergence of DRL since the thermal inertia of IW caused convergence problems in the DRL training. They also pointed out that the inertia of the heating system may have discouraged the occupants from using the app, which in turn may have affected the occupants' psychological feeling of comfort.

5. Discussion

In this section, we discuss some challenges that may be encountered by building designers and managers. Although our work is not an exhaustive review in building environment studies, we, nevertheless, convey ideas about how we might orient ourselves to face these challenges in the future building comfort control and management.

There have only been 33 works up to the current time for which RL has been explored for controlling occupant comfort, much less than those for building energy control. In particular, the studies including comfort factors such as indoor air quality and lighting are relatively rare in comparison to thermal comfort. Furthermore, relatively few of the reviewed works incorporate occupancy patterns and/or occupant feedback into the control loop which are crucial for occupant-centric building operation. Moreover, there is a gap with regards to the performance of implementing cooperative MARL.

The majority of our reviewed articles are found after 2010. This is not to say that people have not realised RL before 2010. Rather, it was the learning efficiency (e.g. the curse of dimensionality) that hindered people from implementing desirable experiments. Even though proper discretization of a continuous state space can make the exploration space more tractable, the scale of the problem may still be huge for high dimensional states. Even a long-term training period does not guarantee an optimal policy. However, the breakthroughs in computing power that have occurred since the 2010s have reversed the situation. The changeover to new hardware and computation platforms affords the chances to conduct complex computations, which in turn facilitates function approximation in RL and thus the ability to generate to unseen states and hence a more accurate representation of reality. In the process of designing and constructing smart buildings, considerations of incorporating real-time big data computing and learning platforms to BMSs have become one of the forthcoming challenges. This requires not only that the agent can gather real-time data, but also that building automation systems react in a timely fashion. The integration of computation infrastructures with BMSs is not explicitly formulated in most of our reviewed articles and thus requires more practical studies on this issue.

Further challenges originate from the MARL paradigm (Buşoniu et al., 2010; Ye, Zhang, & Vasilakos, 2017). Firstly, the agents might work independently instead of cooperatively resulting in sub-optimal returns. In an ideal framework, the cooperative agents can communicate immediate actions, rewards or learning experiences. Any interruption of the communication among agents makes the overall goal difficult to achieve. Secondly, the decision-making process for a single agent in MARL is non-stationary due to the dynamic policy changes from other agents. In systems with incomplete information share, the time may affect individual decisions. Thirdly, the exploration strategy becomes complex when the number of agents increases.

In some of our reviewed articles, the learning strategy for an MAS is still limited by applying single-agent RL algorithms to the multi-agent case. Learning depends only on the current agent's action and without being aware of the other agents. This strategy is simple to implement (Mataric, 1994; Sen, Sekaran, & Hale, 1994), but reaching the Nash equilibrium in Eq. (15) is challenging.

Defining the reward function seems to be an additional challenge. Apart from energy consumption, people with different personalities prioritise different comfort factors (Frontczak & Wargocki, 2011; Zalejska-Jonsson & Wilhelmsson, 2013). The magnitude for each factor is also subjective. There are a number of studies where researchers tried to optimise other objectives by setting comfort factors as the constraints. RL implementation in this circumstance is known as a constrained MDP, which is usually solved by linear programming. The dynamic programming techniques that apply in the non-constrained control problem do not hold any more and optimal policies need not exist (Altnan, 1999). Thus, defining the reward function is not an easy task.

Successfully incorporating occupancy schedules and human feedback into the control loop is another challenge and which has drawn only a handful of studies (Cheng et al., 2016; Park et al., 2019; Pedro et al., 2014; Sato et al., 2012; Zhang et al., 2018). As occupants have a strong effect on energy consumption in buildings (D'Oca, Hong, & Langevin, 2018; Park & Nagy, 2018; Park et al., 2019; Yan et al., 2017) including occupancy patterns as well as occupant feedback into the control system is crucial for efficient and occupant-centred building operation. Including human feedback can be particularly troublesome since we must be economical in our requirement of occupant feedback for it not to be too intrusive and time costly (Christiano et al., 2017; Pedro et al., 2014; Zhang et al., 2018).

Another challenge relates to the development of a toolkit, such as OpenAI Gym (Brockman et al., 2016), providing a framework whereby building simulation and advanced RL control using one's favorite machine learning library, such as, TensorFlow, PyTorch, etc can be seamlessly tried and tested in a controlled and reproducible manner. A move in this direction can be seen in the work by Vázquez-Canteli et al (Vázquez-Canteli et al., 2019).

6. Conclusions

The indoor environment affects not only working efficiency and living standards but also influences the occupants' health. Apart from building design, efficient control methods for the indoor environment not only improve the occupants comfort, but can also mitigate CO_2 emissions (Vázquez-Canteli et al., 2019). This paper briefly examines and analyses empirical articles regarding the reinforcement learning control method for occupant comfort in buildings. Based on our analysis, we conclude our findings and formulate future works.

Firstly, the cutting-edge RL technique, which not only can adapt to the dynamic indoor environment of a building, but can also simultaneously adapt to the buildings' occupants, has drawn only limited attention regarding indoor climate oriented smart building controls, even though some studies have empirically tested its feasibility and comparability to other methods. The promising results lead us to a new frontier of occupant-centric building control. We have identified thirty-three empirical articles in this field, which is much less than the studies in building energy control and needs to be extended. The value-based Q-learning method is easy and straightforward to implement and it dominates among learning algorithms. This leaves a question of how policy-based or Actor-Critic algorithms perform in a practical building environment. Secondly, the computation platform and the ways of interaction with the BMS are important for conducting real-time control. Especially in the works with physical tests, the working paradigms are still vague. For example, policy-based and Actor-Critic algorithms require more function approximations and thus the power of computing resources should be updated accordingly. Thirdly, while maintaining proper indoor temperature is the foremost objective to be considered in BMSs, automation of smart buildings is an integrated system and studies about comfort factors like IAQ and lighting are relatively rare in comparison. Forthly, the empirical study of MARL for controlling the indoor environment has been modest. For example, the performance of implementing cooperative MARL still needs to be examined and confirmed through large studies. Fifthly, as occupants have a significant impact on energy consumption of buildings, it is important to include the occupant dimension into the control system. However, only five of the reviewed works considered occupancy patterns and/or human feedback in the control loop. Finally, model-based RL has had little attention in this area, appearing only twice among the reviewed works. The Dyna architecture and simulation-based search should be further explored in building applications as these have already been shown to be effective in other AI applications.

Looking ahead to the future, it is valuable to identify some future works. These include engaging in designing and training non-linear approximation approaches such as deep learning for RL; exploring

algorithms for cooperative agents in MARL; setting up pre-training paradigms; stressing strategies for the exploration-exploitation dilemma; carrying out more studies for including occupancy patterns and/or human feedback into the control loop; exploring model-based RL; finally, further efforts in the direction of creating a seamless framework combining building simulation, advanced RL and other control strategies which can be compared in a standard and reproducible way. We also anticipate some promising practical works in standardising the measurement of indoor comfort, and integrating computation platforms and the ways of interacting with the BMS into smart building systems.

Acknowledgments

The authors are thankful for the financial support from the Micro-data analysis research profile at Dalarna University and the UBMEM project from the Swedish Energy Agency (Grant no. 46068).

References

Altnan, E. (1999). *Constrained markov decision processes.* Chapman & Hall/CRC.

ASHRAE Standard 55 (2017). *Thermal environmental conditions for human occupancy.* ASHRAE Inc.

Baghaee, S., & Ulusoy, I. (2018). User comfort and energy efficiency in HVAC systems by Q-learning. *2018 26th Signal Processing and Communications Applications Conference (SIU),* 1–4. https://doi.org/10.1109/SIU.2018.8404287.

Barrett, E., & Linder, S. (2015). Autonomous HVAC control, a reinforcement learning approach. In A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, ... M. Spiliopoulou (Eds.). *Machine learning and knowledge discovery in databases* (pp. 3–19). Springer International Publishing.

Bellman, R. (1957a). A Markovian decision process. *Indiana University Mathematics Journal, 6*(4), 679–684. https://doi.org/10.1512/iumj.1957.6.56038.

Bellman, R. (1957b). *Dynamic programming.* Princeton, NJ: Princeton Univ. Pr.

Bielskis, A. A., Guseinoviene, E., Drungilas, D., Gricius, G., & Zulkas, E. (2013). Modelling of ambient comfort affect reward based adaptive laboratory climate controller. *Elektronika Ir Elektrotechnika, 19*(8), 79–82. https://doi.org/10.5755/j01.eee.19.8.5399.

Bielskis, A. A., Guseinoviene, E., Dzemydiene, D., Drungilas, D., & Gricius, G. (2012). Ambient lighting controller based on reinforcement learning components of multi-agents. *Electronics and Electrical Engineering, 5*(121), 79–84.

Bonte, M., Perles, A., Lartigue, B., & Thellier, F. (2014). An occupant behaviour model based on artificial intelligence for energy bulding simulation. *Proceedings of the 13th International IBPSA Conference*.

Boodi, A., Beddiar, K., Benamour, M., Amirat, Y., & Benbouzid, M. (2018). Intelligent systems for building energy and occupant comfort optimization: A state of the art review and recommendations. *Energies, 11*(10), 2604. https://doi.org/10.3390/en11102604.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., ... Zaremba, W. (2016). *OpenAI Gym. ArXiv:1606.01540 [Cs].* Retrieved fromhttp://arxiv.org/abs/1606.01540.

Buşoniu, L., Babuška, R., & De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. In D. Srinivasan, & L. C. Jain (Vol. Eds.), *Innovations in multi-agent systems and applications - 1: Vol. 310*, (pp. 183–221). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14435-6_7.

CEN prEN15251 (2005). *Criteria for the Indoor Environment including thermal, indoor air quality, light and noise.*

Chen, Y., Norford, L. K., Samuelson, H. W., & Malkawi, A. (2018). Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings, 169*, 195–205. https://doi.org/10.1016/j.enbuild.2018.03.051.

Chenari, B., Dias Carrilho, J., & Gameiro da Silva, M. (2016). Towards sustainable, energy-efficient and healthy ventilation strategies in buildings: A review. *Renewable and Sustainable Energy Reviews, 59*, 1426–1447. https://doi.org/10.1016/j.rser.2016.01.074.

Cheng, Z., Zhao, Q., Wang, F., Jiang, Y., Xia, L., & Ding, J. (2016). Satisfaction based Q-learning for integrated lighting and blind control. *Energy and Buildings, 127*, 43–55. https://doi.org/10.1016/j.enbuild.2016.05.067.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences. ArXiv:1706.03741 [Cs, Stat].* Retrieved fromhttp://arxiv.org/abs/1706.03741.

Dalamagkidis, K., & Kolokots, D. (2008). Reinforcement learning for building environmental control. In C. Weber, M. Elshaw, & N. Michael (Eds.). *Reinforcement learning*I-Tech Education and Publishinghttps://doi.org/10.5772/5286.

Dalamagkidis, K., Kolokotsa, D., Kalaitzakis, K., & Stavrakakis, G. S. (2007). Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment, 42*(7), 2686–2698. https://doi.org/10.1016/j.buildenv.2006.07.010.

D'Oca, S., Hong, T., & Langevin, J. (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews, 81*, 731–742. https://doi.org/10.1016/j.rser.2017.08.019.

Dounis, A. I., & Caraiscos, C. (2009). Advanced control systems engineering for energy and comfort management in a building environment—A review. *Renewable and Sustainable Energy Reviews, 13*(6–7), 1246–1261. https://doi.org/10.1016/j.rser.2008.09.015.

Eller, L., Siafara, L. C., & Sauter, T. (2018). Adaptive control for building energy management using reinforcement learning. *2018 IEEE International Conference on Industrial Technology (ICIT),* 1562–1567. https://doi.org/10.1109/ICIT.2018.8352414.

Enescu, D. (2017). A review of thermal comfort models and indicators for indoor environments. *Renewable and Sustainable Energy Reviews, 79*, 1353–1379. https://doi.org/10.1016/j.rser.2017.05.175.

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research, 6*, 503–556.

Frontczak, M., & Wargocki, P. (2011). Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment, 46*(4), 922–937. https://doi.org/10.1016/j.buildenv.2010.10.021.

Fu, Q., Hu, L., Wu, H., Hu, F., Hu, W., & Chen, J. (2018). A Sarsa-based adaptive controller for building energy conservation. *Journal of Computational Methods in Sciences and Engineering, 18*(2), 329–338. https://doi.org/10.3233/JCM-180792.

Galasiu, A. D., & Veitch, J. A. (2006). Occupant preferences and satisfaction with the luminous environment and control systems in daylit offices: A literature review. *Energy and Buildings, 38*(7), 728–742. https://doi.org/10.1016/j.enbuild.2006.03.001.

Gambier, A. (2004). Real-time control systems: A tutorial. *Presented at the 5th Asian Control Conference (IEEE Cat. No. 04EX904),* 1024–1031.

Grondman, I., Busoniu, L., Lopes, G. A., & Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems Man and Cybernetics Part C, 42*(6), 1291–1307.

Gu, S., Lillicrap, T., Sutskever, I., & Levine, S. (2016). Continuous deep Q-learning with model-based acceleration. *Presented at the Conference on Machine Learning, Vol. 48*.

Guo, X., Tiller, D., Henze, G., & Waters, C. (2010). The performance of occupancy-based lighting control systems: A review. *Lighting Research and Technology, 42*(4), 415–431. https://doi.org/10.1177/1477153510376225.

Guyot, G., Sherman, M. H., & Walker, I. S. (2018). Smart ventilation energy and indoor air quality performance in residential buildings: A review. *Energy and Buildings, 165*, 416–430. https://doi.org/10.1016/j.enbuild.2017.12.051.

Haq, M. A., Hassan, M. Y., Abdullah, H., Rahman, H. A., Abdullah, M. P., Hussin, F., ... Said, D. M. (2014). A review on lighting control technologies in commercial buildings, their performance and affecting factors. *Renewable and Sustainable Energy Reviews, 33*, 268–279. https://doi.org/10.1016/j.rser.2014.01.090.

Hurtado, L. A., Mocanu, E., Nguyen, P. H., Gibescu, M., & Kamphuis, R. I. G. (2018). Enabling cooperative behavior for building demand response based on extended joint action learning. *IEEE Transactions on Industrial Informatics, 14*(1), 127–136. https://doi.org/10.1109/TII.2017.2753408.

Jouffe, L. (1997). Ventilation control learning with FACL. *Proceedings of 6th International Fuzzy Systems Conference, Vol. 3*, 1719–1724. https://doi.org/10.1109/FUZZY.1997.619799.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *The Journal of Artificial Intelligence Research, 4*, 237–285.

Khalili, A. H., Wu, C., & Aghajan, H. (2010). Hierarchical preference learning for light control from user feedback. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops,* 56–62. https://doi.org/10.1109/CVPRW.2010.5543265.

Klein, L., Kwak, J., Kavulya, G., Jazizadeh, F., Becerik-Gerber, B., Varakantham, P., ... Tambe, M. (2012). Coordinating occupant behavior for building energy and comfort management using multi-agent systems. *Automation in Construction, 22*, 525–536. https://doi.org/10.1016/j.autcon.2011.11.012.

Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. *Presented at the Advances in Neural Information Processing Systems, Vol. 12*, 1008–1014.

Kruisselbrink, T., Dangol, R., & Rosemann, A. (2018). Photometric measurements of lighting quality: An overview. *Building and Environment, 138*, 42–52. https://doi.org/10.1016/j.buildenv.2018.04.028.

Li, B., & Xia, L. (2015). A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings. *2015 IEEE International Conference on Automation Science and Engineering (CASE),* 444–449. https://doi.org/10.1109/CoASE.2015.7294119.

Li, D., Zhao, D., Zhu, Y., & Xia, Z. (2015). Thermal comfort control based on MEC algorithm for HVAC systems. *2015 International Joint Conference on Neural Networks (IJCNN),* 1–6. https://doi.org/10.1109/IJCNN.2015.7280436.

Li, N., Cui, H., zhu, C., Zhang, X., & Su, L. (2016). Grey preference analysis of indoor environmental factors using sub-indexes based on Weber/Fechner's law and predicted mean vote. *Indoor and Built Environment, 25*(8), 1197–1208.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., ... Wierstra, D. (2016). *Continuous control with deep reinforcement learning. ArXiv:1509.02971 [Cs, Stat].* Retrieved fromhttp://arxiv.org/abs/1509.02971.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *Presented at the Conference on Machine Learning,* 157–163.

Lu, S., Wang, W., Lin, C., & Hameen, E. (2019). Data-driven simulation of a thermal comfort-based temperature set-point control with ASHRAE RP884. *Building and Environment*.

Marinakis, V., Karakosta, C., Doukas, H., Androulaki, S., & Psarras, J. (2013). A building automation and control tool for remote and real time monitoring of energy consumption. *Sustainable Cities and Society, 6*, 11–15. https://doi.org/10.1016/j.scs.2012.06.003.

Mataric, M. J. (1994). Reward functions for accelerated learning. *Presented at the Proceedings 11th International Conference on Machine Learning (ICML-94),* 181–189.

Merabti, S., Draoui, B., & Bounaama, F. (2016). A review of control systems for energy and comfort management in buildings. *2016 8th International Conference on Modelling,*

Identification and Control (ICMIC), 478–486. https://doi.org/10.1109/ICMIC.2016.7804161.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., ... Kavukcuoglu, K. (2016). *Asynchronous methods for deep reinforcement learning. ArXiv:1602.01783 [Cs].* Retrieved fromhttp://arxiv.org/abs/1602.01783.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236.

Mozer, M. C. (1998). *The neural network house: An environmenthat adaptsto its inhabitants.* 5.

Nagy, A., Kazmi, H., Cheaib, F., Driesen, J., & Leuven, K. (2018). *Deep reinforcement learning for optimal control of space heating. ArXiv:1805.03777 [Stat.AP].*

Park, J. Y., Dougherty, T., Fritz, H., & Nagy, Z. (2019). LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. *Building and Environment, 147*, 397–414. https://doi.org/10.1016/j.buildenv.2018.10.028.

Park, J. Y., & Nagy, Z. (2018). Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review. *Renewable and Sustainable Energy Reviews, 82*, 2664–2679. https://doi.org/10.1016/j.rser.2017.09.102.

Pedro, F., Kalyan, V., Pedro, L., & Una-May, O. (2014). Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems. *Journal of Ambient Intelligence and Smart Environments,* (6), 675–690. https://doi.org/10.3233/AIS-140288.

Roetzel, A., Tsangrassoulis, A., Dietrich, U., & Busching, S. (2010). A review of occupant control on natural ventilation. *Renewable and Sustainable Energy Reviews, 14*(3), 1001–1013. https://doi.org/10.1016/j.rser.2009.11.005.

Royapoor, M., Antony, A., & Roskilly, T. (2018). A review of building climate and plant controls, and a survey of industry perspectives. *Energy and Buildings, 158*, 453–465. https://doi.org/10.1016/j.enbuild.2017.10.022.

Ruelens, F., Iacovella, S., Claessens, B. J., & Belmans, R. (2015). Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. *Energies, 8*(8), 8300–8318. https://doi.org/10.3390/en8088300.

Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems.* Cambridge University.

Sato, K., Samejima, M., Akiyoshi, M., & Komoda, N. (2012). A scheduling method of air conditioner operation using workers daily action plan towards energy saving and comfort at office. *Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation (ETFA 2012),* 1–6. https://doi.org/10.1109/ETFA.2012.6489619.

Schmidt, M., Moreno, M. V., Schülke, A., Macek, K., Mařík, K., & Pastor, A. G. (2017). Optimizing legacy building operation: The evolution into data-driven predictive cyber-physical systems. *Energy and Buildings, 148*, 257–279. https://doi.org/10.1016/j.enbuild.2017.05.002.

Schwartz, H. M. (2014). *Multi-agent machine learning. A reinforcement approach* (1st ed.). John Wiley and Sons Inc.

Sen, S., Sekaran, M., & Hale, J. (1994). Learning to coordinate without sharing information. *Presented at the 12th National Conference on Artificial Intelligence (AAAI-94),* 426–431.

Shaikh, P. H., Nor, N. B. M., Nallagownden, P., Elamvazuthi, I., & Ibrahim, T. (2013). Robust stochastic control model for energy and comfort management of buildings. *Australian Journal of Basic and Applied Sciences, 7*(10), 137–144.

Shaikh, P. H., Nor, N. B. M., Nallagownden, P., Elamvazuthi, I., & Ibrahim, T. (2014). A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews, 34*, 409–429. https://doi.org/10.1016/j.rser.2014.03.027.

Silver, D. (2015). *RL course by David Silver.* UCL. Retrieved fromhttp://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359. https://doi.org/10.1038/nature24270.

Song, Y., Wu, S., & Yan, Y. Y. (2015). Control strategies for indoor environment quality and energy efficiency—A review. *International Journal of Low-Carbon Technologies, 10*(3), 305–312. https://doi.org/10.1093/ijlct/ctt051.

Sun, B., Luh, P. B., Jia, Q., & Yan, B. (2013). Event-based optimization with non-stationary uncertainties to save energy costs of HVAC systems in buildings. *2013 IEEE International Conference on Automation Science and Engineering (CASE),* 436–441. https://doi.org/10.1109/CoASE.2013.6654055.

Sun, B., Luh, P. B., Jia, Q., & Yan, B. (2015). Event-based optimization within the lagrangian relaxation framework for energy savings in HVAC systems. *IEEE Transactions on Automation Science and Engineering, 12*(4), 1396–1406. https://doi.org/10.1109/TASE.2015.2455419.

Sun, Y., Somani, A., & Carroll, T. E. (2015). Learning based bidding strategy for HVAC systems in double auction retail energy markets. *2015 American Control Conference (ACC),* 2912–2917. https://doi.org/10.1109/ACC.2015.7171177.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, Mass: MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (second edition). Cambridge, Massachusetts: The MIT Press.

Sycara, K. P. (1998). Multiagent systems. *AI Magazine, 19*, 79–92.

Urieli, D., & Stone, P. (2013). *A learning agent for heat-pump thermostat control.* 8.

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied Energy, 235*, 1072–1089. https://doi.org/10.1016/j.apenergy.2018.11.002.

Vázquez-Canteli, J. R., Ulyanin, S., Kämpf, J., & Nagy, Z. (2019). Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities. *Sustainable Cities and Society, 45*, 243–257. https://doi.org/10.1016/j.scs.2018.11.021.

Veselý, M., & Zeiler, W. (2014). Personalized conditioning and its impact on thermal comfort and energy performance – A review. *Renewable and Sustainable Energy Reviews, 34*, 401–408. https://doi.org/10.1016/j.rser.2014.03.024.

Wang, N., Phelan, P. E., Harris, C., Langevin, J., Nelson, B., & Sawyer, K. (2018). Past visions, current trends, and future context: A review of building energy, carbon, and sustainability. *Renewable and Sustainable Energy Reviews, 82*, 976–993. https://doi.org/10.1016/j.rser.2017.04.114.

Wang, W., Zmeureanu, R., & Rivard, H. (2005). Applying multi-objective genetic algorithms in green building design optimization. *Building and Environment, 40*(11), 1512–1525. https://doi.org/10.1016/j.buildenv.2004.11.017.

Wang, Y., Kuckelkorn, J., & Liu, Y. (2017). A state of art review on methodologies for control strategies in lowenergy buildings in the period from 2006 to 2016. *Energy and Buildings, 147*, 27–40.

Wang, Y., Velswamy, K., & Huang, B. (2017). A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. *Processes, 5*(3), 46. https://doi.org/10.3390/pr5030046.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*Ph.D. Thesis. University of Cambridge.

Wei, T., Wang, Y., & Zhu, Q. (2017). Deep reinforcement learning for building HVAC control. *Proceedings of the 54th Annual Design Automation Conference 2017 on - DAC' 17,* 1–6. https://doi.org/10.1145/3061639.3062224.

Wenqi, G., & Zhou, M. (2009). Technologies toward thermal comfort-based and energy-efficient HVAC systems: A review. *2009 IEEE International Conference on Systems, Man and Cybernetics,* 3883–3888. https://doi.org/10.1109/ICSMC.2009.5346631.

Xu, X., He, H., & Hu, D. (2002). Efficient reinforcement learning using recursive least-squares methods. *The Journal of Artificial Intelligence Research, 16*, 259–292. https://doi.org/10.1613/jair.946.

Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I., ... Feng, X. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings, 156*, 258–270. https://doi.org/10.1016/j.enbuild.2017.09.084.

Yang, L., Nagy, Z., Goffin, P., & Schlueter, A. (2015). Reinforcement learning for optimal control of low exergy buildings. *Applied Energy, 156*, 577–586. https://doi.org/10.1016/j.apenergy.2015.07.050.

Yang, R., & Wang, L. (2012). Multi-objective optimization for decision-making of energy and comfort management in building automation and control. *Sustainable Cities and Society, 2*(1), 1–7. https://doi.org/10.1016/j.scs.2011.09.001.

Yang, R., & Wang, L. (2013). Multi-zone building energy management using intelligent control and optimization. *Sustainable Cities and Society, 6*, 16–21. https://doi.org/10.1016/j.scs.2012.07.001.

Ye, D., Zhang, M., & Vasilakos, A. V. (2017). A survey of self-organisation mechanisms in multi-agent systems. *IEEE Transactions on Systems, Man, and Cybernetics Systems, 47*(3), 441–461.

Ye, W., Zhang, X., Gao, J., Cao, G., Zhou, X., & Su, X. (2017). Indoor air pollutants, ventilation rate determinants and potential control strategies in Chinese dwellings: A literature review. *The Science of the Total Environment, 586*, 696–729. https://doi.org/10.1016/j.scitotenv.2017.02.047.

Yu, Z., & Dexter, A. (2010). Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. *Control Engineering Practice, 18*(5), 532–539. https://doi.org/10.1016/j.conengprac.2010.01.018.

Zalejska-Jonsson, A., & Wilhelmsson, M. (2013). Impact of perceived indoor environment quality on overall satisfaction in Swedish dwellings. *Building and Environment, 63*, 134–144. https://doi.org/10.1016/j.buildenv.2013.02.005.

Zhang, Z., Chong, A., Pan, Y., Zhang, C., Lu, S., & Lam, K. P. (2018). A deep reinforcement learning approach to using whole building energy model for HVAC optimal control. *Presented at the 2018 Building Performance Modeling Conference and SimBuild Co-Organized by ASHRAE and IBPSA-USA.*

Zhang, Z., & Lam, K. P. (2018). Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system. *Proceedings of the 5th Conference on Systems for Built Environments - BuildSys' 18,* 148–157. https://doi.org/10.1145/3276774.3276775.