Applied Energy 235 (2019) 1072–1089

Contents lists available at ScienceDirect

Applied Energy

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

Reinforcement learning for demand response: A review of algorithms and modeling techniques

José R. Vázquez-Canteli, Zoltán Nagy*

Intelligent Environments Laboratory, Department of Civil, Architectural and Environental Engineering, The University of Texas at Austin, Austin, TX 78712, USA

HIGHLIGHTS

- Review of the application of reinforcement learning (RL) for demand response (DR)
- DR is relevant for integrating renewable energy source into the smart grid.
- Considering RL/DR from energy generation and storage, to supply and user satisfaction.
- Typical algorithms and open research directions are discussed.
- Most articles focus on single-agent systems in stationary environments.

ARTICLE INFO

Keywords:
Machine learning
Deep learning
HVAC control
Building energy
Electric vehicles
Smart grid

ABSTRACT

Buildings account for about 40% of the global energy consumption. Renewable energy resources are one possibility to mitigate the dependence of residential buildings on the electrical grid. However, their integration into the existing grid infrastructure must be done carefully to avoid instability, and guarantee availability and security of supply. Demand response, or demand-side management, improves grid stability by increasing demand flexibility, and shifts peak demand towards periods of peak renewable energy generation by providing consumers with economic incentives. This paper reviews the use of reinforcement learning, a machine learning algorithm, for demand response applications in the smart grid. Reinforcement learning has been utilized to control diverse energy systems such as electric vehicles, heating ventilation and air conditioning (HVAC) systems, smart appliances, or batteries. The future of demand response greatly depends on its ability to prevent consumer discomfort and integrate human feedback into the control loop. Reinforcement learning is a potentially model-free algorithm that can adapt to its environment, as well as to human preferences by directly integrating user feedback into its control logic. Our review shows that, although many papers consider human comfort and satisfaction, most of them focus on single-agent systems with demand-independent electricity prices and a stationary environment. However, when electricity prices are modelled as demand-dependent variables, there is a risk of shifting the peak demand rather than shaving it. We identify a need to further explore reinforcement learning to coordinate multi-agent systems that can participate in demand response programs under demand-dependent electricity prices. Finally, we discuss directions for future research, e.g., quantifying how RL could adapt to changing urban conditions such as building refurbishment and urban or population growth.

1. Introduction

The building sector accounts for more than 40% of the global energy use and 30% of greenhouse gas emissions [1]. Among the top ten CO_2 emitting countries, the main reasons for this increase of energy consumption are population growth, rapid urbanization, the increase of the ownership of personal appliances, and the lower average occupancy rates in the residential sector [2]. At the same time, integration of information and communication technologies and internet of things approaches are leading current cities towards the concept of smart cities, allowing them to achieve greater energy savings and become more efficient, livable, and sustainable [3].

Renewable energy resources can improve the energy autonomy of residential consumers and reduce CO_2 emissions [4]. However, high penetration levels of renewable energy resources can cause instability problems in the electrical grid due to their limited predictability,

* Corresponding author.
E-mail address: nagy@utexas.edu (Z. Nagy).

https://doi.org/10.1016/j.apenergy.2018.11.002
Received 2 April 2018; Received in revised form 30 October 2018; Accepted 2 November 2018

Nomenclature

AI	artificial intelligence
ANN	artificial neural network
BRL	batch reinforcement learning
CPP	critical peak pricing
DG	distributed generation
DHW	domestic hot water
DR	demand response
EV	electric vehicle
HVAC	heating ventilation and air conditioning
HEV	hybrid electric vehicle
PCM	phase change material
MA	multi-agent
MC	Monte Carlo
MDP	Markov decision process
RL	reinforcement learning
RTP	real-time price
RES	renewable energy sources
SA	single-agent
TOU	time of use
TD	temporal difference
A	actions
α	learning rate
γ	discount factor
$P^{a}_{ss'}$	transition probability
π	policy
S	states
Q	state-action value
Q^{*}	optimal Q-value
r	reward
V^{π}	state value

controllability and variability [5]. Demand response (DR) can enable consumers to reduce their energy consumption through load curtailment, shift their energy consumption over time, or generate and store energy at certain times to provide the grid with more flexibility. In exchange, consumers typically receive a reduction of their energy bill [6].

In OECD countries, the use electricity in buildings for cooling accounted for approximately 3.5–7% of the total energy use in 2010; and an additional 15% of the global space and water heating is produced by electricity [7]. Thus, heating ventilation and air conditioning (HVAC) and domestic hot water (DHW) supply are relevant areas for providing DR capabilities. HVAC systems can contribute to load curtailment events by modifying the temperature set points, participating in load shifting by pre-heating or pre-cooling the buildings [8] (passive energy storage), or by directly storing thermal energy in an energy storage system (active energy storage). Further, learning thermostats with DR capabilities are available to residential consumers to help them save energy by allowing energy retailing companies to save energy costs during peak-demand events through consumer participation.

In residential buildings, lighting and home appliances, without including cooking devices, accounted for 15–25% of the household total energy consumption in 2010 [7]. The integration of smart appliances at the residential level will help in the development of effective DR programs. Home appliances, e.g. dishwashers, washing machines, dryers, can be scheduled to perform their tasks during a time window specified by the user. Then, they are switched on and off in response to changes in the electricity prices.

In developing countries, the demand for new air conditioning devices is expected to increase significantly in the coming years [9]. This is an opportunity to commercialize smart air conditioning devices, or smart thermostats with DR capabilities, which could additionally lead to lower capital investments in the electrical infrastructure of these countries.

Another important field that can have a great impact on the way energy is generated and stored are electric vehicles (EVs). The increasing number of EVs in use, their electrical storage capacity, as well as their inherent connectivity hold a great potential for integrating them in the future. In combination with additional energy storage devices, and distributed generation systems, such as residential rooftop photovoltaic systems, EVs are essentially another grid management resource that can be used for DR at the residential level [10].

In the industrial sector, demand response-based smart and microgrid systems have a significant potential that has not been completely appreciated yet. New regulatory frameworks would be required to allow electricity markets to more appropriately benefit from the flexibility that industrial facilities can provide. Furthermore, the implementation of Auto-DR technology may be a solution to manage demand in large industrial facilities such as data-centers [11].

1.1. Motivation for this review

Although DR is a promising approach to increase demand flexibility, the potential peak reduction from DR programs in the US was only 6.6% of the peak demand in 2015 [12]. The reason for this is that electricity is a resource whose value for consumers is much higher than its price. For example, Centolella et al. showed that in a one-hour power outage, residential consumers would be willing to pay between $0.73 and $2.51 per kWh, i.e., about one order of magnitude higher than the actual price [13]. Commercial and industrial consumers would be willing to pay even more than residential consumers. Therefore, electricity consumers are generally unwilling to sacrifice much of their comfort or satisfaction for a lower electricity bill. As a result, the future success and scalability of DR depends on its ability to generate greater economic savings than dissatisfaction to the consumers [14–16]. The main factors that cause dissatisfaction among residential consumers are:

- Requiring consumers to modify their preferred electricity consumption patterns, e.g. forcing them to delay the usage of home appliances.
- Potentially undesired temperature set-points in buildings
- The effort consumers need to make to acquire information about electricity prices and take decisions accordingly about their consumption patterns.

Thus, the technical framework required to implement effective DR programs must be both automated and able to sense and minimize user discomfort. Artificial intelligence (AI) can have a great contribution in the integration of DR programs by automating energy systems, while learning from human behavior to minimize user discomfort and the level of required human-controller interaction.

The contribution of this paper is to provide a review of the studies that focus on the application of Reinforcement Learning (RL), an agent-based AI algorithm, for demand response applications. Given its adaptability and capacity to learn preferences of the user through interaction and without an explicit mathematical model, RL is an algorithm with a great potential of applicability for complex, real-world applications, specifically DR. We consider distributed generation (DG) at the building scale, with solar PV, HVAC systems, EVs and storage devices used in building energy systems. We also review articles about topics that study the applications of RL to control these energy systems but do not directly implement them under DR programs.

Besides summarizing the state-of-the art, our intention with this paper is to highlight the interdependences of demand response

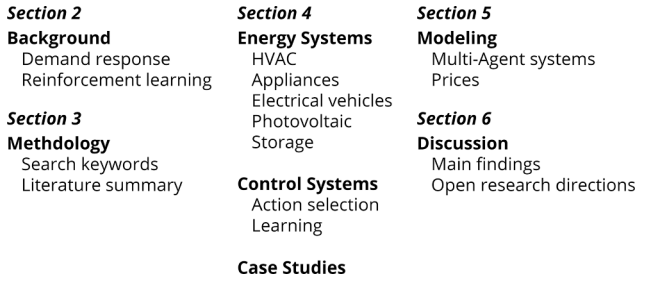

Fig. 1. Organization of the article.

schemes, potential control architectures and various versions of RL algorithms, as well as modeling techniques to determine research gaps and potential future research directions.

1.2. Previous reviews

There are several general reviews of DR in the literature. Aghaei and Alizadeh analyzed studies on DR in smart grids equipped with renewable energy resources [17]. Rajasekhar and Naran, and Law et al. reviewed diverse control algorithms and architectures that have been applied to DR [18,19]. Vardakas et al. reviewed multiple DR pricing schemes and algorithms [20], whereas Li and Wen analyzed various types of building energy models used in building control and operation studies [21].

Other reviews have focused on analyzing the use of different algorithms in specific energy systems such as HVAC or energy storage. Yu et al. reviewed control algorithms, including RL, for the integration of thermal energy storage in buildings [22]. Wang and Ma reviewed several supervisory control strategies for building HVAC systems, and highlighted the advantages and easy implementation of model-free algorithms such as RL [23]. Salehizadeh and Soltaniyan reviewed the applications of the fuzzy Q-learning algorithm for modelling the electricity market considering renewable power penetration [24]. Dusparic et al. compared some methods, including RL, that have been investigated for DR applications [25]. However, the lack of an extensive literature review which integrates all components of DR, ranging from energy generation, storage and control to user satisfaction, motivates the need for our paper.

1.3. Organization

This paper is organized as follows (see Fig. 1). Section 2 provides the background to our review by introducing demand response programs, and reinforcement learning. Section 3 details the methodology for the survey and summarizes the literature. Based on the review, Section 4 then discusses control algorithms for different energy systems, and those studies that contain a case study in which the controller was implemented in a physical system. Section 5 focuses on multi-agent systems, the way electricity prices are modelled, and how they affect the learning process. Then, Section 6 discusses the findings in each field and potential directions for future research. Finally, Section 7 concludes the paper.

2. Background

2.1. Demand response

There are two main categories of DR programs: incentive-based and time-based (Fig. 2). In incentive-based programs, consumers voluntarily participate in a scheme in which the system operator can directly turn off some appliances to reduce the energy consumption of consumers during the periods of peak demand. By contrast, time-based programs are generally based on dynamic pricing, and aim to flatten the curves of

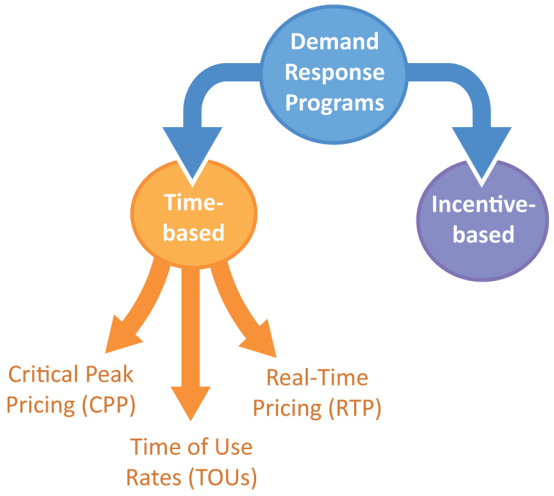

Fig. 2. Common types of demand response programs.

demand by offering the consumer electricity prices that vary in time [26].

Many studies have highlighted the importance of dynamic pricing at the distribution level to incentivize consumers to participate in DR programs [27,28]. Some common types of dynamic pricing mechanisms used in time-based DR programs are time of use rates (TOU), critical peak pricing (CPP), and real-time pricing (RTP) [6,29]. Other studies suggest that time-based programs are more suited for residential consumers, while incentive-based programs are more appropriate for industrial consumers [30]. In brief, the advantages of DR approaches are [5,6,31,32]:

1. Improved grid stability due to increased demand flexibility.
2. Shift of peak demand towards periods of peak renewable energy generation.
3. Lower thermal costs and electricity prices. Since the peak to average ratio of the demand decreases, less peaking plants need to be operated.
4. Reduction of the investments in generation, transmission, and distribution assets, which are sized to meet peak demand.
5. Lower capacity reserves requirements.
6. Reduced energy bills for consumers.

Two examples of implementations of demand response programs in the U.S. are the Energy-Smart Pricing Plan^{SM} in Illinois from 2003 to 2006 [33], and the Critical Peak Pricing experiment in California [34]. The programs showed that consumers did increase their demand elasticity during times of very high electricity prices. In both cases, the DR implementations were not automated, and the participants were informed at least a day ahead and had then make their decisions. Thus, in addition to being a manual program, the participants had to frequently make conscious decisions on electricity demand, which limits the success of the program to only periods with very high electricity prices.

2.2. Reinforcement learning

Reinforcement Learning (RL) is an agent-based AI algorithm in which the agents learn the optimal set of actions, i.e., the optimal policy through their interaction with the environment (see Fig. 3). We review it here briefly. For a thorough introduction, the interested reader is referred to standard textbooks [35].

RL can be formalized using a Markov Decision Process (MDP). An MDP contains four elements: a set of states S, a set of actions A, a reward function r: $S \times A$, and transition probabilities between the states P: $S \times A \times S' \in [0, 1]$. The policy π maps states to actions as π:

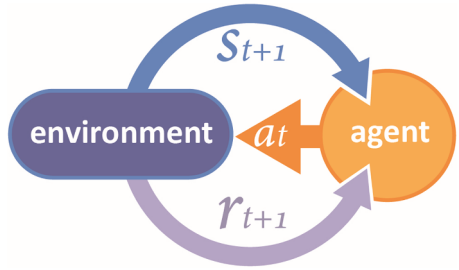

Fig. 3. Agent-environment interaction in reinforcement learning.

$S \rightarrow A$. As in eq. (1), the value function $V^{\pi}(s)$ of a state s

$$V^{\pi}(s) = \sum_{a} \pi(s|a) \sum_{s',\mathrm{r}} P^{a}_{ss'}[r + \gamma V^{\pi}(s')] \tag{1}$$

is the expected return for the agent when starting in state s and following the policy π. $R^{a}_{ss'}$ is the reward received for taking the action a while being in state s and transitioning to the state s'. While $\gamma \in [0, 1]$ is a discount factor allowing to balance between an agent that only considers immediate rewards ($\gamma = 0$) or strives towards long term rewards ($\gamma = 1$).

The way of solving an MDP, i.e. determining the optimal policy, depends on whether the probability transitions $P^{a}_{ss'}$ and the reward function r, i.e., the dynamics of the system, are known. If $P^{a}_{ss'}$ and r are known, the learning problem is essentially reduced to a planning problem, and the optimal solution can be found through iterative approaches: either policy iteration or value iteration [36]. However, RL is also applicable when the model dynamics ($P^{a}_{ss'}$ and r) are unknown and must be determined or estimated through interaction of the agent with the environment. One can distinguish between two approaches. In the model-based approach, the model is first learned, and then used in a planning procedure as described above. This method has been proven to be successful in, for instance, the training of an aerobatic helicopter [37]. In the model-free approach, the agent learns to associate the optimal action for each state without explicitly determining transition probabilities between the states [38]. Q-learning is the most widely used model-free RL technique due to its simplicity [39,40]. In simple problems, a table can represent the transitions in which the state-action values, or Q-values, are stored. Every entry to the table represents a state-action pair, and all Q-values are updated as:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a) + \gamma \max_{a} Q(s', a) - Q(s, a)] \tag{2}$$

where $\alpha \in [0, 1]$ is the learning rate, which explicitly defines to what degree new knowledge overrides old knowledge: for $\alpha = 0$, no learning occurs, while for $\alpha = 1$, all prior knowledge is lost.

We can distinguish between on-policy and off-policy learning. Q-learning is an off-policy algorithm, which means that it makes the Q-values converge to Q^{*}, the optimal policy. Off-policy algorithms do not need to follow a specific policy to update the Q-values, and this allows them to learn from historical data that was obtained without selecting actions in any specific manner. By contrast, Sarsa is an on-policy algorithm, in which Q-values are updated as:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a) + \gamma Q(s', a') - Q(s, a)] \tag{3}$$

As a consequence, the Q-values converge toQ^{π}, the Q-values of the specific policy that is being. Therefore, on-policy algorithms like Sarsa require the controller to perform the action selection following the policy for which the Q-values are being found. This usually accelerates convergence but does not necessarily allow to learn from historical data or to find the optimal policy. Mathematically, the only difference between these two algorithms is that in Q-learning the Q-values are updated using the maximum Q-value in the new state, while in Sarsa they are updated with the Q-value of the action that is selected in the new state.

To achieve convergence of the Q-values, it is necessary to explore actions for which the Q-values are not necessarily the highest yet. Therefore, there is a trade-off between exploring actions that at a given iteration seem to be suboptimal and exploiting the actions that seem to lead to the optimal policy. Exploration allows convergence of the Q-values, while exploitation helps receive greater rewards. The way actions are selected to manage this trade-off between exploration and exploitation is known as action-selection, and the most widely used methods are the ε-greedy policy, and the soft-max action selection. The ε-greedy policy consists of taking the action with the greatest Q-value with probability $(1 - \varepsilon)$, and selecting a random action with probability ε. On the other hand, in the soft-max action selection method, the probability of choosing an action is related to its Q-value following

$$Pr(a_i|s) = \frac{e^{\frac{Q(s,a_i)}{\tau}}}{\sum_{b=1}^{n} e^{\frac{Q(s,a_b)}{\tau}}} \tag{4}$$

This probability uses a Boltzman distribution, in which τ is a positive parameter called the temperature, and n is the number of possible actions that can be taken.

A more general Q-learning algorithm is Q-learning with eligibility traces, or Q(λ). In Q(λ), all Q-values are updated every time-step based on the current reward, which leads to faster convergence speeds. Q-learning is a special case of this algorithm that assumes $\lambda = 0$, in which only the Q-value associated with the current state-action pair is updated. In this case, Q-values are updated based on previous estimates of the same Q-values, which accelerates convergence. For $\lambda = 1$, each Q-value is no longer estimated based on previous estimations of the same Q-value, but on the discounted sum of all the received rewards. This is another special case of Q(λ) known as Monte Carlo (MC) value iteration and is particularly advantageous when a problem cannot be modelled as a perfect MDP. For any other value of $\lambda = (0, 1)$, Q(λ) can combine the advantages and disadvantages of standard Q-learning and MC in a different proportion.

When the state-action space is very large, the speed of convergence is severely reduced, which is a problem known as the curse of dimensionality. Furthermore, Q-learning is a discrete algorithm, but in many problems states and actions are continuous. In order to avoid the discretization of the states and actions, and address the problem of curse of dimensionality, the Q-table can be replaced by a function estimator such as Artificial Neural Networks (ANN) or other linear and non-linear regression techniques [41]. To replace the Q-table by a function estimator, it is necessary to train the function estimator with a set of states, actions and Q-values. For this purpose, Batch Reinforcement Learning (BRL) with fitted Q-iteration is used, an algorithm based on Q-learning that aggregates the experience before making the updates and training the function estimator [42]. As a result, the order in which experience is obtained does not negatively affect the learning process. Another common BRL method is experience replay, which does not require any function estimator but randomly sampling the batched experiences instead [43].

Actor-critic methods are another form of RL algorithms that have separate memory structures to represent the state-action space, and the state-value space respectively. They can learn stochastic policies explicitly, which is an advantage in non-Markovian or in stochastic processes [35,44].

One of the most important characteristics of RL algorithms is their easiness to acquire human feedback and learn from it. In the built environment, the comfort of the occupants, i.e., thermal comfort, can be used as a reward for the RL controller. This comfort can be either estimated or obtained from the occupants, i.e., occupants changing the temperature set-point on a thermostat can be a sign of discomfort. Another example is the scheduling of smart appliances or EVs to control their charge and discharge process. Failing to meet the user needs would lead to negative rewards that are part of the learning process. Additionally, RL algorithms can be trained off-line using past

experiences, which is very useful when large amounts of data are available and the system to control is too complex to develop a model.

Traditional RL focuses on a single agent that interacts with an environment. However, many real-world applications require coordinating several agents. This makes the learning process more difficult because each agent sees a non-stationary environment that also reacts to the other agents. Furthermore, as the number of agents and dimensions increases, the curse of dimensionality worsens. These non-stationary problems with moving learning targets are hard to tackle [45]. RL in multi-agent systems is a theoretical field still in its infancy, and most results of convergence and stability are for two agents. W-learning [46] is an example of a multi-agent RL algorithm that can be used competitively or cooperatively (collective W-learning).

3. Methodology

This literature review analyzes in detail those articles focused on DR because we believe this is a particularly important topic due to its direct economic savings for both utility companies and consumers. We also included articles that apply RL techniques to maximize human comfort and reduce the peak energy consumption or the energy consumption itself. These factors are important to understand the advantages and difficulties of implementing RL in an urban setting, and by association in DR programs. These fields are related to each other via the human factor, i.e., comfort, desire to consume less energy, or to pay a lower energy bill. Furthermore, RL algorithms that are used to maximize user comfort or minimize the energy consumption can be adapted and applied to DR just by adding a few additional states (i.e. the prices of electricity) and penalties (i.e. cost of electricity).

The literature review was performed in the Web of Science search engine using:

$$TS = A \ \ and \ \ \{B \ \ or \ \ (C \ \ and \ \ D)\} \tag{5}$$

where the parameters A, B, C and D are to the search terms shown in Table 1. We explicitly added the term *Q-learning* because, it is frequently used in the articles as a placeholder for *Reinforcement Learning* due to its popularity. However, we did not observe that adding this term biased our results towards this specific RL algorithm. We found 194 articles (cut-off date: 10/23/2018), which we manually filtered to remove review articles, some duplicates and other studies not related to our field of interest for this review. We included a few additional articles that were not available in Web of Science, and the final selection contained 105 articles, which are summarized in Table 2. As Fig. 4 illustrates, there was a significant increase in publications after 2012, which was particularly remarkable in the field of electric vehicles.

In Table 2, we classified the articles based on the energy systems they focus on, and analyzed them according to the RL algorithms they use as well as their ability to address the problems of speed of convergence, and curse of dimensionality. We also analyzed if the studies modelled the systems as single-agent (SA) or multi-agent (MA) systems and classified them based on their objectives. Note that systems with many independent agents were classified as single-agent, and studies in which a central agent controls a group of subsystems were classified as a single-agent too. Additionally, we classified the articles on DR based on whether electricity prices were modeled as independent variables or as variables that depended on the energy demand, and therefore, on the actions of the agents, and whether the prices are stochastic or deterministic. We also analyzed whether the articles investigated the performance of the controllers under environments with non-stationary transition probabilities, and whether they tracked renewable energy resources (RES), such as wind power or solar. Articles in which electricity prices were modeled using historical real-time electricity prices were classified as non-stationary because these prices do not necessarily follow pure MDPs. However, it is important to note that the articles that focus on multi-agent systems have a higher degree of non-stationarity, which is discussed further in Section 6.3. For completeness, Table 2 also shows whether the authors pre-trained the RL controllers with existing data or simplified models of the real system.

4. Energy systems and reinforcement learning methods

In this section, we discuss the literature based on the controlled energy systems, and the algorithms they use for action-selection and knowledge acquisition.

4.1. Energy systems

We have identified four major groups of energy systems that have a significant potential for DR applications: HVAC and DHW systems, smart appliances, EVs and HEVs, and distributed generation with energy storage. The fact that all these energy systems can not only have an impact on the electrical grid, but are also interrelated with human comfort and behavior, has been one of the reasons that has led some of the authors to use RL to control them. Self-adaptability, model-free nature, the ability to learn from historical data are some other important features of RL that explain why it has been used to control these energy systems.

4.1.1. Heating ventilation air-conditioning and domestic hot water

Participation of buildings in DR requires considering both human comfort and proper management of the electrical loads. While some authors consider occupants' comfort as a constraint that must always be prioritized, others focus on achieving a trade-off between comfort and DR capabilities. HVAC systems can participate in DR events by pre-heating or pre-cooling the indoor spaces to achieve some degree of load shifting (passive energy storage in the thermal mass of the building), or by regulating the amount of thermal energy stored in specialized thermal energy storage systems or in DHW storage devices.

The likely very first application of RL to control and automate the built environment was implemented by Mozer in 1998 [48]. In his Neural Network House, he controlled the HVAC, DHW, and lighting systems to minimize user discomfort and energy costs. Later, Anderson et al. simulated a hybrid RL/Proportional-integral control for a heating coil with potential applications in HVAC systems [47]. Du and Fei also used RL to minimize the tracking error of the set-point temperature in an HVAC system [55] and, in 2003, Henze and Shoenmann proposed a RL controller for the charge and discharge process of an electrical-driven ice storage system in order to minimize its energy costs [49].

Several studies, many of them in the field of DR, have focused on reducing the cost of the energy consumed by building energy systems. Liu and Henze focused their work on reducing the energy costs of a passive thermal storage inventory [53]. As they concluded that the training time required was unacceptably long, in [51,52] they investigated the effects of adding prior simulated knowledge in the training stage. Sun et al. minimized the day-ahead energy cost of an HVAC system connected to several rooms [88]. Costanzo et al. implemented a RL controller to minimize the energy cost of an air

Table 1
Search queries are composed of the following terms as shown in eq. (5). The symbol "*" is used to search for terms in both singular and plural forms.

A	B	C	D
"Reinforcement learning"	"Demand-side management"	Heating	Building*
Q-learning	"Demand response"	Cooling	House*
	"Electric vehicle*"	"Electricity price*"	Residential
	HVAC	Comfort	Home*
		Energy	Household*
		Photovoltaic	
		PV	
		Solar	

Table 2
Summary of the papers reviewed.

Ref	Date	Energy system	Learning algorithm	Action selection	Pre-training	Q-function estimator	Agent	Objectives	Non-stationary transition probabilities	RES integration	Electricity prices	
[47]	1997	Heating coil	Q-learning	ε-greedy	No	N/A	SA	Temperature tracking error	N/A	N/A	N/A	N/A
[48]	1998	HVAC, DHW, lighting	Q-learning	ε-greedy	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[49]	2003	HVAC	Q-learning	ε-greedy	No	N/A	SA	Energy cost	No	No	Deterministic	Independent
[50]	2003	Solar PV, Battery	Q-learning	Soft-max	No	N/A	SA	Energy cost	No	Yes	Stochastic	Independent
[51]	2006	Thermal storage	Q-learning	Soft-max	Yes	N/A	SA	Energy cost	No	No	Deterministic	Independent
[52]	2006	Thermal storage	Q-learning	Soft-max	Yes	N/A	SA	Energy cost	No	No	Deterministic	Independent
[53]	2007	Thermal storage	Q-learning	Soft-max	No	ANN	SA	Energy cost	No	No	Deterministic	Independent
[54]	2007	HVAC	TD(λ)	ε-greedy	No	RBF	SA	Energy, comfort	N/A	N/A	N/A	N/A
[55]	2008	HVAC	Actor-critic	N/A	No	ANN	SA	Temperature tracking error	N/A	N/A	N/A	N/A
[56]	2010	HVAC	Q(λ)	ε-greedy	Yes	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[57]	2010	Smart appliances	Q-learning	ε-greedy	No	N/A	SA	Energy cost, satisfaction	No	No	Stochastic	Independent
[58]	2011	EV	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	No	Stochastic	Independent
[59]	2011	Battery	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	Yes	Deterministic	Independent
[60]	2012	Battery	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	No	Stochastic	Independent
[61]	2013	HVAC	Policy iteration	Tree-search	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[62]	2013	HVAC	Q-learning	Greedy	No	N/A	SA	Energy cost, satisfaction	No	No	Deterministic	Independent
[63]	2013	HVAC, lighting	Actor critic	N/A	No	RBFN	SA	Energy, comfort	N/A	N/A	N/A	N/A
[64]	2013	EV	Sarsa(λ)	ε-greedy	No	N/A	SA	Energy cost	No	No	Stochastic	Independent
[65]	2013	EV	Q-learning	Soft-max	No	N/A	MA	Energy cost, satisfaction	Yes	No	Deterministic	Dependent
[66]	2013	EV	Q-learning	ε-greedy	Yes	N/A	SA	Energy cost, satisfaction	No	No	Stochastic	Independent
[67]	2013	EV	Q-learning	N/A	No	N/A	SA	Energy cost	No	No	Stochastic	Independent
[68]	2013	EV	W-learning	N/A	No	N/A	MA	Energy cost, satisfaction	Yes	No	Deterministic	Dependent
[69]	2013	EV, appliances	BRL	Soft-max	No	Extremely Rand. Trees	MA	Peak demand	Yes	Yes	N/A	N/A
[70]	2013	DG, storage	Actor critic	N/A	Yes	ANN	SA	Energy cost	Yes	Yes	Stochastic	Independent
[71]	2013	Smart appliances	Q-learning	N/A	No	N/A	SA	Energy cost, satisfaction	No	No	Deterministic	Independent
[72]	2013	Smart appliances	Q-learning	ε-greedy	No	N/A	SA	Energy, satisfaction	N/A	N/A	N/A	N/A
[73]	2014	Water heaters	BRL	Soft-max	No	Extremely Rand. Trees	SA	Energy cost	No	No	Deterministic	Independent
[74]	2014	Electric water heaters	Fuzzified Q-learning	Random / greedy	No	N/A	SA	Peak demand, satisfaction	Yes	No	N/A	N/A
[75]	2014	HEV	TD(0)	ε-greedy	No	N/A	SA	Energy cost	N/A	N/A	Deterministic	Independent
[76]	2014	EV	Q-learning	ε-greedy	No	N/A	MA	Energy cost, satisfaction	Yes	No	Stochastic	Dependent
[77]	2014	EV	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	Yes	Stochastic	Independent
[78]	2014	EV	W-learning	N/A	No	N/A	MA	Energy peak, satisfaction	Yes	No	N/A	N/A
[79]	2014	DG, storage	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	No	Stochastic	Dependent
[80]	2014	Battery	Dual Q-learning	N/A	No	ANN	SA	Energy cost	Yes	No	Stochastic	Independent
[81]	2014	Generic loads	Multi-player RL	ε-greedy	No	N/A	MA	Energy cost, satisfaction	Yes	No	Stochastic	Dependent
[82]	2014	Electrical storage	Q-learning	N/A	No	N/A	SA	Energy cost	Yes	No	Stochastic	Independent
[83]	2014	Smart appliances	Q(λ)	Soft-max	No	N/A	SA	Energy cost, satisfaction	No	No	Stochastic	Independent
[84]	2015	HVAC	Q-learning	ε-greedy	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[85]	2015	HVAC	Q-learning	ε-greedy	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[86]	2015	HVAC	Q-learning	ε-greedy	No	N/A	MA	Energy cost, comfort	Yes	No	Stochastic	Dependent
[87]	2015	HVAC	Q-learning	"PAC" alg.	No	N/A	SA	Comfort	N/A	N/A	N/A	N/A
[88]	2015	HVAC	Q-learning	Greedy	No	N/A	SA	Energy cost, comfort	No	No	Deterministic	Independent
[89]	2015	HVAC	Monte-Carlo	ε-soft	No	N/A	SA	Energy cost	No	No	Deterministic	Independent
[90]	2015	HVAC	BRL	Soft-max	No	Extremely Rand. Trees	SA	Energy, comfort	N/A	N/A	N/A	N/A
[91]	2015	HVAC	BRL	ε-greedy	No	Extremely Rand. Trees	SA	Energy cost	Yes	No	Deterministic & stochastic	Independent
[92]	2015	Ventilated facade	Sarsa(λ)	ε-greedy	Yes	N/A	SA	Energy	N/A	N/A	N/A	N/A
[93]	2015	EV	W-learning	N/A	No	N/A	MA	Energy peak, satisfaction	Yes	No	N/A	N/A
[94]	2015	EV	W-learning	N/A	No	N/A	MA	Energy peak, satisfaction	Yes	Yes	N/A	N/A
[95]	2015	EV	BRL	ε-greedy	No	Kernel-based approx.	SA	Energy cost, satisfaction	Yes	No	Stochastic	Independent
[96]	2015	EV	BRL	Soft-max	No	Regression	SA	Energy cost	No	No	Stochastic	Independent
[97]	2015	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[98]	2015	Battery	Dual Q-learning	N/A	No	ANN	SA	Energy cost	Yes	No	Stochastic	Independent

(*continued on next page*)

Table 2 (*continued*)

Ref	Date	Energy system	Learning algorithm	Action selection	Pre-training	Q-function estimator	Agent	Objectives	Non-stationary transition probabilities	RES integration	Electricity prices	
[99]	2015	Battery	Q-learning	ε-greedy	No	N/A	SA	Energy cost	No	Yes	Stochastic	Independent
[100]	2015	Battery	Q-learning	ε-soft	No	N/A	SA	Peak demand	Yes	Yes	N/A	N/A
[101]	2015	Battery, heat tank	Q-learning	ε-greedy	Yes	N/A	SA	Energy cost, comfort	No	No	Deterministic	Independent
[102]	2015	Battery, PV	Q-learning	ε-greedy	Yes	N/A	SA	Energy cost	Yes	Yes	Deterministic & stochastic	Independent
[103]	2015	Battery, PV	TD(λ)	ε-greedy	No	N/A	SA	Energy cost	Yes	Yes	Deterministic	Independent
[104]	2015	PV/T, thermal storage, heat pump	BRL	ε-greedy	No	ANN	MA	Energy, comfort	Yes	N/A	N/A	N/A
[105]	2015	DG, storage	Q-learning	ε-greedy	No	N/A	SA	Energy cost	Yes	No	Stochastic	Dependent
[106]	2015	Smart appliances	Q-learning	N/A	No	N/A	SA	Energy cost	Yes	Yes	Stochastic	Independent
[107]	2015	Smart appliances	Q-learning	ε-greedy	No	N/A	SA	Energy cost, satisfaction	No	No	Stochastic	Independent
[108]	2016	HVAC	BRL	ε-greedy	No	Extremely Rand. Trees	SA	Energy cost	Yes	No	Stochastic	Independent
[109]	2016	DHW	BRL	N/A	Yes	N/A	SA	Energy, satisfaction	N/A	N/A	N/A	N/A
[110]	2016	Blinds, lighting	Q-learning	ε-greedy	No	N/A	SA	Energy, satisfaction	N/A	N/A	N/A	N/A
[111]	2016	EV	Q-learning	Pursuit alg.	No	N/A	SA	Energy cost, satisfaction	No	No	Deterministic	Independent
[112]	2016	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[113]	2016	HEV	Q-learning	N/A	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[114]	2016	HEV	Q-learning	ε-greedy	No	N/A	SA	Energy cost	N/A	N/A	N/A	N/A
[115]	2016	Battery	Q-learning	ε-greedy	No	N/A	SA	Energy	N/A	N/A	N/A	N/A
[116]	2016	Battery	Q-learning	N/A	No	N/A	SA	Energy cost	No	Yes	Deterministic	Independent
[117]	2016	DG, EV, appliances	Q-learning	ε-soft	No	N/A	SA	Peak demand, satisfaction	No	No	Stochastic	Independent
[118]	2016	Appliances	Q-learning	N/A	No	N/A	MA	Energy cost, satisfaction	Yes	No	Stochastic	Dependent
[119]	2017	HVAC	BRL	N/A	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[120]	2017	HVAC	BRL	Soft-max	Yes	ANN	SA	Energy	N/A	N/A	N/A	N/A
[121]	2017	HVAC	BRL	ε-greedy	No	Extremely Rand. Trees	SA	Cost	No	No	Stochastic	Independent
[122]	2017	HVAC	Actor-critic	N/A	No	Recurrent neural net.	SA	Energy, comfort	N/A	N/A	N/A	N/A
[123]	2017	DHW, heat pump	BRL	N/A	No	Extremely Rand. Trees	SA	Energy	N/A	N/A	N/A	N/A
[124]	2017	DHW heaters	Q-learning & Actor-critic	Random / greedy	Yes	N/A	SA	Energy cost	No	No	Deterministic	Independent
[125]	2017	EV	BRL	N/A	No	Kernel-averaging regr.	SA	Cost	No	No	Deterministic	Independent
[126]	2017	EV	W-learning	Soft-max	Yes	N/A	MA	Energy peak, satisfaction	Yes	Yes	N/A	N/A
[127]	2017	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[128]	2017	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[129]	2017	Battery, PV	BRL	N/A	No	Extremely Rand. Trees	SA	Energy	N/A	N/A	N/A	N/A
[130]	2017	Battery, PV	BRL	ε-greedy	No	Echo state network	SA	Cost	Yes	Yes	Stochastic	Independent
[131]	2017	Generic load	Deep transfer Q-learning	ε-greedy	Yes	Deep belief neural network	MA	Energy cost	Yes	Yes	Deterministic	Dependent
[132]	2017	Appliances	Actor critic	Soft-max	No	Linear regression	MA	Energy cost, satisfaction	Yes	No	Deterministic & stochastic	Dependent
[133]	2017	Smart appliances	BRL	Soft-max	No	ANN	SA	Peak reduction	No	No	N/A	N/A
[134]	2018	HVAC	Sarsa(λ)	ε-greedy	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[135]	2018	HVAC	Q-learning	ε-greedy	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[136]	2018	HVAC	BRL	Soft-max	Yes	ANN	MA	Energy cost	Yes	Yes	Deterministic	Dependent
[137]	2018	DHW	BRL	Soft-max	No	Extremely Rand. Trees	SA	Energy cost	No	No	Deterministic	Independent
[138]	2018	DHW	Model-based RL	N/A	No	N/A	SA	Energy, comfort	N/A	N/A	N/A	N/A
[139]	2018	District heating network	BRL	Soft-max	No	Extremely Rand. Trees	SA	Cost, peak demand	No	No	Deterministic	Independent
[140]	2018	HEV	Q-learning	N/A	No	N/A	SA	Energy	N/A	N/A	N/A	N/A
[141]	2018	HEV	BRL	ε-greedy	No	ANN	SA	Fuel consumption	N/A	N/A	N/A	N/A
[142]	2018	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[143]	2018	HEV	Q-learning	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[144]	2018	HEV	BRL	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[145]	2018	HEV	Q(λ)	Pursuit alg.	No	N/A	MA	Energy cost, waiting line, remaining SOC, distance	Yes	No	Deterministic	Independent
[146]	2018	HEV	Monte-Carlo	ε-greedy	No	N/A	SA	Fuel consumption	N/A	N/A	N/A	N/A
[147]	2018	EV	Q-learning	ε-greedy	No	N/A	SA	Energy cost, satisfaction	N/A	N/A	Deterministic	Dependent
[148]	2018	EV	Q-learning	N/A	No	N/A	SA	Energy	N/A	N/A	N/A	N/A

(*continued on next page*)

Table 2 (*continued*)

Ref	Date	Energy system	Learning algorithm	Action selection	Pre-training	Q-function estimator	Agent	Objectives	Non-stationary transition probabilities	RES integration	Electricity prices	
[149]	2018	Generic load	Q-learning	ε-greedy	No	N/A	SA	Demand flexibility, satisfaction	No	No	N/A	N/A
[150]	2018	DG, storage	Fuzzy Q-learning	ε-greedy	No	N/A	MA	Energy	Yes	Yes	N/A	N/A
[151]	2018	DG, storage	Q-learning	ε-greedy	Yes	N/A	SA	Energy cost	No	Yes	Deterministic	Independent

conditioning unit [108], and De Gracia et al. maximized the net electrical energy savings of a ventilated façade with phase change materials (PCM) [92].

Other studies have focused on increasing energy conservation in the built environment. Barrett and Linder reduced the energy consumption of an HVAC system while meeting temperature set-points during periods of occupancy, for cooling and heating a building [84]. In a similar approach, Li and Xia improved the energy conservation and comfort of an HVAC system [85]. Dalamagkidis et al. also aimed to maximize human comfort in buildings in addition to energy conservation, and they accounted for both thermal comfort and CO_2 levels [54], whereas Bielskis et al. accounted for both thermal and visual comfort [63]. Schmidt et al. tested their RL controller in a real building, and achieved a reduction in the energy consumption while maintaining an adequate level of comfort [119]. Yu and Dexter minimized both the energy consumed and the cost of thermal discomfort in a building [56]. As Liu and Henze did [53], they highlighted the need for prior knowledge to reach a more reasonable training time. Urieli and Stone, and Ruelens et al. applied different RL algorithms to control a heat pump with a setback strategy and an auxiliary heater for optimal energy conservation and comfort maximization [61,90]. As future work, Ruelens et al. proposed implementing a similar approach in a scenario with real-time electricity prices, and in a lab environment [90]. In two different studies, Ruelens et al. developed a BRL controller to minimize the operational cost of a cluster of electric water heaters [73], and in a later study they implemented an improved version of this controller in an actual water heater [137]. In a different study, Ruelens et al. used a BRL controller to reduce the cost of operating a heat pump and a thermostatically controlled load [121], whereas Al-Jabery et al. focused on controlling a group of electric water heaters to minimize peak demand and user dissatisfaction [74]. This work was later extended by minimizing the energy costs under TOU electricity rates [124]. Kazmi et al. implemented a hybrid BRL approach to maximize efficiency in DHW systems including user comfort as a constraint [109] and continued this work using a model-based and data-driven RL approach [138], while De Somer et al. maximized the self-consumption of local PV generation by storing the energy in DHW buffers [123]. Yang et al. used RL to control a building energy system with several photovoltaic-thermal panels, geothermal boreholes, and a heat pump [104]. The controller could increase the cumulative net power output of the system with respect to a rule-based controller. Cheng et al. investigated the use of Q-learning in an experimental study to maximize energy savings and user comfort [110]. They controlled the blinds and the lighting system of a building. Li et al. investigated the use of Q-learning for comfort maximization in a two story multi-zone office building [87]. They simulated the controller using EnergyPlus and MATLAB. Vázquez-Canteli et al. minimized the energy consumption of a heat pump combined with a chilled water tank and integrated their control algorithm into a building energy simulator for urban scale analysis [120,136], CitySim [152,153]. Wang et al. used actor-critic RL to maximize occupants' thermal comfort and minimize the energy consumption of the HVAC system in a simulated office building [122]. Claessens et al. used BRL in two simulated case studies for peak demand minimization and to reduce the cost of electricity [139]. Brusey et al. applied RL for comfort maximization and energy use reduction in vehicle cabins [134]. Chen et al. controlled an HVAC and a window system for energy and discomfort minimization in two different climatic zones [135].

4.1.2. Appliances

The use of RL to control smart appliances was introduced by O'Neill et al. in 2010 [57], and their work was further extended by Wen, O'Neill and Maei, in 2014–2015 [83,107]. They proposed a control for an energy management system in which users assign a priority to every task to be completed by any appliance, and a target time when the user prefers the request to be satisfied. They defined different functions to account for the user dissatisfaction when appliances did not meet the target times to complete the tasks. They clustered those appliances that were interdependent, i.e., laundry machine and dryer, and concluded that under the proper assumptions the computational complexity grew linearly with the number of device clusters (making this approach more suitable for large-scale implementation). Liang et al. proposed a RL controller to minimize the cost of scheduled smart appliances under time-varying prices, and compared it with a decentralized heuristic approach [71]. Kaliappan et al. used a RL algorithm to control a set of home appliances without exceeding a maximum power level defined by the electric feeder, and minimizing the discomfort caused by delaying the tasks of the appliances [72]. Liu et al. focused on reducing the electricity cost of shiftable loads [106]. Kim et al. used RL to minimize the dissatisfaction of the consumers, their energy bills, and the costs of the distribution system. In this study, appliances were modelled as generic loads in which the accumulated non-supplied demand generated dissatisfaction on the consumers [118]. Bahrami et al. used an actor-critic network to schedule several interruptible and non-interruptible loads [132]. Sheikhi et al. used Q-learning to shave the peak demand of a set of appliances (dishwasher and washing machine), in addition to EVs, and a combined heat and power unit. To help overcome the curse of dimensionality, they subdivided the problem into several sub-problems of RL that could be solved independently of each other [117,154]. Mahapatra et al. proposed a hardware architecture for the control of smart appliances using BRL [133]. Hurtado et al. also applied Q-learning for DR modelling generic loads and maximized both user comfort and demand flexibility [149].

4.1.3. Electric vehicles and hybrid electric vehicles

Dusparic et al. proposed a control system for the charging process of 9 EVs in a neighborhood in which all the houses were connected to the same transformer [68]. Their decentralized approach utilized three different policies intended to ensure the desired minimum battery charge, avoid overloading of the transformer, and charge the EVs during the periods of minimum load. Taylor et al. extended this work by achieving cooperation between the EVs transferring knowledge among them [78], and Marinescu et al. [93,126], and Dusparic et al. [94] continued this line of research with the implementation of a multi-agent RL control. Dauer et al. aimed to reduce the cost of charging a fleet of EVs, in a double-auction market while complying with the minimum state of the charge (SOC) [65]. Jiang et al. reduced the energy cost of a

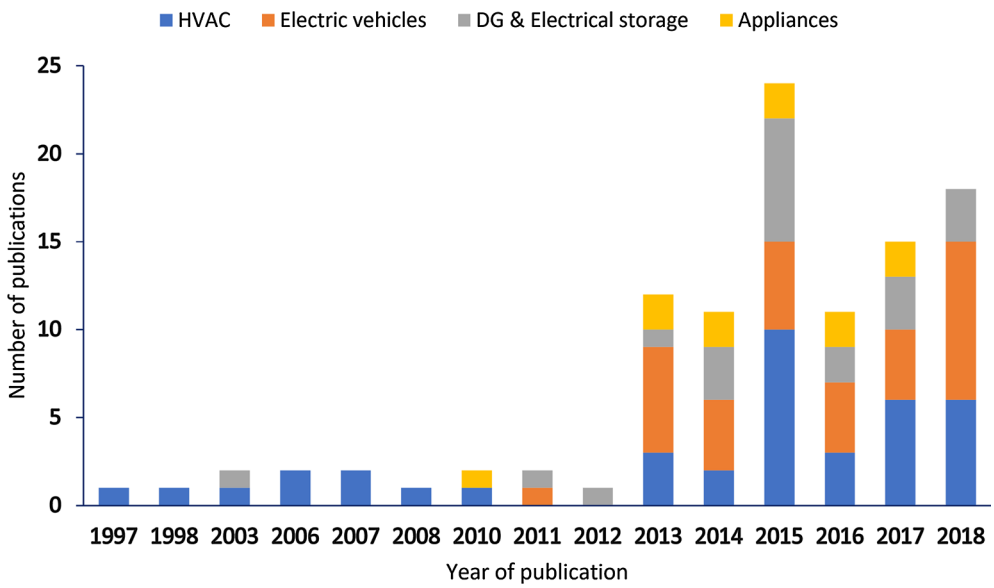

Fig. 4. Articles by topic and year of publication.

fleet of hybrid electric taxis while aiming to reduce the waiting lines at the charging stations in a multi-agent cooperative approach [145]. Di Giorgio et al. minimized the operational costs of EVs by trading electricity in a multi-agent setting without prior knowledge of the electricity prices [66]. Arif et al. investigated different scheduling strategies for EVs with different algorithms and electricity pricing structures [111]. Vandael et al. proposed a RL controller to help a fleet of EVs define a day-ahead consumption plan in order to minimize the cost of the electricity consumed [96].

Valoginanni et al. proposed a RL controller for the maximization of welfare, and energy cost savings [67]. Their controller learned from individual household consumption patterns, and they modelled in detail the consumer satisfaction functions associated with energy usage. Shi and Wong, Dimitrov and Lguensat, and Chis et al. used RL to maximize the profits achievable from EVs during their parking time by controlling the charge and discharge process [58,64,77,95]. In [75,97,112–114,127,128,141–144,146], various RL energy management systems were proposed to reduce the amount of fuel hybrid EVs consumed, while Xiong, Cao and Yu focused on the minimization of the energy losses [140]. Claessens et al. proposed a RL controller for a fleet of 160 EVs, many PV panels, and a cluster of different devices (i.e. non flexible loads and electric boilers) with local wind power generation [69]. The algorithm achieved a reduction of the effective peak power demand of about 50-60% with respect to a non-coordinated approach, and within a learning time of 100 days. Vaya et al. used RL to minimize the charging costs of several hundred thousand plug-in EVs without impacting their end-use [76]. Chis et al. minimized the cost of charging an individual EV using known day-ahead electricity prices and predicting next-day prices using a Bayesian ANN [125]. Ko et al. proposed a centralized controller for a fleet of EVs to minimize their energy cost and user dissatisfaction [147]. Xiong et al. optimized the energy consumption of a battery and a ultra-capacitor for EVs in all climates [148].

4.1.4. Distributed generation and electrical storage

In 2003, Henze and Dodier introduced the use of RL in order to control distributed generation resources: a PV array and a battery to supply a load with electricity at a minimum cost [50]. Berlink et al., and Guan et al. proposed a RL method to maximize the profit associated with the operation of a PV array and a storage devices, and they obtained experimental results [102,103]. In [102], the controller followed a different policy for each season of the year. Jiang and Fei, used Q-learning, in a simulation environment, to control the discharging policy of a battery and accounted for wind power generation [59,99]. In [60,70,80,82,98,100], various RL algorithms were investigated to maximize the economic savings of controlling the charge and discharge process of electricity storage devices. Sekizaki et al. minimized the user discomfort, and the operational costs of a water heater and a battery [101]. They implemented a classifier system, XCS, based on RL and generic algorithms [155]. Qiu et al., and Shi et al. minimized the amount of energy consumed from the electrical grid by controlling a battery under the presence of solar PV panels [115,130]. Mbuwir et al. investigated the use of BRL to maximize the self-consumption of a battery and a PV array on a building and they included a backup controller to guarantee an adequate functioning of the controller [129]. Li and Jayaweera proposed a centralized controller for a group of storage devices and distributed energy resources with the aim of maximizing the profits of the users [79,105]. Wang, Lin, and Pedram minimized the energy cost of a battery by managing its residual energy [116]. Kofinas et al. proposed a decentralized cooperative multi-agent RL algorithm to maximize the self-energy consumption in a microgrid [150], and Tan et al. used a centralized controller based on Q-learning to reduce the operational cost of an electric microgrid [151].

4.2. Case studies with physical systems

Case studies in which RL methods are tested in physical systems are useful to understand the effectiveness of these control systems and their limitations when implemented in the real-world. Most of the articles we reviewed included a case study in which the controller was implemented. However, only five studies tested a RL controller in a physical setting.

Mozer, in his Neural Network House experiment, used RL to control the HVAC, DHW, and lighting systems of an actual residence, a former three-room school house in Boulder [48] with the objective of adapting to the preferences and behaviour of its occupants. Liu and Henze implemented a hybrid RL control with prior simulated knowledge in a test facility [51,52], the Energy Resource Station, operated by the Iowa energy Center (IEC). They suggested that the hybrid RL with prior simulated knowledge is more suitable than the standard RL controller for its implementation in commercial buildings. Costanzo et al. implemented a model-assisted RL controller in an experimental setup consisting of A/C systems, several temperature sensors, a solar irradiation sensor, and a power meter [108]. Kazmi et al. implemented a data-driven and model-based RL controller in a set of 32 houses in the

Netherlands and reducing the energy consumption of their DHW systems by 20% [138]. Ruelens et al. reduced the operating cost of a water heating system, in an experimental setting, using autoencoders to extract meaningful features from raw sensor data and using them as states for their RL controller [137].

4.3. Reinforcement learning methods

In this section we analyze the various RL control methods based on their way of selecting the control actions and the algorithms they use to extract useful knowledge from the environment they interact with.

4.3.1. Action selection

RL algorithms are based on a trade-off between exploring new actions and exploiting those that seem to be optimal. RL is particularly advantageous over other algorithms under situations in which the agent must constantly adapt to a changing environment. As a result, exploration is an important component of RL algorithms, and in this section, we discuss the different action-selection methods that the reviewed articles apply. Most of the reviewed studies apply the ε-greedy policy due to its simplicity. As an example, Costanzo et al. used an adaptive ε-greedy policy whose ε-factor was reduced by half every 4 days [108]. They achieved a near optimal solution after 20 days of training. Soft-max exploration is the second most common approach, which transforms the Q-values into a probability distribution function that is used to select future actions.

In a different approach, Sun et al. investigated the use of a greedy policy to select the actions available within a Lagrangian relaxation framework [62,88]. Arif et al. [111], and Jiang et al. [145] performed an online action-selection process using a pursuit algorithm (PA). This algorithm explores the different actions by following a given probability density function (p.d.f) that is different than the Boltzmann's p.d.f. commonly used in the soft-max method. Li et al. investigated a probably approximately correct (PAC) algorithm [156] to perform the exploration phase [87]. This algorithm allows the agent to learn a near optimal policy within a polynomial time error bound. Zhang et al., in a multi-agent system, applied an ε-greedy policy combined with a cooperative swarm optimization method in order to find the Stackelberg equilibrium [157] among all the actions of the matrices with Q-values [131].

4.3.2. Learning algorithms

Table 3 contains a classification of the articles based on the variant of RL algorithm they use, and the energy systems they control. It is clear that Q-learning, introduced by Watkins and Dayan [39], is the most widely used, irrespective of the application area. Lee and Powell implemented a bias-corrected variant of Q-learning [60], while authors of [64,92,134] used Sarsa. Actor-critic methods [35] were used in [55,63,70,122,124,132], Bielskis et al. combined this method with a radial basis function network (RBFN) to represent the state-action space [63]. Al-Jabery et al. also made use of an actor-critic network and demonstrated that it achieved a better performance than the Q-learning controller [124]. Du and Fei also used an actor-critic controller with two ANNs and trained the critic network using Q-learning [55]. Fuselli et al. investigated the use of actor-critic networks using ANNs, and training them with Particle Swarm Optimization [70]. Bahrami et al. also used an actor-critic method, and demonstrated that, in their problem, it performed better than the Q-learning algorithm [132]. They also analyzed the sensitivity of the algorithm to a parameter that defined the trade-off between pursuing short-term and long-term rewards. Wei et al. developed a variant of the actor-critic algorithm that they called dual iterative Q-learning, and which used two critic networks [80,98]. Wang et al. combined the actor-critic method with a long-short term memory recurrent neural network to maximize comfort and minimize the energy consumption in a simulated office building [122]. Only Rayari, et al. [89], and Yuan et al. [146] used Monte-Carlo value iteration methods.

Other authors have also made use of eligibility traces [54,56,64,83,92,103,134,145], which leads to a faster convergence speed at the expense of increased computational complexity. Batch Reinforcement Learning (BRL) has been used by many authors, in combination with Fitted Q-Iteration. Different regression techniques have been used to approximate the Q-function. Authors of the articles [69,73,90,91,108,121,123,129,137] used extremely randomized trees, which performed better than support vector machines [69]. However, other algorithms such as ANNs [104], echo state neural networks [130], and Kernel-based approximation [95,125] have been used successfully in BRL with Fitted Q-Iteration. Other authors have also used ANNs and other regression algorithms to forecast future states. For instance, Chis et al. used a Bayesian ANN to predict electricity prices one day ahead [64,125].

Some authors have highlighted the importance of using prior knowledge to pre-train the RL algorithm offline before its online execution [50,56]. Liu and Henze investigated the effects of adding prior knowledge by using a hybrid simulation learning control [51,52]. They first trained the controller off-line by simulating a calibrated model of a HVAC system. Then, they implemented it in an experimental environment where the algorithm learned on-line. In a similar approach, De Gracia et al. developed a simplified isothermal model of their PCM ventilated façade, which they used to pre-train their controller under different weather forecasts and control policies before implementation [92]. Costanzo et al. implemented a data-driven approach with an ANN acting as a support model to generate virtual tuples of experience that they could use as prior knowledge to train their algorithm [108]. They implemented the controller in a living lab and achieved convergence within 20 days of training and with a performance within 90% of the mathematical optimum. Vázquez-Canteli et al. used historical control data from a rule-based controller to pre-train their reinforcement learning controller [120]. Zhang et al. used deep transfer Q-learning to transmit prior knowledge among buildings and achieve a faster learning rate [131].

Some researchers have also made use of feature extraction techniques to select the states of the system. Ruelens et al. used an autoencoder, a non-linear feature extraction technique based on an auto-associative ANN, to reduce the dimensionality of the state-space vector and increase the speed of convergence [90,137]. In a different study, Claessens et al. investigated the use of Convolutional Neural Networks for automatic state-time feature extraction for RL applied to residential load control [158]. Urieli and Stone used a different approach to overcome the curse of dimensionality. They explored the effects of the actions during 3 days and used a regression algorithm to fit a transition function that modeled the house [61]. Then, instead of estimating the Q-values through action-selection and value-iteration, they used a look-ahead method based on tree search to find the optimal policy through policy iteration.

Dusparic et al. [68,94], Taylor et al. [78], and Marinescu et al. [93,126] investigated the use of W-learning to coordinate several EVs competing against each other to minimize their energy costs. Zhu utilized a multi-player RL algorithm [159] to control multiple generic loads and minimize the electricity cost and user dissatisfaction [81]. The different agents are semi-honest adversaries who share information with some of the other agents, but also compete in the same market. Taylor et al. extended this work using a cooperative variant of W-learning in which the EVs exchanged knowledge among them [78]. Later, Marinescu et al. continued this work and compared this decentralized algorithm with other control methods [93].

5. Demand response modelling and dynamic response

In this section we focus on analyzing those articles that involved direct DR applications, particularly on how they model electricity prices, whether they have a stationary environment, and whether they

Table 3
Algorithms and systems controlled.

	HVAC & DHW	Appliances	EV & HEV	DG & Storage
Q-learning	[47,48,49,51,52,53,62,74,84, 85,86,87,88,101,110,124,135]	[57,71,72,106,107,117, 118]	[58,65,66,67,76,77,97,111, 112,113,114,117,127,128, 140,142,143,147,148]	[50,59,60,79,82,99, 100,101,102,105,115, 116,117,151]
Q(λ)	[56]	[83]	[145]	
Fuzzy Q-learning	[56]			[150]
Sarsa	[92,134]		[64]	
BRL	[73,90,91,104,108,109,119, 120,123,121,133,137,136,139]	[69]	[69,95,96,125,141,144]	[129,130]
TD(λ)	[54]			[103]
TD(0)			[75]	
W-learning			[68,78,93,94,126]	
Monte-Carlo	[89]		[146]	
Policy iteration	[61]			
Actor critic	[55,63,122,124]	[132]		[70]
Dual Q-learning				[80,98]
Multi-player RL		[81]		

adapt dynamically to renewable energy generation. RL methods only converge when the agents interact with a stationary environment. When the environment changes over time, the knowledge learnt by the agents might become inaccurate, and they need to acquire new knowledge to adapt to these changing conditions. In DR scenarios, there are various reasons why the environment might not be stationary: a difference between the actual and the expected demand, or a mismatch between the actual and the expected renewable energy generation. We refer to non-stationarity when the expectation of these variables under a given set of states and actions changes over time, or more formally, when the transition probabilities of the environment are not time-invariant. For example, agents might not be able to predict accurately a stochastic energy demand, but if the expectation of such demand does not vary over time, the environment will be stationary and stochastic, and the RL algorithm could potentially converge to an optimal stochastic policy.

A major reason for a DR environment to be non-stationary on the demand side is the presence of multiple consumer agents impacting the energy demand simultaneously with their actions. When multiple buildings participate in DR, delaying the energy consumption during the peak hours can produce another peak when the prices are expected to be low [32]. To avoid such rebound effects, households must provide a coordinated response, either centralized or distributed [160]. We discuss the articles that try to address this problem in a multi-agent approach. Fig. 5 depicts the relationships between the different pricing schemes and the possible control approaches. First, it is worth mentioning that a set of agents is not necessarily a multi-agent system: if the agents behave in complete independence of each other, the system is simply a set of multiple single-agents. For instance, a fleet of EVs in which each EV is controlled independently under demand-independent electricity prices, i.e., TOU rates is not considered a multi-agent control approach. The reason is that the actions taken by one EV does not have any impact on the price of electricity other EVs will pay, or on the decisions they will take. However, if we consider demand-dependent electricity prices, i.e., demand-dependent RTPs, the actions of any EV will have an impact on the price of electricity, and as a result on the decisions other EVs will take. In this latter case, the EVs or other types of agents can be controlled using either a centralized single-agent approach (there is one central agent that controls the EVs) or a multi-agent competitive or cooperative approach [161]. We have classified as single-agent controllers the centralized controllers of multiple agents in which a joint action is taken to maximize a joint Q-function [100,149]. The reason is that this is a way of solving a multi-agent problem with a single-agent controller, by allowing a single central agent to select all the actions.

Table 4 contains a classification of the articles based on how electricity prices were modelled. Deterministic prices are easier to learn by a RL controller if the conditions that affect electricity prices are defined as state variables. The reason is that, under the same conditions (i.e. day of the week, hour, demand), the prices are the same and behave as MDPs. On the other hand, stochastic prices have some added randomness and cannot be so accurately predicted by the RL controller. Another relevant factor in modelling electricity prices is whether they depend on the electricity demand of consumers. While under demand-independent prices RL agents can take control actions independently from each other, with demand-dependent prices they should act as a coordinated or competitive multi-agent system to achieve near optimal solutions.

In Table 5, we have classified the controllers used in the articles as a function of their agents and objectives. Although most articles consider the dissatisfaction or discomfort of the users, not all of them include them explicitly as objectives (through penalties or rewards), but rather as constraints.

Most articles focused on single-agents that take actions independent of each other. Although these approaches are valid for small scales, in which only a few buildings participate in DR events, the results from these studies have limitations regarding their usefulness in large-scale scenarios, with many buildings shifting their demand. Authors of [57,71,83,107] applied RL in a multi-objective approach to minimize the electricity cost of home appliances and the dissatisfaction caused by delaying the completion of the tasks. In [57,83,107], they assumed demand-independent RTPs, while the authors of [71] assumed demand-dependent deterministic prices and a decentralized control. Kaliappan et al. minimized the dissatisfaction caused by delaying the usage of appliances while satisfying a maximum level of total power consumption [72]. Claessens et al. implemented a mixed approach to control multiple thermal loads in a district heating network. They aggregated the thermal loads in clusters controlled by BRL as a single-agent problem [139]. Then, they used a proportional integral control to select the control actions of the resulting multi-agent system.

Di Giorgio et al. minimized the operational costs of EVs by buying and selling electricity under stochastic prices. However, EVs did not provide a coordinated response, and no bidding mechanisms were considered [66]. Shi and Wong assumed RTP rates for electricity, and tested their algorithm under both simulated and historical electricity prices [58]. The EVs did not only participate in DR by charging and discharging their batteries, but also in frequency regulation of the electrical grid. Only a few articles did not explicitly include as a penaltythe failure of sufficiently charging the EV in the reward function [58,64]. Instead, they limited the state-action space to avoid taking

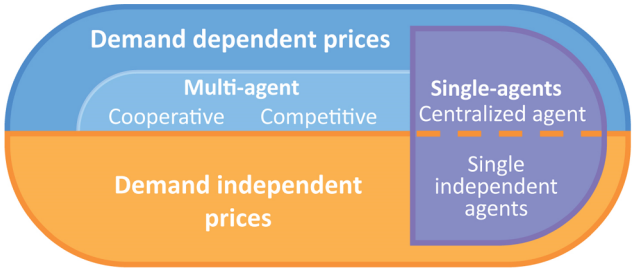

Fig. 5. Control approaches and electricity prices.

actions that would lead to failure in sufficiently charging the vehicles. Chis et al. combined day-ahead and two-day-ahead prices [64], and they updated their work to include a penalty accounting for insufficiently charging the battery [95]. Vandael et al. controlled the electricity purchased by an aggregator in the day-ahead market instead of the individual EVs [96].

5.1. Multi-agent approaches

As Fig. 5 illustrates, under demand dependent prices and multiple buildings, either a multi-agent or centralized needs to be implemented. As shown in Fig. 6, buildings can exchange information with the electrical grid, such as their demand and electricity prices, while they can potentially establish some degree of dynamic coordination among each other (competitive or cooperative), and/or even share information.

Dusparic et al. proposed a novel controller for a fleet of EVs. The prices were directly proportional to the total load and forecasted based on the current prices [68]. The 9 different agents were only aware of their own actions, but they received information on the current load of the transformer, which depended on the current actions of the rest of the agents. As a line for future research they proposed using the multi-agent variant of the W-learning to enable cooperation between the different vehicles [78], for which they developed a prediction-based multi-agent RL method (P-MARL) and implemented it in a non-stationary smart grid scenario. They evaluated P-MARL under situations where agents took decisions independently, simultaneously, and sequentially. A self-organizing map allowed P-MARL to detect changes in the patterns of the non-stationary environment and automatically adjust the prediction models with new data [126]. In [94], Dusparic et al. proposed and implemented this decentralized multi-agent approach not only to reduce or shift the peak demand, but also to directly maximize renewable energy use in a non-stationary environment with up to 90 households (subdivided in groups of 9) and simulated in GridLAB-D [162]. The results of the learning process stabilized after around 20 days of learning data that allowed the agents to successfully collaborate with each other.

Dauer et al. minimized the cost of electricity while meeting the minimum state of the charge (SOC) required to complete the trips [65]. They substituted independent electricity prices by a market platform to which the EVs could submit their bids. They suggested that providing the EVs with the capability of selling surplus stored electricity would be an interesting path to future research. In a multi-agent competitive approach, Claessens et al. implemented a controller for EVs which submitted bids to a virtual auction market [69]. Although the bids represented the priority assigned to the energy consumption at different times, no explicit electricity prices were involved.

Vaya et al. reduced the cost of charging a fleet of EVs using a competitive multi-agent and distributed approach based on a bidding algorithm that learned by interaction with the market. Their RL algorithm was used to define some of the parameters that were then used to calculate the bids submitted to the market [76]. They tested the controller with historical market data from the European Energy Exchange and under realistic traffic patterns. To ensure scalability, they aggregated the bids of the different agents, and took the bidding decisions locally. Although they proved the scalability of their approach, they did not analyze the sensitivity of the performance to the volatility of the market prices and the driving patterns, which remains an open question. Jiang et al. implemented a multi-agent cooperative control to reduce energy costs, the state of the charge at the beginning of each charging period, the distance for driving to charging stations, and the waiting lines to charge hybrid electric taxis [145]. They used demand independent TOU rates.

Sun et al. coordinated several HVAC systems in a double-auction market [86]. They simulated their model in GridLAB [162] and compared the results with empirical data from the AEP gridSMART smart grid demonstration project. They demonstrated that their simulated model-free bidding strategy using RL was a good approximation of the model-based strategy followed by the agents in the AEP gridSMART project.

Kim et al. implemented a competitive multi-agent RL controller for residential appliances with demand-dependent electricity prices [118]. Every consumer used RL to decide when to consume, and the electricity prices were set by a service operator which used RL too. Electricity prices depend on the actions of the consumers and the service provider. Therefore, the different agents are all interconnected through the electricity prices. Bahrami et al. also accounted for stochastic and demand-dependent electricity prices. They focused on minimizing the electricity costs of several buildings, in which each user is aware that the energy consumption of other users influences the electricity prices [132]. Zhang et al. modelled a multi-agent system of several supply and demand agents as a cooperative Stackelberg game in which one of the agents chooses an action first, and shares its information with other agents which select their optimal actions thereafter [131].

Vázquez-Canteli et al. implemented a multi-agent deep RL controller for the control of HVAC systems in multiple buildings in a competitive environment [136]. Buildings share the same electricity prices, which increased with demand. Therefore, buildings learned to coordinate with each other to avoid simultaneous energy consumption. This framework was implemented merging CitySim, a building energy simulator, with TensorFlow. Their future work will focus on implementing improved RL controllers on a larger scale.

6. Discussion

6.1. Research trends and open questions

We have observed a significant increase in the number of publications involving RL after 2012, with a spike in 2015, a year after the publication of the paper "*Playing Atari with Deep Reinforcement Learning*" [163]. This increase in the number of publications has been particularly noticeable in the field of RL applied to EVs, which was almost unresearched before 2013. After 2013, there was also a significant

Table 4
Modelling of electricity prices.

Electricity price	Demand independent	Demand dependent
Deterministic	[49,51,52,53,59,62,73,75,88,89,91,101,102,103,111,116,124,125,137,139,145,147,151]	[65,68,71,131,132,136]
Stochastic	[49,57,58,60,64,66,67,70 77,80,82,83,91,95,96,98,99,102,106,107,108,117,121]	[76,79,81,86,105,118,132]

Table 5
Agents and objectives of RL techniques.

Agents	Objectives					
	Energy consumption	Energy consumption & satisfaction	Energy cost	Energy cost & satisfaction	Peak demand	Peak demand & satisfaction
Single-agent	[92,115,120,123,129, 140,148]	[54,56,61,63,72,84,85,90, 109,110,119,122,122,134, 135,138]	[50,49,51,52,53,58,59,60, 64,67,70,73,75,77,79,80, 82,89,91,96,98,99,102, 103,106,108,114,116,121, 124,125,137,139,151]	[57,62,66,71,83,88, 95,101,107,111,147]	[100,133]	[74,139]
Multi-agent	Cooperative	[150]		[131]	[145]	[78,93,94,126]
	Competitive		[104]	[136]	[65,68,76,81,86,118,132]	[69]

increase of publications in the areas of RL applied to smart appliances, and DG combined with energy storage systems (i.e. solar PV and batteries).

In the field of smart appliances, only a few articles modelled the appliances with a high level of detail, and often included estimated dissatisfaction functions that represented human feedback [57,83,107]. Further research is required to incorporate actual human feedback, either training surrogate models representing humans or implementing RL in a testing facility where occupants can provide their feedback and RL can adapt accordingly [164]. Similarly, more RL algorithms should be tested in physical systems and not only in simulated environments in fields such as HVAC control and EVs. One of the reasons that might be preventing building owners and building energy managers from implementing RL controllers in actual commercial or residential buildings is the lack of experimental results in physical systems proving its capabilities and more importantly, its reliability.

Among the articles that focused on controlling energy systems in the built environment, e.g., HVAC systems, the focus was typically on the behavior of a single building. Thus, analyzing how a RL controller can adapt to changing urban conditions is still an open field for future research. Some of these changing conditions may include building refurbishment policies, addition of new buildings or other shadowing objects, and urban heat island effects.

Most of the studies that focused on DR did not account for a large-scale implementation of the algorithms. As an increasing amount of energy agents participate in DR, electricity prices stop being exogeneous factors and start depending more on the actions of the agents. This creates a moving target problem, in which the transition probabilities of the environment in the RL problem are non-stationary. Multi-agent RL, although still a field in its infancy, seems a promising solution to this problem that should be further researched.

In summary, some potential paths for future research that we have identified are:

- Testing RL control algorithms in physical systems, particularly analyzing their reliability, learning speed and adaptability
- Testing RL control algorithms with actual human feedback
- Studying the adaptation of RL to building retrofit, urban growth, RES integration, or other changing factors at the building level or the urban scale
- Multi-agent RL for DR, particularly analyzing its reliability, economic performance, and scalability

6.2. Algorithms

One of the major challenges that the reviewed studies have identified is the curse of dimensionality, i.e., when the state-action space of the RL controller is very large, and consequently the learning speeds/progress decrease dramatically. Some approaches such as batch RL with fitted Q-iteration use function approximation techniques to address this problem. Costanzo et al., and Claessens et al. highlighted the need for further investigating the pre-processing of data to automatically extract the most relevant features and reduce the dimensionality of the state-space (i.e. autoencoders) [108,139]. However, very few studies have applied feature extraction or dimensionality reduction techniques to minimize the curse of dimensionality [90,91]. This is still an open path for future research and with a great potential for improvement.

The policy iteration method is another technique that has barely been explored in this field [61], and constitutes an important path to future research as well. This algorithm, combined with non-linear regression techniques, fits a transition function of the system to control, and can overcome the curse of dimensionality. Furthermore, data augmentation algorithms could be helpful in improving the datasets used in the training process and overcome the curse of dimensionality. Generative Adversarial Neural Networks (GANNs) could potentially serve this purpose [165,166].

Reinforcement learning is most advantageous when applied in a complex environment due to its model-free nature. However, training the algorithm with some prior experience can reduce learning times significantly. Some studies have developed simplified models of their systems to train their controllers with simulated experience. Other studies used data-driven models to generate this experience, such as Costanzo et al., who used an ANN as a data-driven support model [108], or Urieli and Stone, who used regression [61]. Although they do not take full advantage of the RL model-free nature, these methods that make use of surrogate models should be further investigated as they show great potential for practical applications in terms of reliability and speed of convergence.

Research in RL algorithms is making progress at a fast pace, and new state of the art algorithms should be tested as they can increase speed of convergence, reduce the amount of data required for training, and help overcome the curse of dimensionality. State of the art off-policy algorithms, such as the Deep Deterministic Policy Gradient (DDPG) [167], should be tested in more case studies. DDPG is particularly advantageous for systems for which historical data is available. State of the art on-policy methods, such as the Proximal Policy Optimization (PPO) [168], have demonstrated to be particularly advantageous in systems for which a simplified model is available. Both DDPG and PPO allow the implementation of stochastic action-selection, which is necessary for systems in which the optimal control policy is stochastic. For example, in a city-scale system with multiple buildings, each of which can store or release additional energy at a given time, the optimal control policy could be every building storing or releasing additional energy with a given probability, and not necessarily in a deterministic manner.

In summary, RL is a quickly evolving field and new state of the art algorithms could help solve problems that were previously difficult to tackle. In addition to testing state of the art RL algorithms, we think there are some areas that should be further explored:

- Automated feature extraction techniques to preprocess the data and extract states in an unsupervised way
- Exploring the use some prior knowledge of the environment to

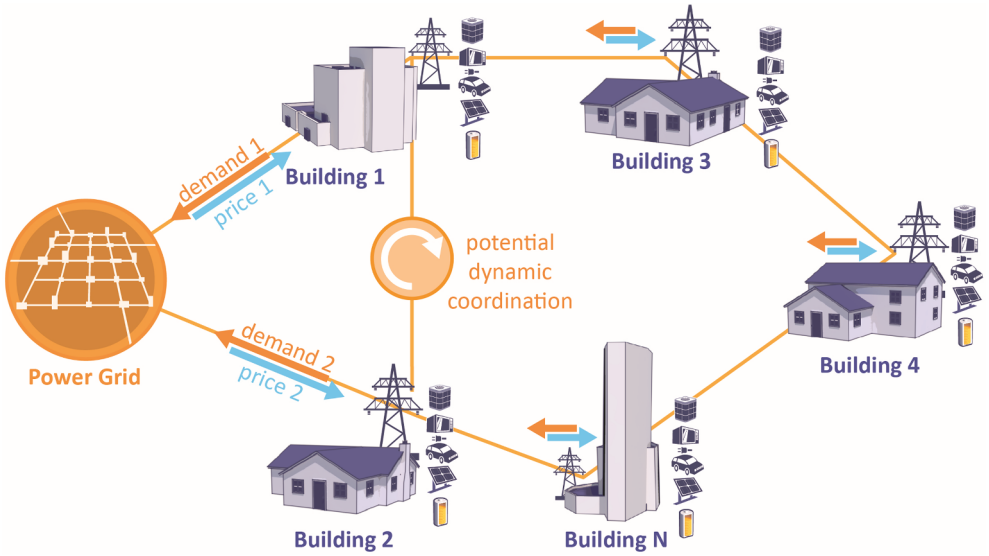

Fig. 6. Multi-agent coordination for demand response.

improve speed of convergence and reliability while maintaining the ability to adapt
- Exploring the use of data augmentation algorithms to improve the datasets used in the training process
- Researching the use of stochastic control policies in problems that might benefit from it, such as multi-agent RL for DR

It is important to note that the RL algorithms to be used should be chosen according to the difficulty of the problem being tackled. Not every problem requires the use of all these techniques, or of "deep learning", but sometimes simpler algorithms are enough to solve a given problem.

6.3. Multi-agent reinforcement learning: an adaptive and scalable solution?

In a multi-agent scenario, electricity prices would not be exogenous, but dependent of the total demand for electricity. This would require a multi-agent controller that avoids the peak demand by shifting instead of shaving. Bahrami et al. addressed this problem using a controller based on an actor-critic network [132]. In this study, the different agents were aware of the interdependencies of their actions due to the existence of demand-dependent electricity prices. Hurtado et al. implemented a joint action learning controller based on Q-learning in which the selection of the actions is performed by a single central agent that knows all the Q-values [149]. Therefore, they approach the multi-agent problem by implementing a central single-agent that takes all the control decisions. Future lines of work should focus on implementing similar approaches to their work in multi-agent scenarios with several households participating, as suggested by Wen et al., and Kaliappan et al. [72,83,107].

Some studies have focused on multi-agent competitive approaches in which, for instance, the agents submit their bids in an auction market. However, as many authors indicate, further research is needed to coordinate agents in cooperative ways. Urieli and Stone [61] highlighted that an important path to expand their work on HVAC control would be to test their controller in a multi-agent smart grid environment in which the agents would affect the energy prices as they try to minimize their own costs. Additionally, testing a RL cooperative and multi-agent controller in an experimental environment with real human feedback would constitute an important breakthrough in this field.

Some of the articles that focused on RL applied to EVs have investigated multi-agent competitive approaches [65,68,69,76,81,86,93,118,132,136]. Many of these studies simulated competitive energy markets to which the different agents can submit their bids, with the exception of Taylor et al. [78], Marinescu et al. [93,126], Dusparic et al. [94], and Jiang et al. [145], who investigated a cooperative controller based on W-learning in which the agents shared information to achieve the common goal of peak demand minimization.

We have also observed a trend towards the use of RL algorithms that make use of ANNs or deep learning techniques, i.e., BRL with fitted Q-iteration. One of the most recent ones uses deep transfer Q-learning among electricity consumers and generators playing a cooperative sequential Stackelberg game in the smart grid [131]. As highlighted by Windham and Treado, distributed and cooperative intelligent multi-agent control systems constitute a disruptive technology if implemented in the smart grid and the built environment [160]. It would allow device manufacturers to embed their intelligent control systems into the equipment from the factory, and build an efficient, scalable and smart infrastructure incrementally. Although there is a trend towards these multi-agent cooperative systems, and many researchers have highlighted their importance, there is still a need for further research and development in this area.

There are various reasons why the controlled environments can have non-stationary transition probabilities. One of them is the non-stationary nature of energy demand, electricity prices or renewable energy generation, which could have changing transition probabilities. Real-time electricity prices based on wholesale prices or the amount of renewable energy generation in the power system can follow these non-stationary processes. To a greater extent, multi-agent coordination (competitive or cooperative), can make the transition probabilities of the environment that is observed by every agent highly non-stationary [126]. Fig. 7 illustrates in a Venn diagram, the number of articles that addressed problems involving either multi-agent coordination, non-stationary transition probabilities, demand dependent electricity prices, or integration of RES. In general, the articles which belong to multiple categories, address problems that are more complex and non-stationary. All multi-agent problems are non-stationary, while not all the non-stationary environments require multi-agent coordination. DR scenarios that used historical real-time electricity prices fall under the category of non-stationary transition probabilities, as they do not necessarily follow pure MDPs. Some articles substituted the objective of cost reduction with demand dependent electricity prices by a peak shaving objective, which can also require multi-agent coordination. We have found that there is still little research combining multi-agent coordination, and integration of RES under non-stationary environments. Only six articles fell under this category, which focused on EVs and appliances [69], EVs [94,126], HVAC systems

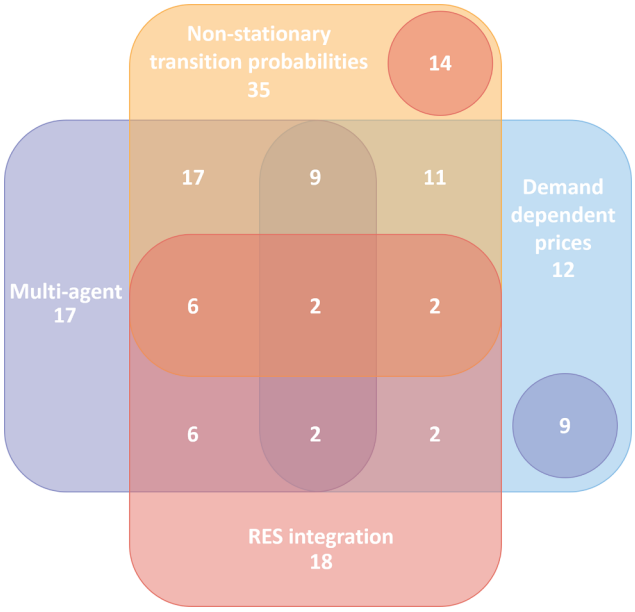

Fig. 7. Venn diagram of the number of reviewed articles that address either multi-agent systems, non-stationary transition probabilities in the environment, demand dependent electricity prices, or integrate renewable energy sources. As it can be seen, very few papers discuss simultaneously multi-agent coordination, and integration of RES under non-stationary environments.

[136], DG and storage [150], and generic electrical loads [131].

6.4. Reinforcement for demand response: a proposed framework

As many authors test their algorithms in diverse case studies under various conditions, readers might find it difficult to understand the relevance of some results, how to reproduce them, or the complexity of the problem being addressed. As a general framework for future research, we suggest that, in addition to defining the basic information for RL, i.e., the states, actions, rewards, and the type of action-selection algorithm used in the RL problem, authors should clearly state whether:

1. The states, actions and rewards are deterministic or stochastic
2. The transition probabilities of the environment are stationary (whether they remain constant during the simulation or experiment), and which state transitions, if any, are non-stationary
3. Agents act independently, simultaneously, or sequentially on the environment [126]
4. Electricity prices are demand dependent or independent
5. The control algorithm is model-free or model-based, off-policy or on-policy
6. Any predictions are used as states, and how accurate they are
7. The algorithm was tested in a physical system or in a simulated environment
8. Actual human feedback was used

Following this basic framework will improve the understanding of the diverse problems being tackled, increase the reproducibility of the results, and highlight for which type of problems certain RL algorithms perform better than others.

7. Conclusion

The future development and implementation of demand response greatly depends on incorporating user feedback and consumption patterns in its control loop. Reinforcement learning is a learning control algorithm that has the potential to achieve this. This article reviewed the research developments concerning the use of reinforcement learning for demand response applications. We analyzed 105 research articles from the fields of HVAC, electric vehicles, home appliances, distributed generation, and electrical energy storage. Although many articles have considered human comfort and satisfaction as part of the control problem, most of them have investigated single-agent systems in which electricity prices do not depend on the energy demand. When electricity prices depend on the demand for electricity, peak demand may be shifted instead of shaved. Therefore, there is a need to further explore the applicability of reinforcement learning in multi-agent systems, which can participate in demand response.

We have found that RL control algorithms have been tested in physical systems in only a small fraction of the articles we have reviewed. Therefore, in order to prove the reliability and adaptability of RL algorithms, more real-world experiments need to be conducted with state of the art RL methods. The use of data augmentation, dimensionality reduction or feature extraction techniques have barely been explored, and the use of data-driven support models should be further investigated too. Future research should also address quantification of how reinforcement learning adapts to varying urban conditions such as building refurbishment, population increase, addition/removal of buildings, or urban heat island effects. We have also observed gaps in the literature regarding the use of RL methods for lighting control, and the use of RL with actual human feedback and interaction in real-world systems.

Finally, we observed that most of the studies are not easily reproducible, and it is rather challenging to compare the performance of the controllers. The main reason for this is that the studies often address similar problems but with different physical properties or dynamics. Therefore, further standardization is needed in both the investigated control problems, and the used methods and simulation tools. Such standardized control problems, as well as integrated software tools that include both building simulation and machine learning features, can help researchers investigate their control approaches, and compare them directly to other approaches. We have also proposed a basic framework to help in this standardization.

References

[1] UNEP. Buildings and climate change, summary for decision-makers; 2009.
[2] Nejat P, Jomehzadeh F, Taheri MM, Gohari M, Muhd MZ. A global review of energy consumption, CO_2 emissions and policy in the residential sector (with an overview of the top ten CO_2 emitting countries). Renew Sustain Energy Rev 2015;43:843–62. https://doi.org/10.1016/j.rser.2014.11.066.
[3] Chourabi H, Nam T, Walker S, Gil-Garcia JR, Mellouli S, Nahon K, et al. Understanding smart cities: an integrative framework. Proc Annu Hawaii Int Conf Syst Sci 2011. p. 2289–97. https://doi.org/10.1109/HICSS.2012.615.
[4] Leibowicz BD, Lanham CM, Brozynski MT, Vázquez-Canteli JR, Castillo N, Nagy Z. Optimal decarbonization pathways for urban residential building energy services. Appl Energy 2018;230:1311–25. https://doi.org/10.1016/j.apenergy.2018.09.046.
[5] Dupont B, Dietrich K, De Jonghe C, Ramos A, Belmans R. Impact of residential demand response on power system operation: a Belgian case study. Appl Energy 2014;122:1–10. https://doi.org/10.1016/j.apenergy.2014.02.022.
[6] Siano P. Demand response and smart grids – a survey. Renew Sustain Energy Rev 2014;30:461–78. https://doi.org/10.1016/j.rser.2013.10.022.
[7] IEA. Transition to sustainable buildings; 2013. http://doi.org/10.1787/9789264202955-en.
[8] Bruninx K, Patteeuw D, Delarue E, Helsen L, D'Haeseleer W. Short-term demand response of flexible electric heating systems: the need for integrated simulations. Int Conf Eur Energy Mark EEM 2013. p. 28–30. https://doi.org/10.1109/EEM.2013.6607333.
[9] McNeil MA, Letschert VE. Future air conditioning energy consumption in developing countries and what can be done about it: the potential of efficiency in the residential sector; 2008.
[10] Mohagheghi S, Stoupis J, Wang Z, Li Z. Demand response architecture-integration into the distribution management system. SmartGridComm 2010:501–6.
[11] Shoreh MH, Siano P, Shafie-khah M, Loia V, Catalão JPS. A survey of industrial applications of demand response. Electr Power Syst Res 2016;141:31–49. https://doi.org/10.1016/j.epsr.2016.07.008.
[12] Federal Energy Regulatory Commission. Assessment of demand response and advanced metering. Staff Report 2016;74:240. https://doi.org/10.1017/CBO9781107415324.004.
[13] Centolella P, Farber-DeAnda M, Greening LA, Kim T. Estimates of the value of

uninterrupted service for the mid-west independent system operator; 2010.
[14] Wang J, Zhong H, Ma Z, Xia Q, Kang C. Review and prospect of integrated demand response in the multi-energy system. Appl Energy 2017;202:772–82. https://doi.org/10.1016/j.apenergy.2017.05.150.
[15] Zeng B, Wu G, Wang J, Zhang J, Zeng M. Impact of behavior-driven demand response on supply adequacy in smart distribution systems. Appl Energy 2017;202:125–37. https://doi.org/10.1016/j.apenergy.2017.05.098.
[16] Park JY, Nagy Z. Comprehensive analysis of the relationship between thermal comfort and building control research – a data-driven literature review. Renew Sustain Energy Rev 2018;82:2664–79. https://doi.org/10.1016/j.rser.2017.09.102.
[17] Aghaei J, Alizadeh MI. Demand response in smart electricity grids equipped with renewable energy sources: a review. Renew Sustain Energy Rev 2013;18:64–72. https://doi.org/10.1016/j.rser.2012.09.019.
[18] Batchu R, Pindoriya NM. Residential demand response algorithms: state-of-the-art, key issues and challenges. Lect Notes Inst Comput Sci Soc Telecommun Eng LNICST 2015. p. 18–32. https://doi.org/10.1007/978-3-319-25479-1.
[19] Law YW, Alpcan T, Lee VCS, Lo A, Marusic S, Palaniswami M. Demand response architectures and load management algorithms for energy-efficient power grids: a survey. Proc – 2012 7th Int Conf Knowledge, Inf Creat Support Syst KICSS 2012. p. 134–41. https://doi.org/10.1109/KICSS.2012.45.
[20] Vardakas JS, Zorba N, Verikoukis CV, Member S. A survey on demand response programs in smart grids: pricing methods and optimization algorithms. Ieee Commun Surv Tutorials. 2015. p. 152–78.
[21] Li X, Wen J. Review of building energy modeling for control and operation. Renew Sustain Energy Rev 2014;37:517–37. https://doi.org/10.1016/j.rser.2014.05.056.
[22] Yu Z, Huang G, Haghighat F, Li H, Zhang G. Control strategies for integration of thermal energy storage into buildings: state-of-the-art review. Energy Build 2015;106:203–15. https://doi.org/10.1016/j.enbuild.2015.05.038.
[23] Wang S, Ma Z. Supervisory and optimal control of building HVAC systems: a review. HVAC&R Res 2007;14:3–32. https://doi.org/10.1080/10789669.2008.10390991.
[24] Salehizadeh MR, Soltaniyan S. Application of fuzzy Q-learning for electricity market modeling by considering renewable power penetration. Renew Sustain Energy Rev 2016;56:1172–81. https://doi.org/10.1016/j.rser.2015.12.020.
[25] Dusparic I, Taylor A, Marinescu A, Golpayegani F, Clarke S. Residential demand response: experimental evaluation and comparison of self-organizing techniques. Renew Sustain Energy Rev 2017;80:1528–36. https://doi.org/10.1016/j.rser.2017.07.033.
[26] Shariatzadeh F, Mandal P, Srivastava AK. Demand response for sustainable energy systems: a review, application and implementation strategy. Renew Sustain Energy Rev 2015;45:343–50. https://doi.org/10.1016/j.rser.2015.01.062.
[27] Dupont B, De Jonghe C, Olmos L, Belmans R. Demand response with locational dynamic pricing to support the integration of renewables. Energy Policy 2014;67:344–54. https://doi.org/10.1016/j.enpol.2013.12.058.
[28] Nguyen DT, Nguyen HT, Member S, Le LB, Member S. Dynamic pricing design for demand response integration in power distribution Networks. IEEE Trans Power Syst 2016;31:3457–72.
[29] Action N, Efficiency E. Coordination of energy efficiency and demand response. Analysis 2010:1–75.
[30] Venkatesan N, Solanki J, Solanki SK. Residential Demand Response model and impact on voltage profile and losses of an electric distribution network. Appl Energy 2012;96:84–91. https://doi.org/10.1016/j.apenergy.2011.12.076.
[31] Hussain I, Mohsin S, Basit A, Khan ZA, Qasim U, Javaid N. A review on demand response: pricing, optimization, and appliance scheduling. Procedia Comput Sci 2015;52:843–50. https://doi.org/10.1016/j.procs.2015.05.141.
[32] Gelazanskas L, Gamage KAA. Demand side management in smart grid: A review and proposals for future direction. Sustain Cities Soc 2014;11:22–30. https://doi.org/10.1016/j.scs.2013.11.001.
[33] Summit Blue Consulting L. Evaluation of the 2006 Energy-Smart Pricing Plan. Final Report; 2007. p. 1–15.
[34] Herter K, McAuliffe P, Rosenfeld A. An exploratory analysis of California residential customer response to critical peak pricing of electricity. Energy 2007;32:25–34. https://doi.org/10.1016/j.energy.2006.01.014.
[35] Sutton R, Barto A. Reinforcement learning: an introduction. Massachusetts: MIT Press Cambridge; 1998.
[36] Littman ML, Dean TD, Kaelbling LP. On the complexity of solving Markov decision problems. Proc Elev Conf Uncertain Artif Intell 1995. p. 394–402. https://doi.org/10.1007/11871842.
[37] Abbeel P, Coates A, Quigley M, Ng AY. An application of reinforcement learning to aerobatic helicopter flight. Adv Neural Inf Process Syst 2007;19:1.
[38] Huys QJM, Cruickshank A, Seriès P. Reward-based learning, model-based and model-free quentin. Encycl Comput Neurosci 2014:1–10. https://doi.org/10.1007/978-1-4614-7320-6.
[39] Watkins C, Dayan P. Technical note: Q-learning. Mach Learn 1992;8:279–92. https://doi.org/10.1023/A:1022676722315.
[40] Peng J, Williams RJ. Incremental multi-step Q-learning. Mach Learn 1996;22:283–90. https://doi.org/10.1007/BF00114731.
[41] Gullapalli V. A stochastic reinforcement learning algorithm for learning real-valued functions. Neural Netw 1990;3:671–92. https://doi.org/10.1016/0893-6080(90)90056-Q.
[42] Ernst D, Geurts P, Wehenkel L. Iteratively extending time horizon reinforcement learning. Mach Learn ECML 2003 14th Eur Conf Mach Learn, vol. 14. 2003. p. 96–107. https://doi.org/10.1007/b100702.
[43] Kalyanakrishnan S, Stone P, Liu Y. Batch Reinforcement Learning in a Complex Domain. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2008;5001 LNAI:171–83. http://doi.org/10.1007/978-3-540-68847-1_15.
[44] Singh S, Jaakkola T, Jordan M. Learning without state-estimation in partially observable markov decision processes. Intl Conf Mach Learn. 1994.
[45] Tuyls K, Weiss G. Multiagent learning: basics, challenges, and prospects. AI Mag 2012;33:41–52. https://doi.org/10.1609/aimag.v33i3.2426.
[46] Action Humphrys M. Selection methods using reinforcement learning. University of Cambridge; 1997.
[47] Anderson CW, Hittle DC, Katz AD, Kretchmar RM. Synthesis of reinforcement learning, neural networks and PI control applied to a simulated heating coil. Artif Intell Eng 1997;11:421–9. doi: 0.1.1.43.3643.
[48] Mozer MC. The neural network house: an environment that adapts to its inhabitants. Am Assoc Artif Intell Spring Symp Intell Environ. 1998. p. 110–4. SS-98-02/SS98-02-017.
[49] Henze GP, Schoenmann J. Evaluation of reinforcement learning control for thermal energy storage systems. HVAC&R Res 2003;9:259–75. https://doi.org/10.1080/10789669.2003.10391069.
[50] Henze GP, Dodier RH. Adaptive optimal control of a grid-independent photovoltaic system. vol. 125; 2003. p. 34–42. http://doi.org/10.1115/1.1532005.
[51] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. Theoretical foundation. Energy Build 2006;38:142–7. https://doi.org/10.1016/j.enbuild.2005.06.002.
[52] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: results and analysis. Energy Build 2006;38:148–61. https://doi.org/10.1016/j.enbuild.2005.06.001.
[53] Liu S, Henze GP. Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory. J Sol Energy Eng 2007;129:215. https://doi.org/10.1115/1.2710491.
[54] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, Stavrakakis GS. Reinforcement learning for energy conservation and comfort in buildings. Build Environ 2007;42:2686–98. https://doi.org/10.1016/j.buildenv.2006.07.010.
[55] Du D, Fei M. A two-layer networked learning control system using actor-critic neural network. Appl Math Comput 2008;205:26–36. https://doi.org/10.1016/j.amc.2008.05.062.
[56] Yu Z, Dexter A. Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. Control Eng Pract 2010;18:532–9. https://doi.org/10.1016/j.conengprac.2010.01.018.
[57] O'Neill D, Levorato M, Goldsmith A, Mitra U. Residential demand response using reinforcement learning. First IEEE Int Conf Smart Grid Commun 2010 2010. p. 409–14. https://doi.org/10.1109/SMARTGRID.2010.5622078.
[58] Shi W, Wong VWS. Real-time vehicle-to-grid control algorithm under price uncertainty. 2011 IEEE Int Conf Smart Grid Commun SmartGridComm 2011 2011. p. 261–6. https://doi.org/10.1109/SmartGridComm.2011.6102330.
[59] Jiang B, Fei Y. Dynamic residential demand response and distributed generation management in smart microgrid with hierarchical agents. Energy Procedia 2011;12:76–90. https://doi.org/10.1016/j.egypro.2011.10.012.
[60] Lee D, Powell W. An intelligent battery controller using bias-corrected Q-learning. Twenty-Sixth AAAI Conf Artif Intell, vol. 1. 2012. p. 316–22.
[61] Urieli D, Stone P. A learning agent for heat-pump thermostat control. Proc. 12th Int'l Conf Auton. Agents Multiagent Syst. (AAMAS 2013). 2013. p. 1093–100.
[62] Sun B, Luh PB, Jia QS, Yan B. Event-based optimization with non-stationary uncertainties to save energy costs of HVAC systems in buildings. IEEE Int'l Conf Autom. Sci. Eng.. 2013.
[63] Bielskis AA, Guseinoviene E, Drungilas D, Gricius G, Zulkas E. Modelling of ambient comfort affect reward based adaptive laboratory climate controller. Elektron Ir Elektrotechnika 2013;19:79–82. https://doi.org/10.5755/j01.eee.19.8.5399.
[64] Chis A, Lunden J, Koivunen V. Scheduling of plug-in electric vehicle battery charging with price prediction. 2013 4th IEEE/PES Innov Smart Grid Technol Eur ISGT Eur 2013 2013. p. 1–5. https://doi.org/10.1109/ISGTEurope.2013.6695263.
[65] Dauer D, Flath CM, Ströhle P, Weinhardt C. Market-based EV charging coordination. 2013 IEEE/WIC/ACM Int Conf Intell Agent Technol IAT 2013, vol. 2. 2013. p. 102–7. https://doi.org/10.1109/WI-IAT.2013.97.
[66] Di Giorgio A, Liberati F, Pietrabissa A. On-board stochastic control of electric vehicle recharging. 52nd IEEE Conf Decis Control 2013. p. 5710–5. https://doi.org/10.1109/CDC.2013.6760789.
[67] Valogianni K, Ketter W, Collins J. Smart charging of electric vehicles using reinforcement learning. AAAI Work. Trading Agent Des. 2013. p. 41–8.
[68] Dusparic IC, Harris A, Marinescu V, Cahill S. Clarke Multi-agent residential demand response based on load forecasting. Technol Sustain (SusTech), 2013 1st IEEE Conf 2013. p. 90–6. https://doi.org/10.1109/SusTech.2013.6617303.
[69] Claessens BJ, Vandael S, Ruelens F, De Craemer K, Beusen B. Peak shaving of a heterogeneous cluster of residential flexibility carriers using reinforcement learning. 2013 4th IEEE/PES Innov Smart Grid Technol Eur ISGT Eur 2013 2013. p. 1–5. https://doi.org/10.1109/ISGTEurope.2013.6695254.
[70] Fuselli D, De Angelis F, Boaro M, Squartini S, Wei Q, Liu D, et al. Action dependent heuristic dynamic programming for home energy resource scheduling. Int J Electr Power Energy Syst 2013;48:148–60. https://doi.org/10.1016/j.ijepes.2012.11.023.
[71] Liang Y, He L, Cao X, Shen Z. Stochastic control for smart grid users with flexible demand. IEEE Trans Smart Grid 2013;4:2296–308. https://doi.org/10.1109/TSG.2013.2263201.
[72] Kaliappan AT, Sathiakumar S. Parameswaran N. Flexible power consumption management using Q learning techniques in a smart home. CEAT 2013–2013 IEEE Conf Clean Energy Technol 2013. p. 342–7. https://doi.org/10.1109/CEAT.2013.

6775653.
[73] Ruelens F, Claessens BJ, Vandael S, Iacovella S, Vingerhoets P, Belmans R. Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. Proc – 2014 Power Syst Comput Conf PSCC 2014 2014. https://doi.org/10.1109/PSCC.2014.7038106.
[74] Al-jabery K, Wunsch DC, Xiong J, Shi Y. A novel grid load management technique using electric water heaters and Q-learning. 2014 IEEE Int Conf Smart Grid Commun. 2014. p. 776–81.
[75] Liu C, Murphey YL. Power management for plug-in hybrid electric vehicles using reinforcement learning with trip information. IEEE Transp Electrif Conf Expo 2014 2014. p. 1–6. https://doi.org/10.1109/ITEC.2014.6861862.
[76] Vayá MG, Roselló LB, Andersson G. Optimal bidding of plug-in electric vehicles in a market-based control setup. Proc – 2014 Power Syst Comput Conf PSCC 2014 2014. https://doi.org/10.1109/PSCC.2014.7038108.
[77] Dimitrov S, Lguensat R. Reinforcement learning based algorithm for the maximization of EV charging station revenue. Proc Int Conf Math Comput Sci Ind 2014. p. 235–9. https://doi.org/10.1109/MCSI.2014.54.
[78] Taylor A, Dusparic I, Galvan-Lopez E, Clarke S, Cahill V. Accelerating learning in multi-objective systems through transfer learning. Proc Int Jt Conf Neural Networks 2014. p. 2298–305. https://doi.org/10.1109/IJCNN.2014.6889438.
[79] Li D, Jayaweera SK. Reinforcement learning aided smart-home decision-making in an interactive smart grid. 2014 IEEE Green Energy Syst. Conf. IGESC 2014 2014. p. 1–6. https://doi.org/10.1109/IGESC.2014.7018632.
[80] Wei Q, Liu D, Shi G, Liu Y, Guan Q. Optimal self-learning battery control in smart residential grids by iterative Q-learning algorithm. IEEE SSCI 2014 - 2014 IEEE Symp Ser Comput Intell – ADPRL 2014 2014 IEEE Symp Adapt Dyn Program Reinf Learn Proc 2014. https://doi.org/10.1109/ADPRL.2014.7010630.
[81] Zhu M. Distributed demand response algorithms against semi-honest adversaries; 2014. p. 0–4.
[82] Zhang Y, van der Schaar M. Structure-aware stochastic load management in smart grids. INFOCOM, 2014 Proc IEEE 2014. p. 2643–51. https://doi.org/10.1109/INFOCOM.2014.6848212.
[83] Wen Z, O'Neill D, Maei H. Optimal demand response using device-based reinforcement learning. IEEE Trans Smart Grid 2015;6:2312–24. https://doi.org/10.1109/TSG.2015.2396993.
[84] Barrett E, Linder S. Autonomous HVAC control, a reinforcement learning approach. Eur Conf Mach Learn Princ Knowl Discov ECML 2015, vol. 2. 2015. p. 3–19. https://doi.org/10.1007/978-3-319-23461-8.
[85] Li B, Xia L. A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings. IEEE Int Conf Autom Sci Eng CASE 2015 2015. p. 444–9. https://doi.org/10.1109/CoASE.2015.7294119.
[86] Sun Y, Somani A, Carroll TE. Learning based bidding strategy for HVAC systems in double auction retail energy markets 2015. Am. Control Conf., vol. 2015–July IEEE; 2015. p. 2912–7. https://doi.org/10.1109/ACC.2015.7171177.
[87] Li D, Zhao D, Zhu Y, Xia Z. Thermal comfort control based on MEC algorithm for HV AC systems; 2015.
[88] Sun B, Luh PB, Jia QS, Yan B. Event-based optimization within the lagrangian relaxation framework for energy savings in HVAC systems. IEEE Trans Autom Sci Eng 2015;12:1396–406. https://doi.org/10.1109/TASE.2015.2455419.
[89] Rayati M, Sheikhi A, Ranjbar AM. Applying reinforcement learning method to optimize an Energy Hub operation in the smart grid. Innov Smart Grid Technol Conf (ISGT), 2015 IEEE Power Energy Soc 2015. p. 1–5. https://doi.org/10.1109/ISGT.2015.7131906.
[90] Ruelens F, Iacovella S, Claessens BJ, Belmans R. Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. Energies 2015;8:8300–18. https://doi.org/10.3390/en8088300.
[91] Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuška R, Belmans R. Residential Demand response applications using batch reinforcement learning; 2015.
[92] De Gracia A, Fernández C, Castell A, Mateu C, Cabeza LF. Control of a PCM ventilated facade using reinforcement learning techniques. Energy Build 2015;106:234–42. https://doi.org/10.1016/j.enbuild.2015.06.045.
[93] Marinescu A, Dusparic I, Taylor A, Canili V, Clarke S. P-MARL: prediction-based multi-agent reinforcement learning for non-stationary environments. Proc Int Jt Conf Auton Agents Multiagent Syst AAMAS, vol. 3. 2015. p. 1897–8.
[94] Dusparic I, Taylor A, Marinescu A, Cahill V, Clarke S. Maximizing renewable energy use with decentralized residential demand response; 2015.
[95] Chis Adriana, Lunden Jarmo, Koivunen Visa. Optimization of plug-in electric vehicle charging with forecasted price. 40th IEEE Int Conf Acoust Speech Signal Process ICASSP 2015. 2015. p. 2086–9.
[96] Vandael S, Claessens B, Ernst D, Holvoet T, Deconinck G. Reinforcement learning of heuristic EV fleet charging in a day-ahead electricity market. IEEE Trans Smart Grid 2015;6:1795–805. https://doi.org/10.1109/TSG.2015.2393059.
[97] Qi X, Wu G, Boriboonsomsin K, Barth MJ. A novel blended real-time energy management strategy for plug-in hybrid electric vehicle commute trips. IEEE Conf Intell Transp Syst Proceedings, ITSC 2015. p. 1002–7. https://doi.org/10.1109/ITSC.2015.167.
[98] Wei Q, Liu D, Shi G. A novel dual iterative Q-learning method for optimal battery management in smart residential environments. IEEE Trans Ind Electron 2015;62:2509–18. https://doi.org/10.1109/TIE.2014.2361485.
[99] Jiang B, Fei Y. Smart home in smart microgrid: a cost-effective energy ecosystem with intelligent hierarchical agents. IEEE Trans Smart Grid 2015;6:3–13. https://doi.org/10.1109/TSG.2014.2347043.
[100] Raju L, Sankar S, Milton RS. Distributed optimization of solar micro-grid using multi agent reinforcement learning. Procedia – Procedia Comput Sci 2015;46:231–9. https://doi.org/10.1016/j.procs.2015.02.016.
[101] Sekizaki Shinya, Tomohiro Hayashida IN. An intelligent home energy management system with classifier system. 2015 IEEE 8th Int Work. Comput. Intell. Appl. 2015. p. 1–8.
[102] Berlink H, Kagan N, Reali Costa AH. Intelligent decision-making for smart home energy management. J Intell Robot Syst Theory Appl 2015;80:331–54. https://doi.org/10.1007/s10846-014-0169-8.
[103] Guan C, Wang Y, Lin X, Nazarian S, Pedram M. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. 2015 12th Annu IEEE Consum Commun Netw Conf CCNC 2015 2015. p. 637–42. https://doi.org/10.1109/CCNC.2015.7158054.
[104] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Appl Energy 2015;156:577–86.
[105] Li D, Jayaweera SK. Machine-learning aided optimal customer decisions for an interactive smart grid. IEEE Syst J 2015;9:1529–40. https://doi.org/10.1109/JSYST.2014.2334637.
[106] Liu Y, Yuen C, Ul Hassan N, Huang S, Yu R, Xie S. Electricity cost minimization for a microgrid with distributed energy resource under different information availability. IEEE Trans Ind Electron 2015;62:2571–83. https://doi.org/10.1109/TIE.2014.2371780.
[107] Wen Z, O'Neill D, Maei HR. Optimal demand response using device-based reinforcement learning. IEEE Trans Smart Grid 2015;6:1–23. https://doi.org/10.1109/TSG.2015.2396993.
[108] Costanzo GT, Iacovella S, Ruelens F, Leurs T, Claessens BJ. Experimental analysis of data-driven control for a building heating system. Sustain Energy, Grids Netw 2016;6:81–90. https://doi.org/10.1016/j.segan.2016.02.002.
[109] Kazmi H, D'Oca S, Delmastro C, Lodeweyckx S, Corgnati SP. Generalizable occupant-driven optimization model for domestic hot water production in NZEB. Appl Energy 2016;175:1–15. https://doi.org/10.1016/j.apenergy.2016.04.108.
[110] Cheng Z, Zhao Q, Wang F, Jiang Y, Xia L, Ding J. Satisfaction based Q-learning for integrated lighting and blind control. Energy Build 2016;127:43–55. https://doi.org/10.1016/j.enbuild.2016.05.067.
[111] Arif AI, Babar M, Ahamed TPI, Al-Ammar EA, Nguyen PH, Kamphuis IGR, et al. Online scheduling of plug-in vehicles in dynamic pricing schemes. Sustain Energy, Grids Netw 2016;7:25–36. https://doi.org/10.1016/j.segan.2016.05.001.
[112] Qi X, Wu G, Boriboonsomsin K, Barth MJ, Gonder J. Data-driven reinforcement learning-based real-time energy management system for plug-in hybrid electric vehicles. Transp Res Rec J Transp Res Board 2016;2572:1–8. https://doi.org/10.3141/2572-01.
[113] Zou Y, Liu T, Liu D, Sun F. Reinforcement learning-based real-time energy management for a hybrid tracked vehicle. Appl Energy 2016;171:372–82. https://doi.org/10.1016/j.apenergy.2016.03.082.
[114] Qi X, Wu G, Boriboonsomsin K, Barth MJ, Gonder J. Data-driven reinforcement learning based real-time energy management system for plug-in hybrid electric vehicles. Transp Res Rec J Transp Res Board 2016;2572:1–8. https://doi.org/10.3141/2572-01.
[115] Qiu X, Nguyen TA, Crow ML. Heterogeneous energy storage optimization for microgrids. IEEE Trans Smart Grid 2016;7:1453–61. https://doi.org/10.1109/TSG.2015.2461134.
[116] Wang Y, Member S, Lin X, Member S, Pedram M, Integrating A. A near-optimal model-based control algorithm for households equipped with residential photovoltaic power generation and energy storage systems. IEEE Trans Sustain Energy 2016;7:1–10. https://doi.org/10.1109/TSTE.2015.2467190.
[117] Sheikhi A, Rayati M, Ranjbar AM. Demand side management for a residential customer in multi energy systems. Sustain Cities Soc 2016;22:63–5.
[118] Kim BG, Zhang Y, Van Der Schaar M, Lee JW. Dynamic pricing and energy consumption scheduling with reinforcement learning. IEEE Trans Smart Grid 2016. https://doi.org/10.1109/TSG.2015.2495145.
[119] Schmidt M, Moreno MV, Schülke A, Macek K, Mařík K, Pastor AG. Optimizing legacy building operation: the evolution into data-driven predictive cyber-physical systems. Energy Build 2017;148:257–79. https://doi.org/10.1016/j.enbuild.2017.05.002.
[120] Vázquez-Canteli J, Kämpf J, Nagy Z. Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. Energy Procedia 2017;122:415–20. https://doi.org/10.1016/j.egypro.2017.07.429.
[121] Ruelens F, Claessens BJ, Vandael S, De Schutter B, Babuska R, Belmans R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. IEEE Trans Smart Grid 2017;8:2149–59. https://doi.org/10.1109/TSG.2016.2517211.
[122] Wang Y, Velswamy K, Huang B. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 2017;5:46. https://doi.org/10.3390/pr5030046.
[123] De Somer O, Soares A, Kuijpers T, Vossen K, Vanthournout K, Spiessens F. Using reinforcement learning for demand response of domestic hot water buffers: a real-life demonstration; 2017. p. 1–6.
[124] Al-jabery K, Xu Z, Yu W, Wunsch D, Xiong J, Shi Y. Demand-side management of domestic electric water heaters using approximate dynamic programming. IEEE Trans Comput Des Integr Circuits Syst, vol. 36. 2017. https://doi.org/10.1109/TCAD.2016.2598563. 1–1.
[125] Chiş A, Lundén J, Koivunen V. Reinforcement learning-based plug-in electric vehicle charging with forecasted price. IEEE Trans Veh Technol 2017;66:3674–84. https://doi.org/10.1109/TVT.2016.2603536.
[126] Marinescu A, Dusparic I, An S. Prediction-based multi-agent reinforcement learning in inherently r r r 2017;12.
[127] Kong Z, Zou Y, Liu T. Implementation of real-time energy management strategy based on reinforcement learning for hybrid electric vehicles and simulation validation. PLoS ONE 2017;12:1–17. https://doi.org/10.1371/journal.pone.0180491.

[128] Liu T, Hu X, Li SE, Cao D. Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle. IEEE/ASME Trans Mechatronics, vol. 22. 2017. p. 1497–507. https://doi.org/10.1109/TMECH.2017.2707338.
[129] Mbuwir B, Ruelens F, Spiessens F, Deconinck G. Battery energy management in a microgrid using batch reinforcement learning. Energies 2017;10:1846. https://doi.org/10.3390/en10111846.
[130] Shi G, Liu D, Wei Q. Echo state network-based Q-learning method for optimal battery control of offices combined with renewable energy. IET Control Theory Appl 2017;11:915–22. https://doi.org/10.1049/iet-cta.2016.0653.
[131] Zhang X, Bao T, Yu T, Yang B, Han C. Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid. Energy 2017;133:348–65. https://doi.org/10.1016/j.energy.2017.05.114.
[132] Bahrami S, Wong VWS, Huang J. An online learning algorithm for demand response in smart grid. 1 1 IEEE Trans Smart Grid 2017;3053. https://doi.org/10.1109/TSG.2017.2667599.
[133] Mahapatra C, Moharana A, Leung V. Energy management in smart cities based on internet of things: peak demand reduction and energy savings. Sensors 2017;17:2812. https://doi.org/10.3390/s17122812.
[134] Brusey J, Hintea D, Gaura E, Beloe N. Reinforcement learning-based thermal comfort control for vehicle cabins R. Mechatronics 2018;50:413–21. https://doi.org/10.1016/j.mechatronics.2017.04.010.
[135] Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy Build 2018;169:195–205. https://doi.org/10.1016/j.enbuild.2018.03.051.
[136] Vázquez-Canteli JR, Ulyanin S, Kämpf J, Nagy Z. Fusing tensorflow with building energy simulation for intelligent energy management in smart cities. Sustain Cities Soc 2019. in press.
[137] Ruelens F, Claessens BJ, Quaiyum S, De Schutter B, Babuška R, Belmans R. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Trans Smart Grid 2018;9:3792–800. https://doi.org/10.1109/TSG.2016.2640184.
[138] Kazmi H, Mehmood F, Lodeweyckx S, Driesen J. Gigawatt-hour scale savings on a budget of zero: deep reinforcement learning based optimal control of hot water systems. Energy 2018;144:159–68. https://doi.org/10.1016/j.energy.2017.12.019.
[139] Claessens BJ, Vanhoudt D, Desmedt J, Ruelens F. Model-free control of thermostatically controlled loads connected to a district heating network. Energy Build 2018;159:1–10. https://doi.org/10.1016/j.enbuild.2017.08.052.
[140] Xiong R, Cao J, Yu Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Appl Energy 2018;211:538–48. https://doi.org/10.1016/j.apenergy.2017.11.072.
[141] Hu Y, Li W, Xu K, Zahid T, Qin F, Li C. Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning. Appl Sci 2018:8. https://doi.org/10.3390/app8020187.
[142] Liu T, Hu X. A Bi-level control for energy efficiency improvement of a hybrid tracked vehicle. IEEE Trans Ind Inform 2018;14:1616–25.
[143] Liu T, Wang B, Yang C. Online Markov Chain-based energy management for a hybrid tracked vehicle with speedy Q-learning. Energy 2018;160:544–55. https://doi.org/10.1016/j.energy.2018.07.022.
[144] Wu J, He H, Peng J, Li Y, Li Z. Continuous reinforcement learning of energy management with deep Q network for a power split hybrid electric bus. Appl Energy 2018;222:799–811. https://doi.org/10.1016/j.apenergy.2018.03.104.
[145] Jiang CX, Jing ZX, Cui XR, Ji TY, Wu QH. Multiple agents and reinforcement learning for modelling charging loads of electric taxis. Appl Energy 2018;222:158–68. https://doi.org/10.1016/j.apenergy.2018.03.164.
[146] Yuan J, Yang L, Chen Q. Intelligent energy management strategy based on hierarchical approximate global optimization for plug-in fuel cell hybrid electric vehicles. Int J Hydrogen Energy 2018;43:8063–78. https://doi.org/10.1016/j.ijhydene.2018.03.033.
[147] Ko H, Pack S, Member S, Leung VCM. Mobility-aware vehicle-to-grid control algorithm in microgrids. IEEE Trans Intell Transp Syst 2018;19:2165–74. https://doi.org/10.1109/TITS.2018.2816935.
[148] Xiong R, Duan Y, Cao J, Yu Q. Battery and ultracapacitor in-the-loop approach to validate a real-time power management method for an all-climate electric vehicle. Appl Energy 2018;217:153–65. https://doi.org/10.1016/j.apenergy.2018.02.128.
[149] Hurtado LA, Mocanu E, Nguyen PH, Gibescu M, Kamphuis IG. Enabling cooperative behavior for building demand response based on extended joint action learning. 1 1 IEEE Trans Ind Informatics 2018;3203. https://doi.org/10.1109/TII.2017.2753408.
[150] Kofinas P, Dounis AI, Vouros GA. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. Appl Energy 2018;219:53–67. https://doi.org/10.1016/j.apenergy.2018.03.017.
[151] Tan Z, Zhang X, Xie B, Wang D, Liu B, Yu T. Fast learning optimiser for real-time optimal energy management of a grid-connected microgrid. IET Gener Transm Distrib 2018. https://doi.org/10.1049/iet-gtd.2017.1983.
[152] Vazquez-Canteli JR, Kämpf J. Massive 3D models and physical data for building simulation at the urban scale : a focus on Geneva and climate change scenarios. WIT Trans Ecol Environ 2016;204.. https://doi.org/10.2495/SC160041.
[153] Walter E, Kämpf JH. A verification of CitySim results using the BESTEST and monitored consumption values. Proc 2nd Build Simul Appl Conf. 2015. p. 215–22.
[154] Sheikhi A, Rayati M, Ranjbar AM. Dynamic load management for a residential customer; Reinforcement Learning approach. Sustain Cities Soc 2016;24:42–51. https://doi.org/10.1016/j.scs.2016.04.001.
[155] Wilson SW. Classifier fitness based on accuracy. Evol Comput 1995;3:149–75. https://doi.org/10.1162/evco.1995.3.2.149.
[156] Strehl AL, Wiewiora E, Langford J, Littman ML. PAC Model-free reinforcement learning; 2006.
[157] Heinrich von Stackelberg Kolev S. Market structure and equilibrium. J Hist Econ Thought 2016:38.
[158] Claessens BJ, Vrancx P, Ruelens F. convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. IEEE Trans Smart Grid 2016:1–12. https://doi.org/10.1109/TSG.2016.2629450.
[159] Zhu M, Martinez S. Distributed coverage games for energy-aware mobile sensor networks. SIAM J Control Optim 2013;51:1–27. https://doi.org/10.1137/100784163.
[160] Windham A, Treado S. A review of multi-agent systems concepts and research related to building HVAC control. Sci Technol Built Environ 2016;22:50–66. https://doi.org/10.1080/23744731.2015.1074851.
[161] Multi-Agent Tan M. Reinforcement learning: independent vs. cooperative agents. Proc Tenth Int Conf Mach Learn. 1993. p. 330–7.
[162] Chassin DP, Fuller JC, Djilali N. GridLAB-D: an agent-based simulation framework for smart grids. J Appl Math 2014;2014:1–12. https://doi.org/10.1155/2014/492320.
[163] Mnih Volodymyr, Kavukcuoglu Koray, Silver David, Graves Alex, Antonoglou Ioannis, Wierstra Daan Riedmiller Martin. Playing Atari with deep reinforcement learning; 2013. p. 1–9. https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf.
[164] Park JY, Dougherty T, Fritz H, Nagy Z. LightLearn : An adaptive and occupant centered controller for lighting based on reinforcement learning. Build Environ 2019;147:397–414. https://doi.org/10.1016/j.buildenv.2018.10.028.
[165] Goodfellow IJ, Pouget-abadie J, Mirza M, Xu B, Warde-farley D. Generative adversarial nets; n.d. p. 1–9.
[166] Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks; 2018. p. 1–14.
[167] Silver D, Lever G, Technologies D, Lever GUY, Ac UCL. Deterministic Policy Gradient Algorithms; n.d.
[168] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms; n.d. p. 1–12.