Designing a generalised reward for Building Energy Management Reinforcement Learning agents

Rubén Mulero Martínez
TECNALIA,
Basque Research and Technology Alliance (BRTA)
Derio, Spain
ruben.mulero@tecnalia.com

Beñat Arregi Goikolea
TECNALIA,
Basque Research and Technology Alliance (BRTA)
Derio, Spain
benat.arregi@tecnalia.com

Iñigo Mendialdua Beitia
Department of Computer Languages and Systems
University of the Basque Country (UPV/EHU)
Donostia, Spain
inigo.mendialdua@ehu.eus

Roberto Garay Martínez
TECNALIA,
Basque Research and Technology Alliance (BRTA)
Derio, Spain
roberto.garay@tecnalia.com

***Abstract*—The reduction of the carbon footprint of buildings is a challenging task, partly due to the conflicting goals of maximising occupant comfort and minimising energy consumption. An intelligent management of Heating, Ventilation and Air Conditioning (HVAC) systems is creating a promising research line in which the creation of suitable algorithms could reduce energy consumption maintaining occupants' comfort. In this regard, Reinforcement Learning (RL) approaches are giving a good balance between data requirements and intelligent operations to control building systems. However, there is a gap concerning how to create a generalised reward signal that can train RL agents without delimiting the problem to a specific or controlled scenario. To tackle it, an analysis and discussion is presented about the necessary requirements for the creation of generalist rewards, with the objective of laying the foundations that allow the creation of generalist intelligent agents for building energy management.**

***Index Terms*—reinforcement learning, reward, generalised, building, energy efficiency, HVAC**

I. Introduction

Building are the most widespread infrastructures created by human beings. The evolution of these infrastructures is directly linked to the advancement in different fields of engineering, from civil to thermal engineering. One of the most important advancements is the inclusion of active elements within the buildings to enhance the comfort levels of occupants, e.g. heating and cooling systems, air recycling systems, heat pumps, light sensors, movement sensors, ... Moreover, the addition of new inter-connected elements [1] thanks to the *Internet of Things* (IoT) paradigm [2] is paving new ways in which the automation of the active elements of the building based on the needs of its occupants is closer than ever. In this regard, the deployment of different interconnected solutions [3] aims to improve the quality of life of the occupants by fulfilling their basic requirements.

Nevertheless, these advancements are increasing the energy requirement of buildings, which causes an increase in greenhouse gas emissions and the *carbon footprint* [4]. For that reason, the reduction of the energy consumption in buildings has become an open challenge for society, public policy and researchers. The aim is the creation of Net Zero Energy (NZE) [5] buildings capable of producing the energy required to maintain their installed active systems without using auxiliary sources. However, this assumption is far from reality due to the current technical or structural limitations which require the help of auxiliary systems to support the building demand.

Artificial Intelligence (AI) is becoming a promising approach in creating Building Management Systems (BMS) in which a *Machine Learning* (ML) algorithm improves the usage of the installed HVAC systems by reducing energy consumption through an intelligent management. These AI systems are fed by data streams gathered by different deployed IoT sensors and labelled by the interpretation of different experts in the field. The main goal is to create a multi-agent management system [6] in which different deployed ML algorithms will be able to interact between physical installed systems and predictive consumption demands to adjust the installed active elements. The creation of suitable ML algorithms is a challenging task because they need to be *trained* using accurate data extracted from the deployed sensors of the building. However, the acquisition of high quality data is complex due to the lack of a good data repository.

For this reason, the usage of Reinforcement Learning (RL) [7], as one of the main three ML paradigms, is becoming a promising approach to train intelligent agents with the ability of creating intelligent decision-making processes to control active systems. These algorithms are based on a trial and error process and use a *reward* signal to learn if the actions performed by the agent are correct or not with the objective of adjusting their future actions. Nonetheless, the creation of these reward functions is not straightforward because it requires a deep understanding of both RL and building research fields to potentially define a clear formulation of the problem to be solved.

In this paper a series of requirements are analysed to support the design of a reward formula that allows the creation of generalist intelligent agents for HVAC systems. The goal is to give some ideas to create a *generalist* reward function that works in different scenarios (e.g. different buildings). Figure 1 depicts an schematic flowchart of the paper development to have a better place of the authors contribution. The left section states the current implementations by several authors whereas the right section states our proposed approach. The remainder of this paper is organised as follows. Section II gives background on what expert rules are and how RL is used for control management in buildings. Section III formulates the reward designs and the considerations to be taken. Finally Section IV discusses and summarises the general conclusions of this work.

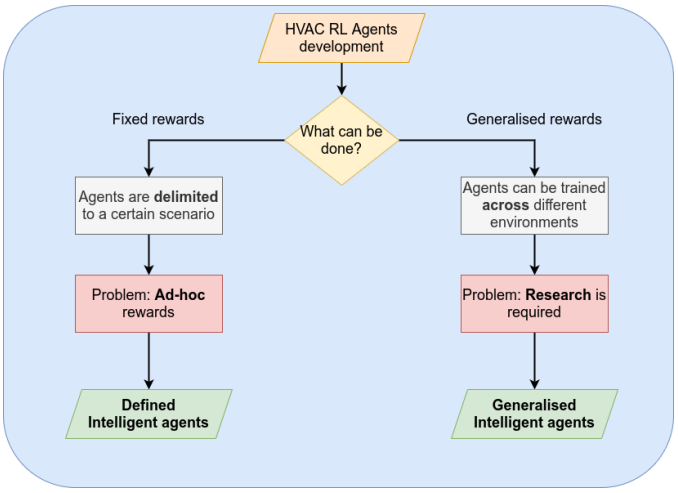

Fig. 1. Schematic flowchart of the paper development

II. State of the art

A. AI and Expert Rule Systems

Traditional approaches rely on using Expert Rule Systems [8] or *rule-based* to govern their operations. An expert in the field configures a set of actions that are triggered when a set of different logical conditions are met. These actions make a direct control in the target system to modify its operational state. The aim of these rule-based systems is to computationally emulate the decision-making process of a human expert to allow to the target machine perform intelligent actions to solve a fixed problem.

Some authors used rule-based systems as the backbone of their systems to perform intelligent control management of different systems [9], [10] or to detect faults or problems in the system life cycle [11]. Likewise, some of these works have been applied to real world buildings [12], [13]. However, even if the created systems are accurate and provide a good performance, the requirement of having an expert in the field to create the needed rules becomes an expensive and a time-consuming task. In some scenarios, the expert needs to create very complicated rules to cover all possible operational cases, which makes this approach unfeasible in some cases.

Artificial Intelligent (AI) solutions are demonstrating the capacity of making intelligent actions to solve several problems. AI solutions rely on using Machine Learning (ML) algorithms by making a *training* process using real (*data-driven*) or simulated (*model-driven*) data from a set of sensors, soft-sensors or simulation environments. Cotrufo et al. [14] tested five ML algorithms to effectively manage electric heaters and avoid the usage of a gas boiler. To do so, they created a model predictive control by training five ML algorithms and compared the obtained results against the set-back strategy (called BAU), to study how ML algorithms are capable of reducing the operation of the gas boiler.

There are some exceptions in which authors make a direct combination of AI and rule-based system to enhance the results of the model actions and avoid undesirable solutions. These solutions try to avoid some common problems like moisture-related problems, occupant comfort problems or to prevent disorders like Legionnaire's disease. Rule-based system acts as an added value to improve the results given by the control algorithms. For example in [15], authors defined an intelligent thermostat which used a trained AI agent. If for some reason the decision-making process of the AI agent did not guarantee the comfort ratio levels of the building occupants, a set-back strategy based on rule-based system was executed to ensure that comfort levels were met.

B. Reinforcement Learning and its growth

In recent years RL is receiving more attention in the literature, and an indication of that is that several reviews have recently been published [16], [17]. As a consequence, RL is being applied to solve different kind of problems such as robotics [18], economics [19] or computer games [20]. Likewise, it is also being applied to control the energy systems of buildings [21], [22].

Yuan et al. [23] proposed to use RL for air-conditioning system optimisation. They presented a new rule-assisted RL-based control method where a fitted Q-iteration algorithm was selected as the batch RL algorithm. Instead of using a simulator, rules were used to optimise the operation of the air-conditioning system. They evaluated the proposed method on a single-storey office building and they compared it with rule-based and proportional-integral-derivative (PID) strategies. Results showed that their proposal got a better comfort level than the compared methods and consumed less energy.

Raman et al. [24] proposed to use Zap Q-learning algorithm for climate control of a commercial building. In this case, besides the temperature comfort, the model was trained to maintain the humidity comfort while minimising energy consumption. They showed that the proposed controller was able to maintain temperature and humidity within the comfort limits. However, the energy saving was smaller compared to a Model Predictive Control (MPC) method proposed by the same authors in a previous work.

Wei et al. [25] proposed to use Deep Q-learning to control building systems. To approximate the Q-value they used a similar neural network structure to the one used in [26]. They proposed two alternatives: 1) to use a single neural network to control the building; 2) to train a neural network for each zone of the building. They compared their proposals with a rule-based approach and Q-learning. The results showed that their proposals maintained the temperature within the desired range and consumed less energy than the baseline methods. The saving was greater when a neural network was trained for each zone of the building. Lissa et al. [27] relied on that same neural network to estimate the Q-values of their deep reinforcement learning algorithm model to manage and control the heating system and the domestic hot water. They proposed a new methodology to define the threshold of indoor temperature comfort, based on the historical temperature of the building. In an experimental study, they showed that their method achieved the desired comfort for indoor temperature and domestic hot water and saved more energy than a rule-based system.

Zhang et al. [28] used asynchronous advantage actor-critic (A3C) to control a radiant heating system in an office building. To train the model, instead of using indoor air temperature as a proxy for thermal comfort, they proposed to use occupants' thermal preferences obtained from a smartphone app. They trained the deep reinforcement model over a simulator and then tested it in a real-life office building. Same authors extended their work in [29] by doing a direct calibration of the presented energy model with a dataset created by their experimental platform to minimise the gap between simulation and real environment. In both works, the obtained results outperforms the results obtained by rule-based approaches concluding that their proposals reduced the heating demand.

Wang el al. [30] implemented an actor-critic-based RL controller applied to the building HVAC control. They used the LSTM recurrent neural network for both policy and value representation. They tested their proposal in a simulated office building and results showed that the proposal maintained the comfort and saved energy.

III. Reward design

In order to maintain a healthy and comfortable indoor climate in buildings, adverse weather conditions are compensated by HVAC systems. Such devices consume thermal and/or electrical energy, involving both a financial expenditure (cost of fuel or energy carrier) and an environmental penalty (linked to resource depletion and greenhouse gas emissions). The reward can be regarded as an explicit formulation of a universal problem in buildings: the trade-off between occupant comfort and energy consumption. The reward function is thus formulated as a combination of an occupant comfort term and an energy consumption term. The present study aims at the formulation of a generalised reward that remains applicable for a wide range of climates, building types and use schedules.

A. Occupant comfort term

The thermal comfort of humans at indoor spaces depends on many parameters governing the heat balance with the environment, such as ambient and radiant temperatures, relative humidity, air speed, metabolic rate, physical activity and clothing. It is ultimately a subjective metric that varies among different cultures and individuals. An extensive literature exists on the topic of thermal comfort. The Predicted Mean Vote (PMV), originally envisaged by Fanger and widely discussed by many later authors [31], aims at predicting the average thermal sensation of a sample group of people as a function of many of the above parameters. The calculation leads to a Predicted Percentage of Dissatisfied (PPD) among healthy adults in air-conditioned buildings, which can also be used as a performance indicator. Still today, the PMV/PPD model remains the reference criterion of international [32], European [33] and North American [34] standards for assessing building performance. The more recently developed adaptive comfort model [35] states that comfort criteria are variable in time, considering both physiological and psychological adaptation. As users set their clothing and expectations in view of the outdoor climate, users might be able to tolerate more relaxed indoor comfort conditions for winter and summer, especially in absence of mechanical cooling. The adaptive control model is now accepted in most standards as an alternative method, and its implementation might help in reducing building energy intensity [36].

For the purpose of the reward, the comfort term requires the definition of a function expressing the occupant comfort as a function of measurable environmental conditions. For practicality, indoor air temperature is often selected as the sole reference parameter for HVAC control, because it is both very influential and easy to measure. More advanced systems estimate an *operative temperature* from additional readings of surface temperatures, relative humidity or air speed. While thermostat-based systems are typically based on a Boolean value (either *within* or *outside* a given comfort range), the declaration of a smooth comfort function would be more realistic and potentially more advantageous for training RL algorithms.

The comfort assessment procedure described above should only apply while the building is occupied. The comfort term should be eliminated from the reward function in the absence of occupancy. Additionally, this comfort term can also be attenuated in the event of low occupancy. For long unoccupied periods, a setback temperature can be advisable for preventing moisture-related problems such as frost, condensation or mould growth. This behaviour might be implemented through an *expert rule* that is triggered when the system gets compromised, or hard-wired into the HVAC control system outside the scope of the RL setting.

B. Energy consumption term

Average energy usages and costs for a given location are typically correlated with climate, available resources, building typology, thermal insulation level and efficiency of HVAC

systems. The consumed energy carrier (e.g. natural gas, electricity) is measured by heat or electricity meters and correspondingly charged by the energy supplier. The widespread deployment of smart meters in recent years provides ready access to energy usage data.

While global and local policies often address the reduction of primary energy consumption (raw fuels at source) and their associated greenhouse gas emissions, individual users tend to focus primarily on the financial cost of energy. The latter should ideally be linked to the associated environmental penalty and is increasingly being used as a proxy to incentivise or discourage consumption through variable price schemes. The knowledge in advance of energy carrier prices provides a greatly increased potential for cost saving and optimisation through RL strategies.

It follows from the above that the energy consumption term of the reward function could be expressed in a currency unit. This would however complicate the formulation of a generalised reward function, which would need to be adjusted for each specific country, with the additional consideration of time-varying exchange rates among currencies. Moreover, even for regions or countries that share the same currency, the relative value can be very different as is closely tied to purchasing power. Lastly, even within a specific country or region, monetary inflation or deflation alters the relative value of the currency over time, so the reward function would also require a periodic adjustment.

Such problems can be overcome if the cost of energy is expressed not as a monetary unit but as a fraction of the Gross Domestic Product (GDP) per capita at Purchasing Power Parity (PPP), an accepted macroeconomic unit. This allows a more realistic valuation of the cost and the self-adjustment of the model responding to external events (e.g. monetary inflation, salary depreciation, variations in fuel prices).

Finally, HVAC devices and systems are subject to deterioration due to wear. Likewise, service intervals are dependent on the usage pattern of these devices. Associated replacement costs and service charges could be included as a penalty in the energy consumption term. For this purpose, operation time and/or the number of on/off cycles can be adopted as chargeable parameters with an associated cost.

C. Weighting of terms

Both the occupant comfort term and the energy consumption term discussed above are time-dependent, as their values depend on instantaneous internal ambient conditions (fundamentally temperature) and power consumption. In turn, the actions adopted by the RL agent at any given instant will affect the current and future values of both of these terms.

The reward function must define the relative importance of the conflicting comfort and energy terms. If such terms are normalised (as is often the case in a RL setting), the reward can be formulated as a weight function. In the form of Eq. 1, the occupant comfort term $C(t)$ has a positive effect on the reward, the energy consumption term $E(t)$ has a negative effect, and both terms are balanced by the weight parameter β.

$$R(t) = \beta C(t) - (1 - \beta)E(t) \tag{1}$$

In this formulation, β can range between zero and one, representing the price that the user is willing to pay for comfort. High values of β prioritise thermal comfort over energy consumption, and vice versa. The extreme cases are $\beta = 1$, where the agent focuses solely on occupant comfort regardless of energy consumption, and $\beta = 0$, which entirely disregards comfort and aims solely at minimising energy consumption.

The formulation of the reward as a weight function allows tuning the model to suit the particular conditions of each use case (e.g. household income, awareness of comfort, environmental concern) while maintaining a consistent generalised formulation for the reward function.

IV. Discussion and Conclusion

The design of rewards for RL is a challenging task because it requires a deep prior understanding of the problem and its formulation into a reward and penalty scheme. As Sutton et al. [7] say, *"the reward signal is not the place to impart to the agent prior knowledge about how to achieve what we want it to do"*, i.e. the agent should learn how to accomplish an specific goal instead of finding a way of how to obtain more reward. In this manuscript, authors argue that RL problems designed in the energy field require the identification of the most meaningful variables to be used and their reformulation in a generalised way.

Previous works by several authors take into account many of the discussed elements (e.g. occupants, consumption, set point penalties and so on). However, most of their proposed solutions contain *ad hoc* rewards focused on their specific use case creating a reward formula with a set of magnitudes that satisfies their outcome (e.g. integrating the power consumption over a specific time interval Δt, or establishing a fixed penalty for violating a given comfort range). This procedure is common because in some scenarios, there is a lack of information from some elements (the current power requirement of the active element, the modification of the elements due to a set-back procedure, the mean variations of the consumption of the installed elements, ...). For that reason, the design of a reward taking into account fixed magnitudes delimits the problem to a single scenario, making it very difficult to reproduce it. Focusing the solution of a problem on a convenient selection of scalar magnitudes prevents the generalisation of the trained agent and the reproduction of the experiment when the target environment is different. It is important to say that each active element inside a building might have a different manufacturer, efficiency or performance curve. These requirements should ideally be modelled as a part of the agent's state rather than introduced into the reward formula.

Considering the previous lines, one challenge is the creation of a generalised reward function that creates rewards based on

balanced normalisation of comfort and energy. These requires a deep knowledge about the limitations of the different factors (occupancy, solar radiation, electricity prices, thermal insulation, ...) and their inclusion into a generalised and normalised formulation. In previous literature, researchers have tried to normalise these factors by defining one by one the upper and lower limits, which can be a distorting factor because some factors do not have a defined boundary. For example, the electricity costs. Depending on the country, electricity prices could be considered *cheap* or *expensive*. There is not a rule which identifies *when* something is cheap or expensive in a use case because this is dependent on several objectives and subjective factors. Thus, in the proposed approach a normalisation procedure is not performed for each involved factors based on fixed boundaries, instead a direct relationship between costs and comfort values is studied to avoid the definition of upper and lower limits of each factor to normalise the reward function.

The main objective of the reward function is to train a RL agent to reduce the energy consumption (and consequently reduce energy bills). However, authors do not always take into consideration the decision between sacrificing comfort or sacrificing energy cost. Most of the proposed works aim to create an intelligent decision-making agent to assess when to activate or not the installed comfort devices by creating an *intelligent scheduling* process. However, there are scenarios in which the comfort of the building occupants should be prioritised over the electricity price, e.g. when the target environment is a hospital. In that case, electricity price is not a priority and the decision making process should try to prioritise the comfort of the occupants rather than creating a schedule to avoid the most expensive electric prices. But, the problem may be the other way around; there could exist a scenario in which the comfort of the occupants is not very relevant and the cost of electricity should be highly prioritised. Thus, the decision between applying a bias between the occupant comfort versus the electricity price should be something that the researcher should decide when the agent is trained. In this regard, a bias factor should be included into the reward function to allow considering the priorities of the assessed scenario (comfort or energy cost) while still maintaining a generalised form.

To overcome the aforementioned difficulties, a reward design based on the following characteristics is proposed:

- The reward function retains a generalised form that is applicable in principle to any type of building, climate and use pattern. The formulation is expressed as a weighted balance between a thermal comfort term and an energy cost term. The only parameter to be adjusted is the weight, which determines the preference of the user between the conflicting aims of comfort and energy efficiency.
- The comfort term is formulated as a bell-shaped function (Figure 2 left), with a theoretical optimum and a gradual decrease of the reward in both directions (e.g. too cold or too hot). The more common use of Boolean models (e.g. a fixed penalty for deviating from a given comfort range) is a legacy from thermostat-based controls that is no longer relevant in a RL setting. The use of a smooth derivable curve is more realistic, fits better with accepted comfort models, and is more appropriate for training RL agents. The comfort term is normalised between 0 and 1 and multiplied by an occupancy factor (thus the comfort term becomes irrelevant for unoccupied periods).
- The energy consumption term is directly proportional to cost (price of energy carrier plus a penalty associated with on/off cycles). A normalisation of this term is proposed (Figure 2 right) by expressing it as a fraction of GDP per capita at purchasing power parity, instead of a currency unit. This should allow self-adjustment to external disturbances such as monetary inflation or variations in fuel price, achieving a more robust model that does not require a periodic recalibration of its parameters. The authors intend to further develop this approach in future work.
- The normalisation of both comfort and energy consumption settles the relation between these terms. While the unity value expresses the maximum achievable comfort, values above 1 are theoretically possible for the energy consumption. Contrasting with other models, clipping above a given maximum value is prevented, as it could lead to excessive expenditures during extreme events (e.g. a short but substantial surge in fuel prices). A system with no upper boundary for cost is more representative of actual economic penalty and thus can be better trained to cope with such events.

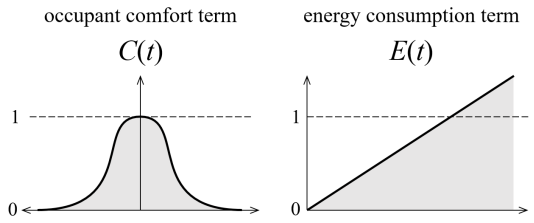

Fig. 2. Schematic diagram of the proposed normalisation scheme for the occupant comfort and energy consumption terms

The result is a reward function with a generalised and interpretable form, which can therefore be applied to any building typology, use schedule or climate, without restricting its application to a specific case. The only input required from the user is a weight parameter that controls the preference among maximising occupant comfort and minimising energy consumption. RL agents trained with a reward of this form should be able to adjust their optimum strategy to any variations in external factors such as weather, occupancy patterns or variations in the cost of fuel carriers, without any external intervention required from the user.

In future work, a complete reward formula will be presented to demonstrate how the presented methodology can be implemented in different scenarios based on the obtained contextual data and user weights definition.

ACKNOWLEDGMENT

The work described in this paper was partially supported by the Basque Government under ELKARTEK project (LANTEGI4.0 KK-2020/00072).

REFERENCES

[1] P. Solic, L. Patrono, C. Zhu, C. Tong, and A. Almeida, "Cross-layer innovations in internet of things," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, p. e4188, 2020, e4188 NA. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4188

[2] L. Patrono, L. Atzori, P. Šolić, M. Mongiello, and A. Almeida, "Challenges to be addressed to realize internet of things solutions for smart environments," 2020.

[3] D. López-de Ipiña, S. Blanco, X. Laiseca, and I. Díaz-de Sarralde, "Eldercare: an interactive tv-based ambient assisted living platform," in *Activity recognition in pervasive intelligent environments*. Springer, 2011, pp. 111–125.

[4] T. Wiedmann and J. Minx, "A definition of 'carbon footprint'," *Ecological economics research trends*, vol. 1, pp. 1–11, 2008.

[5] P. Torcellini, S. Pless, M. Deru, and D. Crawley, "Zero energy buildings: a critical look at the definition," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2006.

[6] A. González-Briones, F. De La Prieta, M. S. Mohamad, S. Omatu, and J. M. Corchado, "Multi-agent systems applications in energy optimization problems: A state-of-the-art review," *Energies*, vol. 11, no. 8, p. 1928, 2018.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[8] P. Jackson, "Introduction to expert systems," 1986.

[9] K.-V. Ling and A. Dexter, "Expert control of air-conditioning plant," *Automatica*, vol. 30, no. 5, pp. 761–773, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005109894901678

[10] D. Ibaseta, J. Molleda, M. Álvarez, and F. Díez, "An expert system for building energy management through the web of things," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2020, pp. 477–485.

[11] J. M. House, H. Vaezi-Nejad, and J. M. Whitcomb, "An expert rule set for fault detection in air-handling units/discussion," *Ashrae Transactions*, vol. 107, p. 858, 2001.

[12] H. Doukas, K. D. Patlitzianas, K. Iatropoulos, and J. Psarras, "Intelligent building energy management system using rule sets," *Building and environment*, vol. 42, no. 10, pp. 3562–3569, 2007.

[13] K. Iatropoulos, H. Doukas, D. Patlitzianas, J. Psarras, N. Tourlis, and S. Louizidis, "An expert model for monitoring building's operations via bems."

[14] N. Cotrufo, E. Saloux, J. Hardy, J. Candanedo, and R. Platon, "A practical artificial intelligence-based approach for predictive control in commercial and institutional buildings," *Energy and Buildings*, vol. 206, p. 109563, 2020.

[15] F. Ruelens, S. Iacovella, B. J. Claessens, and R. Belmans, "Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning," *Energies*, vol. 8, no. 8, pp. 8300–8318, 2015.

[16] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[17] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[18] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[19] A. Charpentier, R. Elie, and C. Remlinger, "Reinforcement learning in economics and finance," *Computational Economics*, pp. 1–38, 2021.

[20] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[21] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Applied Energy*, vol. 269, p. 115036, 2020.

[22] K. Mason and S. Grijalva, "A review of reinforcement learning for autonomous building energy management," *Computers & Electrical Engineering*, vol. 78, pp. 300–312, 2019.

[23] X. Yuan, Y. Pan, J. Yang, W. Wang, and Z. Huang, "Study on the application of reinforcement learning in the operation optimization of hvac system," in *Building Simulation*. Springer, 2020, pp. 1–13.

[24] N. S. Raman, A. M. Devraj, P. Barooah, and S. P. Meyn, "Reinforcement learning for control of building hvac systems," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 2326–2332.

[25] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building hvac control," in *Proceedings of the 54th annual design automation conference 2017*, 2017, pp. 1–6.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[27] P. Lissa, C. Deane, M. Schukat, F. Seri, M. Keane, and E. Barrett, "Deep reinforcement learning for home energy management system control," *Energy and AI*, vol. 3, p. 100043, 2021.

[28] Z. Zhang and K. P. Lam, "Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system," in *Proceedings of the 5th Conference on Systems for Built Environments*, 2018, pp. 148–157.

[29] Z. Zhang, A. Chong, Y. Pan, C. Zhang, and K. P. Lam, "Whole building energy model for hvac optimal control: A practical framework based on deep reinforcement learning," *Energy and Buildings*, vol. 199, pp. 472–490, 2019.

[30] Y. Wang, K. Velswamy, and B. Huang, "A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems," *Processes*, vol. 5, no. 3, p. 46, 2017.

[31] J. Van Hoof, "Forty years of Fanger's model of thermal comfort: comfort for all?" *Indoor Air*, vol. 18, no. 3, pp. 182–201, 2008.

[32] ISO 7730:2005, "Ergonomics of the thermal environment – analytical determination and interpretation of thermal comfort using calculation of the pmv and ppd indices and local thermal comfort criteria," ISO, Tech. Rep., 2005.

[33] EN 16798-1:2019, "Energy performance of buildings – ventilation for buildings – part 1: Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics – module m1-6," CEN, Tech. Rep., 2019.

[34] ANSI/ASHRAE Standard 55-2020, "Thermal environmental conditions for human occupancy," ANSI/ASHRAE, Tech. Rep., 2020.

[35] R. De Dear and G. S. Brager, "The adaptive model of thermal comfort and energy conservation in the built environment," *Int J Biometeorol*, vol. 45, pp. 100–108, 2001.

[36] S. Carlucci, L. Bai, R. de Dear, and L. Yang, "Review of adaptive thermal comfort models in built environmental regulatory documents," *Building and Environment*, vol. 137, pp. 73–89, 2018.