A multi-armed bandit reinforcement learning method for occupant thermal preference learning

Hussein Elehwany[1], Mohamed Ouf [2], Burak Gunay[1]
Nunzio Cotrufo[3], Jean-Simon Venne[3], Junfeng Wen[4]
[1]Carleton University, Department of Civil and Environmental Engineering, Ottawa, Canada
[2]Concordia University, Department of Building, Civil, and Environmental Engineering, Montreal, Canada
[3]Brainbox AI, Montreal, Canada
[4]Carleton University, School of Computer Science, Ottawa, Canada

Abstract

Many commercial buildings depend on fixed thermostat setpoints, which increases energy consumption and may compromise occupants' comfort. In recent years, many occupant-centric control (OCC) methods have been introduced that can learn occupants' thermal preferences and adjust thermostat setpoints accordingly. This study uses reinforcement learning (RL) to develop a smart learning thermostat. The developed algorithm utilizes multi-armed bandit (MAB) RL method leveraging feedback from unsolicited occupant interactions with thermostats and energy use to derive the optimal setpoints. The algorithm aims to balance between the goals of saving energy and enhancing occupant comfort. An occupant behaviour model is developed to simulate the stochastic behaviour of occupants with thermostats in an EnergyPlus simulation environment. The integration of the MAB algorithm with the simulation model was achieved using the Python API in EnergyPlus. The MAB algorithm has a simpler problem formulation than other RL methods as it does not require state variables. This enables the algorithm to learn faster. The MAB algorithm also requires less computational power as it does not require the training of any artificial neural networks (ANN).

Introduction

In recent years, there has been a notable increase in research efforts on occupant behaviour in buildings due to its influence on energy consumption and its stochastic nature that increases the complexity of predicting buildings' performance (Hoes et al., 2009). This resulted in the emergence of occupant-centric controls (OCC) which predicts occupancy trends and occupant behaviours using data from the indoor and outdoor environments (O'Brien et al., 2020). OCC methods are divided into occupancy-centric methods that involve occupants' presence and count; and occupancy behaviour-centric methods that try to learn occupant preferences (Ouf et al., 2021).

Thermal comfort of occupant is influenced by various factors, such as metabolic rate, clothing insulation, temperature, humidity, air movement, and radiant temperature (ASHRAE, 2020). To accommodate these factors, building operators often opt for conservative temperature setpoints to ensure comfort, disregarding any excessive energy cost (Gunay et al., 2018). This behaviour mistakenly assumes that a single setting can satisfy everyone, neglecting the fact that occupants may have varying individual thermal preferences (Fanger, 1970). When occupants are allowed to directly interact with thermostats in commercial buildings, they tend to push setpoints excessively, without allowing enough time for the building to adapt to these setpoints, due to the limited knowledge of most occupants with building dynamics and the HVAC system capabilities. When the occupants are satisfied with the indoor conditions, they would not try to relax the setpoints to a setting that would save more energy while maintaining comfort.

As a result, many studies have been conducted on personalized thermal preference learning to address the conflicting goals of occupant comfort and energy efficiency. Daum et al., (2011) and Konis & Annavaram, (2017) used statistical methods such as logistic regression to learn occupant thermostat preferences. Gunay et al., (2018) trained a Markov logistic regression model using indoor temperatures and occupant thermostat overrides. Awalgaonkar et al., (2019) and Jung & Jazizadeh, (2019) utilized Bayesian model-based methods, while (W. Li et al., 2019) implemented a fuzzy logic-based algorithm to learn thermal preferences of occupants. Under machine learning (ML) methods, Jayathissa et al., (2020) used K-means clustering and Li et al., (2017) employed random forests to predict occupant thermal preferences.

Lately, many studies have adopted reinforcement learning methods for building controls in general and for thermal preference learning specifically. Barrett & Linder, (2015) and Huchuk et al., (2021) applied the tabular Q-learning RL method to control the on/off status of the HVAC machinery. Both studies used the deviation of indoor temperature from setpoints, or comfort range, as a measure of discomfort. Lork et al., (2020) used a deep Q-network (DQN) RL method to learn optimal setpoints. Discomfort was also quantified by the difference between the indoor temperature and the thermostat setpoint. Valladares et al., (2019) utilized DQN to learn optimal setpoints using the predicted mean

vote (PMV) index as feedback. PMV index has been introduced by Fanger, (1970) as a predictor of occupant comfort using several factors, such as temperature, radiation, air speed, vapor pressure and metabolic rate and clothing. Sato et al., (2012) and Zhang et al., (2018) used PMV index and predicted percentage of dissatisfied (PPD), respectively, as feedback for their RL algorithms. Mansourifard et al., (2013) employed a multi-armed bandit (MAB) algorithm to learn optimal thermal conditions using solicited occupant feedback in the form of online feedback surveys. HVAClearn has been developed by Park & Nagy, (2020), using a tabular Q-learning method, to learn optimal thermostat settings. Occupant thermostat interactions were used as feedback for discomfort. It is evident that thermostat overrides by occupants are valid indicators of occupant discomfort (Gunay et al., 2016).

Most of the presented studies used constant range of setpoints to quantify discomfort or the PMV index. Only Park & Nagy, (2020) used unsolicited occupant thermostat interactions showing a gap in the literature. Elehwany et al., (2024) addressed this gap, as well, by deploying a Q-learning algorithm that used unsolicited occupant thermostat overrides as a penalty for the reward signal. The indoor temperature was used as the state variable and the algorithm depended on exploration events to search for energy efficient settings. During the training process, the Q-table learnt the thermal preference at each state, then a set of rules were created to find the optimal heating and cooling setpoints. In this study, an alternative multi-armed bandit (MAB) RL method is proposed. MAB methods do not require state variables or the formulation of a Markov decision process, which allows the algorithm to learn faster using less feedback from the environment. Due to this simple formulation, the MAB algorithm does not require any additional rules, as the optimal setpoints are directly learned during the training process. Moreover, the MAB algorithm does not depend on artificial neural networks (ANN), hence it requires less computation power than some RL methods that involve ANNs.

Methodology

The methodology presents the EnergyPlus simulation environment, and the synthesized occupant thermostat use models used to test the proposed algorithm. Then, the MAB algorithm is introduced.

Simulation Environment

The proposed algorithm is developed and tested in an EnergyPlus simulation environment. A single-zone model of a 4m-by-4m office room is used, with a ceiling height of three meters. The room comprises three internal walls and a single external wall with an 2m-by-1m external window. The internal walls are modelled as adiabatic surfaces, as well as the ceiling and the floor, because the heat transfer between other zones with similar conditions is negligible (Shi, 2018). The thermal properties of the model and the internal loads are shown in Table 1. The zone is ventilated by a variable air volume (VAV) unit with reheat coils. Electric baseboards are used for heating as well. The outdoor ventilation rate is set to nine L/s, complying with ASHRAE Standard 62.1 (ASHRAE, 2019). The weather data of Ottawa, Canada in 2019 is used to simulate the model, with a 15-minutes time step. The rooms are assumed to be occupied on weekdays from 08:00 to 18:00 and the air handling unit (AHU) is available from 06:00 to 19:00. A setback and setup temperatures of 15℃ and 30℃, respectively, are used at unoccupied times. The heating and colling equipment are available year-round. The heating season lasts from November to April and the cooling season lasts from May to October.

Table 1 Envelope properties and loads

Parameter	Value
External wall R-value	3 m^2K/W
External window U-value	2 W/m^2K
External window SHGC	0.5
Infiltration rate	0.25 L/sm^2
Window-to-wall-ratio (WWR)	12%
Occupants	1 person
Internal loads	50W

Occupant Thermostat Use Model

Four occupant thermostat use models are created to test the MAB algorithm with diverse and stochastic occupant behaviours. The thermostat use behaviour is depicted using a discrete time Markov process, where the probability of a thermostat interaction during a 15-minutes span depends on the indoor temperature, T_{in}, at the beginning of the time step (Elehwany et al., 2024). As demonstrated by Gunay et al., (2018) the indoor temperature was more significant in modelling occupant thermostat use behaviour than other factors such as outdoor temperature and indoor humidity. At every time step, the probability, p, of an occupant increasing or decreasing the thermostat setpoint is defined by the exponential probability distribution, in Equation (1).

$$\mathrm{p} = \frac{1}{1 + e^{-(a+bT_{in})}} \qquad (1)$$

In this study, the deadband can vary, however the minimum allowed deadband is ±1℃. An occupant override is set to increase, or decrease, both heating and cooling setpoint by 1℃ and it is assumed that after an occupants interact with the thermostat, they would wait for one hour before interacting again.

Figure 1 shows the occupant thermostat use behaviours used in the study. To address the diversity of occupant preferences, occupant 1 has relatively low thermal preference and occupant 2 has relatively high preference. Moreover, occupant 3 represents a sensitive user with narrow comfort range (2℃), as opposed to occupant 4 who is a tolerant use with a wider comfort range (4℃).

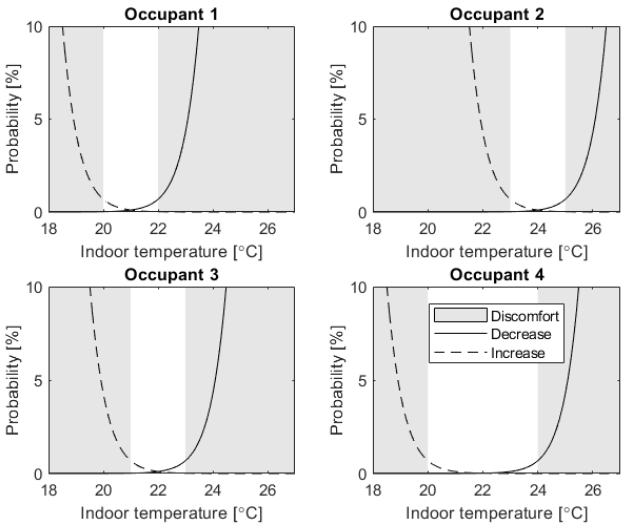

Figure 1 Occupant thermostat use models.

Multi-Armed Bandit Algorithm

Reinforcement learning is an ML branch that depends on continuous interactions between the learning agent and the environment to converge to an optimal solution. In contrary to traditional supervised-learning methods, an RL agent does not instruct the environment with the best actions, instead it evaluates the performance of these actions (Sutton & Barto, 2012).

The multi-armed bandit algorithm is an RL method that excites the environment with actions (A) and receives a reward (R) from the environment, aiming to maximize the average reward over time. This is achieved by calculating the expected reward or value, V, of each action, as shown in Equation (2). Over time the action that has the highest value is the optimal one. With RL, it is necessary to balance between "exploitation", which is using the recent knowledge to choose the best action, and "exploration", which is choosing a non-optimal action to explore possible improvements to the total reward. A common exploration technique is the ε-greedy method, where a random action is chosen with a probability of ε, and the optimal action is chosen with a probability of $1-\varepsilon$. However, the ε-greedy method explores in a random manner treating all non-greedy actions equally. In this study, the upper confidence bound (UCB) method is used. The UCB method considers the uncertainty of the value estimates in the action selection process and the potential of an action to achieve higher rewards (Sutton & Barto, 2012). This is done by adding an uncertainty term to action values, as shown in Equation (4). The less an action is used, the higher its uncertainty, hence it gets a better chance to be explored.

$$V(a) = \frac{sum\ of\ rewards\ with\ action\ a}{number\ of\ time\ action\ a\ is\ used} \qquad (2)$$

In this study, the thermostat setpoints vary from 20℃ to 25℃ and the action space comprises all the possible combinations of heating and cooling setpoints within this range with a minimum deadband of ±1℃. Therefore, the following array is the actions list i.e. [20-22, 20-23, 20-24, 20-25, 21-23, 21-24, 21-25, 22-24, 22-25, 23-25]℃. Each action is a tuple containing the heating and cooling setpoints. The reward signal, R, has two terms, one to reward the deadband between the heating and cooling setpoints, and one to penalize occupant thermostat interactions, as shown in Equation (3).

$$R = \alpha(SP_c - SP_h) - \beta\begin{cases}1 & if\ \omega = true\\ 0 & if\ \omega = false\end{cases} \qquad (3)$$

Where SP_c and SP_h are the cooling and heating setpoint, respectively, and ω is a Boolean indicator for whether there has been a thermostat override by an occupant. α and β are positive constants to tune the weight of the deadband and discomfort terms. After testing several options, it was found that the values 1 and 100 are suitable for the α and β, respectively. The deadband between the heating and cooling setpoints is used as a proxy for energy savings. A wider deadband results in a lower heating setpoint and a higher cooling setpoint, hence greater energy savings. The occupant thermostat interactions are used as an indicator for discomfort. A time step of one hour is chosen for the MAB algorithm, to allow the zone to react to any setpoint changes and for the occupant to interact with the thermostat if needed. At the beginning of each time step, the setpoints are recognized as action, A_t, then the chance of an occupant thermostat override happening is observed for the following hour. The probability of occupant overrides after certain setpoints is higher and that is reflected on the reward value R. Algorithm 1 shows the detailed steps of implementing the MAB algorithm with UCB action selection method.

Algorithm 1

Initialization for all actions:

$V(A) = 0, N(A) = 0$

Repeat at every time step t:

$$A_t = argmax\left(V(A) + c\sqrt{\frac{ln\ t}{N(A)}}\right) \qquad (4)$$

Observe the environment and calculate R_{t+1}

$N(A_t) = N(A_t) + 1$

$V(A_t) = V(A_t) + \frac{1}{N(A_t)}(R_{t+1} - V(A_t))$

N is an array that tracks the count of each action used. The value array, V, is calculated in an incremental manner, as shown in Algorithm 1. The subscript t indicates the current time step, while the subscript $t+1$, in R_{t+1}, represents the reward that is calculated at the next time step after an action is taken. To avoid running into an arithmetic error, Equation (4) is not evaluated when $N(A) = 0$, instead a large number is assigned to the value of the corresponding action to force the algorithm to explore that action. Simulations were executed with each thermostat use model and the MAB algorithm was implemented with the

EnergyPlus Python API. A base case, that does not involve any learning, was carried out for each occupant scenario to compare the energy use and discomfort results with. In the base case, the occupants are allowed to interact with the thermostat, however no preference learning takes place. The heating and cooling setpoints are set to the defaults of 21℃ and 23℃ at the beginning of everyday.

Results

Figure 2 to Figure 5 show the temperature profiles for the scenarios with occupants 1 to 4, during the heating season. Part (a) of each figure depicts the first week of the simulation, showing how the MAB algorithm changed the setpoints randomly due to its limited knowledge and exploratory nature. Part (b) in the figures shows a later week of the simulation, when the setpoints started to converge to the optimal solution.

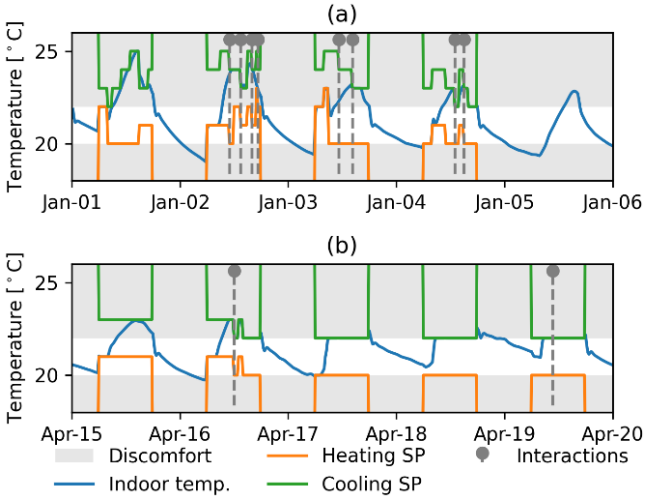

Figure 2 Temperature profiles for occupant 1. The vertical dashed lines indicate thermostat interactions, and the shaded parts show the discomfort regions.

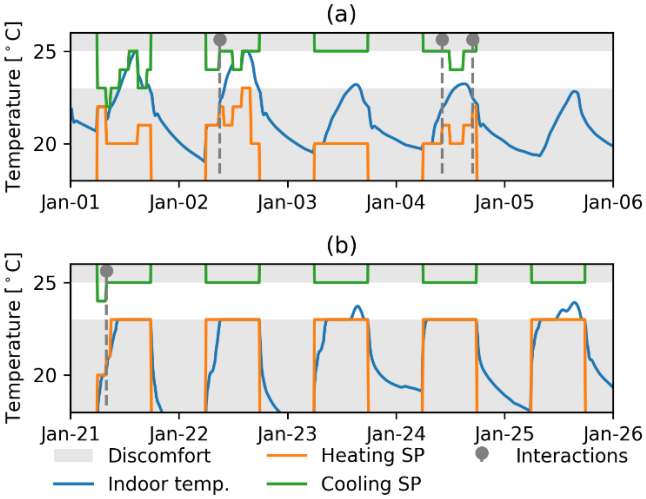

Figure 3 Temperature profiles for occupant 2.

Figure 2(b) and Figure 4(b) show how algorithm converged to the optimal setpoint that satisfies the occupant's comfort range. Occupant 2 required less time to converge than occupant 1 due to the fact that occupant 2 had a high thermal preference (23℃-25℃) and indoor temperatures tend to be lower during the heating season, hence triggering more occupant overrides that accelerated the learning process.

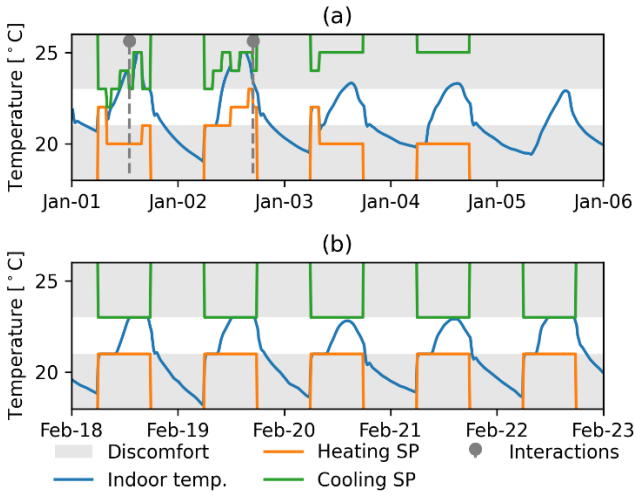

Figure 4 Temperature profiles for occupant 3.

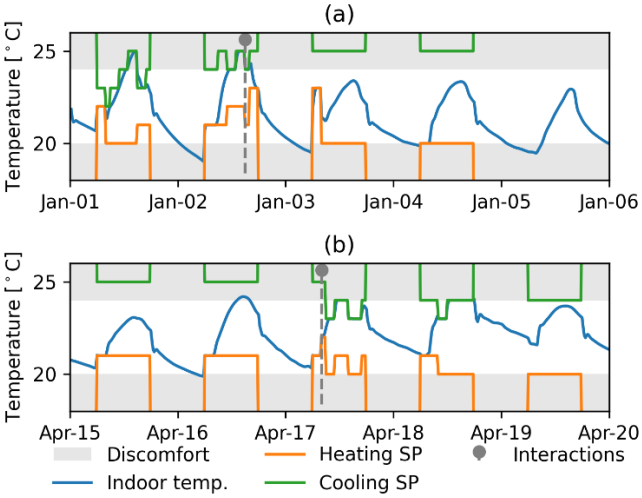

Figure 5 Temperature profiles for occupant 4.

Since occupant 4 had a wide comfort range of 4℃, any action with a cooling setpoint lower than 24℃ would have achieved enough comfort. Therefore, the energy saving term of the reward function played a necessary role in this case because it allowed the MAB algorithm to search for the most efficient setpoints within the occupant's comfort range. In Figure 4(a) and Figure 5(a) the algorithm learnt to choose the most energy saving setpoints (20℃-25℃), during the first week of simulation, because of the energy saving term and the lack of enough thermostat overrides to learn the occupant behaviour. After a couple of occupant interactions with the thermostat, the algorithm adjusts to the

optimal setpoints that satisfy better thermal comfort, as shown in Figure 4(b) and Figure 5(b).

In order to evaluate the energy savings potential of the MAB algorithm, the total energy consumption of the heating and cooling coils, and the baseboard heaters was used. Figure 6 shows the total energy use across different occupant scenarios. The MAB algorithm consumed less energy compared to the base case with occupants 1, 3 and 4. In case of occupant 2, the MAB algorithm consumed more energy than the base case because occupant 2 has a higher thermal preference, 23℃-25℃, than the base case default setpoints, 21℃-23℃. The energy saving potential of the MAB algorithm during the last month tend to be higher than the first month. During the first month the exploratory behaviour of the algorithm lead to higher energy use. This was observed the most with occupants 1 and 4, who have thermal preferences with lower heating demands.

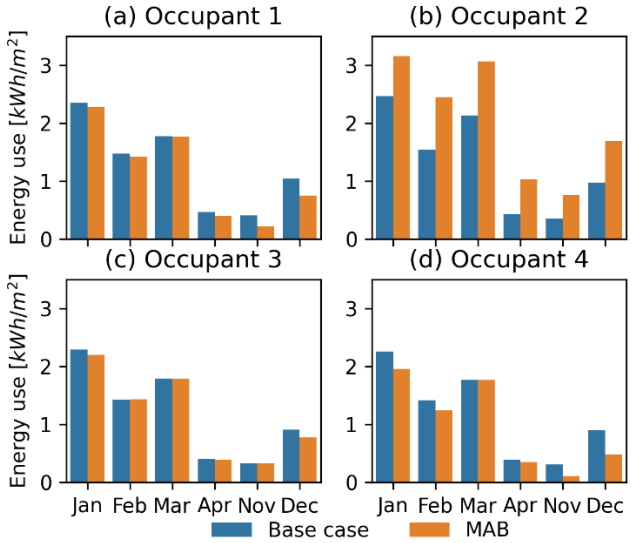

Figure 6 Monthly energy use during the heating season.

The number of occupant interactions with the thermostat, whether to increase or decrease the setpoints, were used estimate occupant discomfort. Figure 7 shows the number of thermostat interactions in every month. With occupants 1 and 2, a significant drop in the number of interactions, compared to the base case, was observed in November and December. This decline was due to the deviation of the thermal preferences of occupants 1 and 2 (20℃-22℃ and 23℃-25℃) from the intermediate default setpoints of the base case (21℃-23℃). On the other hand, occupants 3, showed no significant decrease in occupant interactions as their thermal preferences are similar to the base case (21℃-23℃). Furthermore, the MAB algorithm tends to trigger more occupant overrides due to exploration, however such effect would decrease as the algorithm runs for longer periods. The same result was observed with the thermal preferences of occupant 4 due to their wide comfort range that triggered a few occupant thermostat overrides, whether with the base case or the MAB algorithm. Therefore, the challenge for the MAB algorithm with occupant 4 was to search for the most energy efficient setpoints with limited occupant feedback.

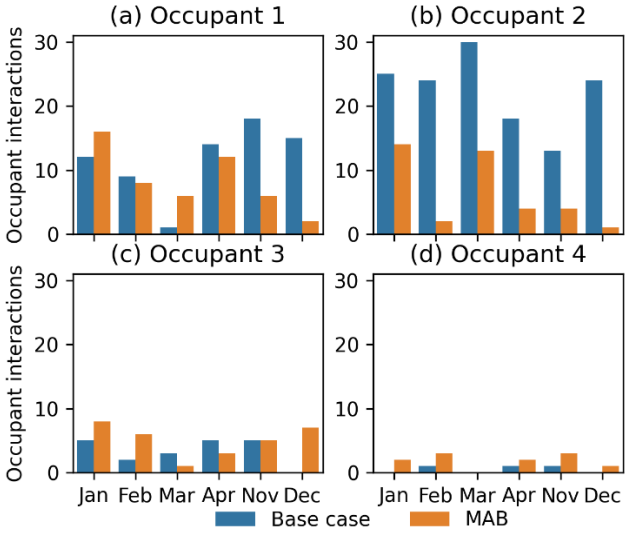

Figure 7 Monthly number of occupant thermostat interactions during the heating season.

Table 2 summarises the results of the energy use and occupant interactions with the thermostat for the heating season. The energy use for occupant 2 was significantly higher than the base case (+54%), due to the high thermal preference of occupant 2. However, after learning the optimum comfort range for occupant 2 a significant drop in the number of thermostat overrides was observed (-96). Furthermore, there was no improvement in the number of thermostat interactions with occupants 3 and 4, however the MAB algorithm achieved some energy saving (up to -16%).

Table 2 Total energy use results and occupant thermostat interactions during the heating season.

Occupant	Energy use [kWh/m^2]		
	Base case	MAB	% change
1	7.5	6.8	-9%
2	7.9	12.2	+54%
3	7.2	6.9	-3%
4	7.0	5.9	-16%
	Occupant interactions		
	Base case	MAB	difference
1	69	50	-19
2	134	38	-96
3	20	30	+10
4	3	11	+8

In order to test the reliability of the algorithm, the simulation was repeated ten times, keeping all initial conditions fixed. Figure 8 shows mean and standard deviation of the action values of occupant 3, over the ten runs. It is observed that the setpoints (21℃-23℃) received the highest value since it is the optimal action for occupant 3. The optimal action was followed by (22℃-24℃), (20℃-22℃), and (20℃-23℃) as

these actions partially satisfy the comfort region between 21℃ and 23℃. In some cases, the actions (22℃-24℃), (20℃-22℃) received higher values than (21℃-23℃), if there was not enough occupant overrides to penalize those actions.

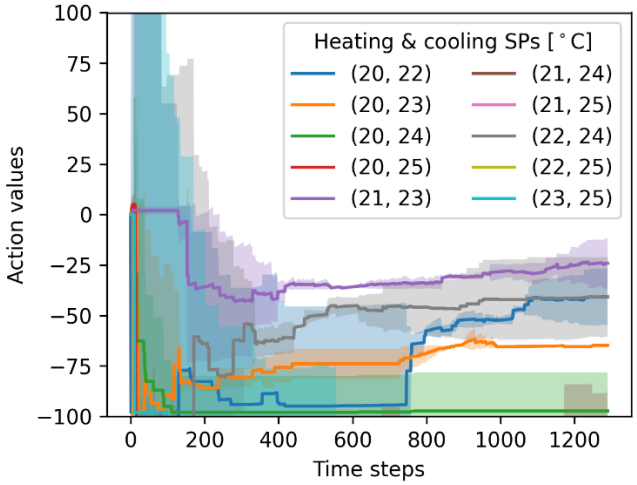

Figure 8 The mean (line) and standard deviation (shading) of action values. Each tuple represents the heating and cooling setpoints (SP) of the corresponding action.

Conclusion

Several studies approached the personalized thermal preference learning problem using RL methods, however a few of them used unsolicited occupant thermostat overrides as feedback for discomfort. This paper developed an MAB algorithm that balances between the goals of energy savings and occupant comfort, taking feedback from thermostat overrides. The advantage of the proposed algorithm is that it does not require state variables, which accelerates the learning process. A UCB action selection method was used to balance between exploration and exploitation.

The MAB algorithm managed to learn the optimal setpoints of four occupant types with varying thermal preferences. The algorithm also achieved up to 16% of energy savings in some cases. A reduction in the number of occupant overrides was observed between the first and last month of the simulation, as the algorithm converged to the optimum solution.

For future work, it is planned to deploy the MAB algorithm in an occupied space to validate the algorithm's capabilities with real occupants. Moreover, the algorithm would be integrated into a multi-agent RL algorithm, where the MAB algorithm will act as a zone-level agent and will communicate with a system-level agent that controls variables such as supply air temperatures and flows.

Acknowledgement

This research is supported by Brainbox AI Inc and the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was also developed thanks to the excellent research networking provided by IEA EBC Annex 79 "Occupant-Centric Building Design and Operation".

References

ASHRAE. (2019). *Standard 62.1: Ventilation for Acceptable Indoor Air Quality*. American Society of Heating, Refrigerating, and Air-Conditioning.

ASHRAE. (2020). *Standard 55: Thermal environmental conditions for human occupancy*. American Society of Heating, Refrigerating, and Air-Conditioning.

Awalgaonkar, N., Bilionis, I., Liu, X., Karava, P., & Tzempelikos, A. (2019). *Learning Personalized Thermal Preferences via Bayesian Active Learning with Unimodality Constraints*. http://arxiv.org/abs/1903.09094

Barrett, E., & Linder, S. (2015). Autonomous HVAC Control, A Reinforcement Learning Approach. In A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, & M. Spiliopoulou (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 9286, pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-319-23461-8_1

Daum, D., Haldi, F., & Morel, N. (2011). A personalized measure of thermal comfort for building controls. *Building and Environment*, *46*(1), 3–11. https://doi.org/10.1016/j.buildenv.2010.06.011

Elehwany, H., Ouf, M., Gunay, B., Cotrufo, N., & Venne, J.-S. (2024). A reinforcement learning approach for thermostat setpoint preference learning. *Building Simulation*, *17*(1), 131–146. https://doi.org/10.1007/s12273-023-1056-7

Fanger, P. O. (1970). *Thermal comfort. Analysis and applications in environmental engineering.* Copenhagen: Danish Technical Press.

Gunay, H. B., Beausoleil-Morrison, I., O'Brien, W., & Bursill, J. (2016). Implementation of an adaptive occupancy and building learning temperature setback algorithm. *ASHRAE Transactions*, *122*, 179–192.

Gunay, H. B., O'Brien, W., Beausoleil-Morrison, I., & Bursill, J. (2018). Development and implementation of a thermostat learning algorithm. *Science and Technology for the Built Environment*, *24*(1), 43–56. https://doi.org/10.1080/23744731.2017.1328956

Hoes, P., Hensen, J. L. M., Loomans, M. G. L. C., de Vries, B., & Bourgeois, D. (2009). User behavior in whole building simulation. *Energy and Buildings*, *41*(3), 295–302. https://doi.org/10.1016/j.enbuild.2008.09.008

Huchuk, B., Sanner, S., & O'Brien, W. (2021). Development and evaluation of data-driven controls for residential smart thermostats. *Energy*

and Buildings, *249*, 111201. https://doi.org/10.1016/j.enbuild.2021.111201

Jayathissa, P., Quintana, M., Abdelrahman, M., & Miller, C. (2020). Humans-as-a-sensor for buildings—Intensive longitudinal indoor comfort models. *Buildings*, *10*(10), 1–22. https://doi.org/10.3390/buildings10100174

Jung, W., & Jazizadeh, F. (2019). Comparative assessment of HVAC control strategies using personal thermal comfort and sensitivity models. *Building and Environment*, *158*, 104–119. https://doi.org/10.1016/j.buildenv.2019.04.043

Konis, K., & Annavaram, M. (2017). The Occupant Mobile Gateway: A participatory sensing and machine-learning approach for occupant-aware energy management. *Building and Environment*, *118*, 1–13. https://doi.org/10.1016/j.buildenv.2017.03.025

Li, D., Menassa, C. C., & Kamat, V. R. (2017). Personalized human comfort in indoor building environments under diverse conditioning modes. *Building and Environment*, *126*, 304–317. https://doi.org/10.1016/j.buildenv.2017.10.004

Li, W., Zhang, J., & Zhao, T. (2019). Indoor thermal environment optimal control for thermal comfort and energy saving based on online monitoring of thermal sensation. *Energy and Buildings*, *197*, 57–67. https://doi.org/10.1016/j.enbuild.2019.05.050

Lork, C., Li, W., Qin, Y., Zhou, Y., Yuen, C., Tushar, W., & Saha, T. K. (2020). An uncertainty-aware deep reinforcement learning framework for residential air conditioning energy management. *Applied Energy*, *276*(June), 115426. https://doi.org/10.1016/j.apenergy.2020.115426

Mansourifard, P., Jazizadeh, F., & Krishnamachari, B. (2013). *Online Learning for Personalized Room-Level Thermal Control: A Multi-Armed Bandit Framework*.

O'Brien, W., Wagner, A., Schweiker, M., Mahdavi, A., Day, J., Kjærgaard, M. B., Carlucci, S., Dong, B., Tahmasebi, F., Yan, D., Hong, T., Gunay, H. B., Nagy, Z., Miller, C., & Berger, C. (2020). Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation. *Building and Environment*, *178*(December 2019), 106738. https://doi.org/10.1016/j.buildenv.2020.106738

Ouf, M. M., Park, J. Y., & Gunay, H. B. (2021). A simulation-based method to investigate occupant-centric controls. *Building Simulation*, *14*(4), 1017–1030. https://doi.org/10.1007/s12273-020-0726-y

Park, J. Y., & Nagy, Z. (2020). HVACLearn: A reinforcement learning based occupant-centric control for thermostat set-points. *E-Energy 2020 - Proceedings of the 11th ACM International Conference on Future Energy Systems*, 434–437. https://doi.org/10.1145/3396851.3402364

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

Shi, Z. (2018). *Building Operation Specialist: A Probabilistic Distributed Fault Detection, Diagnostics and Evaluation Framework for Building Systems*. Carleton University.

Sutton, R. S., & Barto, A. G. (2012). *Reinforcement learning, Second Edition in progress*.

Valladares, W., Galindo, M., Gutiérrez, J., Wu, W.-C., Liao, K.-K., Liao, J.-C., Lu, K.-C., & Wang, C.-C. (2019). Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. *Building and Environment*, *155*, 105–117. https://doi.org/10.1016/j.buildenv.2019.03.038

Zhang, Z., Chong, A., Poh Lam, K., Pan, Y., Zhang, C., & Lu, S. (2018). *A Deep Reinforcement Learning Approach to Using Whole Building Energy Model For HVAC Optimal Control ASHRAE Multidisciplinary Task Group on Occupant Behavior in Buildings View project International Energy Agency Energy in Buildings and Communities Programme, Annex 66, Definition and Simulation of Occupant Behavior in Buildings View project A DEEP REINFORCEMENT LEARNING APPROACH TO USING WHOLE BUILDING ENERGY MODEL FOR HVAC OPTIMAL CONTROL*. https://www.researchgate.net/publication/326711617