EPFL

Thèse n° 10 090

Reinforcement Learning for the occupant-centric operation of building energy systems: Theoretical and experimental investigations

Présentée le 11 novembre 2022

à la Faculté de l'environnement naturel, architectural et construit
Laboratoire d'ingénierie du confort intégré
Programme doctoral en énergie

pour l'obtention du grade de Docteur ès Sciences

par

Amirreza HEIDARI

Acceptée sur proposition du jury

Prof. G. Tagliabue, présidente du jury
Prof. D. Khovalyg, Prof. F. Maréchal, directeurs de thèse
Prof. B. Gunay, rapporteur
Prof. T. Booysen, rapporteur
Dr R. Castello, rapporteur

École polytechnique fédérale de Lausanne

2022

Reinforcement Learning is enough to reach Artificial General Intelligence

— Demis Hassabis, CEO of Google Deepmind

Acknowledgements

My PhD path has been an incredible enrichment for me and would not have been possible without the support from many people.

First, I would like to thank my supervisors, Prof. François Maréchal, and Prof. Dolaana Khovalyg who has guided and supported my academic activities and helped me enriching my knowledge and enlarging my research perspective.

A special “Thank You” goes to Prof. Zoltan Nagy, for inspiring me in the topic of Reinforcement Learning and for kindly answering my questions and providing extremely helpful comments.

I would also like to thank the experts who were involved in the examination of this research project: Prof. Burak Gunay, Prof. Thinus Booysen and Dr. Roberto Castello. Without their passionate participation and input, the validation survey could not have been successfully conducted.

I am also grateful to the owners of case study homes for engaging in our study and letting us to install monitoring equipment in their homes.

A big “Thank you” also to Droople Company, specially Ramzi Bouzerda and Bastien Rojanawisut, that financially supported some of our case studies and kindly provided technical support for IoT monitoring systems.

A unique "Thank you" goes to my family – without my parents, all this would not have been possible.

Amirreza HEIDARI
Fribourg
Switzerland
26 October 2022

Abstract

Occupant behavior, defined as the presence and energy-related actions of occupants, is today known as a key driver of building energy use. Closing the gap between *"what is provided by building energy systems"* and *"what is actually needed by occupants"* requires a deeper understanding and consideration of the human factor in the building operation. However, occupant behavior is a highly stochastic and complicated phenomenon, driven by a wide variety of factors, and unique in each building. Therefore, it cannot be addressed using analytical approaches traditionally used to describe physics-based aspects of buildings. In conventional control systems, referred to as *Expert-based controls* in this study, domain experts distill their knowledge into a set of rules, heuristics (rule-based controls), or optimization models (model predictive controls), and program it to the controller. Since they rely on the hard-coded knowledge of experts, they are limited to expert knowledge. Furthermore, they cannot deal with unexpected situations that were not foreseen by the experts. Given the unexpected variations of occupant behavior over time, and its uniqueness in each building which has limited the experts to globally model it, *Expert-based controls* have a low potential for integrating occupant behavior into building controls. An alternative approach is to program a *human-like learning mechanism* and develop a controller that is capable of *continuously learning and adapting the control policy* by itself through interacting with the environment and learning from experience, referred to as *Learning-based controls* in this study. *Reinforcement Learning*, a Machine Learning algorithm inspired by neuroscience, can be used to develop such a learning-based controller. Given the learning ability, these controllers are able to learn optimal control policy from scratch, without prior knowledge or a detailed system model, and can continuously adapt to the stochastic variations in the environment to ensure an optimal operation. These aspects make Reinforcement Learning a promising approach for integrating occupant behavior into building controls.

The main question that this study deals with is:

How to develop a controller that can perceive and adapt to the occupant behavior to minimize energy use without compromising user needs

In this context, the methodological framework of this dissertation is aimed at

contributing to new knowledge by developing three *occupant-centric* control frameworks:

- *DeepHot:* focused on hot water production in residential buildings;
- *DeepSolar:* focused on solar-assisted space heating and hot water production in residential buildings;
- *DeepValve:* focused on space heating in offices;

In developing these frameworks, special attention is paid to:

1. ***Transferability:*** To be easily transferred to many buildings;
2. ***Data efficiency:*** To quickly learn optimal control when implemented on a new building;
3. ***Safety:*** To impose minimum risk on violating occupant comfort or health;
4. ***Minimal use of sensors and actuators:*** To reduce the initial cost and risk of failure and facilitate filed implementations;

The *DeepHot* and *DeepSolar* are evaluated using real-world weather data and hot water use behavior measured in Swiss residential houses. *DeepValve* is also first evaluated using real-world occupancy data collected from other studies, and then experimentally implemented in an environmental chamber. Comparison of these frameworks with common practice indicated that there is a significant energy saving potential by integrating occupant behavior into building controls, and Reinforcement Learning is a promising method to achieve this goal.

Résumé

Le comportement des occupants, défini comme la présence et les actions des occupants liées à l'énergie, est aujourd'hui reconnu comme un facteur clé de la consommation d'énergie des bâtiments. Pour combler l'écart entre "ce qui est fourni par les systèmes énergétiques du bâtiment" et "ce dont les occupants ont réellement besoin", il faut une meilleure compréhension et considération du facteur humain dans le fonctionnement du bâtiment. Cependant, le comportement des occupants est un phénomène hautement stochastique et compliqué, régi par une grande variété de facteurs, et unique dans chaque bâtiment. Par conséquent, il ne peut pas être abordé à l'aide des approches analytiques traditionnellement utilisées pour décrire les aspects physiques des bâtiments. Dans les systèmes de contrôle conventionnels, appelés contrôles basés sur des experts dans cette étude, les experts du domaine appliquent leurs connaissances dans un ensemble de règles, heuristiques (contrôles basés sur des règles) ou de modèles d'optimisation (contrôles de modèles prédictifs), qu'ils intègrent dans le contrôleur. Comme elles reposent sur les connaissances des experts encodées en dur, elles sont limitées aux connaissances des experts. En outre, ils ne peuvent pas faire face à des situations inattendues qui n'ont pas été prévues par les experts. Étant donnée les variations inattendues qu'ont le comportement des occupants au cours du temps et leur exclusivité dans chaque bâtiment, ce qui a limité les experts dans leur modélisation globale, l'intégration du comportement des occupants dans le contrôle des bâtiments dans le controle basé sur les experts à un faible potentiel. Une approche alternative consiste à programmer un mécanisme d'apprentissage de type humain et à développer un contrôleur capable d'apprendre et d'adapter en permanence la politique de contrôle par lui-même en interagissant avec l'environnement et apprenant selon ses expériences, ce que l'on appelle dans cette étude les contrôles basés sur l'apprentissage. L'apprentissage par renforcement, un algorithme d'apprentissage automatique inspiré des neurosciences, peut être utilisé pour développer un tel contrôleur basé sur l'apprentissage. Grâce à leur capacité d'apprentissage, ces contrôleurs sont capables d'apprendre une politique de contrôle optimale en partant de zéro, sans connaissances préalables ni modèle détaillé du système, et peuvent s'adapter en permanence aux variations stochastiques de l'environnement pour garantir un fonctionnement optimal. Ces aspects font de l'apprentissage par renforcement est une approche prometteuse pour intégrer le comportement des occupants dans les contrôles des

bâtiments.

La question principale à laquelle cette étude répond est la suivante:

Comment développer un contrôleur capable de percevoir et de s'adapter au comportement de l'occupant pour minimiser la consommation d'énergie sans compromettre les besoins de l'utilisateur?

Dans ce contexte, le cadre méthodologique de cette thèse vise à contribuer à de nouvelles connaissances en développant trois cadres de contrôle centrés sur l'occupant:

- *DeepHot:* axé sur la production d'eau chaude dans les bâtiments résidentiels;
- *DeepSolar:* axé sur le chauffage des locaux et la production d'eau chaude assistés par l'énergie solaire dans les bâtiments résidentiels;
- *DeepValve:* axé sur le chauffage des locaux dans les bureaux;

Une attention particulière est accordée à l'élaboration de ces cadres:

- *La transférabilité:* Être facilement transférable à de nombreux bâtiments;
- *L'efficacité des données:* Pour apprendre rapidement le contrôle optimal lorsqu'il est mis en œuvre sur un nouveau bâtiment;
- *La sécurité:* Pour imposer un risque minimal d'atteinte au confort ou à la santé des occupants ;
- *Utilisation minimale de capteurs et d'actionneurs:* Pour réduire le coût initial et le risque de défaillance et faciliter les mises en œuvre déposées;

DeepHot et DeepSolar sont évalués à l'aide de données météorologiques réelles et du comportement d'utilisation de l'eau chaude mesuré dans des maisons résidentielles suisses. DeepValve est aussi d'abord évalué à l'aide de données d'occupation du monde réel recueillies dans d'autres études, puis mis en œuvre expérimentalement dans une chambre environnementale. La comparaison de ces cadres avec la pratique courante indique qu'il existe un potentiel d'économie d'énergie considérable en intégrant le comportement des occupants dans les contrôles des bâtiments, et que l'apprentissage par renforcement est une méthode prometteuse pour atteindre cet objectif.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Occupant behavior: An underexplored issue in building controls

This section discusses:

What is energy-related occupant behavior?
Why does it matter?
Why it is under-explored?

Occupancy patterns can significantly influence the energy use of the buildings [1]. Furthermore, once the occupants are present, they try to achieve a personally comfortable condition by performing different actions on building interfaces such as adjusting the thermostat setpoint, switching lights, opening/closing windows, and adjusting window blinds, which can significantly impact on the energy use of the buildings [2]. The energy-related occupant behavior is defined by human-building interactions that affect the building energy use, which can be divided into occupancy and occupants' actions on the building. Conventional efforts for energy saving in buildings are mostly focused on technological improvements of the building such as installing better-insulated envelopes or more efficient energy systems. However, nowadays, there is an increasing awareness about the pivotal role of occupant behavior on building energy use [2, 3]. *"Buildings don't use energy, people do!"* [4], initiated by Jonda in 2011, is now turned to be an emblematic headline in occupant-centered studies. Previous studies have highlighted the importance of energy-related occupant behavior in several aspects including, but not limited to, the difference in energy use between similar buildings, the gap between simulated and actual energy use, and the potential energy saving by changing occupant behavior [5]. Several studies have tried to explore to which extent occupant behavior can influence on energy

use in buildings. For example, simulation studies in a single occupant office building indicated that the occupant behavior can change the energy use of the office from 50% less to 90% more energy use compared to a standard behavior [6]. Other studies indicated that the occupant behavior in similar residential houses (having the same layout and climatic conditions) can result in different energy consumption of over 300% [7–9]. Simulations of different occupant behavior and schedules in commercial buildings indicated that final energy use can vary from 30% to 150% only due to occupant behavior [10, 11]. Measurements from the real-world buildings also indicated a high variation of electrical loads [12] or hot water energy use [13, 14] mainly due to different occupant behavior. Difference of occupant behavior in buildings can be caused by many variables such as the difference of occupancy patterns, perception of comfort, physiological characteristics of the occupants, household lifestyle, etc., [15].

The importance of occupant behavior is also getting an increasing attention in simulation studies aimed to predict building energy use. While the current simulation tools can model the physical aspects (such as building envelope or thermal systems) with a good accuracy, in some cases, there is a significant discrepancy between the predicted and actual building energy use. A major source of this discrepancy is the occupant behavior, which is over-simplified in the current modeling approaches, for example, by considering static occupancy schedules, lights operation, or temperature settings [2]. With the increased awareness of the importance of occupant behavior for energy use simulations, probabilistic modeling approaches have been applied to represent the stochastic nature of occupant behavior in buildings [16]. Probabilistic models have been used to represent different aspects of occupant behavior such as occupancy [17], lighting control [18], windows action [19], temperature setting [20], and plug-in appliances [21]. The considerable energy saving that can be achieved by only changing occupant behavior is another indicator of the importance of occupant behavior. Behavioral change strategies are recognized as a low-cost and efficient measure to reduce energy use in buildings [22]. Energy awareness campaigns revealed an energy-saving potential of 15% to 20% in residential buildings [23, 24]. Studies on energy engagement in offices also indicated an energy-saving potential of 4% to 10% [25, 26] . The behavioral change is usually stimulated by providing feedback to the users, through mobile or web apps, ambient displays, or even games [27, 28].

Nowadays, the physical aspects of the building such as the envelope, space heating, or energy storage devices can be described by mathematical models. However, occupant behavior is highly stochastic in nature [29], depends on many variables that can not all be quantified or measured [30] and is unique for each person [31], thus cannot be easily modeled similar to the physical aspects of a building. Studies highlight that occupant behavior can be affected by environment-related, time-related, individual, social, and

random factors [2, 3, 5, 32, 33]. Examples of parameters included in each category are shown in Figure 1.1. "*Every individual is essentially unique and different from everyone else*" [31]. Even under similar environmental conditions (environmental factors), every occupant is still affected by many other factors (individual, social, time-related, and random factors), that make the behavior of each occupant unique, and different from the other occupants in the same environment [34]. Every occupant perceives the environment differently, have different preferences, motivations, and habits, and can be constrained by a set of different social or economic barriers, which forms the unique decision-making process of each occupant in performing an action (Figure 1.1). Stochasticity, being

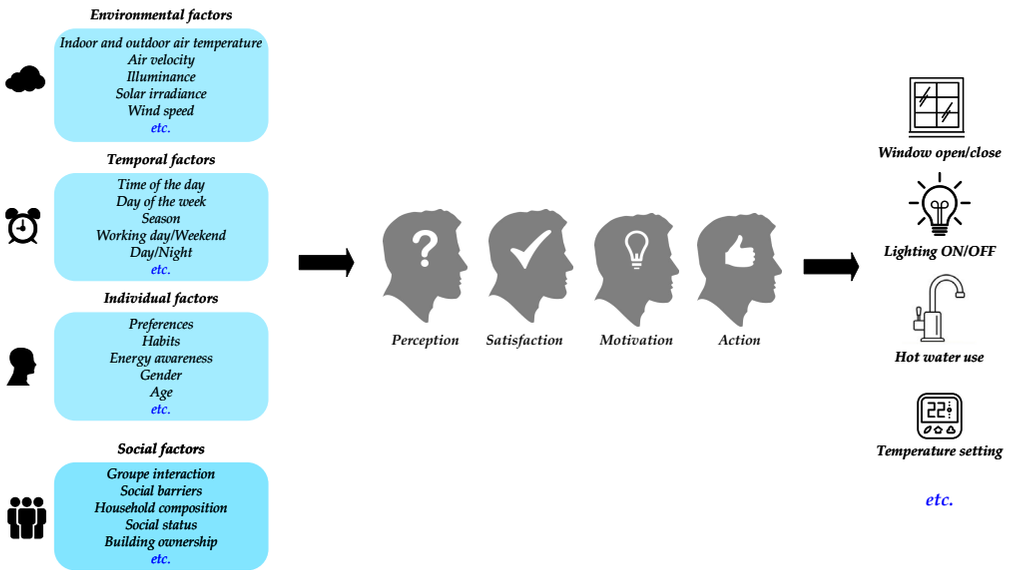


Figure 1.1: Unique decision-making process of each occupant

influenced by many different factors, and uniqueness, make the occupant behavior a very complicated phenomenon that can not be easily understood, predicted, or mathematically modeled by the experts. *Researchers have only started to understand occupant behavior, let alone accurately predicting it* [35]. The field of energy-related occupant behavior is quite new [3], with an increasing number of publications and many unanswered questions. Fabi et al. [32] highlight that much is still unknown about the underlying motivation of occupants in interacting with the building. Due to the complexity of occupant behavior, conventional control systems in buildings are detached from occupant behavior and rely on static and conservative assumptions to control building systems [36]. For example, commercial buildings follow static schedules of occupancy estimated at the design phase, while actual occupancy can significantly vary from assumptions, resulting in significant energy waste. Masoso and Grobler [37] indicated that more energy is used during non-working hours (56%) than during working hours (44%). It is estimated that almost 90% of existing space heating and cooling systems are not controlled optimally [38].

Gunay et al. [39] indicated that if an office thermostat that can learn the arrival and departure time of occupants, and accordingly schedule the temperature setback of the office, it can reduce the space heating and cooling loads by 10%-15%. These studies has raised the necessity of developing controllers that can perceive and react to the occupant behavior, which has initiated the emerging field of *occupant-centric control*. Thus, the next session of this chapter provides an overview of the field of occupant-centric control.

1.2 Introduction to occupant-centric controls

This section discusses:

What is occupant-centric control, and why does it matter?
What are different types of occupant-centric control?

The previous section provided a background about the importance of occupant behavior in buildings. This chapter discusses about occupant-centric control, which is an emerging approach to better integrate the stochastic occupant behavior in building controls with the main aim of energy saving.

Occupant-centric control is a relatively recent area, and therefore there is no standard and widely-known definition for that. By collecting and analyzing the definitions provided by several review papers [40–43] , this study defines occupant-centric controller as follow:

Occupant-centric control is a decision-making algorithm that in addition to the system and environment data, takes into account the occupant-related data, and determines the optimal control action(s) to save energy while maintaining occupants' needs.

This definition is visualized in the Figure 1.2. The concept of occupant-centric control can be applied to any kind of building systems that is influenced by occupant behavior, such as space heating or cooling, hot water production, lighting, ventilation, etc. With the common aim of energy saving, depending on the system and purpose, the additional users' needs might be incorporated such as maintaining occupant comfort, health or improving productivity.

A main aspect of occupant-centric controller is the integration of occupant-related data for optimal decision-making. Required occupant-related data can be different dependent on the system, type of building, purpose of control, etc. Melfi et al. [44] categorized the occupant-related data into the following categories:

- **Occupancy:** a zone has at least one person in it
- **Count:** how many people are in a zone

- **Identity:** who they are
- **Activity:** what they are doing

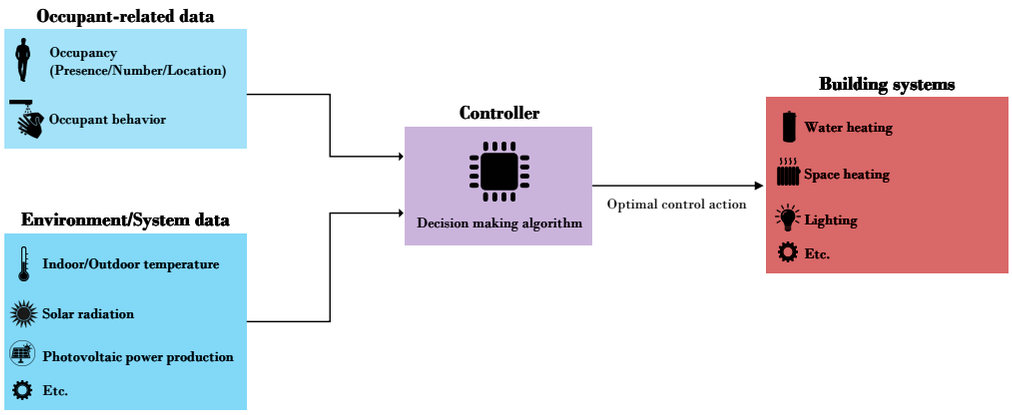


Figure 1.2: Definition of occupant-centric control

In another categorization, Naylor et al [40] included location of occupants as a type of Occupancy data. In this study, to be compatible with the categorization of occupant-centric algorithms that will be discussed later on, the occupant-related data is divided into two categories of *occupancy* (Occupancy, count, location, identity) and *occupant-behavior*, as explained below.

1.2.1 Categories of occupant-related data

Occupancy: Presence/Number/Location/Identity

Different forms of occupancy data include presence of occupants as a binary value, number of occupants, their location and identity. There are different methods to obtain each of these data. Occupant presence is the easiest one to be obtained with the current technologies. To monitor the presence or number of occupants, one of the technological solutions is to use the feed from cameras. While it is known as an accurate method, it is costly, computationally expensive, and its accuracy heavily depends on lighting conditions and space arrangement [45]. An alternative and cheaper method is to use a single or a fusion of PIR motion and environmental sensors such as CO_2, light, acoustic, temperature, relative humidity, etc [46, 47]. To infer the presence or number of occupants from the measured data, the measurements should be first labeled with the ground-truth data, which can be recorded manually by analyzing videos from cameras, or by using highly accurate video-based methods. The labeled data are then used to train a Machine Learning model, which afterward can be deployed on a hardware to directly infer presence

or number of occupants from measurements [48]. Fusion of environmental and motion sensors provides a low-cost and computationally cheap method. But the accuracy can be lower and affected by many environmental parameters. For example, CO_2 concentration has a delayed response time, as CO_2 exhaled by occupants takes time to accumulate to elevated levels, and it is affected by ventilation rates and window openings [49]. This can specifically affect the accuracy of inferring occupant number. Heidari et al.[50] proposed a probabilistic Machine Learning method for occupancy number detection based on sensor fusion. This method provides the estimated number of occupants, alongside with the uncertainty of prediction, which can be used together for an uncertainty-aware occupant-centric control method.

The location sensing provides a higher detailed information for tailoring the building control to an individual level. But on the other hand, it causes more concerns about privacy, and requires a higher technological level. A common method to detect occupant location is to use wearable radio-frequency tags, to transmit signals to receivers located around the sensed space [51–53]. In this method, occupants are expected to wear a tag, which therefore limits the method to places with fixed occupants such as offices [54]. Facilities provided by smart devices, such as Wi-Fi beacons, Bluetooth beacons, GPS, or orientation data from smart devices also can be used to determine occupant location [55, 56]. A dense network of motion or environmental sensors can also be installed to infer occupancy data [57].

Occupant behavior

Occupant behavior-related data includes a vast variety of data. Basically, any occupant-related data that does not fall into the occupancy-related category can be considered occupant behavior-related data. The most common form of occupant behavior-related data in the literature is focused on the sensing of occupant-building interaction. Some examples are monitoring the interaction of occupants with the lighting system [58], hot water appliances [59, 60], space heating and cooling systems through thermostats [39, 61, 62], windows [63] or electric plugs [64]. The measurements of when and how building systems are used by occupants can provide great insight for better design and operation of these systems. Another common type of occupant behavior-related data is the type of activity that occupants perform in the building. The type of activity can be directly detected by a camera or indirectly inferred from different sensors such as acceleration data from a wristband [65] or fusion of ambient, sound, motion, and chair [66]. Occupant feedback about building services is also another type of occupant behavior-related data. This data can be collected in real-time through mobile applications or other specifically designed devices (such as push bottons). A common form of feedback data is the occupant sensation of thermal comfort [61, 67]. However, as experimental

studies also indicated [67], occupants do not keep their interaction with the feedback system over the long term, even if they do not feel comfortable. This shows that directly asking for occupant feedback, even if it is limited to the uncomfortable instances, can not be a good solution in practice. An alternative is to infer occupant feedback indirectly from their interaction with the adjustment devices such as thermostats or light switches [58], assuming that more interactions are due to less comfort.

1.2.2 Categories of occupant-centric control

Occupant-centric control has a vast definition and therefore includes many different approaches. These approaches can be categorized based on different metrics. This study, taking into account the perspective of previous review studies [40–43], presents two categorizations of occupant-centric controls. The first categorization is based on the mechanism of response to occupant data, and the second is based on the type of occupant-related data.

I. Categorization based on the mechanism of response to occupant data

Depending on the mechanism of response to occupant data, occupant-centric controls can be divided into *Reactive* and *Predictive* controls, as discussed below.

Reactive: Reactive controls respond to occupant-related data in real-time. This category of occupant-centric controls usually includes a simple algorithm and hardware setup. Accordingly, they are easier to implement in practice and closer to the commercialization level. A well-established example of these controls are occupancy-based control of lighting systems. In these systems, which is mostly deployed in commercial buildings, lighting is switched ON when a motion is detected by PIR sensors, and switched OFF after a fixed time delay after the last motion event detected. Another example is localized lighting based on the detected location of occupants [68, 69]. Reactive control can be also implemented for appliance power management. For example, by integrating occupancy sensors with power plugs, electrical appliances can be only powered when occupants are detected in the room, eliminating the sleep mode power use [70]. As can be seen in these examples, the real-time response to occupancy data is mainly suitable for the fast-response systems.

Predictive: In the case of slow-response systems, such as space heating and cooling systems, the real-time reactive control might violate the occupant comfort, particularly, upon arrival of occupants at the beginning of working day. To eliminate the risk of comfort violation, the occupant-centric control of slow-response systems should integrate a predictive model of occupancy. In this case, the occupant-centric controller can pre-heat or pre-cool the space according to the predicted occupancy. Prediction of occupant

behavior requires the construction of a model of occupant behavior through observations, which can be done by Machine Learning models [40] or other innovative algorithms [39]. Therefore, the predictive model includes a more complicated decision-making algorithm, requiring more computational power. Examples of this category are control of space heating and cooling by learning occupancy [39, 71], or control of water heating systems by learning occupant hot water use behavior [60]. Figure 1.3 shows the summary of categorization based on the mechanism of response to occupant data.

II. Categorization based on the type of occupant-related data

Another categorization of occupant-related data is based on type of occupant-related data that is used by the controller [41]. As discussed before, occupant data can be divided into occupancy-related data and occupant behavior-related data. Accordingly, the occupant-centric controls can be categorized into occupancy-centric controls and occupant behavior-centric controls as shown in Figure 1.3. For example, controlling lights with a motion sensor is categorized as occupancy-centric control, and controlling hot water production based on hot water use behavior of occupants [60] is categorized as occupant behavior-centric control.

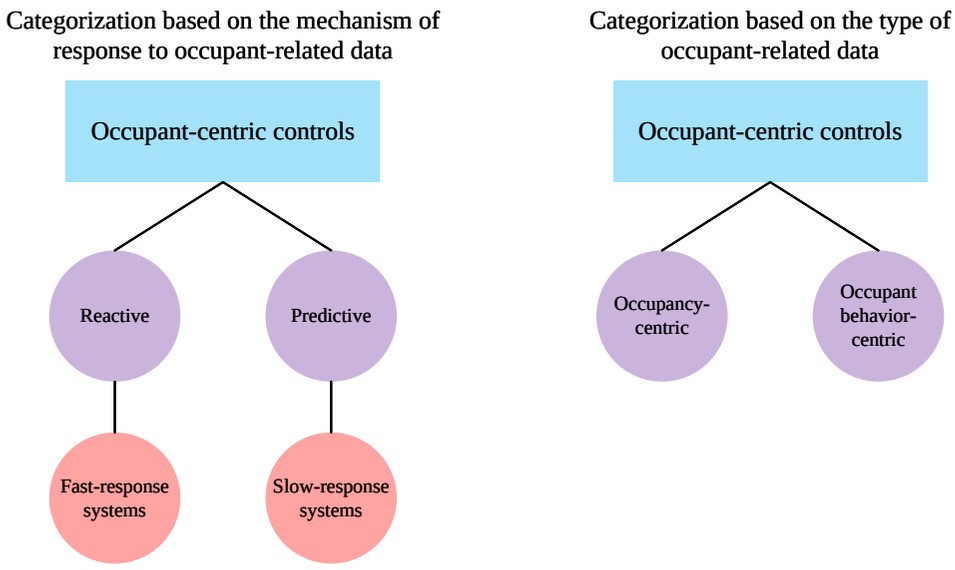


Figure 1.3: Two different categorizations of occupant-centric controls

1.3 The big picture

This section discusses:

What is the overall motivation of this study?
How the thesis is structured?

This section presents the big picture of this thesis. The first subsection states the overall motivation of this study, and the second subsection presents the motivation of each chapter and the structure of this thesis.

1.3.1 Overall motivation

In conventional control systems, domain experts distill their knowledge down to a set of rules and heuristics (rule-based control) or complicated models (model-based control) and program those rules or models into a controller. Since these controllers rely on the hard-coded knowledge of the experts, in this study they are referred to as *Expert-based controls*. Expert-based controls are robust, easy to interpret, and well-known in the industry. But they have certain limitations. First of all, they cannot deal with unexpected situations [72]. If they face a situation that was not expected during their programming, they might fail to achieve requirements. Secondly, they are limited to expert knowledge [73]. If the experts are supposed to hard-program the control solutions into the controller, they should know these solutions in advance. As explained in the last sections, occupant behavior is a highly stochastic phenomenon, that is influenced by a wide range of parameters, evolves during the time, and differs from building to building [2, 3, 5]. Given the complexity and uniqueness of occupant behavior in each building, in programming expert-based controls, experts cannot deal with occupant behavior in the same way as other physics-based aspects of a building that can be mathematically described. Instead, they have to ignore, or over-simplify occupant behavior and follow conservative rules to ensure the comfort of occupants regardless of their stochastic behavior. Examples are conventional hot water systems that totally ignore occupant behavior and always maintain a high temperature in the tank, or space heating systems that over-simplify occupant behavior by following constant occupancy schedules. So there is a significant gap between *what is provided by expert-based controls* and *what is actually needed by the occupants*. To bridge this gap, building controls should perceive and adapt to the unique occupant behavior in each building. An alternative to hard-programming the control solution is to program a human-like learning ability to the controller, referred to as *Learning-based controls* in this study. Reinforcement Learning is a method of Machine Learning that can provide this learning ability for the controller. In this approach, which is inspired by neuroscience and how the brain achieves cognition, the controller autonomously learns the optimal control solution by itself rather than it being programmed by experts. The

controller does not need any prior knowledge of the system and learns the optimal control policy only through interacting and observing the feedback, which is a great advantage over the Model Predictive Control that requires a detailed model of the system that can be time-consuming to develop [74]. Another main potential of the Reinforcement Learning method is adaptability. A Reinforcement Learning controller can continuously learn and adapt to the changes to maintain an optimal operation. For example, it can learn the daily variations of heat pump efficiency and accordingly schedule heating cycles to benefit from high-efficiency hours and minimize energy use. *Not relying on the expert knowledge*, and the *adaptability potential*, make Reinforcement Learning a powerful tool for developing occupant-centric controls. While the occupant behavior is under-explored (with very limited expert knowledge about it), varies over the time, and is unique in each building, a Reinforcement Learning controller can autonomously learn the occupant behavior without being pre-programmed by experts, adapt to the unique occupant behavior in each building, and to the variations of occupant behavior over time. Though, there are also several challenges in applying Reinforcement Learning into occupant-centric control domain. For example, the agent in an occupant-centric control deals with highly stochastic disturbances, behaviour observations come in very infrequently unless discomfort is generated intentionally, and exploration (performing random actions to better explore all possible actions) can generate lots of complaints, etc. Therefore, applying an occupant-centric Reinforcement Learning controller is a major challenge that is that is worthy proper researching in a doctoral degree.

Given the great potential of Reinforcement Learning to deal with occupant behavior on the one hand, and the underexplored challenges on the other hand, this study focuses on *how to use Reinforcement Learning to integrate occupant behavior into building controls* (Figure 1.4). More specifically, this study aims to develop Reinforcement Learning-based occupant-centric control frameworks that are:

1. **Transferrable:** To be easily transferred to many buildings;
2. **Data efficient:** To quickly learn optimal control when implemented on a new building;
3. **Safe:** To impose minimum risk of violating occupant comfort or health;
4. **Utilize minimum number of sensors and actuators:** To reduce the initial cost and risk of failure and facilitate field implementations;

These considerations are taken into account in the design of the hardware layout, formulation of the framework, and training process of three different control frameworks proposed in this study. The next sub-section briefly introduces each framework and presents the overall structure of this thesis.

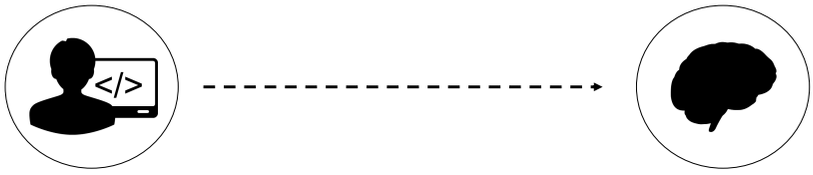


Figure 1.4: Transition from Expert-based to Learning-based controls

1.3.2 Thesis roadmap

With the overall aim of developing self-learning occupant-centric controls, three projects were performed, organized as three different chapters of this thesis.

DeepHot control framework: In the field of occupant-centric control, limited attention is paid to the hot water systems, while they have to follow an energy-intensive operational strategy mainly due to the uncertainty of demand. Therefore, the first project of this thesis is focused on integrating occupant behavior into the control of with-tank water heating systems in buildings. Heat pump is used as the case study water heating system, as it is the most energy-efficient system with a constantly increasing number in the building sector in Europe [75], but is also more complicated to optimally control due to the variations of energy efficiency with the outdoor air and hot water tank temperature [76]. This chapter proposes a Reinforcement Learning control framework, that learns the hot water use behavior of occupants, and accordingly plans heating schedules to reduce energy use while ensuring the comfort and health of occupants. Maintaining the comfort aspect means that when the occupants demand hot water, the temperature of supplied hot water should be above the comfort level of 40 ℃. A common health threatening risk in hot water systems is the growth of Legionella, a bacteria that grows in water between 20 ℃ and 45 ℃, and can be transferred to occupants by breathing in the contaminated water droplets, for example, during taking shower [77]. The proposed framework also learns to periodically over-heat the tank to eliminate the risk of Legionella growth and ensure the health of occupants with minimum energy use. To represent the real hot water use behavior of occupants without imposing any risk on their comfort and health, in this study the hot water use behavior of a residential house in Switzerland was monitored, and the collected data were used in a simulation environment to assess the control framework. The performance of the proposed framework is compared with the conventional rule-based control method.

DeepSolar control framework: In many cases, the hot water system is not stand-alone but is combined with the hydronic space heating system. With the increasing penetration of solar power generation in buildings, the integration of solar panels into the combined space heating and hot water system is becoming more interesting as in such combination the hot water tank can be assumed as low-cost energy storage. However, this combination further complicates the development of optimal control solutions for experts. For programming optimal control solutions, experts should answer many questions such as *when the occupants use hot water?*, *when is the best time to charge the hot water tank?*, *how to get the best use of free solar power?*, *how to take into account the variations of heat pump efficiency?*, etc. Even if a model predictive control framework is developed for such combined systems, due to the unique occupant behavior and system specifications in each building, the control solution cannot be easily transferred to another building. The advantage of the *model-free learning-based approach* over the *expert-based approach* is further highlighted when it comes to more complicated systems. This chapter proposes an occupant-solar-centric control framework for solar-assisted heat pump space heating and hot water production system. The control framework learns the hot water use behavior of occupants as well as stochastic solar power production, and accordingly schedules the heating cycles of the tank, and adjusts the indoor air temperature setpoint to reduce the energy use while ensuring the comfort and health of occupants. Comfort in this case includes both indoor thermal comfort and hot water comfort. Similarly, to represent the realistic conditions, the hot water use behavior of three residential houses in Switzerland are monitored, the solar radiation and other weather data are collected from the nearby weather stations, and the collected dataset was used in the simulation environment to assess the control framework for three different houses. The proposed framework is then compared to the conventional rule-based method.

DeepValve control framework: Space heating in offices usually assume pre-defined static schedules for occupancy, while the occupancy of each office is different from others and varies over time. As a result, for many hours in a week, energy is used to heat a vacant office unnecessarily. Due to the slow response time of heating systems, space heating systems cannot be only activated when occupants are detected, instead should start working for enough time before the arrival of occupants. Thus, the third chapter proposes an occupant-centric control framework for space heating in offices, that learns the occupancy schedule, and the thermal response time of each office, and schedules the heat emission to an office accordingly to ensure occupant comfort with minimum energy use. This control framework is designed to be installed on the heat emission system of each office and thus can be retrofitted to any hydronic heating system (regardless of the layout of the central system) to provide adaptation to the occupancy schedule. This framework is first tested in simulations, using real-world occupancy data collected from

several other studies, to ensure superior performance compared to the rule-based methods. Then, the control framework is experimentally implemented in an environmental chamber to evaluate the adaptability of the framework to the type of heat emission system and response time of the office in a real-world setup. Figure 1.5 visually presents the roadmap of this thesis.

In summary, to explore the potential of Reinforcement Learning for the occupant-centric control, the *DeepHot* and *DeepValve* frameworks are focused on the residential buildings, while the *DeepValve* framework is focused on the office buildings. By covering both residential and office buildings, this research demonstrates the great potential of Reinforcement Learning on a variety of building topologies.

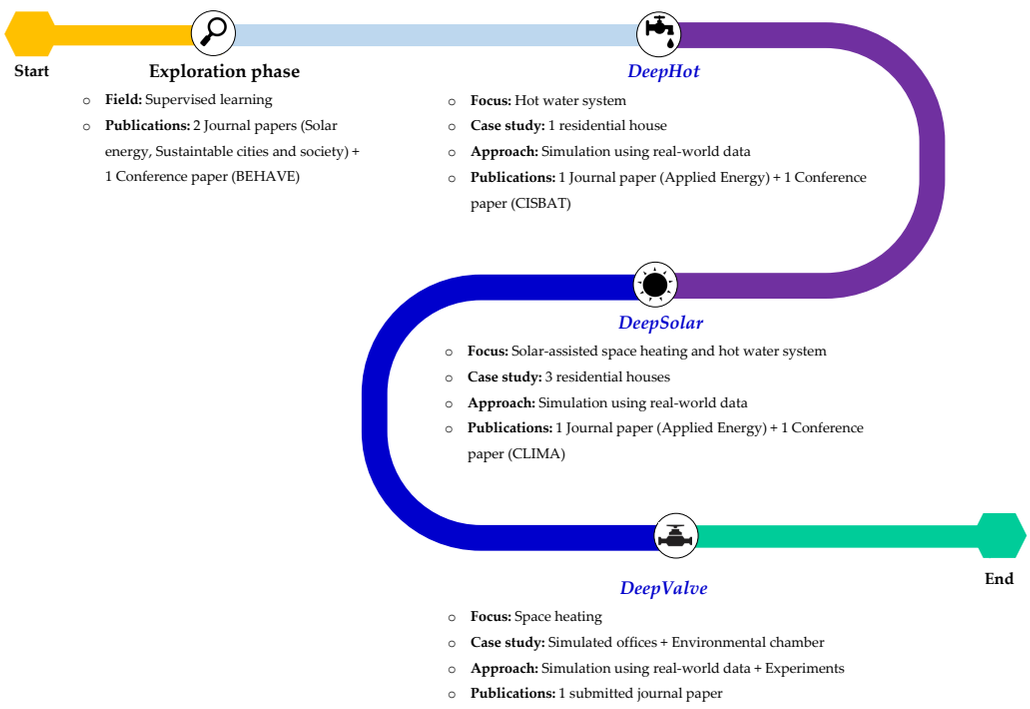


Figure 1.5: Thesis roadmap

Chapter 2

Background knowledge

2.1 Reinforcement Learning in simple words

This section:

Provides a simple introduction into Reinforcement Learning

Discusses different taxonomies of Reinforcement Learning

This section aims to provide a conceptual introduction into Reinforcement Learning (RL), a general understanding of RL, the main concepts, and different methods. Generally, the field of Machine Learning can be divided into 3 categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. In Supervised Learning, a labeled dataset, including inputs and corresponding outputs, is provided to the algorithm. The algorithm then learns to predict the right outputs given the inputs. If the outputs are continuous values, it is called a Regression problem, and if they are categories, it is called a Categorization problem. The term "Supervised" is used since the algorithm is trained by providing the true answer to it (i.e., the true outputs for a set of inputs). On the other hand, Unsupervised Learning is when an unlabeled dataset is provided to the algorithm, and the algorithm is supposed to learn the underlying patterns in the data, for example, to perform clustering. In RL, the algorithm learns how to interact with an environment to maximize a predefined reward. Different from Supervised and Unsupervised Learning methods, in RL there is no dataset, instead, the algorithm learns optimal behavior based on interacting with the environment and observing the feedback. Different Machine Learning algorithms differentiate in terms of how they receive feedback after they make a decision/prediction. In Supervised Learning, the algorithm will immediately know how accurate was its prediction by observing the ground truth data. On the other hand, in Unsupervised Learning no feedback is provided for the

algorithm. RL lies in the middle, as the feedback (reward) is provided with a delay after few interactions with the environment. Figure 2.1 summarizes the comparison between different Machine Learning methods.

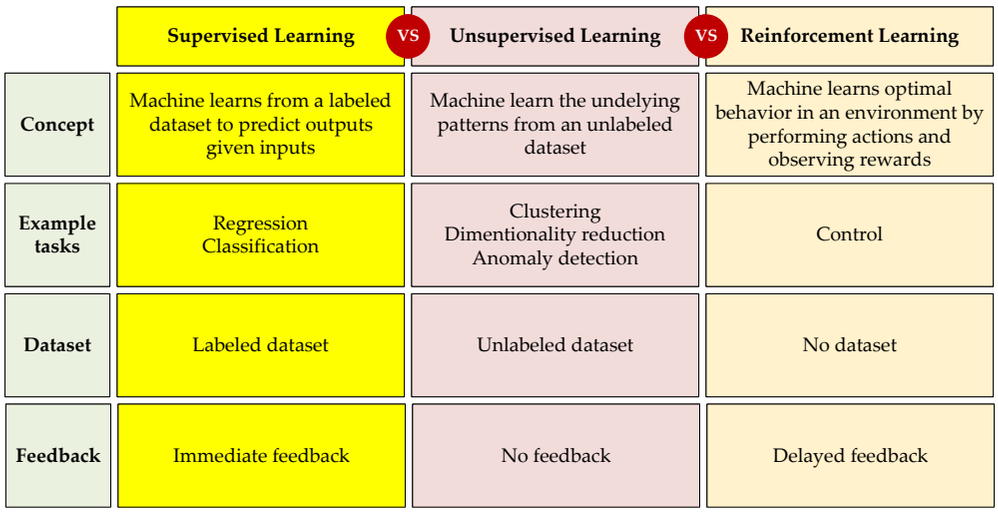


Figure 2.1: Main categories of Machine Learning

2.1.1 Definition and main components of RL

RL can be defined as *a learning technique*, in which an agent interacts with its environment, and uses feedback from the environment to determine the best possible action to maximize a defined reward [78]. This learning technique is inspired by neuroscience and how the human brain works [79]. An RL framework is consisted of 4 different components shown in Figure 2.2, each component is explained below:

- *State:* State is a numerical representation of the current condition of an environment, that is designed to provide relevant information to the decision to be made. Talking about HVAC control, the state can include current indoor air temperature, and predicted outdoor air temperature over the next timestep, that can help the controller to make a better decision [80].

- *Action:* Action is the decision made by the controller to be performed on the environment. In an HVAC control problem, the action can be the indoor air temperature setpoint, turning ON/OFF a heating system, or adjusting fan speed [80].

- *Environment:* Environment is the system to be controlled. It can be represented by two mathematical functions. The first function is *Transition probability*, which determines what would be the next state of environment s_{t+1}, if the action a_t is

performed when the environment is at state s_t. In other words, it is mapping the state and action of the current time step to the state of the next time step [80]. The second function is the *Reward function* which determines the immediate reward of taking action a_t when the environment is at state s_t. It is therefore a mapping from states and action to the reward.

- *Agent:* Agent is the learning controller, which tries to find optimal policy (π), which outputs an optimal control action for each state of the environment. The objective of an RL agent is to maximize the total future reward after a series of states and actions. How the agent learns the optimal policy is dependent on the RL algorithm.

An important assumption in RL is that this sequential decision making process is a Markov Decision Process (MDP). MDP states that the future state s_{t+1} is completely decided by the current state s_t, and is independent from the previous states s_{t-1}, s_{t-2},.... This means that the agent can select next optimal action only having the present state, without the need to remember the whole history [81]. To hold the markovian property, the state vector should be properly designed to include all necessary information for the agent to decide the next optimal action. Even a good algorithm can fail if some pieces of important information are not included in the state vector. The state vector can include the value of a parameter over a few previous timesteps (look-back vector) or the predicted value of a parameter over a few next timesteps. On the other hand, including more parameters in the state vector increases the computational effort and learning time for the agent, which is known as the curse of dimensionality [80]. Therefore, there is a tradeoff in the design of the state vector. The state vector should include necessary information, with minimum number of parameters. Some methods are proposed to include more information while avoiding curse of dimensionality. For example, Ruelens et al. [82] used an autoencoder to compress ten previous indoor air temperature and control signals data into 6 hidden states. Investigation of pre-processing methods to compress a long state vector can be a topic of future studies on RL-based building control frameworks.

The curse of dimensionality also includes in the selection of possible actions. A large number of possible actions can over-complicate the decision making for the agent. A building control problem can include many components (several fans, pumps, valves, etc.) that increases the number of required actions to be included. A possible method to avoid including many actions is to integrate RL with conventional control methods, where the RL acts as a supervisory control and conventional control directly controls the system. An example is to use the RL for activating or deactivating a conventional control method used to track the selected setpoint (*DeepValve control framework in this study*). In this integration, the adaptive potential of the RL is combined with the robustness of the conventional controls.

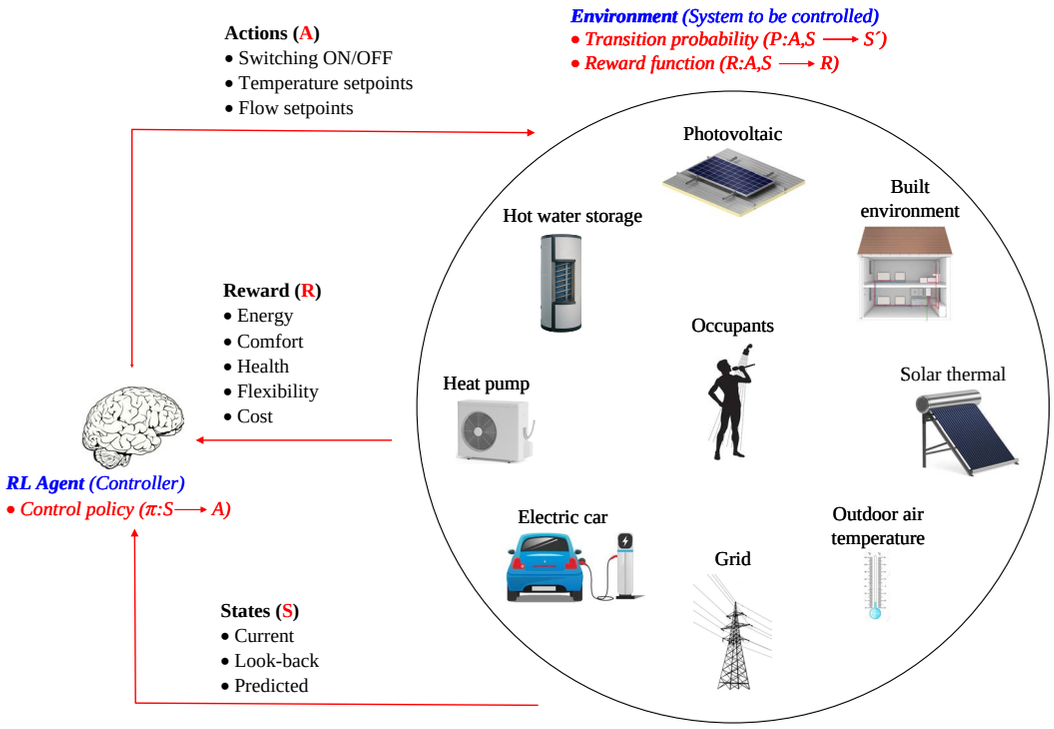


Figure 2.2: Reinforcement Learning in a nutshell

2.1.2 Different taxonomies of RL algorithms

RL includes a wide variety of algorithms, that can be categorized based on different criteria; therefore, there are multiple taxonomies for RL algorithms. The main taxonomies of RL are discussed below and summarized in Figure 2.3.

Model-free RL VS Model-based RL

Depending on *whether the agent has access to the model of the environment or not*, the RL algorithms can be divided into model-based or model-free methods. In model-based methods, the model of the environment, which includes the transition function $p(s_{t+1}|s_t, a_t)$ and reward function $r(s_t, a_t)$, is either given to the agent or should be learned by the agent prior to controlling the environment [80]. This model can be a data-driven or a physics-based model. In model-based RL, the agent knows how the environment will respond to the performed actions and can plan accordingly, which in turn improves the data efficiency of the RL framework [83]. However, developing or learning an accurate model of the environment is time-consuming and requires expertise and labor work [84]. Also, a once accurately developed model of a system can become inaccurate over time due to, for instance, renovation or aging of the system [85]. Model-free RL, on the other hand, does not rely on any model and learns the optimal control policy from interaction with the environment and observing the feedback from the environment [80].

If a model of the environment is available, there are alternative methods such as MPC that can be used. One of the main advantages of RL over alternative control methods is the fact that the RL agent can learn how to control a system from scratch without any prior knowledge, eliminating the cost and effort for model development and providing a higher level of adaptation to the changes [74]. Due to this reason, model-free RL is far more popular than model-based RL [40, 41, 74].

Value-based RL VS Policy-based RL

RL algorithms can also be classified *based on the optimization process to generate the optimal action*. Before explaining this categorization, a few terms should be explained.

- Policy $\Pi_\theta(a|s)$: Is mapping function from state to action. This mapping function can be a neural network with parameters of θ.
- Return (G_t): Is the total summation of discounted rewards, presented in Equation 2.1.

$$G_t = R_{t+1} + \gamma R_{t+2} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2.1}$$

  In which γ is used to reduce the importance of future rewards with respect to the immediate rewards, which is usually desired in practice.

- Value function: There are two types of value functions: State-value function, which measures the goodness of a state as shown in Equation 2.2 or state-action-value function which measures the goodness of performing action a when the environment is in state s, as represented in Equation 2.3.

$$V_\pi(s) = E_\pi[G_t|S_t = s] \tag{2.2}$$

$$Q_\pi(s, a) = E_\pi[G_t|S_t = s, A_t = a] \tag{2.3}$$

Given these definitions, value-based RL methods aim to learn the value/state-value function to determine the optimal action. For example, $Q_\pi(s, a)$ can be represented by a neural network, taking s vector as input and calculating the $Qvalue$ of each action as the output. Then, the action with the maximum expected $Qvalue$ is selected as the optimal action. On the other hand, policy-based RL methods aim to directly learn the parameters of the policy function $\pi_\theta(a|s)$. The policy function can be a neural network with parameters θ, taking s vector as input and generating the optimal action a as the output. On the overlap of these two categories, there is the actor-critic method which learns both value-function and policy-function. The value-based methods are suitable

for problems with a discrete action space while policy-based methods are suitable for problems with a continuous action space. Value-based methods are more data-efficient and less sensitive to hyper-parameters [86].

On-policy RL VS Off-policy RL

The RL algorithms can be categorized based on *how the data required to update the policy are collected*. The policy that is followed to take actions is called *behavior policy*, and the policy that is used to determine optimal action is called *target policy*. If the policy is updated based on the latest generated data, and that policy is then used to generate new data, then the behavior policy and target policy are the same. These kinds of RL algorithms are called *on-policy*. In other words, in on-policy RL, the same policy is used for data (experience) generation and action selection. However, the algorithm can collect experience data in a buffer, and periodically select some of the collected data to update the behavior policy. Therefore, the behavior policy is updated periodically and the data used to update the behavior policy are not necessarily the latest data. Therefore, the behavior policy is not the same as the target policy. These algorithms are called *off-policy*.

On-policy RL: Behavior policy=Target policy

Off-policy RL: Behavior policy $\neq$ *Target policy*

Online RL VS Offline RL

If the behavior policy is used to directly interact with the environment and generate data, that is then used to update the target policy, the RL algorithm is an *online RL*. Another approach is when the RL agent does not have access to the environment to interact with and to collect the new data from, and, instead, it is provided with a fixed dataset of transitions to learn the optimal policy. This approach is called *offline RL*, and it follows a similar training principle of supervised learning.

Online RL: Behavior policy used to generate data

Offline RL: There is no behavior policy

The four explained taxonomies are summarized in Figure 2.3. It should be noted that these 4 categories are not exclusive, for example, an algorithm can be off-policy, model-free, offlie, and value-based.

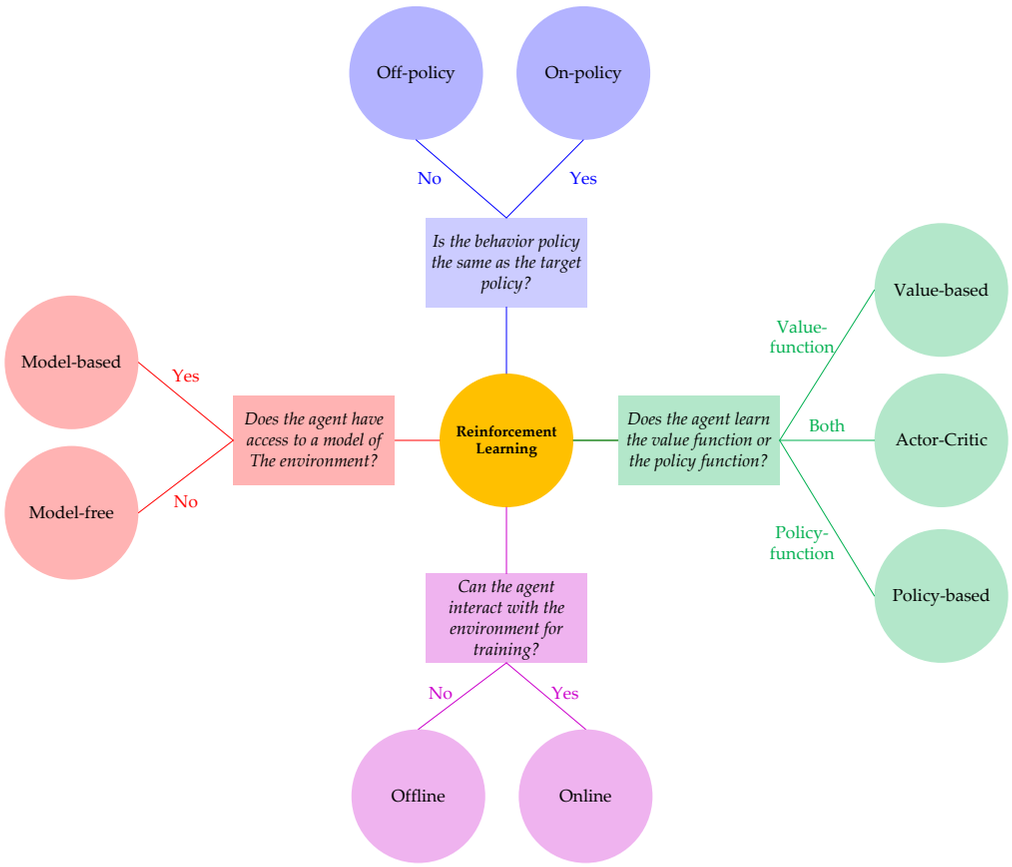


Figure 2.3: Different taxonomies of Reinforcement Learning algorithms

2.2 Introduction to Double Deep Q-Learning

This section discusses:

What is Q-learning?
What is Deep Q-learning?
What is Double Deep Q-learning?

This section provides the required theoretical understanding of the double deep Q-learning, which is a value-based RL method used in this study. To understand this method, first, the concept of Q-learning is explained, then the deep Q-learning method is described and the limitation of this method is explained, and, finally, the double deep Q-learning method is explained, which is meant to solve the limitation of simple deep Q-learning method.

2.2.1 Q-learning

Q-learning, as the name implies, is about learning the $Q(s_t, a_t)$ function. $Q(s_t, a_t)$ function is the value of performing action a_t when the environment is at state s_t. This value function tends to combine the immediate and long-term impact of performing action a_t into one single value. Equation 2.4 represents the Q function.

$$Q(s_t, a_t) = r(s_t, a_t, s_{t+1}) + \gamma.max_a Q(s_{t+1}, a) \tag{2.4}$$

When the environment is in the state s_t, by performing the action a_t, the environment transits to the next state s_{t+1} . Since the agent observes these transitions, the immediate reward $r(s_t, a_t, s_{t+1})$ can be directly quantified. But the Q function also needs to evaluate the goodness of the performed action over a long time. When the environment transits to s_{t+1}, the long-term effects of the performed action are not observed, and the only way to somehow estimate these long-term effects. The long-term goodness of the performed action is estimated by asking how good the new state of the environment s_{t+1} is. The goodness of this new state s_{t+1} can be measured by calculating the Q value of all possible actions in the new state and taking the maximum $max_a Q(s_{t+1}, a)$. Depending on the relative importance of the long-term rewards versus the immediate reward, which depends on the problem, the long-term reward is multiplied by a discount factor $\gamma \in [0, 1]$ to adjust its importance relative to the immediate reward.

Different Q-learning methods are all about how to learn this Q-function, which is used to select the best action. To gain a better understanding, we start by explaining the traditional learning methods and follow the evolution of methods to the double deep Q-learning method used in this study.

2.2.2 Tabular Q-learning

The most traditional way of learning Q-function is called tabular Q-learning. In this method, the discretized states and actions are sorted as the columns and rows of a table, as shown in Figure 2.4 The value of performing action a in state s is the intersection of the state and action in this table.

	a_1	a_2	a_3	a_4
s_1	1	4	3	6
s_2	2	7	10	4
s_3	3	10	8	12
s_4	15	9	7	6

Figure 2.4: Tabular Q-learning example

The learning process starts with initializing the Q values in the table by some guessed values. After performing the action a_t when the environment is in the state s_t, the immediate reward $r(s_t, a_t, s_{t+1})$ is observed. The future reward $max_a Q(s_{t+1}, a)$ is also calculated by looking at the row of the next state s_{t+1} and taking the maximum value of all possible actions. The Q value of the state s_t and action a_t is calculated by the Equation Q function and replaces the previous value. This process continues, and after many interactions, the Q values converge to the true values. While this process is simple and easy to be implemented, it is limited to environments with discrete states and actions. For continuous states, discretization of states will result in a high dimensional table which is not efficient for learning.

2.2.3 Deep Q-learning

Since tabular Q-learning is not efficient in dealing with high dimensional or continuous states, an alternative is deep Q-learning. In deep Q-learning, a Neural Network is used for mapping states to Q-values. In this method, a neural network takes the states as the inputs and estimates the Q value of each action at that specific state. To train this Neural Network, the estimated Q value of each action should be compared with a ground truth value. The ground truth value is calculated as Equation 2.5.

$$Q^*(s_t, a_t) = r_t + \gamma . max_a Q(s_{t+1}, a) \tag{2.5}$$

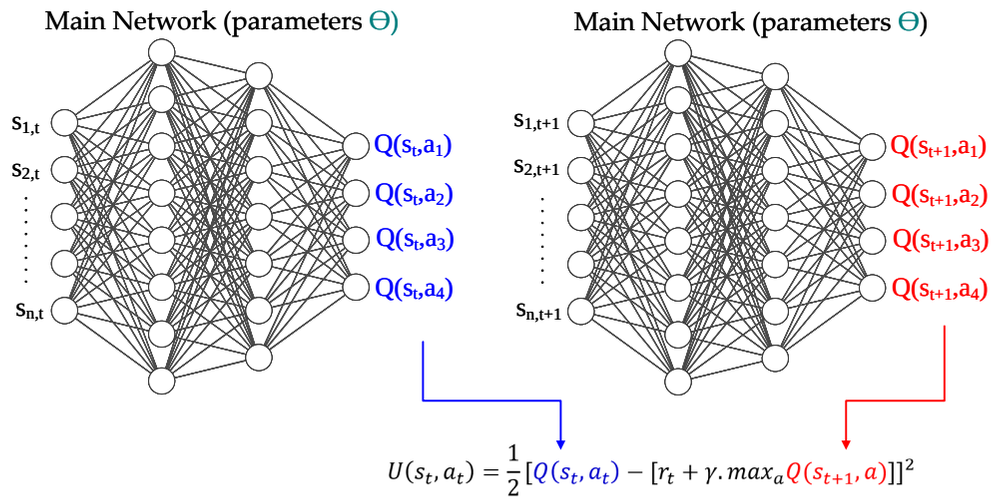
The ground truth Q value is composed of two terms, immediate reward r_t and the discounted future reward $\gamma . max_a Q(s_{t+1}, a)$. The immediate reward is directly calculated after taking action on the environment and observing the next state. But similar to the tabular Q-learning, the future reward cannot be directly calculated and should be somehow estimated. Deep Q-learning also follows a similar method to tabular Q-learning for estimating future rewards. In this method, the same neural network that was used to estimate the Q value at the state s_t (with the same parameters θ that were used to estimate $Q(s_t, a_t)$) is also used to estimate the Q value of all actions over the next state s_{t+1}. Then, the maximum of these Q values is used as an indicator of the future rewards. The ground truth value is therefore composed of the actual reward observed by performing an action and the discounted future reward estimated by the same Neural Network. Having the estimated Q value $Q(s_t, a_t)$ and the ground truth Q value $Q^*(s_t, a_t)$, the typical training process of Neural Network is used to minimize the loss function and update network parameters θ. The loss function used to train the Neural Network is shown in Equation 2.6.

$$U(s_t, a_t) = \frac{1}{2}[Q(s_t, a_t) - Q^*(s_t, a_t)]^2 = \frac{1}{2}[Q(s_t, a_t) - [r_t + \gamma . max_a Q(s_{t+1}, a)]]^2 \tag{2.6}$$

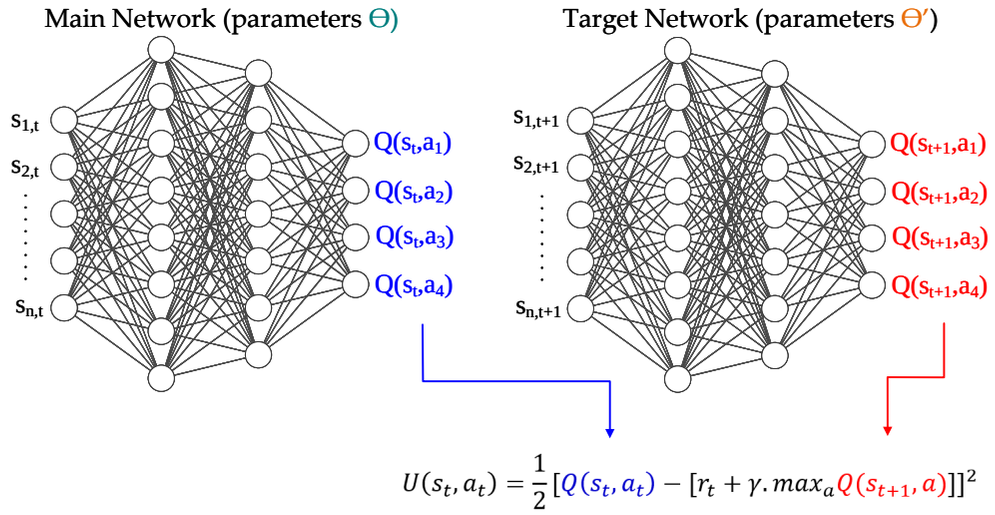
This training process of deep Q-learning is visually presented in Figure 2.5 (a). A potential issue with this method is that the estimated Q value ($Q(s_t, a_t)$) and part of the ground truth Q value ($\gamma.max_a Q(s_{t+1}, a)$) are calculated with the same Neural Network (same parameters θ). Therefore, the estimated and ground-truth values can move in the same direction. This can, in turn, result in the overestimation of the Q value [87].

2.2.4 Double Deep Q-learning

To solve the overestimation issue, an alternative method is introduced, which is called double deep Q-learning. In this method, two neural networks are used, one for estimating the Q value ($Q(s_t, a_t)$) known as the main network, and the other for estimating the future rewards ($\gamma.max_a Q(s_{t+1}, a)$) known as the target network. The parameters of the main network are continuously updated, but the parameters of the target network are periodically updated based on the parameters of the main network. Therefore, the network used to calculate the estimated Q value and the ground truth Q value are different. This is known to provide better stability and solve the issue of overestimation. The process of training a double deep Q-learning is visually presented in Figure 2.5 (b). Further details about Double Deep-Q learning method are available in the publication by Marszal-Pomianowska et al. [88].



(a) Deep Q-Learning



(b) Double Deep Q-Learning

Figure 2.5: Training process of (a) Q-Learning method (b) Double Deep Q-Learning method

Chapter 3

DeepHot: An occupant-centric control framework for hot water systems

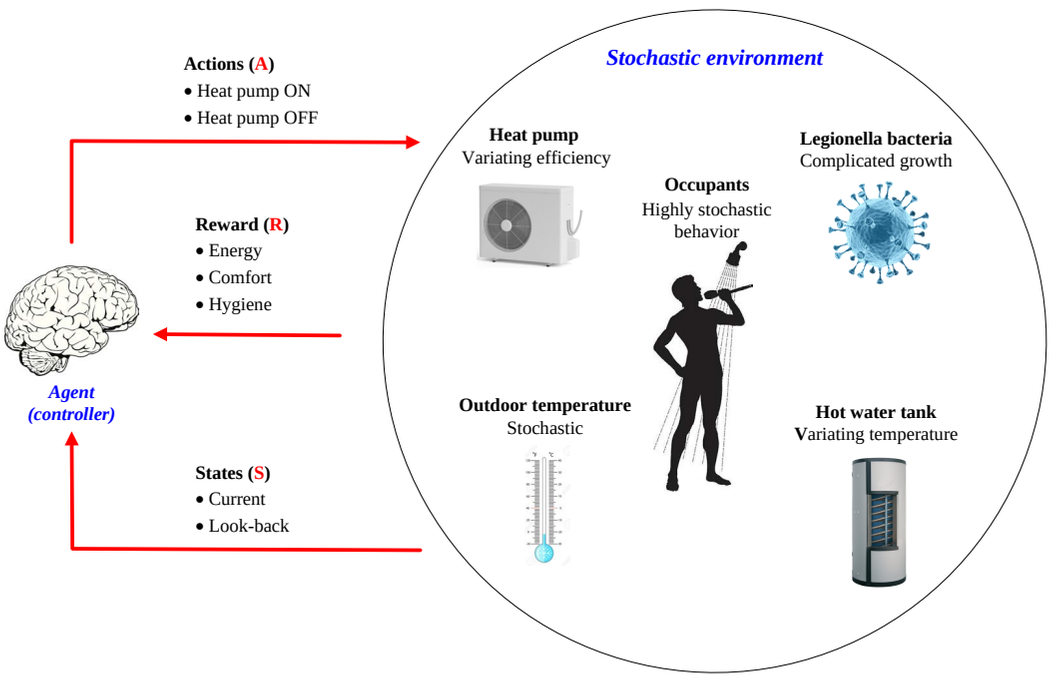


Figure 3.1: *DeepHot* control framework in a nutshell

3.1 Abstract

Occupants' behavior is one of the most significant sources of uncertainty for optimal scheduling and operation of building energy systems. Consequently, the conventional control method of water heating systems follows a conservative operational approach

to ensure the occupants' comfort regardless of their stochastic behavior. Due to the conservative operational approach, the energy use for water heating has not changed significantly over the past decades and can reach up to 70% of total energy use in modern buildings with low space heating demand. This study introduces a control framework based on Reinforcement Learning, which can autonomously learn and adapt to the stochastic occupant behavior and environmental conditions to ensure a balance between water hygiene, comfort, and energy use in water heating systems. To ensure transferability, a model-free approach is implemented. Also to achieve a fast convergence while being model-free, an off-site training stage integrating a stochastic hot water use model is included in the training methodology. As the second step, the control framework is implemented on an actual hot water use dataset collected over 29 weeks from a residential house in Switzerland. The performance of the proposed control framework is compared to the conventional rule-based controller that is commonly used in hot water systems. Despite the unusual hot water use behavior of occupants during the COVID-19 pandemic, results indicate that the proposed control framework could successfully learn and adapt to the occupant behavior and achieve 23.8% energy saving while maintaining the occupants' comfort and water hygiene. The adaptive nature of the proposed control framework provides a significant potential in reducing the discrepancy between supply and demand in hot water systems.

3.2 Introduction

Despite the improved efficiency of water heating systems, the energy use for hot water production has not changed considerably over the generations of buildings [89], while the energy use for space heating and cooling has reduced significantly. Consequently, the share of hot water energy demand in the total heat requirement can be up to 70% of total energy use in modern low energy buildings [90]. The hot water systems not only account for an increasing share in the building energy demand, but also they can provide several other potentials to the building energy system. Some examples are renewable energy integration [91, 92], demand response scenarios [93], and load shifting [94]. Thus, hot water systems are increasingly becoming a vital factor for efficient energy management in buildings.

Hot water usage in households is strongly correlated with Occupant behavior [95], which is highly stochastic, differs from building to building, and varies over time in the same building [96, 97]. Consequently, the highly stochastic nature of hot water use behavior is a major challenge for developing energy-efficient control methods for hot water systems. Due to this challenge, conventional control methods follow a conservative approach to make sure enough hot water is available whenever it is required. The most

conventional control method for hot water systems is a rule-based control, also known as the two-point control, in which the hot water system is switched ON when the tank temperature is below a lower threshold (usually 65 °C) and is switched OFF when the tank temperature is above a higher threshold (usually 75 °C) [97, 98]. Although it is a simple and easy-to-use controller, it is static and detached from the Occupant behavior, and therefore energy-intensive due to over-preparing hot water. A more advanced control method is the Model-Predictive Control (MPC) in which the predictions of stochastic phenomena such as renewable energy availability and environmental conditions can be used for optimal operation of energy systems in the built environment [99]. Predictions of Occupant behavior can also be integrated as another stochastic parameter in the MPC control framework. However, as there is no physical model for Occupant behavior, either stochastic models or data-driven models should be used. Stochastic models are usually developed based on the data of several buildings of a specific type and, therefore, mimic the hot water use behavior of these specific buildings [100]. However, the hot water use behavior of the target building can be very different from the buildings used to generate stochastic models. Besides, the hot water use behavior of occupants can change over time, for example, by the change in occupancy due to special situations such as COVID19-imposed home office working. On the other hand, data-driven models are usually based on supervised learning, which can learn the hot water use behavior of a specific building by using data collected from the same building to predict the future demand [97]. However, training of a supervised learning model requires enough historical data of hot water demand, which limits its application to buildings with an existing dataset. In addition to the difficulties to predict Occupant behavior, MPC requires an accurate model of the system, which is time-consuming to develop and needs to be updated (tuned) for every other system with different parameters [101]. Furthermore, an accurately developed model could become fairly inaccurate over time due to the aging or renovation of the water heating system, and subsequently sub-optimal policies would continue to be executed because of the obsolete dynamic model [102].

Given the importance of occupant behavior in the energy use of buildings, an ideal control system should be able to integrate the stochastic occupant behavior into the control loop [103]. With recent advances in the Internet of Things (IoT) technologies (cheap sensors and microelectronic boards, efficient online platforms, etc.) on the one hand and vast progress in Machine Learning methods on the other hand, the implementation of new control systems utilizing occupants-related data looks ever more realistic now. Reinforcement Learning (RL) has recently gained increasing interest as a data-driven control method for the built environment, as it can continuously adapt to the stochastic occupant behavior, the time-varying environmental conditions, and the system aging with no need for a rigorous mathematical representation of the system [104]. Subsequently, RL

potential has been demonstrated for various applications in the built environment, such as optimal control of space heating and cooling [105], lighting [106], windows [107], air handling units [104], solar energy integration [108] and water heating systems [102].

A very common application of RL in the built environment is to optimize the control of indoor air temperature. In these studies, the agent usually tries to balance the energy consumption and occupants´ comfort. Brandi et al. [109] proposed a double deep Q-learning framework for optimal control of indoor air temperature using a water-based space heating system in an office building. In this study, two important aspects of RL are discussed in detail, which are the design of state-space, and the comparison of static versus dynamic deployment. In the static deployment the learning process of the agent only happens during the training phase, while in the dynamic deployment the agent continuous to learn during the deployment phase. While the dynamic approach may enhance the performance, it requires more computational resources. Zou et al. [104] developed an RL framework for optimal control of air handling units. The agent tries to minimize the energy consumption while preserving the comfort of occupants quantified by Predicted Percentage of Discomfort (PPD). To represent the system, Neural Network models of air handing units were developed using two years of operational data recorded by the building automation system. The application of RL for occupant-centric control has been mostly focused on indoor air temperature control, and little attention has been paid to the other domains of occupant-building interaction. Park et al. [106] investigated the application of RL for occupant-centric control of lights in offices. The agent tried to balance visual comfort and energy consumption. A device called Lightlearn was specifically developed for the experiments to allow both manual and automatic switching of lights. This allows monitoring the interactions of occupants with the lighting system over the learning phase. Han et al. [107] proposed an RL framework to optimize window opening/closing. The agent in this case tried to make a balance between thermal comfort and indoor air quality.

Performed literature review shows that very few researchers have addressed the application of RL in hot water systems. Kazmi et al. [102] proposed a model-based RL control framework to balance comfort and energy use in heat pump water heating systems. In particular, they used a model-based heuristic method that incorporates the storage tank state and occupant hot water use behavior into the optimal control problem. The models for heat pump, storage tank, and occupant behavior prediction are probabilistic data-driven models that are trained with historical data. In another study, Kazmi and Ali [108] proposed an RL framework based on deep Q-learning for optimal operation of photovoltaic-assisted domestic hot water production systems. The proposed framework tried to maximize the self-consumption of photovoltaic panels by shifting the consumption into the production period. Correa-Jullian et al. [110] evaluated the application of tabular Q-learning for the optimal operation of a heat pump water

heater integrated with the solar thermal panels and heat recovery chiller. The proposed framework tried to determine the operational schedules of the solar field and heat recovery chiller to make a balance between energy efficiency and comfort indicators. Table 3.1 presents a summary of the above-mentioned and few other studies that have investigated the application of RL on different aspects of the built environment.

Table 3.1: Summary of recent studies on the application of Reinforcement Learning-based controls in the built environment

Reference	Year	Purpose	Method	States	Actions	Reward terms	Baseline method	Comparison to the baseline
[111]	2016	Optimal schedule for domestic hot water production	Model-based heuristic method	Temperature distribution inside the storage tank	Heat pump ON/OFF	Energy consumption, thermal comfort	Rule-based	20%-27% energy saving
[108]	2017	Maximizing self-consumption of PV panels integrated to hot water system	DQN	Hot water demand, storage tank temperature, heat pump and PV panels power	Heat pump ON/OFF	Self-consumption, thermal comfort	Rule-based	18% to 72% increase in self-sufficiency
[102]	2018	Optimal control of hot water system	Model-based heuristic method	Tank temperature, Heat pump state (ON/OFF), predicted demand, predicted ambient temperature	Sequence of control actions	Lost occupant comfort; energy usage; exploration bonus	Rule-based	20% energy saving
[112]	2018	Optimal control of electric with-tank water heater	fitted Q-iteration	Day number, quarter of day, temperature along the height tank	Heater ON/OFF	Cost of consumed electricity	Rule-based	Reduced total energy cost, by 15%

[106]	2019	Occupant-centric control of lighting system	Iterative probabilistic-based model	Occupancy, switch position, indoor light level, day period	Switch ON/OFF	-1,+1 or 0 depending on different combinations of occupancy, switch position and comfort	Schedule-based and occupancy-based control scenarios	Better comfort (based on the survey results)
[110]	2020	Operation schedule of solar thermal hot water system	Tabular Q-learning	Temperature of multiple points in system, solar radiation, weather conditions	operational status ON/OFF signal to solar field and chiller pumps	Solar field energy gain; system energy use	Thermal comfort (supply temperature) Rule-based	21% higher of a defined performance metric
[104]	2020	Optimal control of air handling units	DDPG	Environmental conditions, system operating parameters (temperature, humidity, etc)	Fan speed, Valve state, Damper position	Energy use, thermal comfort (predicted percentage of dissatisfied)	Rule-based	27% to 30% energy saving
[113]	2020	Load shifting in a cooling supply system	DQN and DDPG	Predicted electricity prices and weather data, previous consumer loads and indoor air temperatures	Chiller setpoint temperature and flow, valves´ position	Energy costs compared to the baseline, thermal comfort (penalty for temperature violations)	Rule-based	20% lower operation costs

[109]	2020	Optimize indoor temperature control	Double DQN	Indoor and outdoor temperature-related terms, water supply and return temperatures, supplied heating, remaining time to pre-defined occupancy	Supply water temperature setpoint	Energy consumption, thermal comfort (penalty for indoor air temperature violations)	Combination of rule-based and climatic-based logics	5% to 12% energy saving
[114]	2020	Maximize self-consumption of PV generation connected to heat pump and battery	Model-based fitted Q-iteration	Hour of the day, day number, storage tank heat content, battery state of charge, PV production	Start heating cycle for heat pump(with 3 different target temperature), Charge, discharge or idle for battery	Cost of injecting the locally produced power into the grid	Rule-based	Increase of self-consumption by 14%
[115]	2020	Optimal setpoint temperature for air conditioning	Q-learning	AC Power, AC set-point, AC power status, indoor temperature, outdoor temperature, outdoor humidity, time ON, time OFF	Indoor air set point (from 15 different possible options)	Energy consumption, thermal comfort, uncertainty in power and temperature prediction, operation smoothness	Rule-based	7.69% comfort improvement and 3.59% energy saving
[107]	2020	Optimal control of windows opening/closing	Q-learning and SARSA	Indoor and outdoor temperature, wind speed, solar radiation, outdoor air quality index, window position (open/close)	Switch (changing from open to close or vice versa) or inaction	Thermal comfort, air quality, energy	Actual Occupant behavior	Higher reward than actual Occupant behavior

[105]	2021	Optimal control of space cooling and heating simultaneously	multi- task DDPG	Outdoor and indoor temperature for each zone, retail price	Setpoint for each point	Total energy consumption cost, temperature violation penalty	Rule-based	6% to 10% energy cost reduction
[116]	2021	Optimal control of indoor and domestic hot water temperature, and PV production	Deep Q-learning	Outdoor temperature, indoor temperature, DHW tank temperature, PV production, hour of the day	Space heating ON, domestic hot water ON, system OFF	Comfort (assumed as acceptable temperature interval for DHW and indoor air), energy not covered by PV	Rule-based	1.5% to 16.6% energy saving, 10.2% load shifting for PV
[117]	2021	Optimal control of a cooling system with ice-storage tank	Double Q-learning	Power use of chillers, storage stage of charge, ambient temperature (current and forecasted), electricity price, hour, day, month, electricity price forecast, cooling demand	3 different possible combinations of supplying, charging or inactive for chillers and tank	Electricity costs compared to the baseline	Rule-based (direct load coverage)	8% to 12.5% energy cost saving

This literature review on the application of RL in water heating systems identified the following limitations:

- **Model-based:** Most of the studies have relied on the modeling of the system and predictions of occupants' behavior. Although modeling of the system increases the data efficiency as the agent does not need to learn the system model from scratch, it reduces the transferability of the control framework to the other buildings, poses a risk of model inaccuracy and predictions error, and increases the computational load [118].
- **Hygiene aspect:** The risk of Legionella growth is not integrated into any of the proposed frameworks, while it is an essential issue to be considered in the optimal operation of hot water systems. Energy-saving efforts by lowering hot water tank temperature can result in a higher risk of Legionella. For instance, the increasing number of Legionella infection cases in Switzerland has elaborated the Legionella control efforts [119]. Subsequently, ignorance of the hygiene aspect will limit the widespread implementation of a control framework.
- **Direct implementation on the target building:** At the beginning of the learning phase, the agent does not have enough experience with its environment and can perform non-optimal actions causing user dissatisfaction. However, the previous studies have directly started the learning process on the target houses.

The highly adaptive nature of RL makes it a great choice for hot water systems where the occupant behavior plays a vital role. Currently, there is limited knowledge and practice on the use of RL-based controls for hot water systems. The aim of this research is to fill the knowledge gap by developing an RL control framework with the following main features:

- **Model-free:** It does not require a complex thermodynamic model to represent the actual system. Rather, it learns the system and occupant behavior through the interactions with the environment. Therefore, the controller can be implemented to the systems with no prior knowledge of affecting parameters such as heat pump capacity, tank size, efficiency, number of occupants, etc. This feature significantly improves the transferability of the proposed framework to different buildings.
- **Off-site training:** A drawback of the model-free approach is a long learning period as the control framework needs to learn the models from scratch. To reduce the learning period without relying on any model, this framework includes an off-site training stage where an agent gains prior knowledge in the safe simulation environment before implementation on a target building. Over the off-site training,

a stochastic model is integrated into the framework to mimic the occupant hot water use behavior.

- **Hygiene supervision:** The hygiene aspect of the hot water tank is also included in the framework to make sure that energy-saving efforts by the agent do not increase the risk of Legionella growth. The agent should learn to regularly overheat the tank, taking into account the demand profile, to eliminate the bacterial growth in the tank while saving energy.
- **Double deep Q-learning:** All the reviewed papers on hot water systems with model-free methods either have implemented tabular or deep Q-network. However, it is known that both methods can overestimate the value of an action and continue to generate a non-optimal policy. The proposed framework in this research is based on double deep Q-learning, which is known to solve the issue of overestimation [120].

This research addressed a novel research topic on occupant behavior-centric building design and operation, which is the focus of a current research project of the international energy agency (IEA-EBC Annex 79) [103].

3.3 Methodology

Reinforcement Learning (RL) is a machine learning method in which an agent learns to choose an optimal action at each state of the environment to maximize a reward over time. Through interactions with the environment and learning from mistakes, the agent becomes increasingly more intelligent at decision-making under uncertainty. As shown in Figure 3.2, at each time t, the agent receives the current state of the environment (S_t), then chooses an action (A_t) from a set of possible actions. Caused by this action, the environment moves to a new state (S_{t+1}), and the agent receives a reward (R_{t+1}) that indicates the goodness of the performed action. This transition experience ($S_t, A_t, S_{t+1}, R_{t+1}$) is then stored as a single experience in the memory to be used for training the agent.

The goal of an RL agent is to maximize the sum of the discounted future rewards, known as the return function defined by Equation 3.1. The RL agent not only takes into account the immediate effect of performing an action but also considers the discounted future impact of the action. This feature makes it suitable for thermal systems with a slow response time due to their thermal inertia.

$$R = \sum_{t=0}^{\infty} \gamma^t R_t \tag{3.1}$$

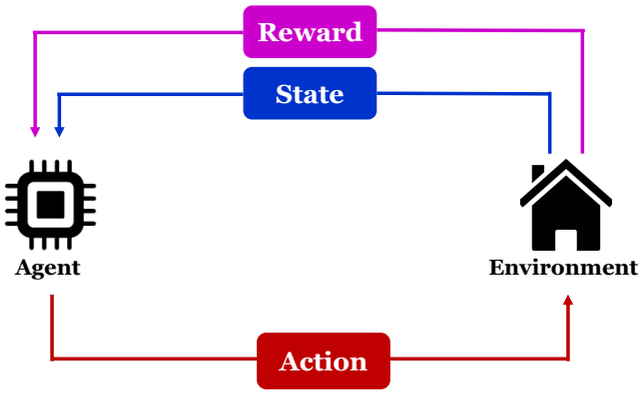


Figure 3.2: Interactions of agent and environment in a Reinforcement Learning framework [121]

The most common method of RL is deep Q-learning [106] in which a neural network takes the state of the environment as the input and estimates the value of performing each action as the outputs. In the conventional deep Q-learning method, the action value is calculated based on Equation 3.2:

$$Q(S_t, A_t) = R_{t+1} + \gamma max_a Q(S_{t+1}, A) \tag{3.2}$$

in which $Q(S_t, A_t)$ is the value for the pair of S_t and A_t, R_{t+1} is the immediate reward, γ is the discount factor and $max_a\ Q(S_{t+1}, A_t)$ is the estimated maximum value for the next state which shows how good is the next state. Therefore, in this method, estimating the value of performing an action in the current state relies on estimating the value of the next state using the same network. It is indicated that this initial estimation of Q-value can result in an overestimation and selecting non-optimal actions by DQN. To solve this issue and also provide better stability over the training phase, in 2015 Hasselt et al. [122] proposed Double Deep Q-Network (DDQN). In DDQN, the term $Q(S_{t+1}, A)$ in Equation 3.2 is calculated using another neural network (called Target Network) which its parameters are updated based on the main network but with less frequency. Since the DDQN method solves the problem of overestimation by conventional DQN, the control framework is developed based on the DDQN method in this work.

3.3.1 State, action, and reward design

Proper setup and sufficient training are two key factors to obtain a good performance in RL. State vector should provide enough information to the agent for decision making.The reward function also should be properly formulated to impose the desired balance between the contrasting objectives when it is minimized. While the state vector should include

enough information, a high dimensional vector should also be avoided as it can increase the error and computational time. Figure 3.3 shows the state and action at a time step.

State

At each time step (indicated by the "current time" in Figure 2) the state vector includes the following elements:

- **M previous hours of demand intervals:** The agent needs to learn the hot water use behavior of the occupants to make the right decision that ensures the occupants' comfort. Several studies have shown that an array with a sequence of previous demands is the best feature for predicting future demands [92, 97, 123, 124]. To include the previous demands in the state vector, the demand data is converted into the demand intervals of 5 liters (e.g., a demand of 3 liters is in the first interval and therefore would be 1). Considering that the effect of demands within the same interval on the tank temperature is almost the same, categorizing the continuous demand data into intervals makes it easier for the agent to learn the demand pattern. An array including the demand intervals over the M previous hours is included in the state vector. A sensitivity analysis is done to select the proper value of M;
- **N previous hours of ambient temperature:** As the outside air temperature affects the heat pump Coefficient of Performance (COP), the agent should be able to learn the outdoor air temperature and shift the heating schedules as much as possible to the hours with higher ambient temperatures. Similarly, N previous hours of the outdoor air temperature are also included in the state vector to enable the agent to learn variations of ambient air temperature;
- **Storage tank temperature $T_{storage}$:** The current tank temperature is also included in the state vector to inform the agent how much energy is currently stored in the tank;
- **Hour of the day:** Studies show that the hot water use behavior is highly correlated with the time of the day [92, 97, 123, 124]. To further assist the agent to predict the upcoming hot water demand, the upcoming hour (for which the decision is going to be taken) as an integer between 1 to 23 is included in the state;
- **Type of the day:** Studies highlighted that the hot water use behavior of the working days are similar to each other and different from the weekends [97, 123]. Therefore, the type of the day is informed to the agent as 1 for working days and 0 for weekends.

Actions

The decision to be taken by the agent is whether to turn ON or OFF the heat pump for the upcoming hour. Including a limited number of actions can reduce the learning time significantly.

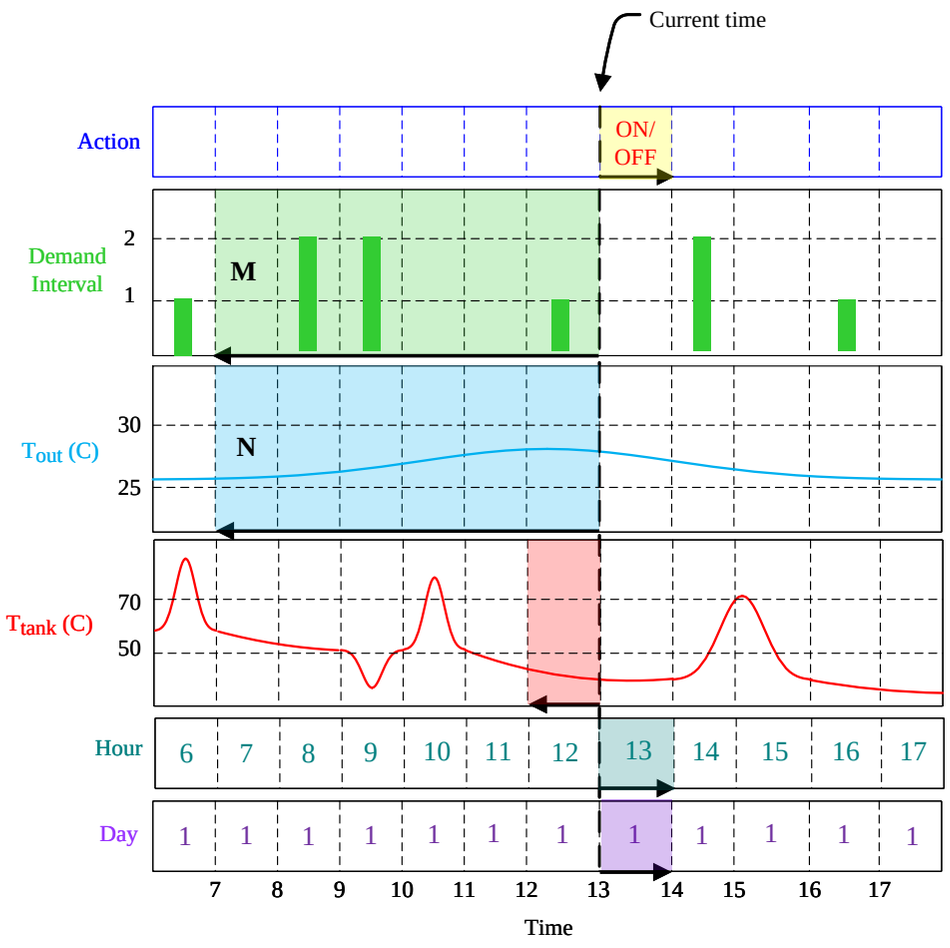


Figure 3.3: Representation of state and action design at each time step

Reward

The reward function is defined in Equation 3.3, which includes the competing terms multiplied by coefficients to weigh the importance of each term.

$$R = \begin{cases} \text{if } Demand > 0\text{:} \\ -a \times P_{hp} - b \times max(40 - T_{tank}, 0) - c \times (Hours_{from\ overheat} - 24, 0) \\ \text{if } Demand = 0\text{:} \\ -a \times P_{hp} - c \times (Hours_{from\ overheat} - 24, 0) \end{cases} \tag{3.3}$$

In this equation, P_{hp} is the electricity consumption of the heat pump (kWh), T_{tank} is the

storage tank temperature (℃), and $Hours_{from\ overheat}$ is the total number of hours from the last time when the tank was heated above 60 ℃. The first term is called the *Energy term*, which gives a negative reward to the agent proportional to the energy consumption. The second term is called the *Comfort term*. The temperature of 40 ℃ is known as the highest temperature that occupants may request after mixing cold and hot water at the end uses [98, 125, 126]. As the desired temperature differs for each fixture and can vary over time [127], we considered 40 ℃ as a lower limit that can satisfy all the type of demands. The comfort term gives a negative reward to the agent if water is supplied at a temperature lower than 40 ℃. As shown in Equation 4.12, when there is no demand, the comfort term is excluded from the total reward. This allows the agent to save energy by lowering the tank temperature when there is no demand predicated for a long time in future. The third term is the *Hygiene term*, as a health-related term to make sure that the agent is aware of the risk of Legionella growth in the tank and sterilizes the tank periodically. This term is based on the recommendation to heat the hot water tank at least once a day at 60 ℃ for 11 min [98, 128]. Kenhove et al. [129] developed a model to quantify the concentration of Legionella in the hot water system and integrated the model into a rule-based controller to sterilize the tank when the concentration of Legionella bacteria in the tank is estimated to be higher than a threshold. The resulted heating schedule showed that the tank was heated above 60 ℃ once a day, which further confirms the validity of the recommendation in [98, 128]. To include the overheating rule in the reward function, the hygiene term counts the number of hours from the last overheat and if it passes from 24 it gives a negative reward proportional to the hours passed. Therefore, the agent needs to overheat the tank above 60 ℃ at least once a day to keep this term zero. The coefficients a, b and c adjust the relative importance of each term.

3.3.2 Proposed sensing layout

The agent should receive certain data from the environment at each time step to extract the required states and calculate the reward. To implement the proposed control framework in practice, a set of sensors shown in Figure 3.4 should be installed on the heat pump. An important consideration in developing the proposed framework is to rely on the minimum required number of sensors, which in turn increases the economic feasibility of this solution and reduces the probability of malfunctioning due to the fault of sensors. Only four sensors are required to implement the proposed framework, including an air temperature sensor to measure the temperature close to the evaporator, a power meter to measure power use of the heat pump, a water flow sensor at the tank outlet to monitor hot water use behavior, and a temperature sensor at the middle of the tank to monitor tank temperature. The use of the temperature sensor in the middle of the tank is a conservative approach to further ensure the occupants' comfort. It means that the comfort limit of 40

℃ is considered for the middle part of the tank, while water at the tank outlet is at a higher temperature. Thus, a temperature sensor at the tank outlet is not required.

In this study, the power and tank temperature sensors are considered in the energy model of the system implemented in TRNSYS. The air temperature data from a nearby weather station (downloaded from Agroscope, a competence center for agricultural research in Switzerland [130]) is used in the TRNSYS model as the air temperature sensor data. Hot water use behavior during the off-site training is modeled using a publicly available stochastic model [100]. During the on-site training and test, the real-world measured data from the case study building is used to represent an actual behavior. The data collection layout to monitor the hot water use behavior of occupants is presented in the case study section.

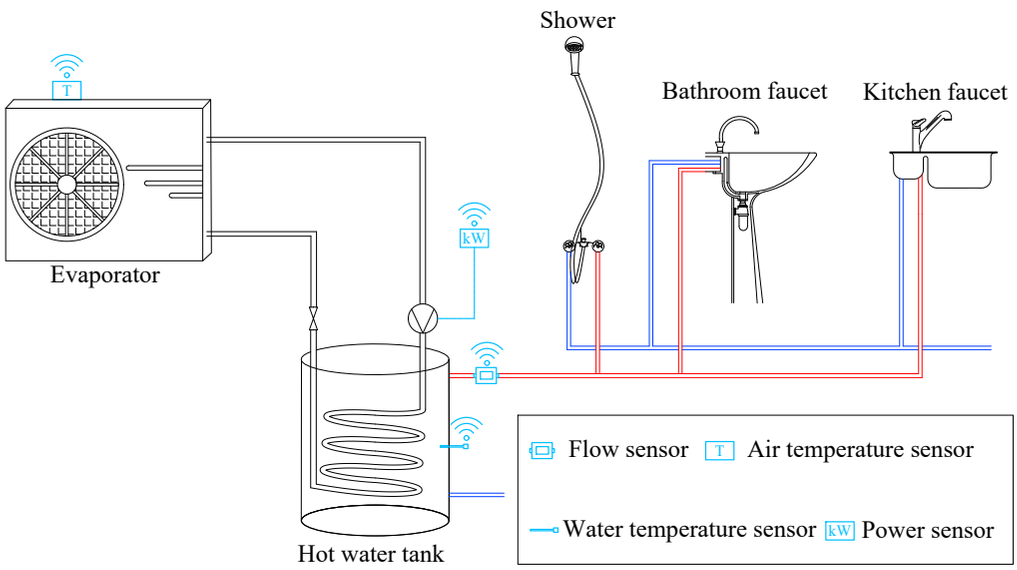


Figure 3.4: Required sensors to implement the proposed framework in practice

3.3.3 Training and deployment

The proposed framework unfolds over three sequential stages, as shown in Figure 3.5.

- **Off-site training:** This stage aims to provide prior experience to the agent in the lab (off-site) before being trained on the target building (on-site). This stage is considered due to two main reasons: first and foremost, once the agent gains prior experience with similar data, it will converge faster to the optimal operation

policy on the target building. Secondly, it will reduce the probability of disturbing occupants' comfort over the on-site training phase. The RL agent during the training should be able to directly interact with the environment and learn the optimal behavior over the interactions. To provide these interactions, the transient model of an air-source heat pump connected to a storage tank is developed in TRNSYS. The agent model developed in Python (Tensorforce library) can send actions to the TRNSYS model and receive back the next state and calculate the reward function accordingly. Details of the agent-environment interaction are discussed in the next sections. To simulate the environment, hot water demand data and outdoor air temperature should also be provided to the TRNSYS model of an air source heat pump. To provide the hot water demand data, a stochastic hot water use model developed by Ritchie et al. [100] is used to generate 6 years of hourly hot water use data. This stochastic model is recently developed based on actual measurements from 77 households in South Africa. It was reported that the simulated hot water profiles closely match the actual measurements and were representative enough to be used for modeling the hot water use behavior in residential households [100]. The outdoor air temperature data is provided collected from a nearby weather station (from Agroscope). To improve the transferability of the framework over Switzerland, for each year of the off-site training phase the outdoor air temperature of a different city is used in the TRNSYS model. Six different Swiss cities are selected as shown in Figure 3.6. Cities are selected from the north, south, east, west, and center of the county to ensure that the agent gains experience with different climatic conditions of the country and can adapt quickly when implemented across Switzerland.

- **On-site training:** After gaining prior knowledge in off-site training, the agent needs to learn the specific system characteristics and occupant behavior in the target building. Accordingly, a training period on the target building is also necessary to obtain an optimal control performance. Twenty-seven weeks are considered for the on-site training to ensure that the agent will have enough time to adapt to the target household. The duration of 27 weeks is considered as a conservative choice because it is expected that the agent will converge in a shorter time. The minimum required duration can be selected by observing the variations of reward value. A sufficient on-site training period will ensure that the agent has obtained enough experience with occupants' behavior in the target house. Similar to the off-site training, the target system is modeled in TRNSYS to enable the interactions with the agent. However, it should be noted that the role of the TRNSYS model in this stage is different from the off-site training. In this stage, the TRNSYS model represents the target system, while in the off-site training it is a part of the framework to train the

model before being implemented on the target system. Therefore, the parameters of the model used in this stage are different from the off-site training. To represent the actual hot water use behavior of occupants, the demand data at this stage is based on the measurements from the target house, as described in the case study section.

- **Deployment:** Over the training stages, the agent is both controlling the system and learning from its mistakes to improve its policy. Over the deployment stage, however, the agent is no longer learning but only controlling the system. The deployment stage needs much less computational power. Therefore, the agent can be implemented on an inexpensive single-board computer (such as a Raspberry pi) embedded in the heat pump. To test the performance of the proposed framework, two weeks of deployment is considered in this study using the actual data of hot water demand and outdoor air temperature of the target house.

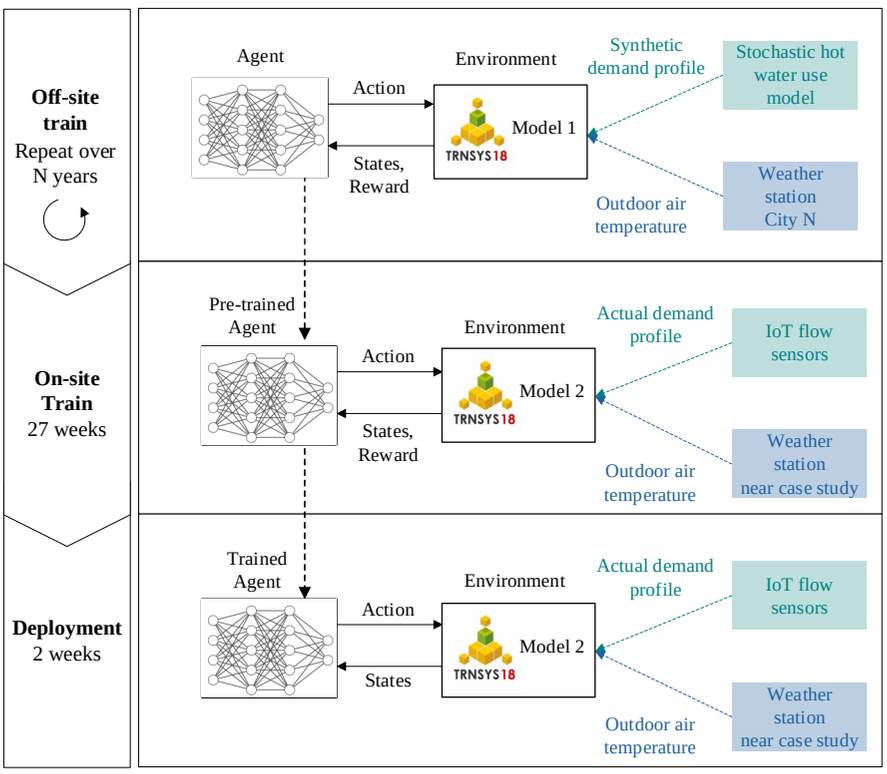


Figure 3.5: Different stages of the proposed framework

3.3.4 Agent setup

The agent model is developed in Python based on the Tensorforce library [131], which provides very customizable classes for modeling the agent and environment. Table 3.2

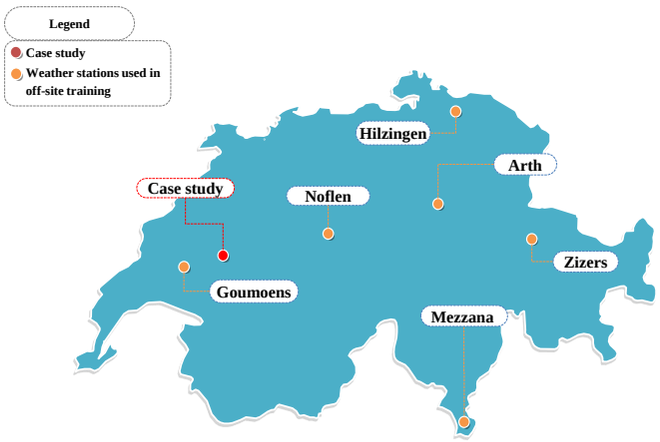


Figure 3.6: Location of the selected cities to use their temperature data in off-site training

shows the selected hyper-parameters after tuning. One of the important aspects of training an RL framework is the trade-off between exploration and exploitation. To maximize the rewards, the agent should select actions that are expected to achieve a higher reward in that state, which is called exploitation. On the other hand, it is desired that the agent also performs the actions that have not been experienced before in that state, which is called exploration. One of the well-known methods to balance the exploration/exploitation trade-off is the $\epsilon - greedy$ method, in which a small probability of ϵ is specified and agent performs exploration when a random number would be higher than ϵ. In this study, it is desired that the agent only performs exploration during the off-site training phase because the execution of random actions on the target house can disturb the occupants and reduce their satisfaction. Therefore, a linear decay is established for exploration, where the ϵ linearly decays from 0.9 to 0.001 at each time step over the first two weeks. The update frequency is set as 12 hours, so the agent updates its policy two times a day. The memory size is also considered as 20 weeks to keep enough history of occupant behavior.

3.3.5 Environment setup

An air-source heat pump connected to a storage tank is simulated in TRNSYS to represent the environment. The heat pump evaporator is placed outdoors to evaluate the agent adaptability in a more complicated situation where the COP varies with the outdoor air temperature. The proposed framework without any change can also be used when the evaporator is located indoor, which is easier for the agent to control. The parameters of the models used in off-site and on-site training are different. This is because the model

Table 3.2: Specifications of agent

Parameter	Value
Type	Double deep Q-network
Number of layers	2
Number of nodes in each layer	64
Nodes type	Dense
Activation function	tanh
Learning rate	0.0003
Batch size	24
Update frequency	12
Exploration	Linear decay
Memory size	3360

used in off-site training is a part of the framework, while the model used in on-site training is representative of the actual system. Using different model parameters for off-site and on-site training (as listed in Table 3.3) can further highlight the transferability of the proposed control framework.

Table 3.3: Main parameters of TRNSYS model

Parameter	Off-site training model	On-site model
Heat pump total cooling capacity (kW)	1.5	1.8
Heat pump sensible cooling capacity (kW)	1.36	1.63
Heat pump compressor power (kW)	0.5	0.6
Heat pump heat rejection rate (kW)	2	2.4
Storage tank volume (liters)	300	350
Thermal conductivity of the walls (W/m.K)	0.7	0.7

3.3.6 Agent-Environment interaction

At each time step, the agent should be able to perform an action on the simulated air-source heat pump and receive back the next state to calculate the reward. A Python function is developed to make it possible. This function can run the TRNSYS simulations from Python using the desired parameters. This function makes it very easy to automate repetitive simulations that is needed in this study. As illustrated in Figure 3.7, at the first time step the interactions start with the “Reset ( )” function that outputs a pre-defined initial state to start the process. Based on the given state, the agent selects an action and performs it on the TRNSYS model. This represents moving one step forward over time.

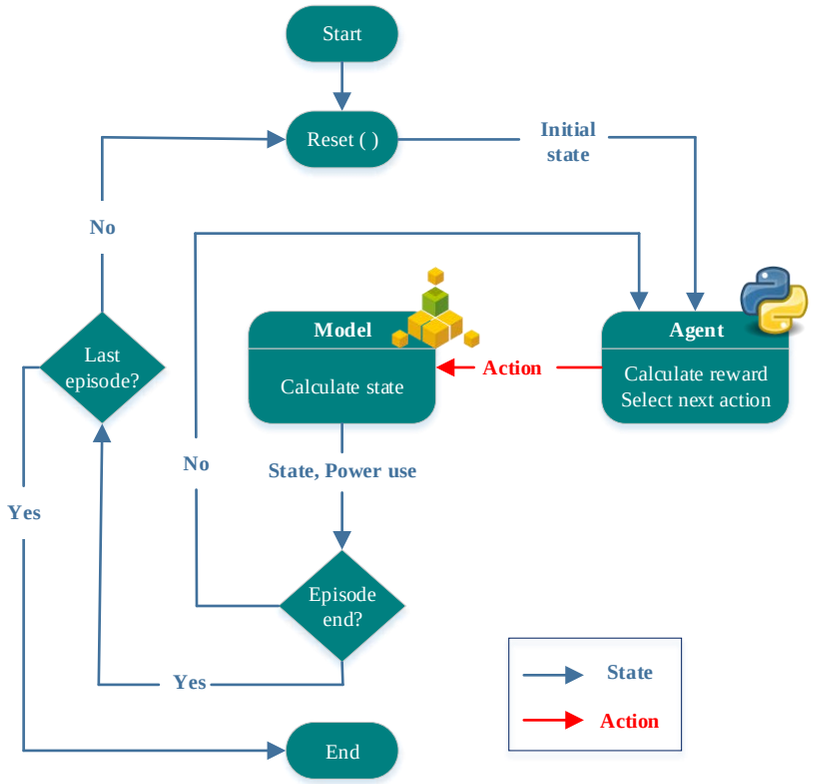


Figure 3.7: Flow of information during the interactions between the agent and environment

At each step, the outdoor air temperature and hot water demand of that specific hour are imported to the TRNSYS model from a dataset. The TRNSYS simulation is performed for a time-step of one hour. This simulation timestep outputs the state parameters and the additional parameters required for calculating the reward (such as the power use over the last hour). If the episode is not ended, the state is given to the agent to choose the next action. The episode length is set as 168 hours (1 week).

3.3.7 Baseline controller

The performance of the proposed framework is compared to a rule-based (two-point) controller, which is the most common control approach for with-tank water heating systems. The two-point controller turns ON the heat pump when the tank temperature falls below 65 ℃ and turns it OFF when the tank temperature exceeds 75 ℃. The set-points are selected based on the common practice [96].

It should be noted that the user adjusts the desired flow and temperature at the end-use

by mixing hot and cold water streams. Consequently, if hot water is supplied with a higher temperature, the user would mix a lower flow rate of the hot stream with a higher flow rate of the cold stream to achieve the desired flow rate and temperature after mixing. This should be considered in the comparison between the proposed and baseline control framework. To do so, it is assumed that both cases need to provide the same flow rate after mixing, with the same desired temperature of 40 ℃. Considering the cold water temperature of 10 ℃ the required flow rate of hot water to produce the desired water flow at 40 ℃ after mixing is calculated at each time step based on the supply temperature of hot water. In other words, it means that if a control method produces hot water with higher temperature, the supply flowrate will be lower accordingly.

3.3.8 Case study description

To monitor the actual hot water use behavior of occupants, a detached residential building in Switzerland is selected as the case study. The family living in this house is composed of two adults and three children. The hot water demand of this family is monitored for 29 weeks, from 28 August 2020 until 19 March 2021. The monitoring campaign is coincident with the COVID-19 pandemic. Consequently, occupants' schedule and hot water use behavior have been different from the typical weeks before mid-March'2020. To further evaluate how the occupancy schedule and hot water use behavior have been changed over the pandemic, survey responses about the occupancy are collected at the end of the monitoring campaign. The typical daily presence of occupants before and during the COVID-19 pandemic is shown in Figure 3.8 and Figure 3.9, respectively. As can be seen in Figure 3.8, before the COVID-19 pandemic most of the occupants used to leave home around 8 A.M. and come back around 6 P.M., which follows a typical occupancy pattern expected for a residential building. On the weekends, the occupants typically used to stay at home. However, as shown in Figure 3.9, during the COVID-19 pandemic (and therefore during the monitoring weeks) all the occupants used to work from home. This will change the routine of hot water demand expected for a residential building and can significantly increase the stochasticity of hot water demand. For example, the occupants are more flexible with showering time than before. Figure 3.10 shows the presence of occupants over different weeks of study. One of the children was present only one week and the other one only 8 weeks during the 29 weeks of the monitoring campaign. Varying the number of permanent occupants between 3 and 5 further increases the stochasticity of demand in the case study. In the survey, a question was also asked about whether the occupants' hot water use habits have changed during the COVID-19 pandemic compared to the normal weeks or not. The occupants' feedback shows that their usage of kitchen and toilet faucets is increased both in frequency and duration. Their showering frequency is reported to be the same as before, while its duration is reported

to be longer. Therefore, both the stochasticity and amount of hot water use during the pandemic have been different from the normal periods. This makes it harder for the agent to learn and predict the occupants' behavior. The outdoor air temperature of this case study is also obtained from a weather station of Agroscope located about 200 meters away from the building.

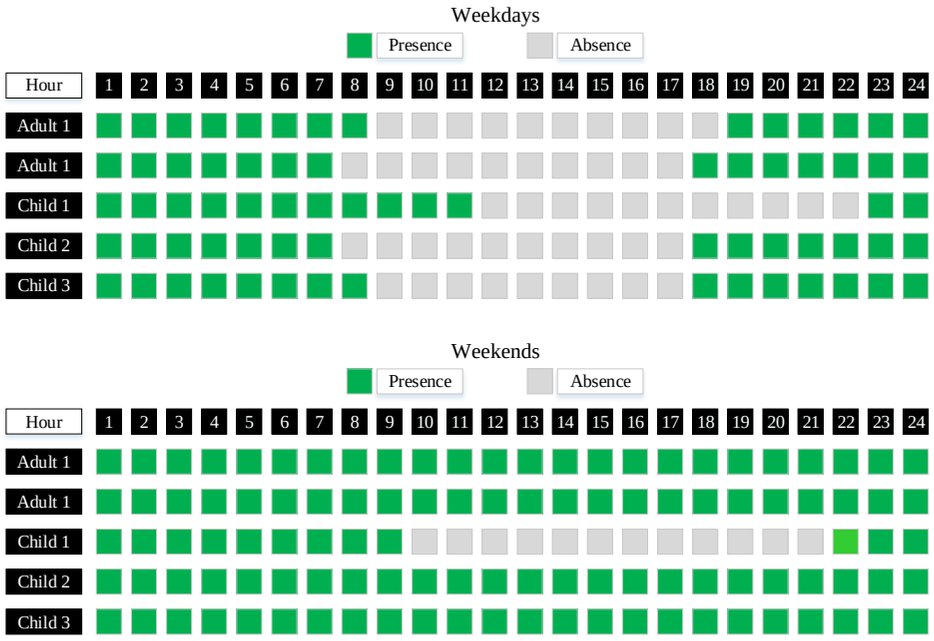


Figure 3.8: Daily schedule of occupants before COVID-19 pandemic

3.3.9 Monitoring campaign

To implement the proposed control framework, the hot water use behavior can be monitored using a single water flow sensor at the tank outlet, as shown in Figure 3.4. In this research, a more comprehensive monitoring layout is implemented to monitor the hot and cold water usage at each of the end-uses, such as shower, faucet, etc. This detailed monitoring is not necessary for this framework but was implemented to collect a more comprehensive dataset that can be used for other studies. There are particular challenges with this kind of monitoring layout. First of all, sensors should continuously work for several months while there is no plug available near to most of the end uses and wiring in some places such as shower is not easy. Secondly, long-term data storage using locally placed data loggers is expensive. Finally, a sensor placed in the shower needs to be water-resistant (withstand high humidity and water splashes). To address all these challenges, an IoT monitoring campaign was implemented with close collaboration with Droople company, a startup company in Switzerland [132]. Figure 3.11 shows the

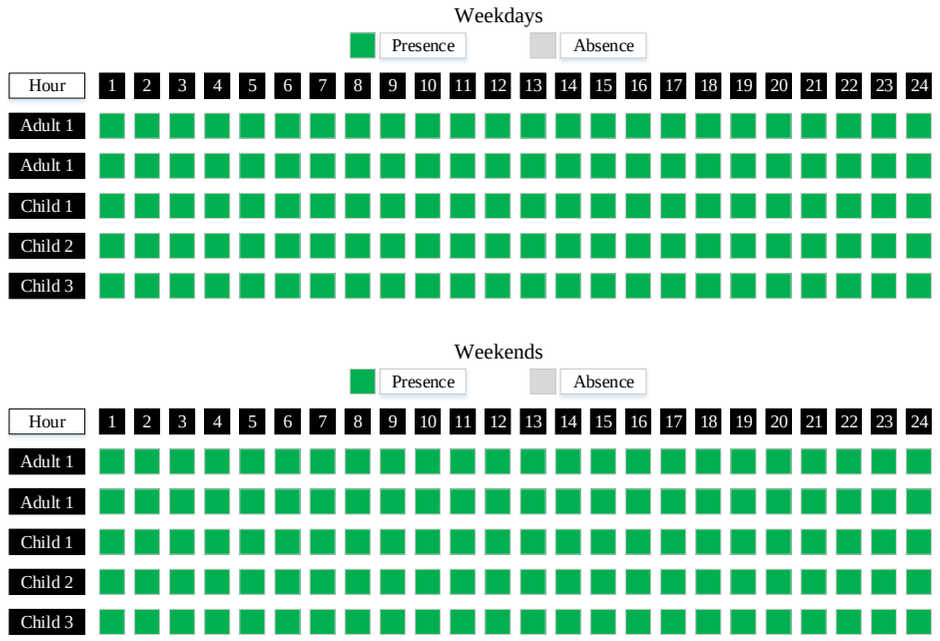


Figure 3.9: Daily schedule of occupants during COVID-19 pandemic

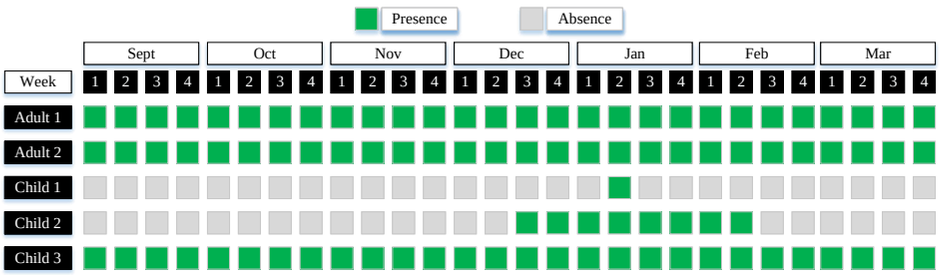


Figure 3.10: Presence of occupants over the weeks of monitoring campaign

data streaming architecture of the IoT sensors used in this research. A compact and cost-effective hall-effect flow sensor is used to measure the flow and send pulses to the hardware, called iLink. iLink is a board with a waterproof casing that works based on the LoRaWAN, a low-power and wide area network designed to connect battery-operated IoT nodes wirelessly. This allows the iLink modules to monitor the hot water use behavior for several months operating with a small battery. The data are then sent to the gateway, which acts as a bridge between the LoRaWAN and WiFi networks. The gateway connected to a power plug sends data to the cloud server for storing and monitoring in real-time. Figure 3.12 shows two examples of the installation of the sensors at the end uses. In total 11 sensors are installed, including one pair under the kitchen sink, 4 pairs under 4 faucets, and 1 in the shower.

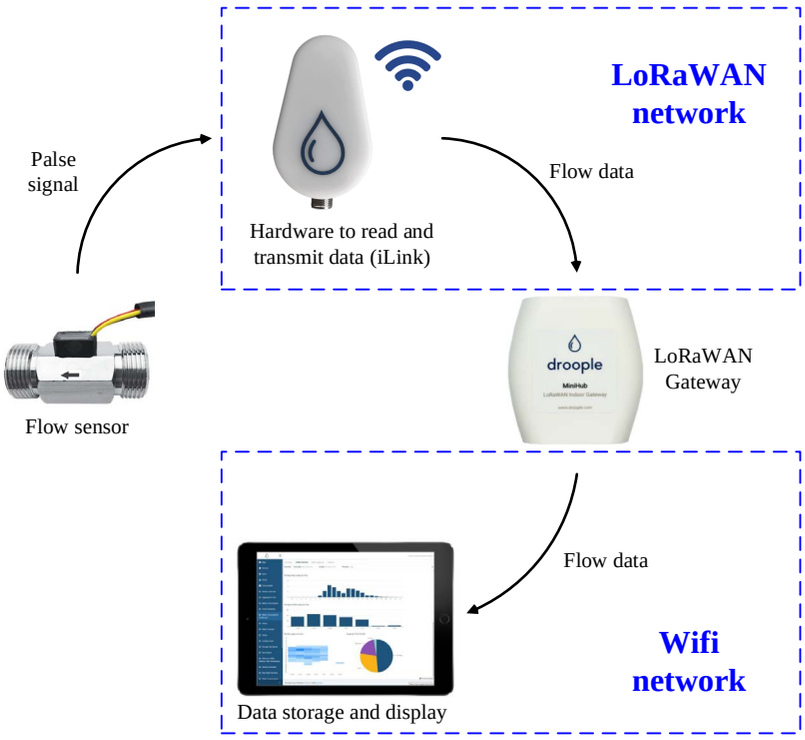


Figure 3.11: Data streaming architecture from Droople IoT sensors for monitoring hot water use behavior at each end-use

3.3.10 Implementation of the framework at the design stage or as a retrofit

The proposed framework can be implemented both at the design stage of the heat pump or as a retrofit to an existing heat pump water heating system. The only difference is the communication of the action signal to the water heating system. Suppose the control framework is going to be implemented at the design stage of the water heating system. In that case, the action signal is an ON/OFF signal similar to the conventional controllers. However, if the framework is going to be implemented as a retrofit to an existing water heating system it should communicate to the existing controller. Most of the tank water heating systems use a rule-based or two-point controller, as shown in Figure 3.13 (a). This controller takes the temperature sensor readings as an input to determine the ON/OFF signal for the heat pump. To retrofit the proposed framework, it might be desired not to replace the built-in controller but to integrate the proposed controller to the existing one to control the heat pump indirectly. In this case, the proposed controller should simulate the sensor behavior as the input to the two-point controller to impose the desired ON/OFF control action indirectly to the heat pump. As shown in Figure 3.13 (b), for the temperatures in the “Hysteresis” range, the ON/OFF action of the two-point controller is not only dependent on the current temperature but also on the previous temperatures.

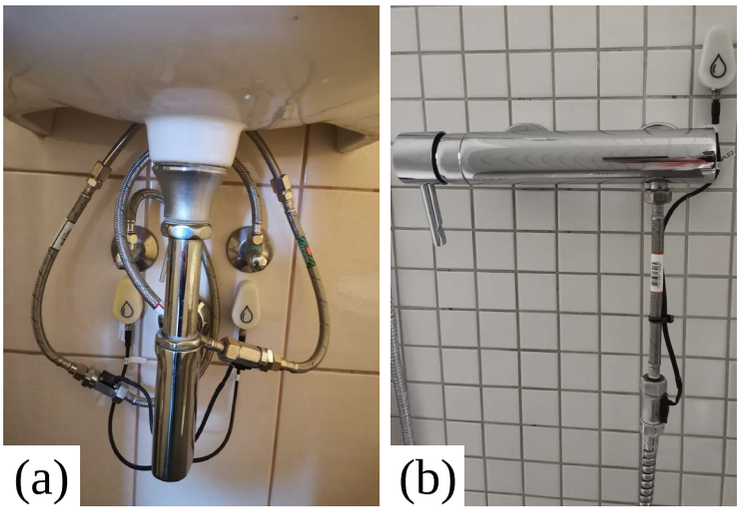


Figure 3.12: Installation of sensors at the different end uses: (a) a bathroom faucet (b) a shower

However, if a synthetic temperature above the "Turn-off" temperature limit (indicated in Figure 3.13) is sent to the two-point controller, this controller will certainly turn OFF the heat pump regardless of previous temperatures. Similarly, for a synthetic temperature below the "Turn-on temperature", this controller will certainly turn ON the heat pump. Accordingly, for a retrofit scenario the agent should be installed between the temperature sensor and the two-point controller, as shown in Figure 3.13 (b). The agent then takes the temperature sensor readings as an input and selects the optimal action. Then, it translates the action to a synthetic voltage signal and sends it to the two-point controller to impose an ON/OFF action.

3.4 Results and discussion

This section first discusses a primary analysis of the hot water use behavior during the monitoring campaign and sensitivity analysis of the framework parameters to select the best configuration for the framework. Finally, the detailed performance results of the selected configuration are presented.

3.4.1 Evaluation of hot water use behavior of occupants during COVID-19 pandemic

A better understanding of the routines in hot water use behavior helps properly design the control framework, particularly to define the number of previous demands to be considered in the state vector. Previous studies on residential buildings [97, 123, 124] have shown that hot water use behavior of the weekdays are similar to each other and

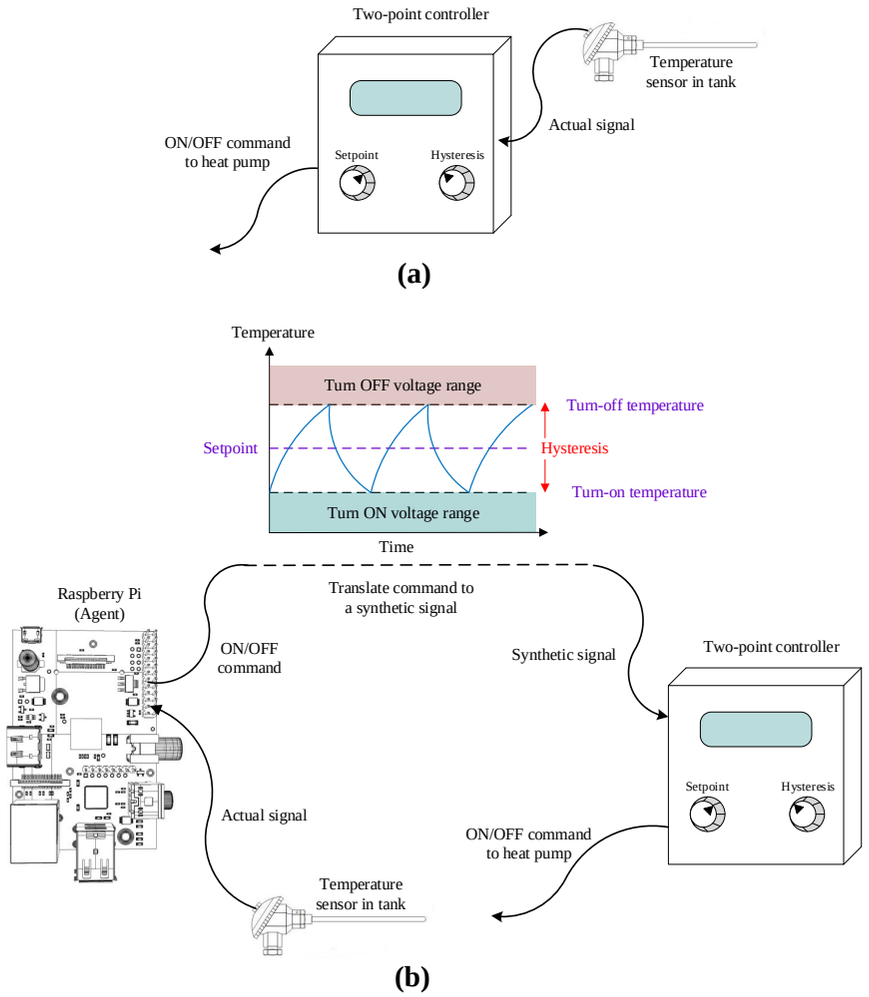


Figure 3.13: Integration of the proposed framework as a retrofit to the conventional water heating systems: (a) a conventional two-point controller, (b) Integration of the proposed framework into the conventional controller

different from the weekends. However, in this study the monitoring has taken place during the COVID-19 pandemic, and the occupants' schedule and habits have been different from the "normal" behavior. Therefore, in this section the recorded hot water use data are analyzed to better understand the Occupants' behavior and the differences between the pandemic period and the prior (normal) period described in the previous studies. Figure 3.14 shows the boxplots of hot water demand indicating how demand is distributed over the weekdays and the day hours. For this diagram, all the non-zero demands are separated and then categorized into days of the week (Figure 3.14 (a)) or hours of day (Figure 3.14 (b)). The average values over different days of the week are very similar, mainly because the occupants have always been present at their homes (as shown in Figure 3.9). The distribution of demands over the day hours indicates three peaks at 8, 10, and 21

o'clock. This is while previous studies reported two peaks, typically between 6-9 and 18-22 o'clock over the day [97, 123]. This is mainly due to showering in the morning and cooking or showering at the night. A third peak in the hourly average hot water demand at 10 o'clock can be explained by the flexible schedule of occupants staying at home.

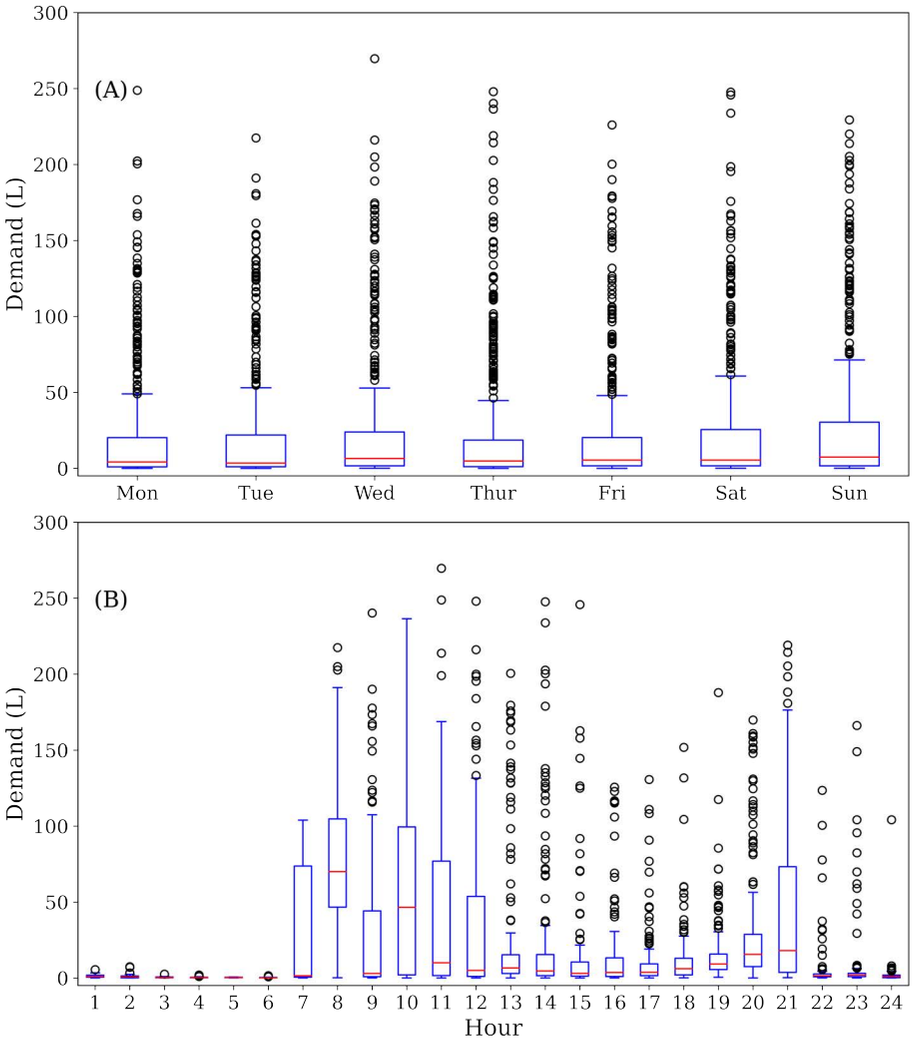
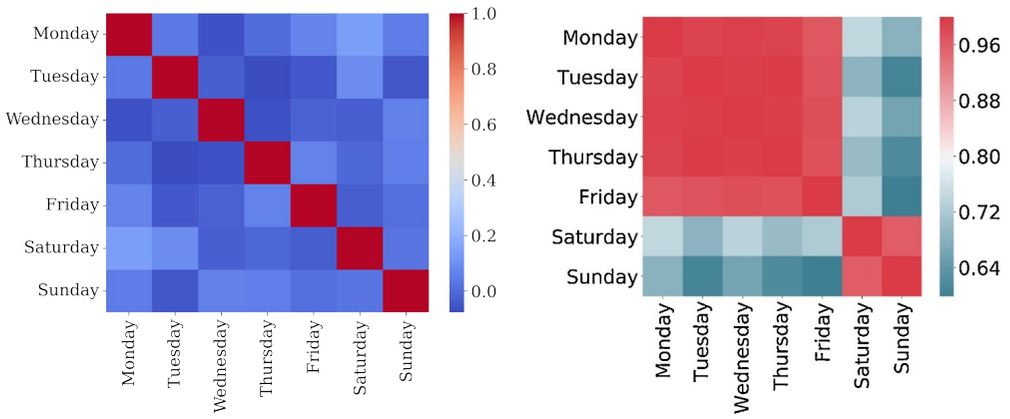


Figure 3.14: Hot water demand of the monitored household over the weekdays (A) and over the day hours (B)

To quantify the similarity of the demand patterns between different days in a week, a Pearson correlation analysis is performed, as shown in Figure 3.15 (a). Previous studies that used correlation analysis [97, 123] have reported that the weekdays have a high correlation between each other and less correlation with the weekends, which consequently have separated the correlation matrix into two sections of weekend days

and weekdays. An example of a correlation matrix for data collected during the normal days (before COVID-19 pandemic) reported in [97] is shown in Figure 3.15 (b). However, comparison of Figure 3.15 (a) and (b) shows that the occupant behavior in our case study household does not follow the same trend as normal days, and there is no longer a strong correlation between the weekdays as well as between the weekend days.



(a) (b)

Figure 3.15: Correlation matrix between the days of the week (a) the data collected in this research during COVID-19 pandemic and (b) the data collected in [97] before COVID-19 pandemic

Autocorrelation analysis can quantify the relationship between the hot water demand of a specific hour with previous hours. This information can provide a good insight for adjusting the number of prior demands to be included in the state vector. Figure 3.16 shows the Pearson autocorrelation of hot water demand data describing a negative correlation (values [-1;0]) or a positive correlation (values [0;1]). The values on this figure show how the variations of total hot water consumption of each hour is correlated with the other hours of the day. The highest autocorrelation is observed at 168 hour and 24-hour time lag, indicating that the demand at a specific hour has the highest correlation with the demand at the same hour a week before and a day before, respectively. Although the data have been recorded during the COVID-19 pandemic, the highest autocorrelation with the demand of a prior week and a prior day is in line with previous studies reporting pre-pandemic normal situations [97, 124]. It can be therefore concluded that although during the pandemic occupants do not follow the normal schedule of residential houses, there is still a repetition in their hot water use behavior similar to the normal situation.

Previous studies on prediction of hot water demand by supervised learning [97, 124, 125] have reported that temporal factors, such as the hour of the day and the day of the week, are important features for prediction. Temporal features can also provide useful information for an RL agent. Figure 3.17 shows the correlation between the hot water demand pattern and three features available in the dataset. For example, the correlation

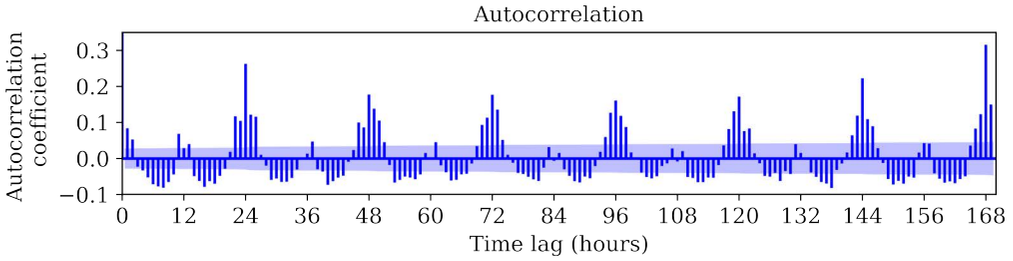


Figure 3.16: Autocorrelation coefficient of hourly hot water demand data

factor between Demand and Hour shows that how a vector including the hourly hot water demands of a day is correlated with a vector including the hour of the day (e.g. [1,2,3,...,24]). A correlation coefficient close to 1 shows a linear relationship between two variables. Therefore, that variable can be a valuable input to predict the other. As shown in Figure 3.17, hot water demand has a positive correlation with both temporal features of the hour and day number, and the value of correlation is higher than the correlation with outdoor temperature. The magnitude of the positive correlation factors between demand with an hour and day number is quite small, because the amount of hot water demand is not direclty a linear function of the hour or day number. For example, the amount of demand does not increase with the increment of hours. However, temporal features are important for learning the routines of occupants, for example, at what time the peaks of demands are expected to happen. A very useful characteristic of the agent is to automatically adjust the relative importance given to each feature and neglect the features which are not useful. Therefore, if one or both of the temporal features do not provide useful information, the agent learns to neglect them over the training phase.

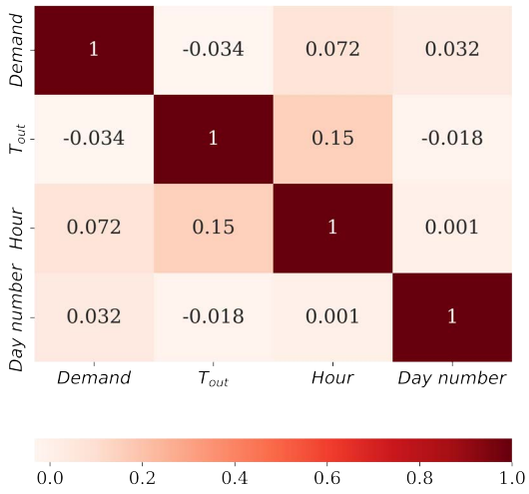


Figure 3.17: Correlation between demand and other features in the dataset

3.4.2 Sensitivity analysis of hyper-parameters

The proper selection of hyper-parameters can significantly improve the performance of the RL framework. To select the best hyper-parameters, a sensitivity analysis is performed on few main parameters such as the weight factors *a, b, c* of the reward function and the number of previous hours (*demand lags*) of demand to be used in the state array. Each set of hyper-parameters is called a scenario in this chapter. All scenarios for the sensitivity analysis are trained for 27 weeks and deployed for 2 weeks to ensure the convergence of each scenario and thus a fair comparison. Some of the scenarios do not converge if fewer train weeks are considered. The number of previous hours to be considered for outdoor air temperature is fixed to 6 hours for all scenarios. Since both the train and deployment stages should be repeated for each scenario, the number of evaluated scenarios is limited to 6 to keep the analysis manageable. To compare the performance of evaluated scenarios, two metrics to quantify energy-saving and comfort preservation outcomes of the frameworks are defined in Equations 3.4 and 3.5.

$$Energy\ saving = \frac{E_{baseline} - E_{RL}}{E_{RL}} \times 100 \tag{3.4}$$

$$Comfort = 1 - \frac{Violated\ demand(L)}{Total\ demand(L)} \times 100 \tag{3.5}$$

$E_{baseline}$ and E_{RL} are the total energy use of the RL and the baseline controllers in *kWh*. Violated demand is the amount of demand supplied with a temperature below 40 ℃. The comparison is made for the deployment stage in which all the scenarios are converged. Figure 3.18 shows the selected parameters and performance metrics of each scenario over the deployment phase. Due to the high weight of the reward function in the reward function, all scenarios properly respect the hygiene aspect by over-heating the tank at least once per day. This shows that the agent takes the health of occupants as a priority. Therefore, evaluated scenarios are only compared in terms of energy and comfort metrics. The first three scenarios aim to adjust the relative importance of the reward terms. Then the last three scenarios aim to adjust the number of previous hours of demand intervals to be included in the state. The first scenario is executed with $a = 1, b = 4, c = 1$, and $demandlags = 6$. This scenario provides energy saving of 16.25% compared to the baseline controller while preserving the occupants' comfort all the time. To identify the increase in energy-saving potential without the disturbance of occupants' comfort, the relative importance of the comfort term *b* is reduced to 3 in scenario 2. It results in a higher energy saving of 29%. However, the occupants' comfort is violated by 8%. The average temperature of the instances when the occupants' comfort is violated is 39.1 ℃ which is very close to the comfort threshold of 40 ℃. Therefore, this scenario could be acceptable for the buildings where the occupants are not very sensitive

to slight temperature violations but prefer higher energy savings. Since it is desired to always maintain occupants' comfort, the comfort weight b is increased from 3 to 3.5 in scenario 3. In this scenario, the comfort of occupants is always preserved, while energy saving increases to 17.6%. Therefore, the combination of weighting factors defined in scenario 3 ($a = 1, b = 3.5, c = 1$) is selected to preserve comfort and hygiene. After fixing the combination of the factors for the reward, the number of hours of the previous demand intervals in the state (demand lags) is increased to 12, 24, and 48 as scenarios 4, 5, and 6, respectively. The number of hours of 12, 24, and 48 are selected to let the agent observe the previous profile over half-day, one-day, and two-days ago. It should be noted that increase of demand lags on one hand provides more information to the agent but on the other hand increases the dimensionality of the state vector. Increasing the demand lags from 6 to 12 has significantly increased the energy saving (from 17.67% to 23.8%), while preserving the occupants' comfort. Further increase of the demand lags to 24 and 48 slightly increases the energy-saving potential while the occupants' comfort is violated somewhat. Therefore, a demand lag of 12 can be considered as optimum. Scenario 4 having the highest energy-saving potential among scenarios that totally preserve occupants' comfort is selected as the best performing scenario. Next sections include the detailed performance of this scenario. It should be pointed that an automated process can be developed to optimize hyperparameters in RL, which is expected to significantly improve the performance. However, it was out of the scope of this study.

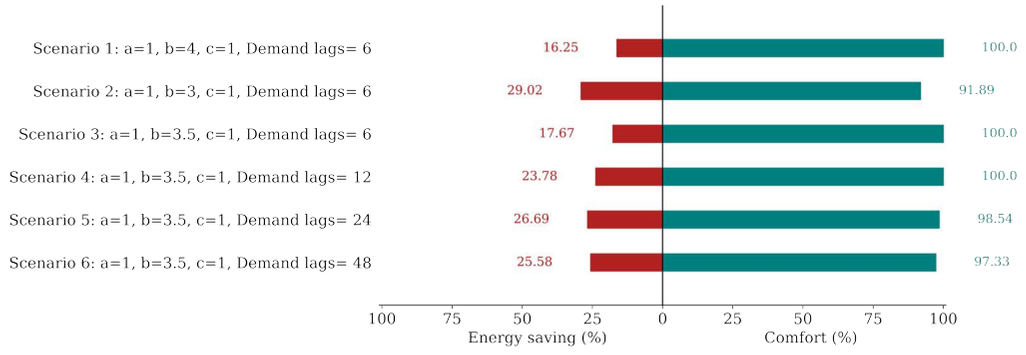


Figure 3.18: Energy saving and comfort index by different scenarios

3.4.3 Performance of the selected scenario

The evolution of the reward over the training episodes is an important indicator of the goodness of the training process. The higher the total reward the better the performance of the agent. The convergence of the total reward over the training episodes should be supervised to ensure that the agent reaches an optimal policy and the duration of the training stage is sufficient. Figure 3.19 shows the evolution of the total reward and

the constituting terms over the train and deployment stages on the target house. For evaluating the variations of total reward, the potential range of variations should be taken into account. The total reward function in this study is consisted of three terms: energy reward, comfort reward, and hygiene reward. The comfort and hygiene terms are supposed to reach zero at the optimal policy. However, the energy expenditure to heat water is not avoidable. Therefore, the energy term is not expected to reach zero even at optimal conditions. Fluctuations of the energy term over different episodes (different weeks) are also affected by the amount of hot water demand over each week. Thus, small variations of this term are caused by the demand variations and do not necessarily indicate a non-convergent policy.

The evolution of the total reward shown in Figure 3.19 indicates that the agent has converged to an optimal policy at the end of the training period, with the comfort and hygiene terms stable at near zero. There are some fluctuations between weeks 15 and 21 of the training stage. Over this period, the energy reward is slightly increased, but the comfort and hygiene rewards are decreased. Therefore, it shows that the agent has been trying to save more energy by reducing the tank temperature during this period. However, after receiving negative rewards on hygiene and comfort terms, it has learned from its mistake and increased the energy expenditure again in week 21 to respect hygiene and comfort terms. It should be noted that considering the possible range of variations of the reward terms, the fluctuations over the on-site train period are negligible and should not be interpreted as unstablility. In the worse case which has happened at week 16 the comfort term has reached -2. Considering the formulation of the comfort term, it shows that the average violation of comfort temperature in this case has been less than 1 ℃, which is negligible. Similarly, the hygiene reward over this week has reached -0.5. It shows that the delay in daily overheating has been less than 1 hour on average. Although the results show some fluctuations in the reward terms, the variations are very small and it can be considered that the reward value has been almost stable from the first week of training on the target house. This stability in the reward value shows that the agent has already obtained significant experience during off-site training and does not violate the comfort and hygiene terms from the very beginning week of on-site training. This indicates a fast convergence over the on-site training stage, which shows the effectiveness of the intensive off-site training (6 years of experience for agent). It also highlights that the statistical hot water usage model to represent occupant behavior over the off-site training stage has been very useful. Statistical models of other types of occupant behavior can be used in the same methodology to apply RL in other occupant-centric problems. It also highlights the substantial adaptability of the agent, as the behavior of occupants over the COVID-19 pandemic has been very different from the statistical-based normal behavior that the agent has observed over the off-site training stage.

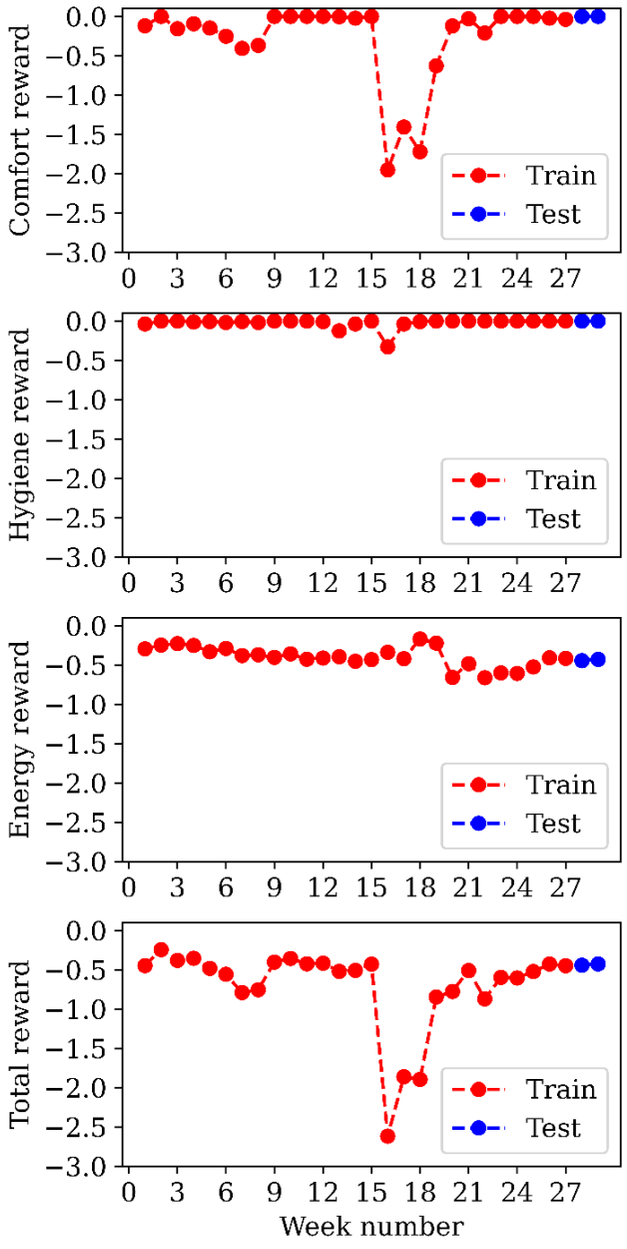


Figure 3.19: Evolution of the reward over the train and deployment stages

To better observe the agent performance, the operation of the selected scenario over the deployment stage is shown in Figure 3.20. It is important to monitor if the agent has learned and adapted to the occupant behavior, if it properly follows the hygiene aspect, and if it considers the advantage of higher outdoor temperature for heating the water. Comparison of the control signal (action) versus demand shows that the agent has properly learned the occupant behavior, by turning ON the system during or even before the demand and turning it OFF when no demand is expected. Therefore, the tank temperature has never exceeded the lower limit of 40 ℃ and the occupants' comfort has always been preserved. Comparison of the tank temperature versus demand also indicates that the overheating schedule is properly adapted to the demand, as most of the time agent has overheated the tank before or during the demand to ensure that the energy used for overheating the water will not be wasted by the heat losses. The cumulated number of hours from the last superheat is consistently below 24, indicating that the tank is adequately sterilized and the achieved energy saving is not achieved at a cost of higher Legionella risk. Since the health of the occupants has a higher priority than their comfort, the hygiene term in Equation 4.12 punishes the agent for every single hour exceeding the 24 hours threshold from the last superheat. Consequently, the agent is very considerate about the health-related term, and non of the evaluated scenarios violate this aspect. Analysis of the control signal with respect to the outdoor air temperature indicates that most of the ON signals occur during the hours of high outdoor temperatures. However, as comfort and hygiene are of a higher priority than energy saving, the ON signals do not perfectly match the instances of elevated outdoor temperatures. There are even several instances when the agent heats water even though the outdoor air temperature is low.

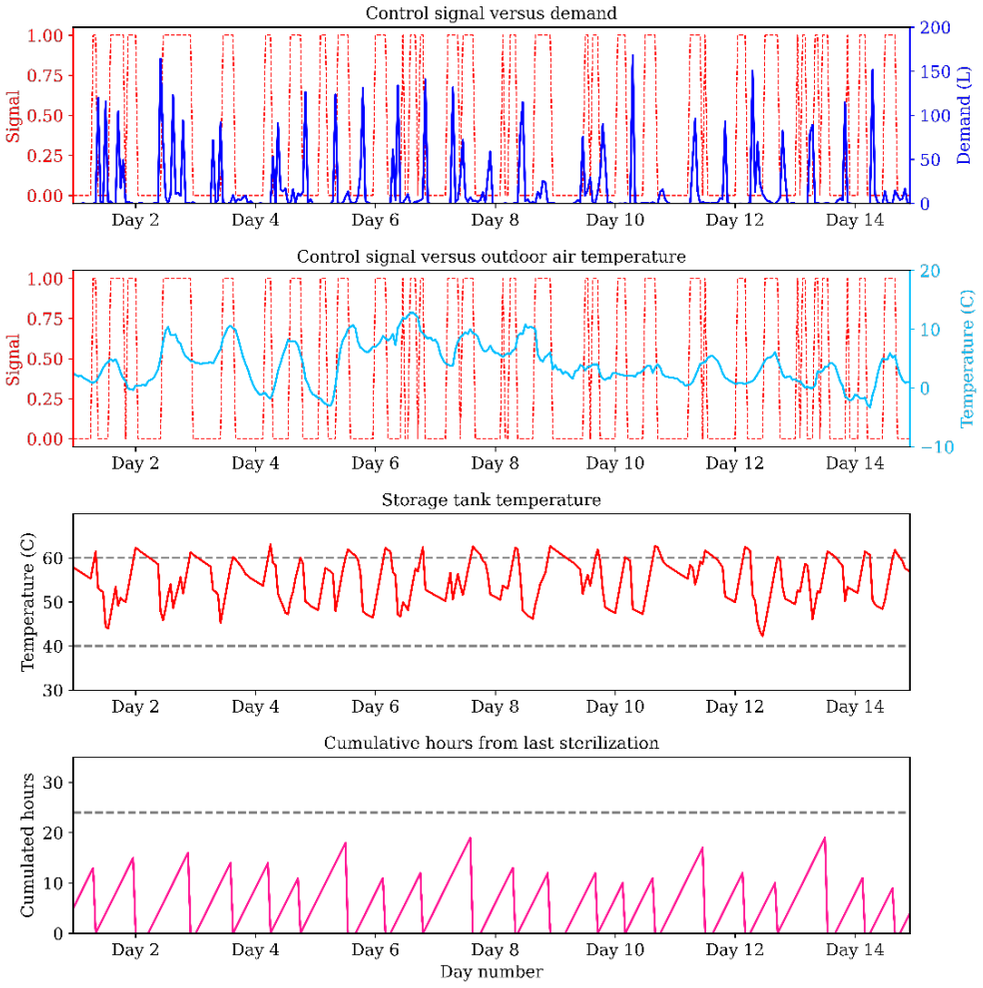


Figure 3.20: Performance of the RL agent during the deployment stage

For comparison, the performance of the conventional two-point controller is also evaluated and presented in Figure 3.21. By considering only the tank temperature, the two-point controller turns ON the heat pump only when the temperature falls below the threshold. It is not necessarily the best time for heating the tank due to the COP variations of the heat pump. Therefore, the main advantage of the proposed RL-based control framework over the conventional method is in adapting to the behavior of occupants which enables the system to save more energy while preserving the occupants' comfort.

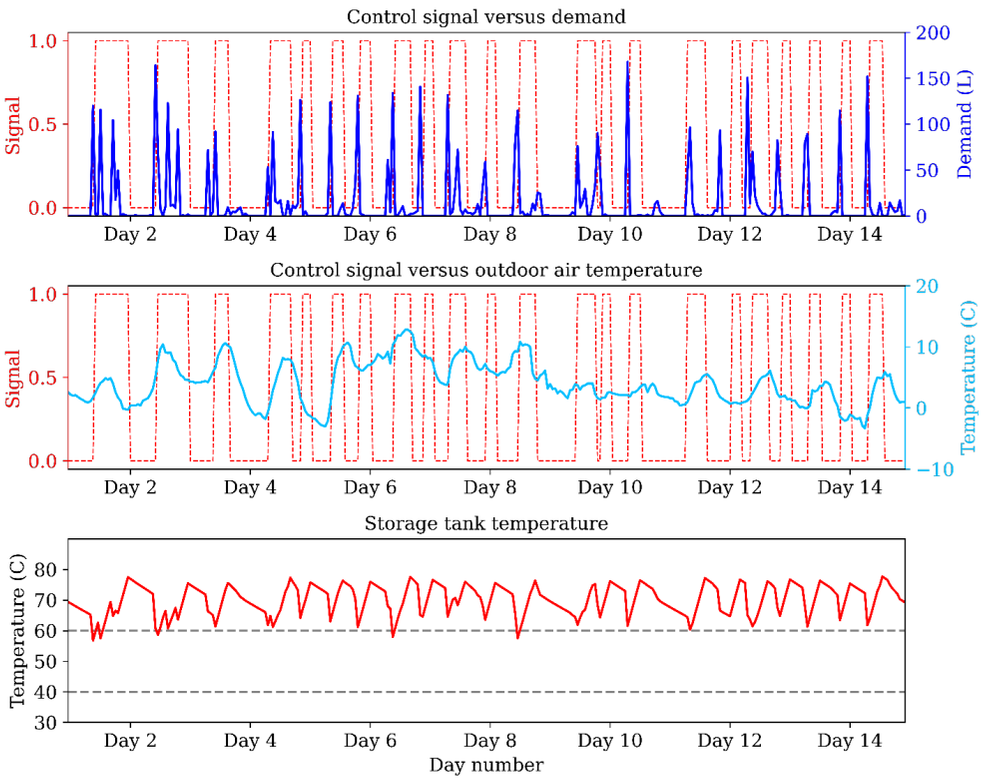


Figure 3.21: Performance of the rule-based controller during the deployment stage

3.5 Suggestions for future work

There is ample room for further research on this topic. A great potential of the RL is that different dimensions can be easily added to the framework by including corresponding terms in the states and reward. The proposed framework can be further developed by:

- Including the time-varying electricity price in the states and the energy cost in the reward function to optimize the heat pump operational cost;
- Including the number of ON and OFF switchings in the states and the heat pump age in the reward to ensure a smooth operation and a longer lifetime of the heat pump;
- Integration of a deep learning demand prediction model, trained on enough data to be transferrable to different buildings and ensure the generalizability of the framework;
- Combination of renewable energy (photovoltaic or solar thermal panels) to the system and the associated terms in the states and reward function for using the flexibility of hot water tank for optimal integration of renewable energy resources;

- Developing an off-site training stage that can ensure enough experience for the agent and eliminate/minimize the need for an on-site training stage;
- Comparison with a more sophisticated baseline controllers; For example, a rule-based controlled can be modeled that maintains the tank temperature as low as possible (for example 45 ℃), and only overheats the tank above 60 ℃ once per day during the off-peak period. It is expected that the Reinforcement Learning method still outperforms the baseline model due to its adaptation potential.

3.6 Conclusion

This chapter proposes an occupant-centric control framework based on the model-free Reinforcement Learning (RL) for heat pump water heating systems. The proposed framework balances water hygiene, a critical health-related aspect of water heating systems, occupants' comfort, and energy consumption. The training stage is separated into two different stages in the proposed framework, including an off-site training using a stochastic hot water use model and an on-site training. The off-site training aims to provide the agent an initial experience with the system, occupants' behavior, and climatic conditions. Thus, it can ensure fast convergence and preservation of the occupants' comfort on the target system. A stochastic hot water use model is included at this stage to represent a realistic hot water use behavior. The agent is trained off-site for 6 different climatic conditions of Switzerland using 6 years of weather data (2014-2020). For on-site training, the actual hot water use behavior of a single-family residential building is monitored for 29 weeks. The monitoring campaign taken place during the COVID-19 pandemic (28 August 2020 -19 March 2021) when almost all occupants were working from home. This has resulted in an unusual hot water use behavior compared to pre-pandemic times. The proposed framework does not rely on any model of the system, which ensures its easy transferability to other residential buildings. The following conclusions can be drawn from this study:

- Statistical analysis on the monitored demand data reveals the impacts of the COVID-19 pandemic on the hot water use behavior. There is no strong differentiation between the weekdays and weekends patterns compared to the "normal" (pre-pandemic) patterns reported in previous studies [97, 124]. Furthermore, there are 3 peaks in the average hourly pattern, while only 2 peaks are reported for the normal, pre-pandemic cases [97, 124].
- The proposed framework with the selected choice of hyper-parameters can provide an energy saving of 23.8% over two weeks of deployment compared to the

common rule-based control method. The evolution of the total reward during the training and the deployment phase on the target building shows a fast convergence and preservation of the occupants' comfort from the very beginning of training, despite the unusual hot water use behavior due to the pandemic. It indicates the effectiveness and importance of the off-site training with a stochastic hot water use model.

- Analysis of scenarios with different combinations of hyper-parameters indicates an optimum for the number of prior hours to be considered in the demand lags, thus, the inclusion of more hours does not necessarily enhance the performance.
- Comparison of control actions (ON/OFF) versus water demand shows that the agent has properly learned the occupant behavior to ensure their comfort, which indicates the adaptive potential of the proposed framework.

The findings of this study, performed during the COVID-19 pandemic period with the altered behavior of occupants, highlighted the importance of adaptive control for building energy systems. Future studies on the current topic are therefore required to investigate the potentials of RL for building energy systems.

Chapter 4

DeepSolar: An occupant-centric control framework for solar-assisted space heating and hot water systems

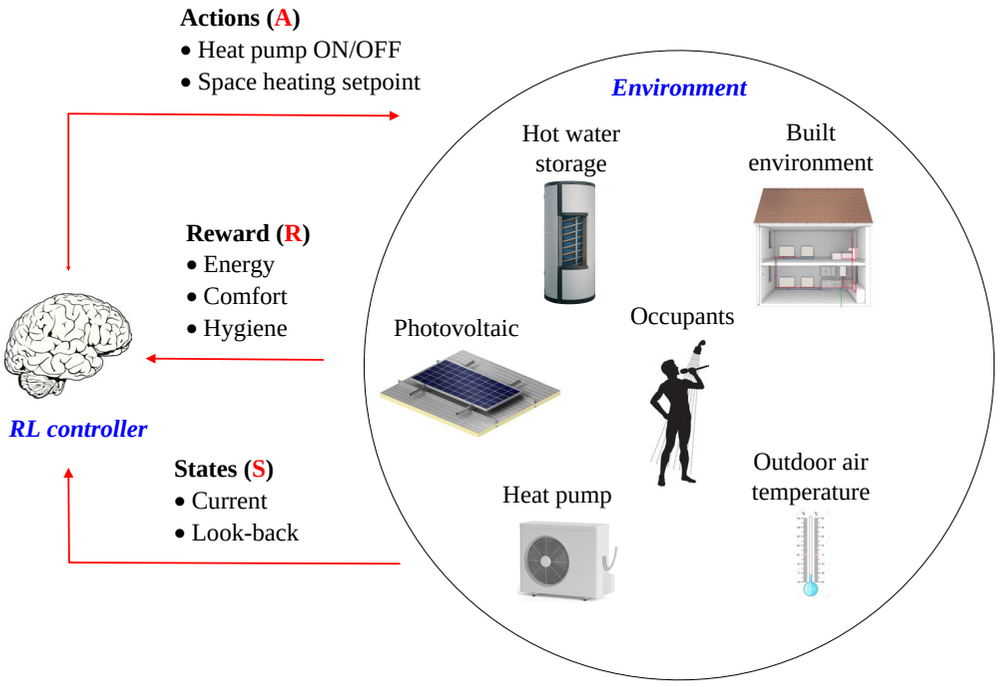


Figure 4.1: *DeepSolar* control framework in a nutshell

4.1 Abstract

To optimally control energy systems in residential buildings, several stochastic parameters should be considered including renewable energy production, outdoor air conditions, and occupants' behavior. However, these parameters are hard to model and predict accurately and even some of them (such as occupant behavior) are unique in each specific building. This increases the complexity of developing a generalizable optimal control method that can be transferred to different buildings. Rather than hard-programming human knowledge into the controller (in terms of rules or models), a human-like learning mechanism can be programmed to the controller so it can autonomously learn the optimal control policy in each specific building. This research proposes a model-free control framework based on Reinforcement Learning that autonomously learns how to make a balance between the energy use, occupant comfort and water hygiene in a solar-assisted space heating and hot water production system. To optimally control the building, this framework tries to adapt to the stochastic hot water use behavior of occupants, solar power generation, and weather conditions. A stochastic-based off-site training procedure is proposed to give a prior experience to the agent in a safe simulation environment, and further ensure occupants comfort and health when the algorithm starts learning on the target house. To make a realistic assessment without interrupting the occupants, weather conditions and hot water use behavior are experimentally monitored in three case studies in different regions of Switzerland, and the collected data are used in simulations. Two rule-based control methods are modeled as baseline. Results indicate that the proposed framework could achieve an energy saving from 7% to 22% without violating comfort or compromising the health of occupants, which is achieved mainly by adapting to solar power generation.

4.2 Introduction

Occupant behavior is a major driver of energy use in buildings [133]. Occupants influence the building energy use by their presence, activation or dis-activation of energy devices and adjustment of desired setpoints [134]. The role of occupant behavior is specifically important for indoor conditioning and hot water production systems [135]. Occupant behavior is considered as a major source of uncertainty for optimal operation of building energy systems [136]. Modeling the occupant behavior may, therefore, help to better understand and integrate it into the control of energy systems in buildings [137, 138]. However, occupant behavior can be affected by many different parameters, including environment-related, time-related, and random factors, which makes it extremely stochastic and complex [139, 140]. Even when an advanced modeling method is developed to predict occupant behavior, it is challenging to quickly apply that

model to a similar, but distinct building [136]. This is due to the uniqueness of occupant behavior in each building. Consequently, it is challenging to develop a holistic and transferable model of occupant behavior to be used by the controller without any prior data of that specific occupants. Current controls of building energy systems are detached from occupant behavior and follow a conservative and energy-intensive approach.

Besides occupant behavior, integration of renewable energy sources into the buildings forms another source of uncertainty for their optimal operation. The share of renewable energy sources in the building sector is projected to be doubled by 2030 [141]. While this increasing share would reduce CO_2 emissions, the fluctuating and stochastic nature of renewable energy sources increases the complexity of optimal control. Due to the intermittent nature of renewable energy sources, injecting the surplus power into the grid also complicates the grid operation and can pose problems (e.g. voltage fluctuation) [142]. One way to cope with the fluctuating supply is to make the local electricity demand flexible and responsive to the supply, for example, by maximizing the self-consumption [143]. Demand flexibility can be provided through several methods, such as flexible thermal generators, electrical or thermal energy storage, demand-side measures, or even grid-connected electric vehicles [144]. Among these options, storing the surplus energy as heat (power-to-heat) is considered to be particularly promising because both the cost of generating heat from electricity and the cost of heat storage are relatively low [145]. Air-to-water heat pumps emerge as a favorable power-to-heat option that can provide a great opportunity for solar energy integration in the building sector. This is because, first of all, the number of heat pumps as an energy-efficient technology is steadily increasing in the building sector. For example, the number of installed heat pumps in Germany has almost doubled over the last 6 years [146]. Secondly, hot water storage of a heat pump is cost-effective energy storage that can provide the same level of self-consumption as of electric storage, but at half of the levelized electricity cost [147]. Furthermore, the thermal mass of the building itself can serve as an additional heat storage for heat pumps, making it possible to further increase the flexibility without additional investments [143]. Buildings, therefore, can be seen as free batteries for the grid.

To incorporate the energy flexibility of residential heat pumps, their operation should be responsive to the stochastic occupant behavior, climate conditions (that affect the heat pump efficiency), and solar power production. The most conventional heat pump controllers today are *rule-based controllers*, which follow a set of rules defined at the design stage. These methods are computationally inexpensive and can be easily programmed on a cheap hardware. However, rule-based controllers totally neglect the stochasticity of the environment and follow a static operational strategy which is usually far from optimal strategy. A more advanced control method is Model Predictive Control (MPC), which uses a model of the system to make predictions about the future

outputs. It solves an optimization problem at each time step to determine the next actions that drive the predicted output as close as possible to the desired reference. MPC has shown a promising performance when applied to complex air conditioning systems [148–151]. However, there are several limitations to the application of MPC in practice. First of all, the performance of MPC and other model-based control methods is highly dependent on the accuracy of the developed model and prediction of the stochastic parameters. However, developing an accurate model of the system is extremely time-consuming and, therefore, not practical in most cases [152]. Moreover, even if an accurate model is developed, it can become fairly inaccurate over the time due to, for example, aging or modification of the system. Being dependent on an accurate model also makes the MPC building-specific, limiting the transferability to the other buildings and widespread adoption in the building sector [85]. To optimize the developed model at each time-step, MPC requires a considerable computational power which further limits its implementation in practice [153].

An alternative to hard-programming the expert knowledge as rule-based or model-based control methods is to program a human-like learning mechanism to the controller, and let it learn the optimal control strategy in each building by itself. With recent advances in the Internet of Things (IoT) technology on the one hand, and vast progress in Machine Learning methods on the other hand, the development of controllers which can learn by themselves is ever more realistic [85]. Among Machine Learning methods, Reinforcement Learning (RL) has recently gained popularity as a model-free control method [154]. In RL, the learning controller, known as agent, interacts with its environment and uses feedback from the environment to select the best possible action given the current state [155]. RL is gaining increasing attention for the built environment applications due to its three main advantages. First of all, it can be model-free, which therefore does not require a complicated and costly model of the system. It is a big advantage specifically when the system is complex [136]. Secondly, it is computationally efficient (after training), even when the state-space has a high dimension [153]. Finally, an RL agent can continuously adapt to the changes in the environment to maintain an optimal control policy. It makes RL an ideal method for integrating time-varying parameters such as solar energy potential, environmental conditions, or even occupant behavior into the controller. The RL agent treats occupant behavior as an unknown factor and learns and adapts to it over the time [136].

In recent years, RL has been investigated for a diverse set of applications in buildings. Park et al. [58] proposed a device called Lightlearn, which uses RL for occupant-centric control of lights in offices. The device was installed in five different offices for eight weeks. The performance of the proposed solution was compared with conventional occupancy-based and schedule-based methods in case of energy use and comfort of the

people. Results showed that the occupant-centric control based on RL successfully made a balance between occupant comfort and energy use and provided energy saving compared to both conventional methods. RL is also studied for other applications such as thermal storage inventory [156], natural ventilation [157] or integrated lighting and blind control [158]. However, regarding the big share of thermal conditioning energy use in buildings, most of the studies on RL have been focused on air conditioning systems. Zou et al. [159] developed an RL model for optimal control of air handling units to minimize the energy use, while preserving the comfort of occupants. The operational results indicate that the agent has learned how to adapt to the occupancy schedule to save energy, for example, by pre-cooling the spaces before the start of occupied hours. Schreiber et al. [153] proposed the application of RL for load shifting of a cooling network under the dynamic pricing. The cooling network included a chiller that supplied cooling to 3 different sites. The RL agent in this system was supposed to regulate the cooling supply to each site, to shift the power consumption to periods with lower electricity prices or lower outdoor air temperature while keeping the indoor air temperature violations in an acceptable range. Brandi et al. [154] implemented double deep Q-learning to control the operation of a water-based space heating system in an office building. In this study, the static deployment (where the agent is no longer trained over the deployment phase) is compared to the dynamic deployment (where the agent continues training even over the deployment phase). It was shown that the RL agent with carefully designed state-space is capable of providing the required adaptability even in case of static deployment. Comparison with the rule-based method showed that the RL-based controller could provide 5% to 12% energy saving with an enhanced comfort. Valladares et al. [160] evaluated the potential of deep Q-learning for controlling the indoor air temperature and air quality (CO2 concentration) while reducing energy use. Two different case studies were evaluated, a laboratory room having around 2–10 occupants and a classroom with up to 60 students. The trained agent was tested in an experimental setup using IoT sensors and actuators. The proposed method was then compared to the conventional rule-based control. Results show that the proposed framework could provide a better comfort (measured by Predicted Mean Vote (PMV) index) and 10% lower CO2 levels than the current control system while using about 4–5% less energy.

There are only a few studies that have taken hot water production into account, while it accounts for a big share of buildings' energy use, and is usually integrated into the space heating systems. Kazmi et al. [72] proposed a model-based RL control framework to balance comfort and energy use in heat pump water heating systems. In particular, they used model-based heuristics that incorporate the state of hot water tank and occupant behavior into the optimal control problem. The models for heat pump, storage tank, and occupant behavior prediction were probabilistic, data-driven models developed based on

historical data. Thirty-two net-zero buildings in the Netherlands using heat pumps and storage tanks were studied. It was shown that the proposed RL control approach reduces energy use for hot water production by roughly 20% with no loss of occupant comfort. Heidari et al. [161] proposed an RL-based control framework to learn and adapt to the occupants' hot water use behavior, and make a balance between energy use, comfort and water hygiene. The proposed framework was tested over data collected in a Swiss residential house. While the monitoring campaign was during COVID-19 pandemic with an abnormal occupant behavior, the proposed framework could quickly learn the occupant behavior and provide 24% of energy saving over the conventional rule-based method.

Regarding the increasing interest in integrating solar energy into buildings, a number of studies have also focused on solar-assisted space heating and hot water production. Correa-Jullian et al. [162] proposed a condition-based control approach based on tabular Q-learning for the optimal control of a solar-assisted water heating system. The Reinforcement Learning agent in this system was supposed to determine the operational schedules of the solar field and heat recovery chiller according to the energy efficiency, comfort levels, and participation of renewable energy sources. The results showed that the Reinforcement Learning-based operation performed better than the nominal operation schedule when solar radiation was low. On the other hand, nominal operation yielded a higher performance when the solar radiation was highly available. Ali and Kazmi [163] proposed an RL-based control framework for Photovoltaic-assisted (PV-assisted) domestic hot water production systems. The control approach tried to maximize the self-consumption of PV production by shifting the consumption into the periods of PV power production. However, temperatures above 50 ℃ were awarded equally so preventing the over-consumption of PV power for overheating the water. Comparison of the RL-based control with the rule-based control over 6 different case studies showed that the RL-based control successfully increased the self-consumption of PV production. Lissa et al. [164] proposed a framework for optimal control of PV-assisted space heating and hot water system. The proposed framework aimed to reduce energy use by optimizing the operation of the heat pump and maximizing the PV self-consumption while keeping the comfort of occupants. To monitor the comfort aspect, higher and lower temperature limits were considered for indoor air and hot water temperatures. The limits of indoor air temperature were based on the hourly average temperatures recorded in the case study building, and the limits for hot water temperature are 40 ℃ and 55 ℃. It was indicated that as the indoor heating is a slow process, the agent can better follow the comfort limits. However, as the water heating is a faster process, there is a higher probability of surpassing the comfort limits. The evolution of reward term showed that after the first month of training, the agent learned to keep the occupant comfort and the occupants no longer experienced high deviations from comfort limits. The proposed framework could

provide 8% to 16% energy saving compared to the rule-based controller.

4.2.1 Objectives and contributions

This chapter proposes an RL-based control framework for PV-assisted space heating and hot water production. This framework can learn and adapt to the stochastic parameters, namely hot water use behavior of occupants, PV power production, and outdoor air temperature, and accordingly make a balance between energy use, comfort, and water hygiene. Very few studies have investigated RL for the entire system of solar energy, space heating, and hot water production. This study intends to further broaden the current knowledge by investigating the following aspects:

- **Model-free:** This framework does not use any model, such as a data-driven or thermodynamic model of the system, and rather learns the required knowledge from scratch. The model free nature of this framework facilitates the transferability of the control framework to the other residential buildings with different system specifications;
- **Integration of water hygiene:** Legionella is a waterborne bacteria that grows in warm water between 25 ℃ and 47 ℃ [165] and pose health risks to the occupants. According to the literature review, the hygiene aspect of water is never investigated in previous studies on RL. This is while the hygiene aspect, mainly Legionella, is the main barrier for reducing water temperature to save energy [166]. This study integrates water hygiene into the control framework by integrating a Legionella growth model. This will help the agent to properly adjust hot water temperature for reducing energy use without endangering the health of occupants;
- **Stochastic-based off-site training:** To speed-up the convergence and to minimize the risk of violating comfort or hygiene aspects on the target house, a stochastic-based off-site training phase is designed to provide enough experience to the agent in the safe simulation environment before being implemented on the target house. The off-site training phase integrates a stochastic hot water use model and trains the agent over a variety of system sizes, geographical locations, and hot water use behavior to ensure the agent has obtained a generalized experience and can quickly adapt to different houses. Off-site training is done in simulation, which is a safe environment where the agent can learn from scratch and even try random actions without any consequences on the real occupants;
- **Investigating the adaptation potential to different hot water use behaviors:** Hot water usage is a very stochastic parameter that the agent should take into account. To evaluate the adaptation potential, real-world hot water use behavior is monitored

in 3 Swiss residential houses. As the monitoring campaign was performed over the COVID-19 pandemic, it allows investigating the adaptation potential to an abnormal situation different from what the agent has observed during the off-site training. Also, the behavior of 3 cases was found to be very different, which allows to further investigate the adaptation potential of the agent to different occupant behaviors;

- **Investigating the generalization potential of the knowledge gained in off-site training:** A well-designed training procedure should provide a generalizable knowledge to the agent. If the knowledge gained in off-site training is generalizable, it can minimize or at the best case eliminate the need for an on-site training on the real house. Since the on-site training of the agent on the cloud can be challenging and costly, in practice, it would be much easier if an agent could be only trained on simulations and directly deployed on the target environment. This chapter investigates two scenarios. The first scenario is the direct deployment of the agent, where the agent is directly deployed on the target house after off-site training, without any on-site training on that specific house. The second scenario is long-time deployment, where after a short-time on-site training the agent is deployed for a long time to see if there is a need for sequential trainings or one initial training is enough. These scenarios can provide insight for elimination or reduction of training phase on the target house by a generalizable off-site training, which will facilitate the practical implementation of RL in residential buildings;

The remaining of this chapter is organized into four sections. The first section presents the methodology of the research. The second section gives a brief overview of the case study houses and the monitoring campaign. The results of the study are outlined in the third section. Finally, the fourth section concludes the chapter.

4.3 Methodology

The methodology section presents the layout of the energy system, the monitoring campaign in the case study houses, the Legionella concentration model, the proposed RL control framework as well as baseline control scenarios.

4.3.1 Case study description

4.3.1.1 System configuration

The proposed framework is focused on a residential energy system including space heating, hot water production and PV power generation. There are many alternative

configurations for this system, such as integrated or separated thermal storage for space heating and hot water production. However, as the aim of this study is to prove the potential of RL for optimal operation of these systems, one common configuration of the system is examined as an example. The proposed framework can be easily adjusted to other configurations. The configuration used in this study is shown in Figure 4.2.

The heating system is an air-to-water heat pump, a favorable power-to-heat option with increasing number in building sector. Combined with a hot water storage tank, it can provide a great opportunity for solar energy integration. Heat pump has a variable Coefficient of Performance (COP) depending on outdoor air and hot water temperature. This dependency makes it more challenging for the RL agent to schedule heating cycles optimally. Secondly, hot water tank is considered as an energy storage, because it is more cost-effective than electric storage [147], provides both functionalities of energy storage and hot water provision, and is available in many buildings. While the space heating can be integrated or detached from energy storage, in this configuration it is considered to be integrated to storage to provide further energy flexibility. In this case, the surplus solar energy can be stored in the tank also for the space heating purpose. PV panels are considered to be grid-connected, so the surplus power can be also injected to the grid.

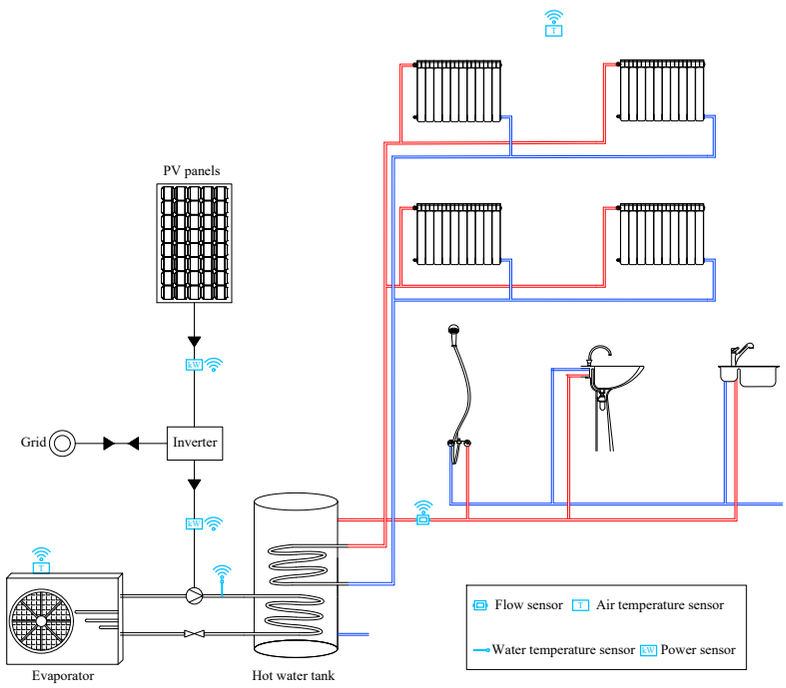


Figure 4.2: Configuration of system to be controlled by RL

4.3.1.2 Monitoring campaign

Hot water demand is less predictable than space heating demand, can be very different between similar buildings [14] and can impose a fast change in the hot water tank temperature which causes the violation of user comfort [167]. Thus, for the proposed framework, the most challenging task for the agent is to learn the hot water use behavior of occupants in each building. This study intents to evaluate the performance of framework over the actual hot water usage measurements. In this research, a cost-effective, low-power and water-proof monitoring system is implemented to monitor all the assets. Then the flow rate of all end-uses are summed to obtain the main flow rate. For this specific framework, monitoring all the end-uses is not necessary and a single sensor on the tank outlet can provide the required demand data. The detailed monitoring in this research was to provide a high-resolution dataset for future research.

Three residential houses in Switzerland are monitored for 20 weeks. Geographical location of buildings is indicated in Figure 4.8. Monitoring period of houses 1 and 2 was entirely during cold season (28 August 2020 to 15 January 2021), while for house 3 it also includes the hot season (23 March 2021 to 10 August 2021). The third house is to analyse how the agent will adapt to a period where PV power production is high, but energy demand is low (as there is no space heating demand in this period). The heated area and number of adults and children in each building are shown in Table 4.1. As shown in this table, the case study buildings are selected to include a variety of family compositions, which allows to further evaluate the adaptation potential of the agent to different houses.

The case studies were equipped with heat pump. But since they were occupied residential buildings, in this phase of early evaluation it was not desired to test the proposed framework directly on the actual systems as it could result to discomfort and dissatisfaction of tenants who were volunteer in this study. Rather, the real-life collected data can provide a more realistic evaluation in simulation environment, without violating the comfort of occupants.

Table 4.1: Area and number of occupants in case study houses

	heated area (m^2)	Adults number	Children number
House 1	160	2	3
House 2	120	2	2
House 3	150	2	1

4.3.2 Legionella concentration model

Legionella is a water-born bacteria that grows in water between 25 ℃ and 47 ℃ and can be transferred to humans by breathing in the contaminated water droplets. Infection with this bacteria results in a respiratory illness, known as Legionnaires' disease (LD) [166]. Hot water systems are responsible for the most number of infection cases, as they can provide the desirable temperature regime for the growth of Legionella [168].

While there are several disinfection methods, such as chemical methods, one of the most conventional methods is thermal disinfection [168]. With a temperature of 60 ℃ Legionella cells die in only 2 minutes [169]. Therefore, as a common practice, the hot water tank temperature is constantly kept above 60 ℃ to ensure Legionella can not grow in the tank. The high temperature of hot water tank reduces the heat pump COP, increases the heat loss, and also increases the risk of scalding at the point-of-use. This conservative operational approach is because the controller does not have any sense about the real-time risk of Legionella in the tank. This framework aims to quantify the risk of Legionella for the agent in real-time, so it can overheat and disinfect the tank only when it is needed. Legionella growth is a complicated process that depends on many different factors such as temperature, PH, and existence of nutrients [170]. It is therefore complicated to develop a model for accurate calculation of Legionella concentration. Few mathematical models are developed to estimate the Legionella concentration only based on the variations of water temperature [165, 166, 171]. Assuming that the hot water tank has not been initially contaminated with Legionella and biofilm, and also the network water is properly treated, these models can be used to provide the real-time estimation of Legionella concentration only based on temperature. Controlling the hot water tank temperature by considering Legionella concentration can make a shift from energy-intensive conservative control approaches into energy-efficient while safe methods. However, little attention has been given to the integration of Legionella risk assessment into the control systems. Kenhove et al. [172] integrated a model of Legionella concentration into the rule-based controller, where the controller heats the tank when the estimated concentration passes a threshold. Based on the literature review, there is no study on the integration of Legionella growth into RL-based control frameworks. Different from the rule-based control which only overheats the tank when a threshold is passed, an RL agent can learn how to proactively plan overheating cycles while minimizing energy use. For example, when the RL agent is informed about Legionella risk, it can overheat the tank when there is a surplus of PV power, when the heat pump COP is higher, or when a demand is expected to happen in near future.

Estimation of Legionella concentration in this study is based on the model proposed by Amerongen et al. [165]. In this model, for the temperature range of 25 ℃ and 47 ℃, the

doubling time (the number of hours required for Legionella concentration to get doubled) is calculated as:

$$DO = 0.5702 \times T_{tank}^2 - 43.3 \times T_{tank} + 829 \tag{4.1}$$

Where T_{tank} is the hot water tank temperature (℃) and DO is doubling time (hours). Using this equation for doubling time, and considering the effect of inlet and outlet water streams, the following equation can be used to calculate the concentration of Legionella:

$$C = \frac{(C_{initial} + \frac{C_{initial}}{DO}) \times V_{tank} + C_{network} \times Demand - C_{initial} \times Demand}{V_{tank}} \tag{4.2}$$

Where $C_{initial}$ is the concentration of Legionella at the beginning of timestep (CFU/L), the $C_{network}$ is the concentration of Legionella in network water (CFU/L), $Demand$ is the hot water demand (L), V_{tank} is the tank volume (L), and C is the concentration of Legionella at the end of that timestep. Regarding that in the hot water tanks the same amount of consumed hot water is replaced by the cold network water, the term $C_{network} \times Demand$ is the amount of Legionella entering the tank from network water, and $C_{initial} \times Demand$ is the amount of Legionella exiting the tank. For the temperature above 60 ℃, the reduction in concentration is calculated as:

$$C = \frac{(C_{initial} - 0.999 \times C_{initial}) \times V_{tank} + C_{network} \times Demand - C_{initial} \times Demand}{V_{tank}} \tag{4.3}$$

For the temperatures below 25 ℃ or between 47 ℃ and 60 ℃, the concentration of Legionella is assumed to be constant. It is a conservative assumption to further ensure the health of occupants, because for a temperature above 50 ℃ the disinfection still happens but with a lower rate [169].

4.3.3 Reinforcement Learning control framework

A variety of RL algorithms have been developed so far. These algorithms can be divided into two main categories of policy-based and value-based methods. Policy-based methods are suitable for problems with a continuous action space (such as robotic applications), while value-based methods are suitable for environments with a discrete action space, where the agent implicitly finds a policy by learning the optimal value function [86]. It is shown that value-based methods learn faster, as they include a limited number of possible actions and are less sensitive to hyper-parameter tuning [173]. One of the most widely used value-based RL algorithms is deep Q-learning. Deep Q-learning tries to estimate the value of each action, known as Q values, and select the action with the highest estimated

value. These values are calculated based on the following formula:

$$Q^{new}(s_t, a_t) = Q^{old}(s_t, a_t) + \alpha.(r_t + \gamma.max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (4.4)$$

Where $Q^{old}(s_t, a_t)$ is the old estimated value, $\alpha[0, 1]$ is the learning rate, r_t is the immediate reward, γ is the discount factor, and $max_a Q(s_{t+1}, a)$ is the estimated future reward. As the original deep Q-learning use the same network for the estimation of future values, it can lead to the overestimation of value for some specific actions and therefore a non-optimal action can be selected. To solve this issue, a modified technique called double deep Q-learning is recently developed. The main characteristic of this technique is the presence of two networks to counteract the overestimation of the Q-values. The second network is an exact copy of the first one, but its weights are only updated every τ steps. This network is used to calculate the target Q-values ($Q(s_{t+1}, a)$) [88]. Therefore, this research uses double deep Q-learning algorithm to develop the control framework. Tensorforce library [131] is used to program this framework in Python. In spite of other RL libraries that are mainly developed for computer games, Tensorforce is developed with a modular design with customizable agent and environment that can be easily used for other domains.

4.3.3.1 State, actions and reward design

The RL agent observes the state of the environment, then selects an action based on the observed state, and tries to maximize a reward. The proper setup of state, actions and reward is an important aspect to design a robust RL framework. State parameters should provide all necessary information for the agent to predict future immediate reward, and also should be possible to be measured by sensors in practice [154]. The following parameters are included in the state vector:

- **History of hot water demand:** As one of the most important aspects of this framework, the agent is supposed to learn and predict future hot water use behavior of occupants. Studies have shown that there are some routines in hot water use behavior of occupants in residential buildings, and therefore future hot water use is correlated with the historical demand [13, 161, 174, 175]. Therefore, a look-back vector of previous hot water demands is included in the state vector to enable the agent to forecast future demands. The length of this vector (the number of previous hours to be included) for this parameter and also other parameters of state will be determined based on the sensitivity analysis. It should be noted that in the *DeepHot* framework, the history of hot water demand was discretized into the 5 liters demand intervals, but in the *DeepSolar* framework the demand is simply normalized into [0-1] interval. Based on the lessons learned, the normalization approach performs

better than the discretization.

- **Demand ratio:** It would be useful to the agent to estimate how much hot water would be used in total by the end of today from now. Considering the routine in occupant behavior, the total hot water demand of today can be close to yesterday. Thus, the total demand at time *t* to the end of the day, is expected to be close to the total demand at time *t* to the end of the day for the previous day. The following equation quantifies the ratio of consumption up to the time $t = H$ of today, over the total consumption of previous day. In simple words, this ratio tries to inform the agent that how much more hot water demand is expected for today by looking at the total demand of yesterday. This is indeed a very simple estimation and the relative importance given to this estimation can be adjusted by the agent.

$$DR = \frac{\sum_{h=0}^{H} Demand_{Day=D}}{\sum_{h=0}^{24} Demand_{Day=D-1}} \tag{4.5}$$

  Where H is the current time of day, D is the day number, $Demand$ is the volumetric demand (L), and DR is the Demand Ratio.

- **Outdoor air temperature:** The outdoor air temperature affects the space heating demand and also heat pump COP. A look-back vector of outdoor air temperature lets the agent learn the variations of outdoor air temperature and heat pump COP. The agent can then take advantage of hours with higher COP to charge the hot water tank.

- **Indoor air temperature:** Indoor air temperature is important for the agent from two aspects. First of all, it affects the occupants' comfort and should be carefully adjusted. Secondly, as the building thermal mass is also a potential energy storage, it is indicating the current level of stored energy in the building thermal mass.

- **PV power production:** Another important functionality of this framework is to learn the variations in PV power production and optimally schedule the future actions. The look-back vector of PV power production enables the agent to learn its variations.

- **Heat pump outlet water temperature:** The heat pump outlet water temperature informs the agent about the rate of energy delivery to the tank and indoor air.

- **Legionella concentration:** For optimal adjustment of the hot water tank temperature and overheating cycles, the agent should know the current estimated concentration of Legionella in the tank (CFU/L). This lets the agent prevent unnecessary thermal disinfection of the tank, and only overheat the tank when it is necessary or when surplus of PV power production needs to be stored in the tank.

- **Hot water tank temperature:** The agent should know the current tank temperature to properly adjust it above the comfort level once needed. Also, it shows the current energy stored in the tank.
- **Hour of the day:** Many of the stochastic parameters, such as occupants hot water use behavior, solar energy and outdoor air temperature are strongly correlated with the hour of the day. To further assist the agent to learn and predict these parameters, hour of the day for the upcoming hour is also provided to the agent. Different from other parameters, this is not a look-back vector but is associated to the upcoming timestep.
- **Day of the week:** There is a significant difference between the hot water use profile of working days and weekends. Also, the hot water use profile of each day is found to be highly correlated with the profile of the same day over the last week [13, 161, 174, 175]. Accordingly, to learn and predict the hot water demand it would be helpful for the agent to know what is the current day of week. To reduce the number of states, the day number is provided to the agent as a single integer.

A visual representation of state parameters at a specific time step is shown in Figure 4.3. The length of look-back vector indicated for each parameter is symbolic in this figure as it will be determined over the sensitivity analysis.

Given actions to the agent should also provide enough flexibility to maintain an optimal operation. The possible actions in this study are selected according to the comfort limits and hygiene aspects. As shown in Figure 4.4, the comfort limits for indoor air temperature in winter are between 20 ℃ and 24 ℃ based on *ISO7730* [176]. It is assumed that the agent can select a setpoint to overwrite the existing thermostat. The possible setpoints are 21 ℃ and 23 ℃, with a dead-band of 2 ℃. The option of 21 ℃ is an *energy-saving* choice that maintains occupant comfort without overheating the indoor air. On the other hand, the option of 23 ℃ provides the opportunity of storing surplus PV power in the building thermal mass and thus can be seen as an *energy-storing* choice. In this study only these two options are considered for the agent because the occupants are always present in the building. It is assumed that in practice, a backup controller or a manual interface on the thermostat can be used to turn OFF the heating system once the occupants are away.

In case of hot water tank, the multiplication of Legionella at each temperature range, as well as the comfort limit for hot water are shown in Figure 4.4. While the required temperature of mixed water at each point-of-use is different, 40 ℃ is assumed as the minimum required supply temperature for simplicity [14]. In this research, 40 ℃ is considered as the minimum comfort level for the tank temperature measured at the middle

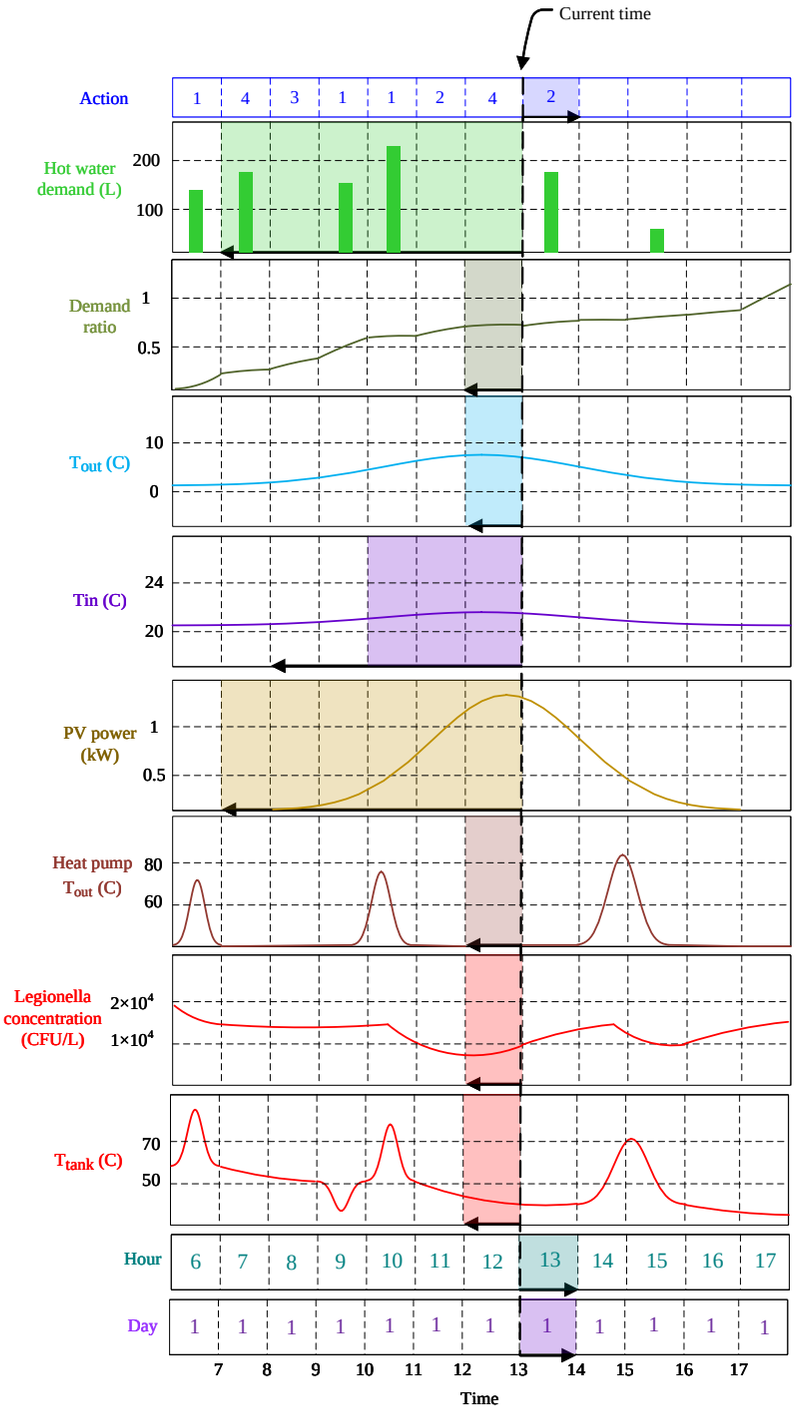


Figure 4.3: Visual representation of states and actions

of the tank. This is a conservative assumption that will further ensure the comfort of occupants, because in practice the hot water is supplied from the top of the tank which has a higher temperature due to the stratification of tank. Since the range of possible

temperatures for hot water is quite wide, discretization of setpoints would result in many different actions. To limit the number of potential actions while providing enough flexibility for tank temperature, possible actions are considered as *turning ON* and *turning OFF* the heat pump. This would give the possibility to the agent to adjust any temperature with only two actions. On the other hand, the agent should properly learn how the tank temperature varies based on the ON/OFF actions.

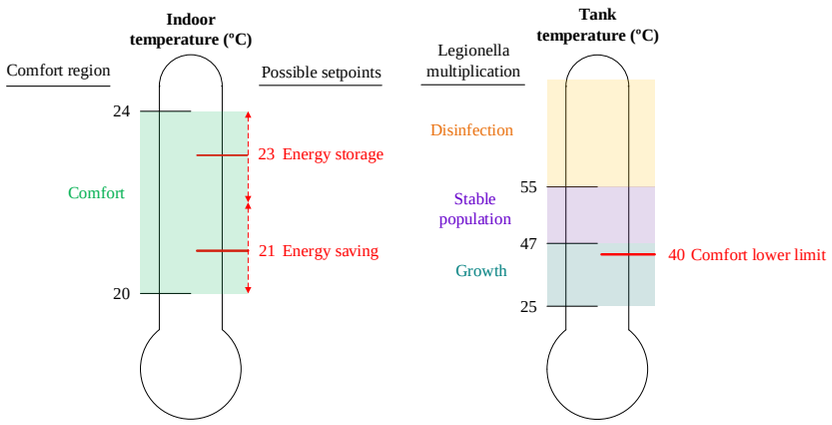


Figure 4.4: Temperature ranges for comfort limits of indoor air and Legionella multiplication and comfort limit for hot water tank

All the possible actions for the agent are presented in Figure 4.5. Actions related to the hot water tank are separated from the ones related to the space heating, meaning that the agent can not simultaneously change the indoor air setpoint and heat pump status, and should prioritize between them. While it is possible to combine the tank and space heating actions, for example one action representing turning ON the heat pump and selecting a setpoint of 21 ℃ for indoor air, initial evaluations in this study indicated that such combined actions make it more complicated for the agent to learn the relationship between performing each action and the associated impact on the environment.

The reward function should be well designed to clearly reflect the aims and priorities as simple as possible. This control framework intends to minimize the energy usage of heat pump, and maximize the self-consumption of PV power, while maintaining the occupants comfort and water hygiene. The reward function is composed of four different terms as follow:

- **Energy term:** The energy term penalizes the agent for (1) any energy usage of heat pump and (2) the surplus of PV power not used by the heat pump. This term is defined as

$$R_{energy} = -a \times \mid HP_{power} - PV_{power} \mid \tag{4.6}$$

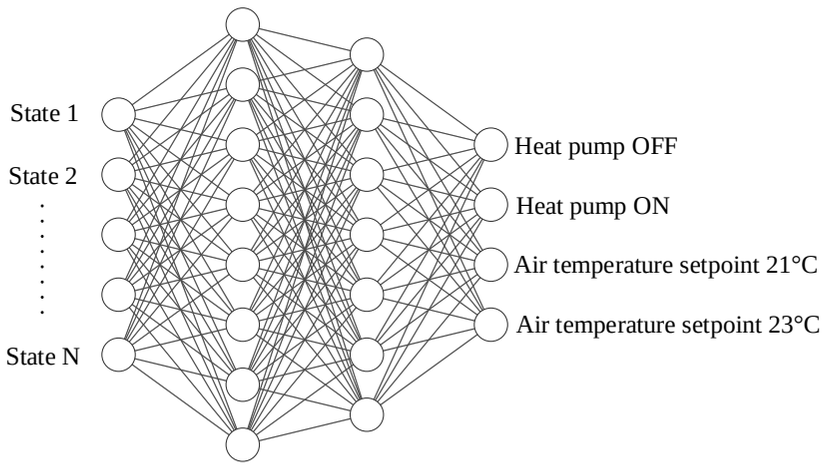


Figure 4.5: Possible actions for the agent

Where HP_{power} and PV_{power} are the power usage of heat pump (kW) and power production of PV panels (kW), accordingly. a is the weighting factor to adjust the importance of the energy term compared to the other terms.

- **Hot water comfort term:** This term penalizes the agent if the temperature of hot water tank falls below the comfort level of 40 ℃.

$$if\ T_{tank} \geq 40,\ R_{DHWcomfort} = 0\ else\ -b \tag{4.7}$$

Where T_{tank} is the hot water tank temperature and b is the weighting factor. It is also possible to penalize the agent proportional to the temperature deviation from the comfort level, but here a constant number is used to speed up the learning process.

- **Indoor air temperature comfort term:** This term penalizes the agent if the indoor air temperature is out of comfort limits. While the possible setpoints are inside of comfort region, still a comfort violation can happen if the hot water tank temperature is not high enough to provide required heat for radiators. This term is defined as

$$if\ 20 \leq T_{indoor} \leq 24,\ R_{Indoor\ comfort} = 0\ else\ -c \tag{4.8}$$

Where T_{indoor} is the indoor air temperature and c is the weighting factor.

- **Hygiene term:** This term penalizes the agent if the estimated concentration of Legionella in the tank exceeds the maximum acceptable level. This term is defined as

$$if\ C \leq C_{max},\ R_{Hygiene} = 0\ else\ -d \tag{4.9}$$

C is the current concentration of Legionella (CFU/L), and C_{max} is the maximum acceptable concentration (CFU/L) specified for residential buildings (5×10^5 CFU/L) [165].

The total reward, which is going to be maximized by the agent, is the summation of these rewards as

$$R_{total} = R_{energy} + R_{DHWcomfort} + R_{Indoorcomfort} + R_{Hygiene} \tag{4.10}$$

4.3.3.2 Training procedure

To train the RL agent, it is required to establish an interaction between the agent and the environment, which lets the agent to perform actions on the environment, receive back the next state and calculate the subsequent reward. The interactive procedure in this research is established by coupling the agent developed in Python with the dynamic model of the system developed in TRNSYS. Figure 4.6 presents how the agent and environment interact with each other. At each timestep, the agent writes the selected action to the input file of TRNSYS, runs the TRNSYS model for one timestep, and then reads the subsequent parameters of state from TRNSYS output file. If the episode is not ended, the state is again used by the agent to select the next action. And if it is the last timestep of the episode, the state is reset and sent to the agent. As the agent tries to maximize the reward over the period of each episode, the reset is to set the timestep counter as zero and inform the agent that the episode is ended. The length of an episode is considered as one week.

This study propose a multi-step training procedure in which the agent is first trained off-site on a safe virtual environment, then trained on-site on the target house and finally is deployed on that house. The overall procedure is presented in Figure 4.7. During the off-site training phase, the agent is interacting with the virtual model of the system for 10 years. An important consideration in the off-site training phase is to provide a generalizable knowledge to the agent and avoid overfitting to a specific case. It lets the agent to quickly adapt to different houses with different system sizes, located in different weather conditions, and with different occupant behavior. To mimic the hot water use behavior of occupants, a stochastic hot water use model driven by actual data [177] is used in this stage. Actual weather data from multiple weather stations in Switzerland are also collected. Then for each year of the off-site training phase, the solar and weather data of a different city is used as presented on Figure 4.8. In addition, a different set of system sizes (e.g. heat pump capacity, hot water tank volume, radiators and PV panels area, etc) are used in each year. The pre-trained agent is then saved to be used for on-site training on each of the target houses. It should be noted that the simulation model is only used to

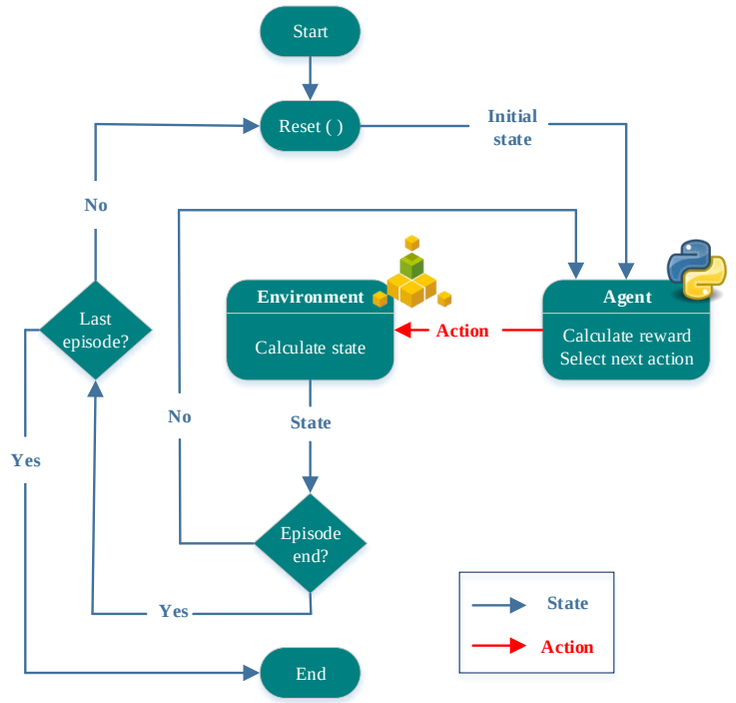


Figure 4.6: Procedure of interactions between the agent developed in Python and system model developed in TRNSYS

provide an initial experience for the agent, so it does not need to be an exact model of the target system.

On the on-site training phase, the pre-trained agent is again trained with the actual hot water demand and weather data of the target house. While the off-site training phase might be enough for the agent, the on-site training on the target house is to ensure that the agent observes and adapts to the specific conditions of the target house. In this phase, the actual hot water demand data that are measured experimentally in case study houses is used to represent a real occupant behavior. Detailed description of monitoring campaign is provided in the next sections. After the agent is trained for several weeks on the target house, it is deployed on this house. It means that the agent is no longer learning, but only controlling the system. This phase is computationally efficient and can be done on a low-price hardware such as a Raspberry Pi.

While TRNSYS models are used in all phases, it should be noted that the model in off-site training phase is a virtual model to be used in a laboratory, while the model used in on-site train and deployment phases is to represent an actual building.

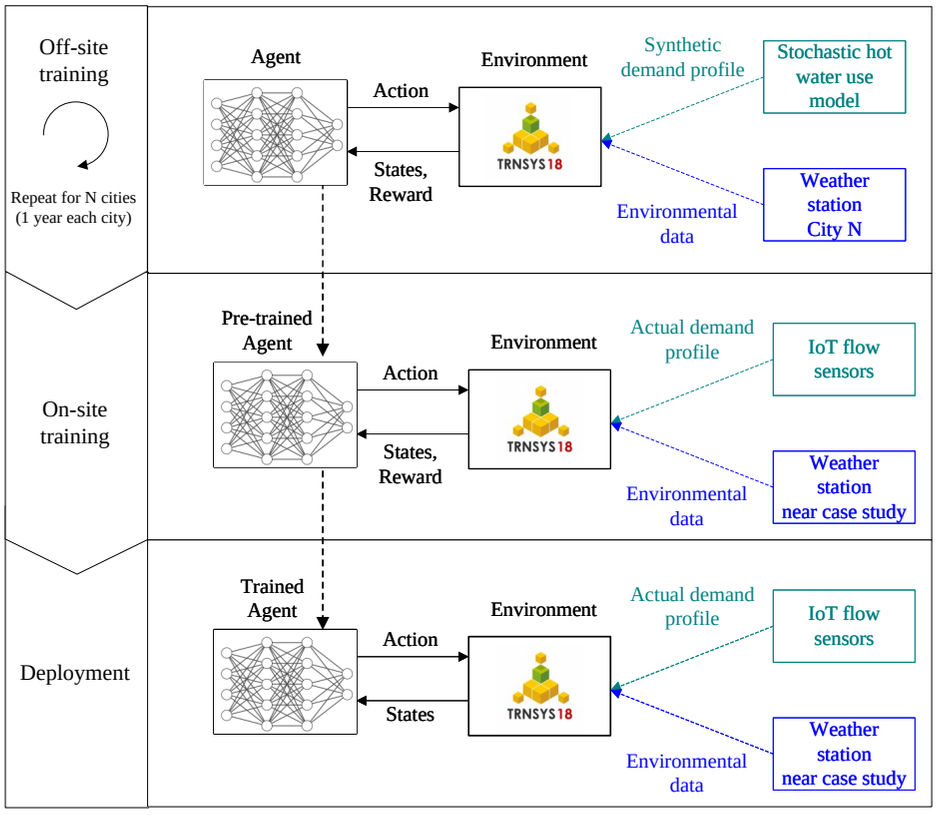


Figure 4.7: Training procedure

4.3.3.3 Different training scenarios

To get the full potential of RL it should be continuously trained, which enables it to adapt to all changes during the life-time of the system. This is, however, costly (in case of using cloud services) and computationally expensive. This research aims to gain a good level of adaptation by intensive off-site training to minimize the need for on-site training. Due to the complex and error-prone setup required for on-site training, it would be easier in practice if the on-site training phase would be reduced or even be totally eliminated. One of the aims of this study is to assess if the stochastic-based intensive off-site training can reduce or totally eliminate the need for on-site training on the target house. To this aim, three different scenarios for training phase are evaluated. These scenarios include:

- **On-site training and Short-time Deployment (RL-OSD):** After off-site training, the agent is trained on-site on the target house, and then deployed for a short period of 1 month;
- **On-site training and Long-time Deployment (RL-OLD):** After off-site training, the agent is trained on-site on the target house, and then deployed for a long period

of 8 months;

- **Direct Deployment (RL-DD):** After off-site training, without any further on-site training, the agent is directly deployed on the target house for a short period of 1 month;

For a better understanding, these three scenarios are visually presented in Figure 4.9.

4.3.4 Baseline control methods

In order to better highlight the advantage of a learning controller, it can be compared to the conventional rule-based controllers that only follow static rules while ignoring the variations of occupant behavior, solar energy or weather conditions. Two following rule-based controllers are also modeled in this study:

- **Rule-based controller with Conventional setpoints (RC):** A rule-based method which uses the setpoints of common practice. In this method, setpoint for indoor air temperature is considered as 21 ℃ with a deadband of 2 K, which is a recommended setpoint for healthy and comfortable air temperature [178, 179].

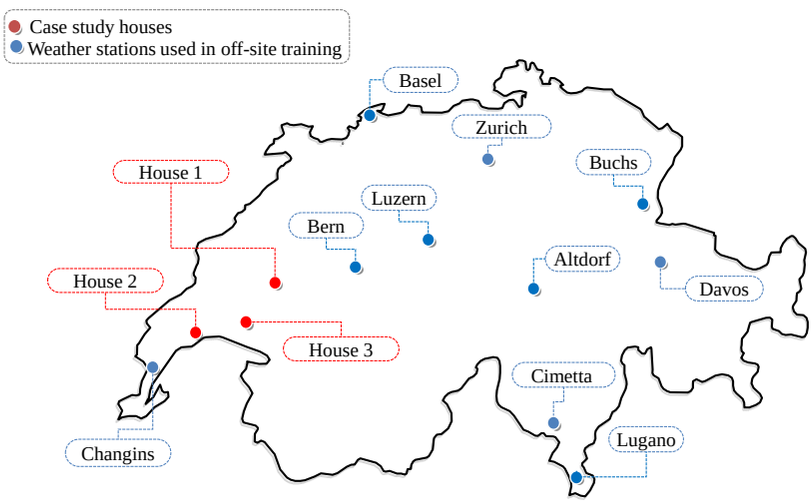


Figure 4.8: Location of cities used in off-site training phase as well as case study houses on the Swiss map

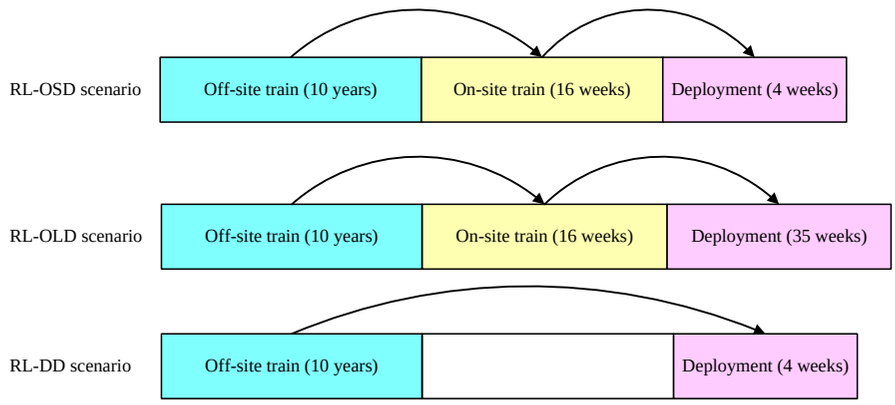


Figure 4.9: Visual presentation of different training and deployment scenarios

Also, the setpoint for the hot water tank is 60 ℃ with a deadband of 10 K, which is a commonly used setpoint to follow hygiene requirements in storing hot water [180, 181];

- **Rule-based controller with Energy saving setpoints (RE):** A rule-based method with similar setpoint air temperature to the RC method, but with the setpoint temperature of 50 ℃ for hot water tank to save energy;

Due to the hygiene aspects, the RE scenario is not common in practice. However, in this study this scenario considered to illustrate that the energy saving of proposed control framework is not only achieved by lowering the setpoint temperatures, but rather by learning how to optimally schedule the heating cycles. There are many other alternative control methods, such as using a heat curve, that are today applied in the buildings. These methods are similar in the sense that they follow static rules, which are detached from occupant behavior or renewable energy. Similar results are expected if a comparison is made between the learning agent and other rule-based controllers.

4.3.5 System sizes

Table 4.2 shows the specifications of modeled systems used in the off-site training and target houses. The agent is supposed to be able to adapt to a new building, with different area and different system sizes than what it has observed during the off-site training phase. The heated area and heat pump capacity in case study buildings are bigger (House 1), smaller (House 2), and almost similar (House 3) to the off-site training phase. Area of PV panels is equal to the available area for tilted roofs calculated based on [182]. The rated heating of Heat pump is also proportional to the heated surface area, and is sized based

on the capacity per area of a real-world similar installation presented in detail in [147]. The same tank size is considered in all houses for simplicity.

Table 4.2: System sizes used in off-site training and different case studies

	off-site training	House 1	House 2	House 3
Total heated area (m^2)	140	160	120	150
Heat pump rated heating (kW)	6	7	5	6
Heat pump compressor power (kW)	0.95	1.1	0.8	0.95
Tank size (L)	500	500	500	500
PV panels type	Monocrystalline	Monocrystalline	Monocrystalline	Monocrystalline
PV panels total area (m^2)	10	11	8	12
Panles slope	45	45	45	45

4.4 Results

In summary, the results of this study are presented in 5 sections as below:

- **Dataset overview:** Provides an overview of collected datasets during monitoring campaign;
- **Hyper-parameters:** Describes the hyper-parameters selected for the proposed framework;
- **Reward evolution:** Evaluates the convergence of the proposed framework;
- **Visual assessment:** Some operational parameters (e.g. air temperature, water temperature, hygiene, etc) are visualized to provide a detailed and hourly presentation of the agent performance;
- **Quantified assessment:** Quantification metrics are used to summarize and compare the agent performance (such as total energy use) with respect to the conventional methods;

4.4.1 Overview of datasets of different houses

Figure 4.10 shows the hourly variations of hot water demand, PV power production and outdoor air temperature in three case study houses. It can be seen that there is a good

diversity in hot water use behavior of case study houses. The Houses 1, 2 and 3 can be categorized as high volume (up to 250 L/h), low volume (mostly below 50 L/h), and medium volume (up to 150 L/h) consumers. There is a good variation also in the trend of PV power production in case study houses. Hourly variations of PV power on the first and second case studies show a decreasing trend, with higher values during the training phase compared to the deployment phase. On the third house, the date of monitoring campaign has been different from the first and second case studies, with the training phase starting from cold weeks and the deployment phase during the warmer weeks. Therefore, the trend of PV power production is increasing in this house, with higher hourly production during the deployment phase compared to the training phase. Variations of hourly outdoor air temperature also show a similar trend, with a decreasing trend on the first and second case studies and an increasing trend on the third case study. The deployment phase of the first and second case studies is during the cold weeks, when both space heating and hot water production is required, while the deployment phase of the third house is during the warm weeks, when only hot water production is required. The agent is supposed to learn that during the warm weeks there is less energy demand, and the variations of the hot water tank temperature only depends on the hot water demand. The overview of datasets shows that there is a very good diversity between the case studies, and between train and deployment phases. These variations provide a great opportunity to examine how the agent can generalize its knowledge and adapt itself to different situations, such as different hot water use behaviors.

To better explore the diversity in hot water use behavior between the case study houses, boxplots of their hourly hot water use data are also presented in Figure 4.11. Datasets from other residential buildings [13] show that hot water use pattern usually has two major peaks, one in the morning and the other in the evening. Regarding that the monitoring campaign in this study has been during COVID-19 pandemic, the monitored data over these three houses show some differences with the normal pattern of residential buildings. For example, the peak of average demand for the Houses 2 and 3 is located at the middle of the day, while in the normal situation occupants are at work on this time and no peak is expected. Also, the hot water use pattern in House 2 shows a quite uniform demand between 7 A.M. and 9 P.M., which indicates that the occupants have been spending most of their time at home. These differences indicate that the hot water use behavior over the case studies is more stochastic and less predictable than the normal behavior that the agent has observed during the off-site training. This abnormal occupant behavior on case study houses is valuable, because it lets to evaluate the adaptation potential of the agent to a behavior never observed before. It also shows if the agent can still perform well if the occupant behavior is significantly different from the normal behavior.

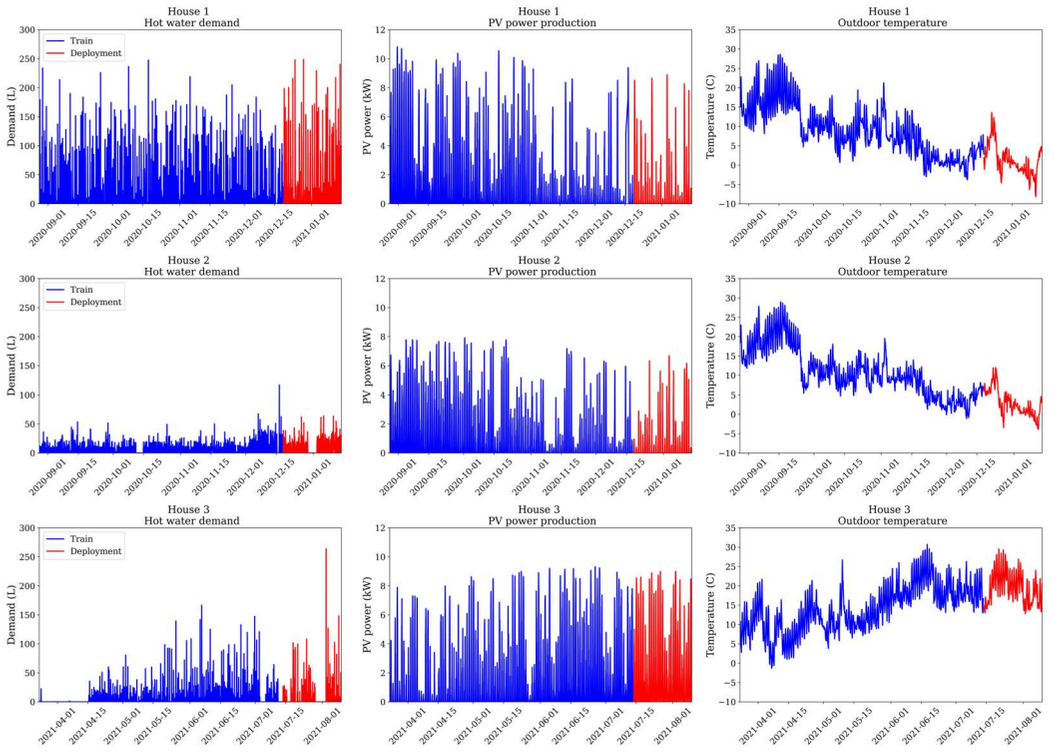


Figure 4.10: Hourly hot water demand, PV power production and outdoor air temperature on the case study houses

4.4.2 Hyper-parameters

The RL framework include a number of hyper-parameters that should be selected based on the specific problem and desired objectives [183]. The main hyper-parameters in this framework include specifications of agent (e.g. Learning rate, Batch size, Update frequency, Memory), weights of the reward function, and also the length of look-back vector for some specific states that are expected to have a higher importance for the target system to be controlled. The look-back vectors that are of specific interest in this study are the number of previous hours of hot water demand, PV power production and indoor air temperature to be included in the state vector. For each set of hyper-parameters all the phases of off-site training, on-site training and deployment should be repeated which takes a long time on a normal computer. Therefore, only a few of hyper-parameters could be evaluated in the sensitivity analysis phase. The hyper-parameters for the agent are selected based on the experience from our previous study [13], as presented in Table 5.1.

One of the important aspects of RL is the trade-off between exploration and exploitation [154]. To maximize the reward, the agent tries to select actions that has previously experienced and are expected to return a higher reward, which is called exploitation. On the other hand, it is still possible that the action with expected highest

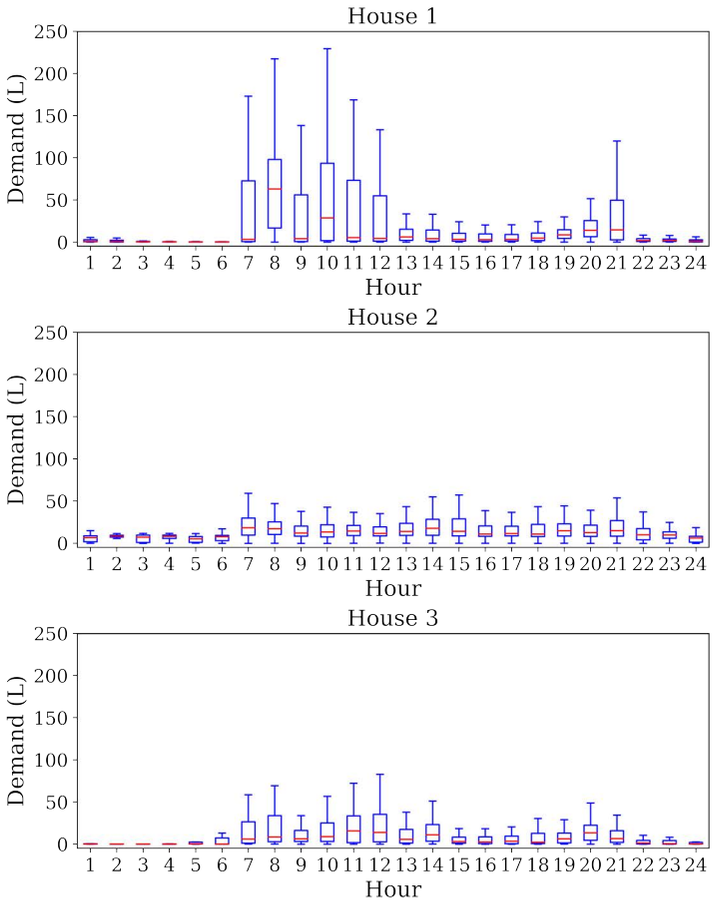


Figure 4.11: Boxplots of hourly hot water demand in case study buildings

reward would not be the best action, so it is required that sometimes the agent randomly selects an action during the training phase to better explore the environment, which is called exploration. One of the commonly-used methods to make a balance between exploration and exploitation is the ϵ-greedy method, in which a small probability of ϵ is specified and the agent performs exploration when a random value between 0 to 1 would be higher than specified value for ϵ. In this study, it is desired that during the off-site training phase the agent performs higher exploration (more random actions) at the beginning and then gradually reduces the exploration to near zero. Therefore, a linear decay is established for exploration, where the ϵ linearly decays from 0.9 to 0.0001 at each time step over the first 12 weeks.

Weights of the reward function are selected based on the relative importance of each term in the reward. The selected weights are indicated in Table 4.4. A weight of 1 is selected for energy term, because it is multiplied by the net energy usage, which is in the range of 0-4kW. The agent is supposed to reduce energy usage, without violating the comfort and hygiene aspects. Higher weights are selected for the hygiene and comfort

Table 4.3: Selected parameters for the agent

Hyper-parameter	Value
Agent type	Double deep-Q network
Learning rate	0.003
Batch size	24
Update frequency	4
Memory	48×168
Discount factor	0.9

terms to highly penalize the agent if any of these aspects are violated. The weight of hot water comfort is a bit higher than the weight of space heating. This is because the hot water use behavior is more stochastic, and the hot water use can change the tank temperature quite fast. Thus, the agent should be more conservative towards the comfort of hot water use.

Table 4.4: Selected weights for reward function

Weight	Associated term	Value
a	Energy	1
b	Hot water use comfort	20
c	indoor air temperature comfort	10
d	Hygiene	10

4.4.3 Reward evolution

The evolution of reward over the training phase should be monitored to evaluate if the agent has found an optimal control policy to minimize reward function. Figure 4.12 presents the weekly-averaged reward over the off-site training, as well as on-site training on each of the houses. It should be noted that energy reward in this framework is not avoidable. Therefore, depending on the heat pump capacity, variations of reward function up to -5 are due to the power use of heat pump. Considering the weights presented in Table 4.4, reward values lower than -10 (more negative values) indicate that the comfort or hygiene terms are also violated. As can be seen from the first diagram, there are 5 periods during the off-site training phase, where the value of reward reaches to -10 or below. In these periods, the agent has been trying to minimize the energy reward by turning OFF the heat pump, but due to a low hot water tank temperature it has violated comfort or hygiene terms. After each violation and receiving a high penalty, it has learned

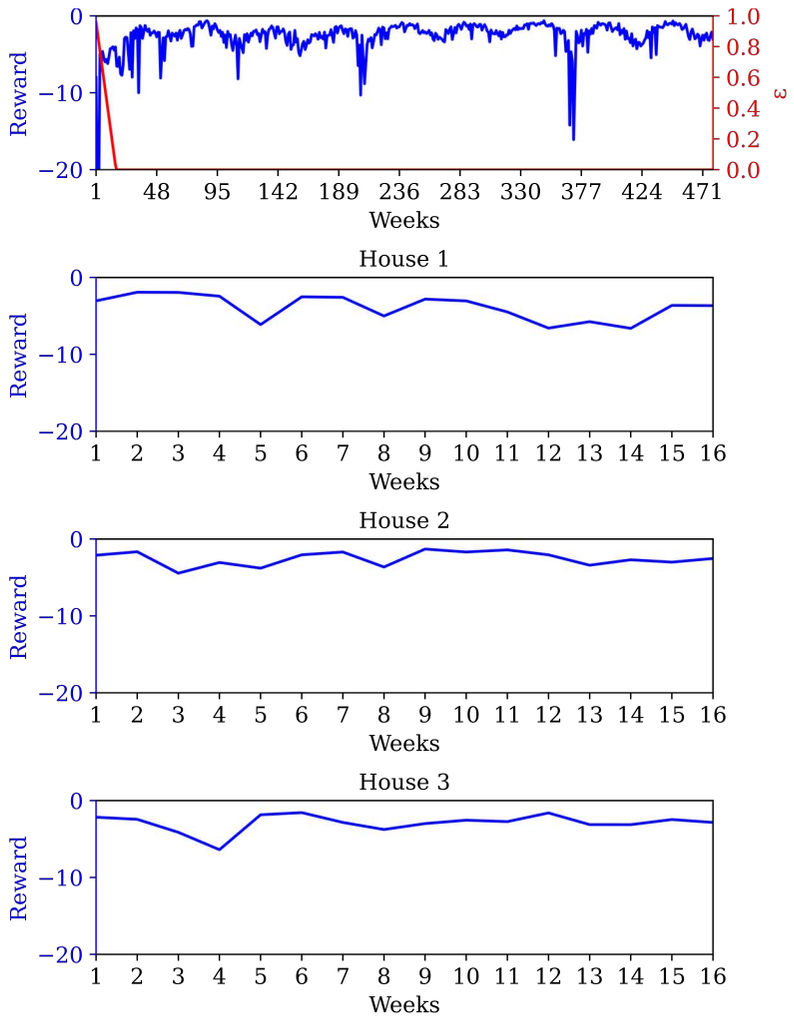


Figure 4.12: Evolution of reward over the off-site training stage and on-site training stages in each house

that it should increase energy usage to avoid the violation of other terms. After the last violation around week 377, reward value is almost stable. The value of reward function during the on-site training on the target houses is always above -10, and shows a good stability. This indicates that the agent has gained enough experience during the intensive off-site training phase, which has guaranteed an optimal policy since the first week of training phase on each target house. The fast convergence on the target houses, in spite of abnormal hot water use behavior, shows that the variations included in the off-site training phase (variations of the system sizes, hot water use pattern, weather conditions, etc) have provided a generalizable knowledge for the agent, and ensured the transferability to the other houses.

4.4.4 Visual assessment of the proposed framework

4.4.4.1 Performance of the agent during the off-site training

As shown in Figures 4.7 and 4.8, off-site training phase was performed for 10 years, each year on a different city and with different system sizes. It is interesting to have a closer look at the off-site training phase to see if the agent could preserve the occupant comfort with such variations. Figure 4.13 presents the boxplots of hot water tank and indoor air temperatures during the off-site training phase. It can be seen that there is a higher variance in hot water and indoor air temperatures over the first year, which is due to the lack of experience by the agent, as well as performing random actions during the exploration phase. From the second year, the hot water and indoor air temperatures show a lower variance and are closer to the comfort limits. It indicates that in only few hours the occupant comfort is slightly violated. Also, the average hot water and indoor air temperatures are higher over the first year. It shows that at the beginning the agent has been trying to preserve occupants comfort by spending more energy, but from the second year it has learned to further reduce temperatures and save more energy while respecting occupant comfort. Overall, from this figure it can be seen that although several parameters (weather, solar radiation, occupant behavior and system sizes) vary from year to year in the environment, the agent performance is stable since the second year. This indicates the adaptation potential of RL to the potential variations that can happen from building to building in a wide-spread implementation.

4.4.4.2 Performance of the RL-OSD

A major capability of the RL agent is adaptation to stochastic parameters, which in this problem are mainly PV power production and hot water demand. To visualize the adaptation potential of the agent, Figure 4.14 presents the control signal versus PV power production and hot water demand. As can be seen in this Figure, the agent mostly turns ON the heat pump when PV power is available. This is more clear in case of House 3. In this house, the deployment phase has been during the summer, with a higher PV power production and lower energy demand, which enables the agent to harvest most of the required energy from PV panels. Hot water tank temperature is also visualized to assess how the agent has adapted to the hot water use behavior to preserve the comfort aspect. It can be seen that the agent has successfully learned the hot water use behavior, because even in case of high volume demands, e.g. in House 1, the agent has always kept the hot water tank temperature above the comfort limit.

Previous studies have usually used the self-consumption of PV power as an evaluation metric for their proposed control approach [184]. However, it should be noted that in this study a higher self-consumption can be caused by the higher energy use of heat pump,

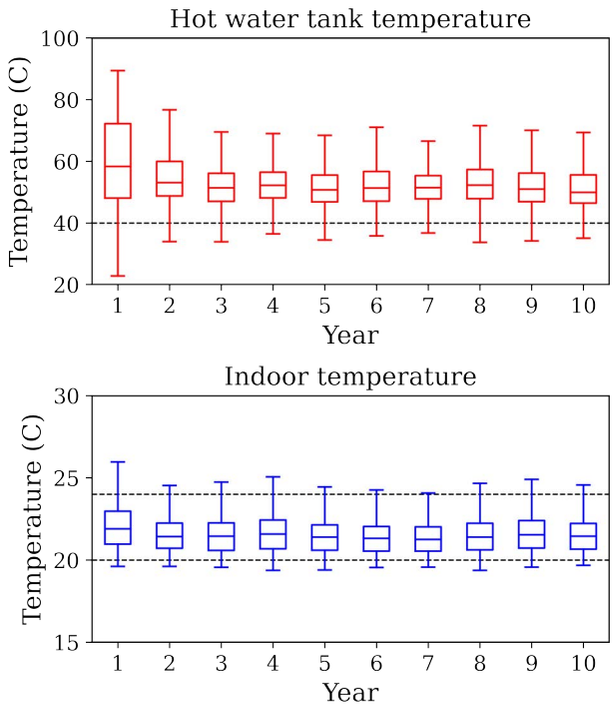


Figure 4.13: Boxplots of yearly hot water and indoor air temperature versus comfort limits

e.g. by operating with a lower COP, which is not desired. Rather, in this study the share of PV power production in total power consumption of heat pump is used for comparison. It should be noted that this share only indicates how much of heat pump energy use is directly covered by PV panels. The surplus power injected to the grid is not considered in this diagram. This is to better observe how the proposed control framework can get a better use of PV power production and reduce the interaction with grid. As shown in Figure 4.15, in all Houses, RL-OSD has obtained a higher share of energy consumption from PV panels. This share is much higher in case of House 3, as the deployment phase in this house has been during the summer, with much higher PV power production and lower energy demand. These results indicate that the proposed framework would provide a higher energy saving in regions with a high solar radiation.

To evaluate the comfort aspect during the deployment phase, boxplots of indoor air and hot water tank temperatures by different control methods are shown for House 1 in Figure 4.16. Due to the high number of plots only one house is presented, and the other houses show a similar performance. Indoor air temperature has a narrow comfort range which can be violated easily. Therefore, all of the methods show some violations of comfort. RL-OSD shows more violations than the rule-based methods, but the violations are less than 2 ℃ and happen in few hours, which therefore can be ignored. In case of the hot water tank temperature, similarly, RL-OSD shows very slight violations that can

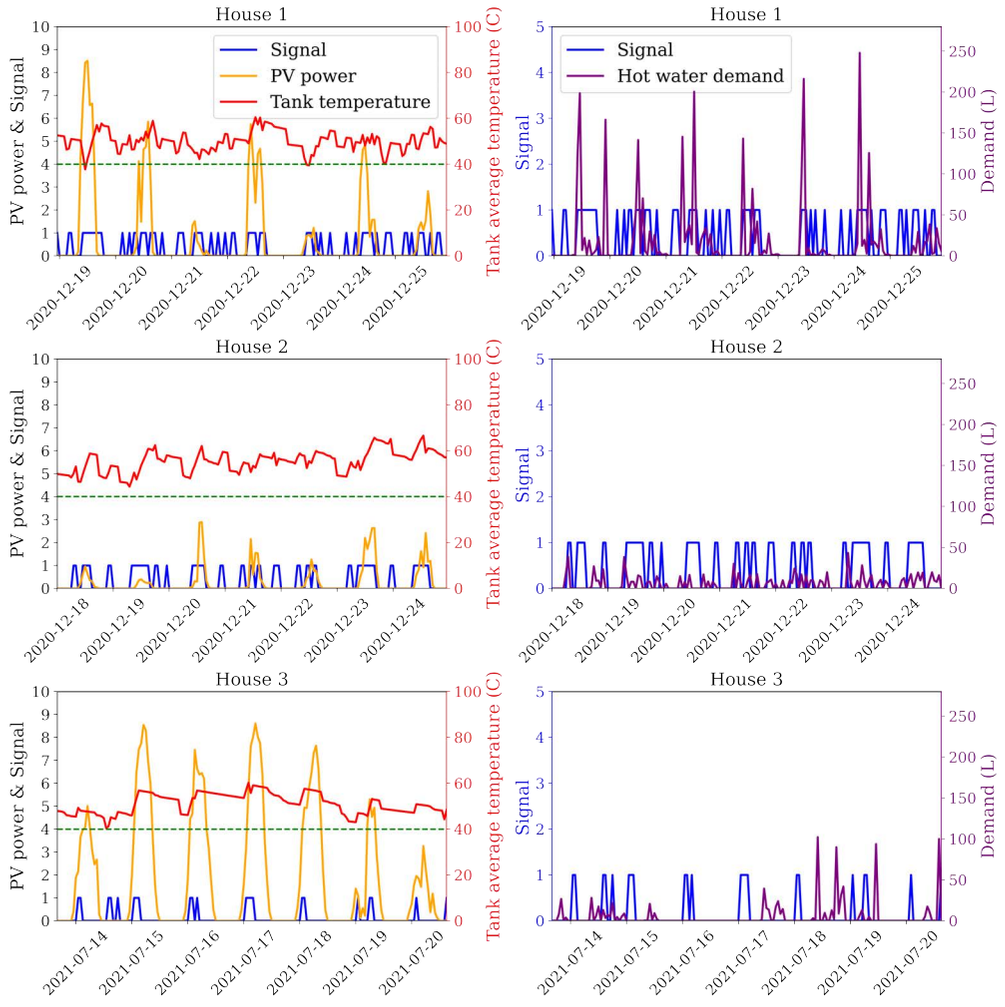


Figure 4.14: Adaptation of control signal to the PV power production and hot water demand in RL-OSD scenario

be ignored. Interestingly, violations by RL-OSD are even less than RE, which is due to the fact that RL-OSD tries to save energy by adapting to the occupant behavior, while RE tries to do so only by lowering the hot water tank temperature, regardless of occupant behavior.

Boxplots of Legionella concentration over the deployment phase are shown in Figure 4.17. As expected, both of rule-based methods maintain a lower concentration than the RL-OSD method, because they are over-conservative. While RL-OSD method is less conservative, it has always respected the hygiene aspect as the maximum concentration is less than 4500 CFU/L, which is much less than the risky limit of 500,000 CFU/L placed for single-family residential houses [165]. It shows that RL-OSD has learned to maintain hygiene aspect while avoiding over-necessary heating of the tank.

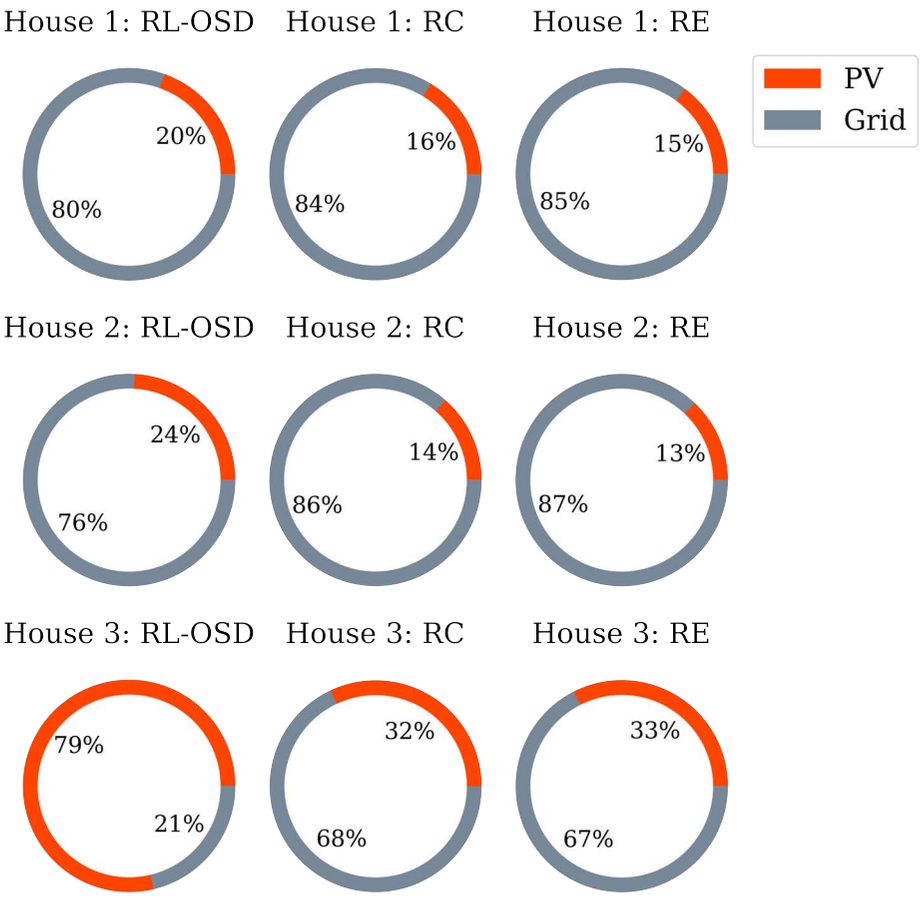


Figure 4.15: Contribution of PV power production in heat pump power consumption in RL-OSD scenario

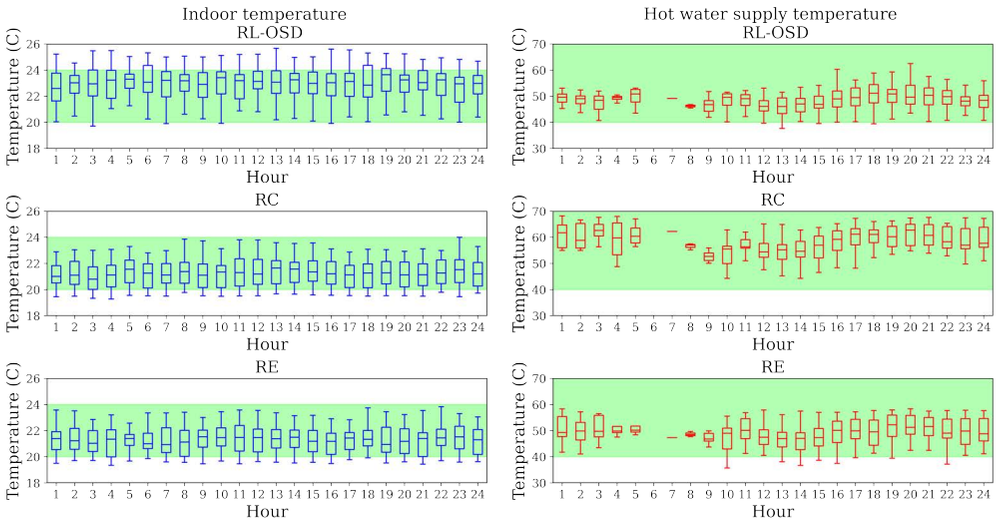


Figure 4.16: Boxplots of indoor air and hot water tank temperatures by three control methods in House 1 in RL-OSD scenario

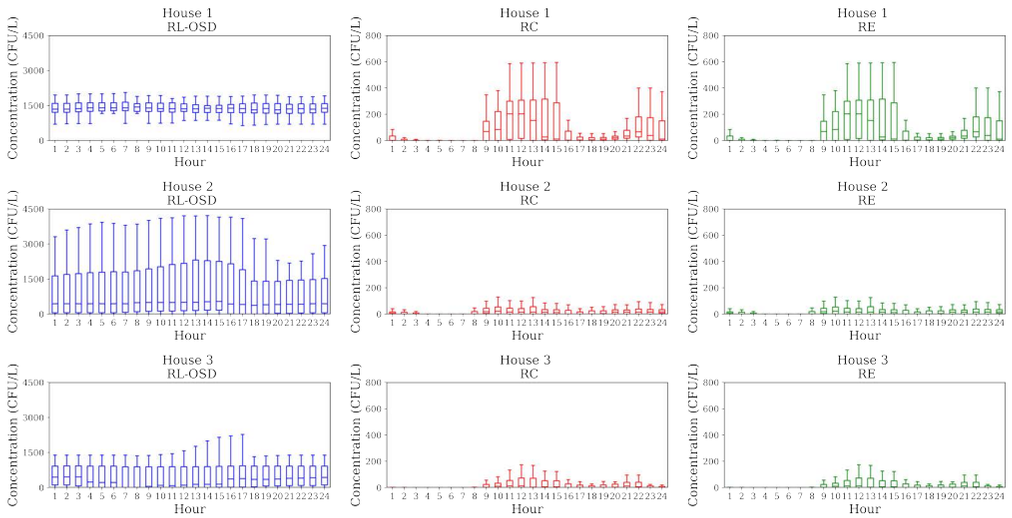


Figure 4.17: Boxplots of Legionella concentration in tank by three control methods over three case studies

4.4.4.3 Performance of RL-OLD

Figure 4.18 presents the performance of the agent over the long-time deployment (RL-OLD scenario). As can be seen, there are a lot of variations in hot water use behavior of occupants over this period, including a sudden decrease for one month, and an absence period. In addition, this period includes a good diversity in outdoor air temperature as it includes cold months at the beginning and hot months at the end. These diversities are valuable to assess how the agent will adapt to the possible changes in environment over a longtime deployment. As shown in this Figure, although there are significant variations in hot water use behavior, the agent has always kept the hot water tank temperature above comfort temperature of 40 ℃. There is an increase in the temperature of pressurized hot water tank from the middle of May (2021-05). This is because in this period there is a higher PV power production, a lower demand for space heating, and at the same time a sudden decrease in hot water demand. Therefore, hot water tank temperature is increased as the agent is trying to get the best use of PV power production by storing the surplus energy in the hot water tank. In this study an upper limit is not considered for the tank temperature, so the agent is free to store surplus energy by reaching a high temperature. A tempering valve should be installed at the tank outlet to mix hot and cold water to prevent the risk of scalding at the point of use. Indoor air temperature is also within the comfort limit, with slight violations of less than 2 ℃. Legionella concentration is also always below the risky limit, while it is higher during the cold season and lower during the warm season, when extra energy is stored by over-heating the tank.

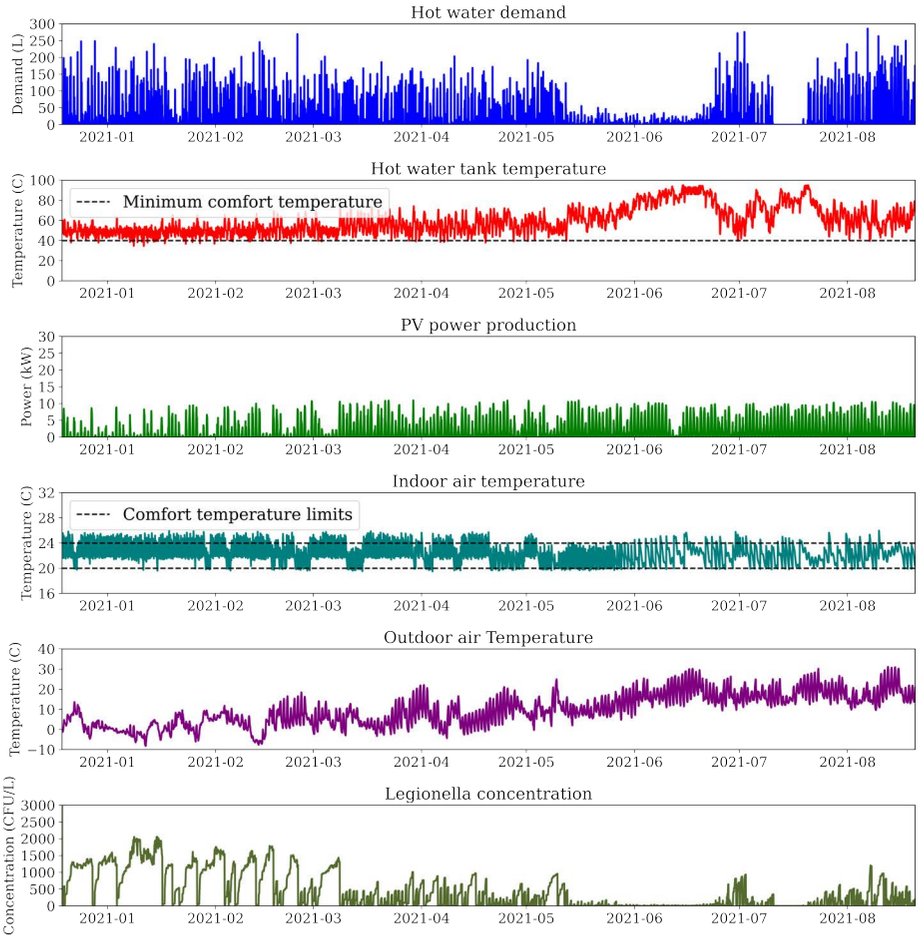


Figure 4.18: Performance of the RL-OLD agent during long-time deployment on House 1

4.4.4.4 Performance of RL-DD

Figure 4.19 compares the performance of the agent which is also trained on the target house (RL-OSD), versus the agent which is only trained off-site and has never observed the behavior of that specific house (RL-DD). The training phase of RL includes some randomness, mainly due to the exploration phase. Therefore, even if two agents are trained with exactly same specifications, they can have slightly different performances. So the slight differences between these two agents should not be associated with the lack of an on-site training phase in RL-DD. The two agents, thus, show very similar performance on Houses 1 and 2. The only significant difference is on House 3, where the RL-DD agent has kept a higher hot water tank temperature than the RL-OSD agent. This is because the RL-OSD has observed and learned the specific behavior of occupants on House 3, and is better adapted to their behavior than the off-site trained agent which has only observed the stochastic-based hot water use behavior. These diagrams show that

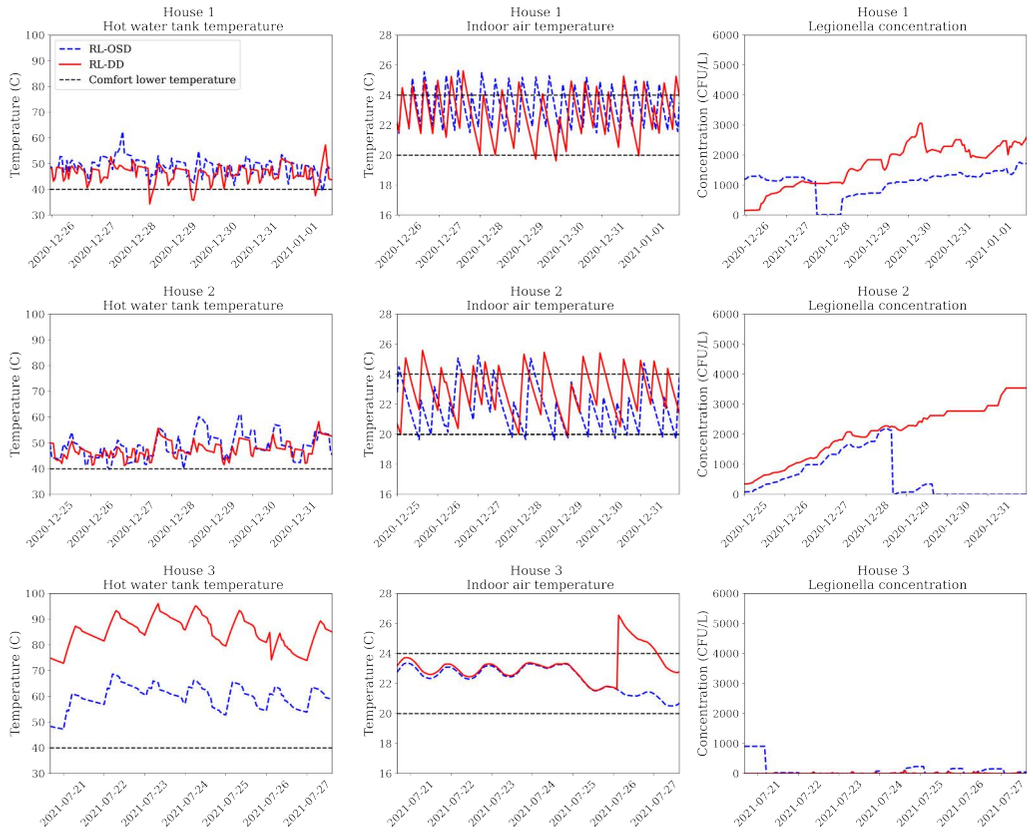


Figure 4.19: Performance of the RL-DD agent during direct deployment

while the RL-DD agent has never seen the specific parameters of the target house (weather conditions, occupant' behavior, etc.), it can still maintain the comfort and hygiene aspects.

4.4.5 Quantified assessment of the proposed framework

To compare the energy performance of the proposed RL framework with the conventional rule-based methods, the net energy use from grid by each control framework is calculated as Equation 4.11.

$$Energy\ use = \sum_{h=0}^{H} E_{hp} - \sum_{h=0}^{H} E_{pv} \tag{4.11}$$

Where $\sum_{h=0}^{H} E_{hp}$ and $\sum_{h=0}^{H} E_{pv}$ are the summation of hourly energy use by the heat pump and hourly energy production by PV panels over the intended period, respectively. This metric can be negative if the total energy production of PV panels are higher than the total energy used by the heat pump in that period.

4.4.5.1 RL-OSD versus baseline methods

Table 4.5 presents the performance metrics of RL-OSD versus RC and RE methods during the deployment phase. RL-OSD consumes the least energy in all the Houses. In House 3, all the methods have a negative energy consumption, which indicates that the total energy production of PV panels has been more than the total energy consumption of the heat pump and thus the difference is injected to the grid. The RL-OSD method in this house has a higher surplus of PV power production, which shows a better performance of this control method. To quantify the comfort aspect of the RL-OSD method, the percentage of total hot water demand which was met with a temperature less than the comfort limit is also indicated in this table. At the worst case, which has happened in House 1, 8% of total demand is violated. The average temperature of these violations is 38.9 ℃, which is very close to the comfort limit of 40 ℃. In case of space heating, although RL-OSD has violated the comfort limits during a few hours, the average of violations is very close to the comfort limits. In Houses 2 and 3, this average is in comfort limits because some of the violations has been less than 20 ℃ and some other more than 24 ℃. It can be therefore considered that in all the houses RL-OSD has properly maintained the occupant comfort. The average COP of heat pump by RL-OSD is always equal or lower than by RE. It proves that the energy saving by RL-OSD is not achieved just by lowering the hot water tank temperature (and therefore increasing the COP), but by properly scheduling of heating cycles to profit more from PV power production.

4.4.5.2 RL-OLD versus baseline methods

Table 4.6 presents the metrics of RL-OLD scenario. These metrics show that over the long-time, even without any other on-site training, the agent has provided an energy saving while maintaining the occupant comfort and water hygiene. This scenario was presented to prove the performance of the trained agent over a long time deployment without any further on-site training. But if it is technically possible in practice, sequential or continuous training of the agent will probably provide a higher energy saving.

Table 4.6: Summary of performance of Long-time deployment scenario (RL-OLD) with other control methods

	RL-OLD	RC	RE
Energy use (MWh)	5.17	7.6	5.4
Violation of DHW comfort (%)	5.3	0	9.2
Average temperature of DHW comfort violations (℃)	38.7	0	38.5
Number of space heating comfort violations (Hours)	1015	548	532
Average temperature of space heating comfort violations (℃)	23.8	19.8	19.9

4.4.5.3 RL-OSD versus RL-DD

Table 4.7 represents the energy use and comfort metrics of RL-OSD versus RL-DD. The performance of two scenarios are quite similar in different houses. In House 1, the RL-OSD has a bit higher energy use, but in turn has provided a better comfort. So it can be concluded that the off-site training step has provided a generalizable experience for the agent and eliminated the need for further on-site training. However, in case a significant change of occupant behavior happens during the operation, the RL-DD can fail to adapt to the new conditions and provide comfort.

4.4.6 Conclusion

There are several stochastic parameters such as occupant behavior, renewable energy potential, and weather condition, that increase the complexity of developing an optimal control method for residential energy systems. Among them, occupant behavior is of significant concern, as it is highly stochastic, unique in each building, varies by time, and therefore very challenging to model and predict. This study proposes a data-driven and model-free control method based on Reinforcement Learning, that can learn these stochastic parameters by itself, and maintain an optimal operation. The agent in this framework also takes into account the hygiene aspect of hot water to save energy without compromising the occupants' health. The goal of the learning agent is to save energy while maintaining the health and comfort of occupants. The energy system evaluated in this study was a PV-assisted air-source heat pump for space heating and hot water production, though the methodology and proposed framework are easily adjustable to other systems. A two-step training method is proposed, including an off-site phase integrating stochastic hot water use behavior to provide an initial experience for the learning agent, and an on-site phase to learn and adapt to the behavior of the target house.

Table 4.5: Comparison of performance between RL-OSD and rule-based control methods in three case studies during the deployment phase

	House 1			House 2			House 3		
	RL-OSD	RC	RE	RL-OSD	RC	RE	RL-OSD	RC	RE
Energy use (MWh)	1.24	1.6	1.34	0.62	1.22	0.8	-0.97	-0.78	-0.9
Violation of DHW comfort (%)	8.1	0	0	5	0	0	1.7	0	0
Average temperature of DHW comfort violations (°C)	38.9	-	-	39	-	-	38	-	-
Number of space heating comfort violations (Hours)	153	0	0	84	0	0	29	0	0
Average temperature of space heating comfort violations (°C)	24.3	-	-	22.2	-	-	23.5	-	-

Table 4.7: Comparison of performance between RL-OSD and RL-DD in three houses during the deployment phase

	House 1		House 2		House 3	
	RL-OSD	RL-DD	RL-OSD	RL-DD	RL-OSD	RL-DD
Energy use (MWh)	1.24	1.03	0.62	0.63	-0.97	-0.94
Violation of DHW comfort (%)	8.14	17.1	5	3.9	1.7	0
Average temperature of DHW comfort violations (°C)	38.9	37.9	38.9	30.8	38	-
Number of space heating comfort violations (Hours)	153	142	84	126	29	136
Average temperature of space heating comfort violations (°C)	24.3	24.2	22.2	24.2	23.5	24.9

The framework was evaluated for three houses in different regions of Switzerland. For these case study houses, weather and solar radiation data were collected from nearby weather stations, and hot water use data was experimentally monitored to evaluate the framework on a realistic occupant behavior. The following main conclusions can be drawn from this study:

- The proposed framework (RL-OSD scenario) achieved 7% to 22% energy-saving compared to an energy-saving rule-based method (RE), and 22% to 47% compared to the common practice rule-based method (RC), without violating the occupant comfort and water hygiene.
- The agent properly learned the variations of PV power production in each building and adapted the heating cycles to the PV power production to get the best use of free solar energy (As can be seen in Figure 4.14).
- Evaluation of direct deployment scenario (RL-DD) indicated that the stochastic-based intensive off-site training provides a generalizable knowledge for the agent, and therefore it could still outperform rule-based methods even without any on-site training on the target houses. As expected, the agent that was also trained on the target houses (RL-OSD scenario) indicated slightly better performance. It shows that, if enough computational power is available, the stochastic-based off-site training can be further extended by including many possible conditions that can happen in reality (e.g., a sudden change in occupant behavior and weather conditions, change of system components, etc.), which makes it possible to directly implement the trained agent on several houses without any need for on-site training. It will significantly facilitate the transferability of the proposed framework to other buildings.
- Evaluation of long-time deployment scenario (RL-OLD) indicated that the agent could provide a satisfactory performance over long time, and further sequential or continuous training is not necessary.

With the increasing complexity of residential energy systems, rather than hard-programming the expert knowledge as a rule-based or model-based control method, it is possible to let the agent learn the optimal control method by itself in each specific building. In this study, experimentally measured data was used in simulations to provide a realistic while safe environment to perform a primary test of the agent performance.

Chapter 5

DeepValve: An occupant-centric control framework for space heating in offices

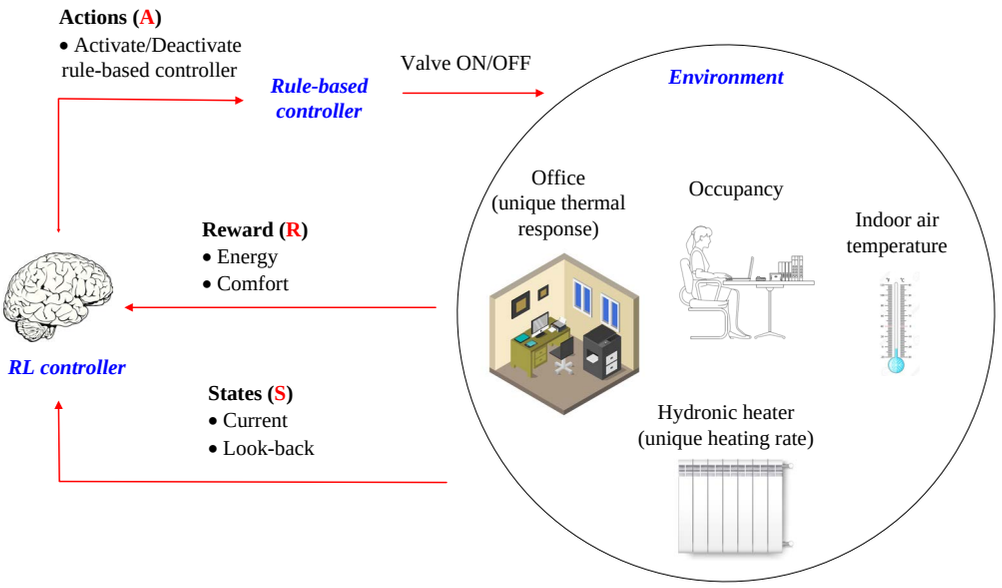


Figure 5.1: *DeepValve* control framework in a nutshell

5.1 Abstract

Stochastic occupancy profile and unique thermal response time of each office are two main challenges to adapt space heating schedule to occupancy profile in offices. Space heating controls in offices are usually programmed to follow static and conservative schedules detached from actual occupancy, which usually results in energy waste by heating vacant offices unnecessarily. A control solution should be developed to be easily

transferred to many offices and continuously adapt the space heating to the occupancy profile. This study introduces *DeepValve*, a novel occupant-centric control framework based on Reinforcement Learning that can simultaneously learn the unique occupancy behavior and unique thermal response behavior of each office, and accordingly schedule office heating to save energy while maintaining comfort. Algorithm, hardware setup, and training methodology of *DeepValve* are specifically designed to ensure an easy transfer to many offices to save energy. The performance of the proposed framework is tested in two steps, first step with simplistic simulation tests and second step with realistic experimental tests in an environmental chamber by applying different thermal response behavior in each test. It was observed that the agent came up with interesting strategies to save energy, for example, turning OFF the system at a proper time before occupants' departure. Results indicate that the agent can quickly adapt to new offices and save significant energy while maintaining occupant comfort.

5.2 Introduction

An ideal controller for space heating should heat the building only when it is needed [47]. To this aim, the controller should be able to continuously adapt the space heating to the occupancy schedule. Adaptation to occupancy behavior in a fast-response system such as lighting can be easily provided by incorporating an occupancy detection sensor and a simple logic to turn ON the lights when an occupant is detected [40]. However, there are two main challenges in using the same approach for space heating systems. Space heating systems have a slow response time, depending on several parameters, such as thermal inertia of the building, heat losses, and the heating rate of air conditioning systems, which is unique in each building [185]. Given the slow response time of heating system, to maintain the occupant comfort the heating system should predict the occupancy and start to heat the building in advance. This requires a model that can properly predict the occupancy behavior. However, the occupancy behavior is highly stochastic and complicated, can be affected by environmental, temporal, and random factors, and is unique in each building [186], which makes it challenging to develop a generalizable model to predict occupancy behavior in many offices. Current control systems can be called expert-based controls, in which a group of domain experts hard-program their knowledge as a set of rules and heuristics (rule-based methods (RBC)) [147] or optimization models to be solved dynamically over time (Model Predictive Control (MPC)) [187, 188]. Expert-based systems are limited to expert knowledge [189]. Uniqueness of occupant behavior and thermal response in each building make it challenging for the experts to program generalizable rule-based or model-based controls that can be easily transferred to many offices and optimally adapt heating schedule to occupancy behavior.

An alternative to the expert-based approach is the *learning-based* approach, which is inspired by neuroscience and how the brain works [79]. In this approach, instead of hard-programming the solution to the controllers a human-like learning mechanism is programmed to the controller, that enables it to autonomously learn the optimal control strategy from scratch [189]. Reinforcement Learning (RL) is a method of Machine Learning that can provide this learning ability [80]. In RL, the learning controller (so-called agent) learns the system model and subsequently the optimal control policy only by interacting with the system and observing the impacts of its actions [155, 190]. With the increasing complexity of energy systems in buildings due to intermittent renewable energy generation on site, increasing use of heat pumps with variable efficiency, electric car charging, energy storage, grid interactions, etc. [80], developing accurate models for expert-based methods (such as MPC) is becoming more complicated and time-consuming than ever. This is while an RL agent does not require any prior model of the system, can learn the system model by itself from scratch, and continuously adapt to the changes [159]. Secondly, a learning controller can learn a unique control strategy for each building, depending on the unique occupant behavior, which can not be easily achieved by expert-based controllers [191–193].

Researchers have used RL to provide learning ability to different building systems, such as solar thermal systems [162], air handling units [159], lighting [58], space heating [154, 192] and hot water systems [72, 191, 194], and indicated promising performance compared to expert-based systems. Among them, some studies evaluated RL for developing occupant-centric controllers. An occupant-centric controller is defined as a controller that perceives occupant behavior through sensory data and accordingly adapts the control actions [41]. Depending on the type of occupant-related data that is taken into account by the controller, occupant-centric control can be categorized into occupancy-centric and occupant behavior-centric. Occupancy-centric controls tend to adapt the control actions to the occupancy state or number of occupants, whereas occupant behavior-centric controls tend to adapt to any other kind of occupant-related data, such as hot water demand, indoor temperature preferences, etc [41]. Occupant-centric controls can also be categorized depending on the temporal aspect of the algorithms. If the algorithm responds to occupant behavior in real time, for example by turning on the lights when a motion is detected, it is categorized as reactive control. If the algorithm relies on the prediction of occupant behavior to take actions in advance, it is categorized as predictive control [40]. Given that stochastic occupant behavior is a major source of uncertainty for optimal control of buildings, the field of occupant-centric control is gaining increasing interest and indicates promising results [80]. For fast-response systems, such as lighting, a reactive occupant-centric control based on simple logics can be used. However, for slow-response systems, such as thermal systems, a predictive

occupant-centric control is required to take actions in advance [40].

Considering that RL can learn occupant behavior to adapt future actions, it has a good potential to be used for occupant-centric controls. There are yet few studies on RL for occupant-centric control, while the number of publications is increasing in recent years [80]. Jung et al. [65] proposed an RL control framework for personalized control of indoor air temperature based on occupants' activity and physiological data. A wristband is used to wirelessly collect physiological and acceleration-related data from the occupants. A CNN-based deep learning model is used to recognize the occupant activity (e.g., sitting, walking) based on the acceleration data from a wristband. The RL model then takes into account the activity type and physiological data to decide the best setpoint temperature to balance between comfort and energy use. The proposed method reduced the thermal discomfort of occupants by 10.9%. Soares et al. [195] used RL to maximize the self-consumption of PV power production by storing the excess heat in a hot water tank or batteries. The proposed algorithm learns the stochastic hot water use behavior of the occupants, uses predictions of local PV production, and considers the dynamics of the system to increase self-consumption while maintaining the comfort of the occupants. Marantos et al. [196] proposed an RL-based smart thermostat that takes into account the presence of occupants, their number and activity, the indoor and outdoor weather data and energy use to select the indoor temperature setpoint. The aim of the proposed controller was to make a balance between comfort and energy use. Zhang and Lam [67] proposed an RL model to control the supply water temperature to the radiant system regulated by a three-way valve that mixes the supply and return water. The agent tried to reduce the heating provided by the radiator while maintaining the occupants' comfort. The behavior of occupants considered in this study included the occupancy status and the thermal comfort feedback, both included in the state vector. The thermal comfort feedback collected using a mobile app was converted to Predicted Percentage of Dissatisfied (PPD) and provided to the agent. Lee et al. [197] proposed an RL framework to control the indoor air temperature in a room with a variable air volume system. It was a simulation-based study where a room and a heating system were modeled in EnergyPlus. Two aspects of occupant behavior were considered, namely occupants' presence and their desired air temperature setpoint. Both parameters were simulated based on stochastic models and integrated into the EnergyPlus model. Fazenda et al. [198] proposed an RL control framework which learns the schedule and desired setpoint temperature of occupants, and controls the heating system accordingly to make a balance between comfort and energy use. System models were developed in MATLAB, and stochastic models were used to represent occupants' behavior. Park et al. [58] proposed a device called *Lightlearn*, which uses RL for occupant-centric control of lights in offices. The device was designed to enable both manual and automatic switching ON/OFF lights,

which therefore allowed learning of occupant interactions over time. It was installed in five different offices for eight weeks. The performance of the *Lightlearn* solution was compared with occupancy-based and schedule-based methods in terms of energy use and comfort. Kazmi et al. [72] proposed a model-based RL control framework to balance comfort and energy use in heat pump water heating systems. In particular, they used model-based heuristics that incorporate the vessel state and occupants' behavior into the optimal control problem.

The RL studies for occupant-centric building controls usually consider comfort and energy use aspects. Another important aspect that can be also taken into account by the RL agent is the health of occupants, which can be influenced by regulating air temperature, water temperature, air quality, etc. Heidari et al. [191] proposed an RL control framework for water heating systems that learns the hot water use behavior of occupants and plans the heating cycles accordingly. The algorithm aimed to save energy while maintaining the occupant comfort and water hygiene in the storage tank. Water hygiene, as the health-related aspect, is about the growth of Legionella bacteria in the tank which should prevented by periodical over-heating of stored water. The trained agent was tested using the real-world hot water use and weather data collected over 29 weeks from a residential house in Switzerland. In another study, Heidari et al. [192] proposed an RL control framework for occupant-solar-centric control of a photovoltaic-assisted space heating and hot water production system. By learning the hot water use behavior and solar power generation, the algorithm plans the heating cycles of hot water tank and select the indoor air temperature setpoint to save energy while maintaining comfort and hygiene. The proposed framework was tested using the real-world hot water use and weather data measured in 3 residential houses in Switzerland. Yang et al. [199] proposed an RL framework for the simultaneous control of indoor air quality and air temperature. The framework aimed to minimize the energy cost while maintaining CO_2 level and indoor temperature within the desired healthy ranges. The study was simulation-based, and the number of occupants was modeled as a random number.

Research on the use of RL for occupant-centric building control has indicated promising results, but it is still at the early stage. Several gaps should be addressed in this field. Following are the main limitations associated with the previous studies:

- **Simulation-based:** Most of the studies in RL are simulation-based, in which both building and occupant behavior are simulated. Building models are usually developed using building simulation software, and occupant behavior is modeled using stochastic or random models. This is mainly because the hardware setup required for establishing an RL framework in practice is usually complicated and requires a coordinated work between Internet of Things (IoT), controls, and

energy engineers. A simulation model is a simplified representation of an actual building, and an agent performing well in a simplified simulation environment does not necessarily have a good performance in a more complicated actual building. Similarly, a stochastic model of occupant behavior may not represent all the complexities of real occupants, such as changes in number, habits, or preferences.

- **Case-specific:** Most of the control frameworks proposed by previous studies are developed for a specific case and little/no attention is paid to transferability. This is because in most of them, the algorithm and hardware setup was developed for a specific layout of energy system, or the framework was only trained and then tested on the target case study without evaluating the transferability potential.
- **Complicated hardware setup:** In some studies, data from several sensors were needed to provide the required information for the agent. Relying on the input from several sensors is a drawback in practice and limits the widespread implementation of the proposed methods. Because it will increase the risk of failure due to the malfunctioning of sensors, a higher initial and maintenance cost, and a higher labor work for installation.

5.2.1 Research Scope

Space heating in office buildings usually follows static schedules that are set once to the thermostat and rarely updated afterward. The indoor temperature is usually set to a comfort temperature for the expected occupancy hours and set to a lower (setback) temperature for the rest of the hours such as nighttime, holidays, and weekends. This approach has to assume a static schedule for the office occupancy, while the actual occupancy of an office is dynamic and can vary from day to day. Consequently, it commonly happens that the office is heated when the occupants are not present. An ideal control approach should schedule indoor heating, taking into account the unique occupancy behavior and thermal response of each office. Due to the uniqueness of occupancy and thermal response in each office, rule-based or model-based controls can be developed only for a specific office, but they lack generalization potential. To transfer expert-based controls to different offices, specific parameters such as thermal characteristics of room should be updated for every other office. This study aims to develop a novel control framework that can simultaneously learn the unique occupancy schedule and the unique thermal response time of each office, and accordingly adapt the heating schedule to save energy while maintaining the thermal comfort. This study includes the following main novel aspects:

- **Simultaneous learning of occupancy and thermal response:** To adapt the space heating schedule to occupancy, a possible approach is to separately predict the

arrival and departure time of the next day and then schedule the office heating accordingly [39]. In this approach, the optimal time for starting and ending of space heating should be determined by experts, depending on the thermal response time of each specific office. *DeepValve* simultaneously learns the occupancy and thermal response time of each office to optimally adapt the heating schedule without any expert knowledge. This can provide a plug-and-play occupant-centric controller for offices.

- **Easy Transfer to many offices:** A major objective of this study is to design a control framework that can be easily transferred to many offices and convert their heating into occupant-centric. To this aim, several considerations are taken into account in developing this framework. First, the framework is only focused on scheduling the heat emission system in each room, independent of the type and layout of the main heating system. The framework can be implemented on any hydronic-based heating system (floor heating, ceiling heating, radiator, etc) as the most common system in Europe offices [200]. It is worth mentioning that the same methodology can be also applied to the air-based systems. In that case, an additional constraint about the ventilation requirements should be taken into account. Second, the framework is formulated to rely on a minimum number of sensors and actuators. It will also reduce the installation cost, and the risk of failure due to hardware malfunctioning. Finally, training methodology is designed to provide generalizable knowledge to the agent and ensure quick adaptation to a new office.

- **Experimental assessment on different setups:** After ensuring good performance in simulations, the framework is experimentally implemented on different setups representing different offices. Each setup includes a different hydronic heat emission system (ceiling or floor heating), with a different thermal response time adjusted by the air change rate. While most of the previous studies have only implemented RL in simulations [80], the experimental tests in this study can firmly prove the performance of the proposed framework in a real-world case with higher complexity than a simulated environment. Given the limited experimental studies on RL, the detailed description of experimental setup also provides a guideline for future studies.

5.3 Methodology

5.3.1 Physical Layout of the Control Framework

The physical layout of the proposed framework is shown in Figure 5.2. It is designed to rely on a minimum number of sensors and actuators and to be compatible with any

hydronic heating system. The framework only needs to collect two types of data from a single thermal zone of the office: the indoor air temperature and the state of occupancy as a binary value. The indoor air temperature can be easily obtained with an air temperature sensor, and the occupancy status can be obtained using any kind of reliable occupancy detection method. Previous studies proposed several occupancy detection methods, such as using a Bluetooth module to detect cellphone devices [201], infrared thermopile array sensor [202] that can detect both stationary and moving people, WiFi usage data [48], environmental sensors (air temperature, CO_2, etc) [203], smart power meters [204], or a combination of these methods [205]. Figure 5.2 shows the use of an infrared thermopile array sensor, as a non-intrusive, simple, and robust method to obtain occupancy status [202]. But occupancy detection method is not the focus of this study, and any method that can provide a binary value with acceptable accuracy can be used. For big offices, several air temperature sensors and occupancy detection devices can be used, and their data can be combined (e.g., averaging air temperature data and taking the maximum of occupancy status binary value). A controller reads and stores the data from the mentioned sensors to be used by the algorithm. The proposed control framework can be implemented on a cloud service or on a local computer. Figure 5.2 shows a local computer connected to the controller to read data and write values to the controller. To regulate the heat emission to the office, an electric valve can be installed at the inlet of the heat emission system. The emission system can be any type of hydronic heating systems such as radiator, radiant floor heating, or radiant ceiling heating. The controller can open or close the electric valve, depending on the decision made by the control framework. As can be seen in Figure 5.2, the hardware layout of the proposed framework is simple and easy to be applied in many offices. The control framework is designed for one single thermal zone. If there are multiple radiators available in the zone, a single valve can be installed on the main inlet of hot water before divisions.

5.3.2 Conceptual layout of the control Framework

The conceptual block diagram of the proposed framework is shown in Figure 5.3. This framework is consisted of the RL model and the RB model. The RL model is based on Double Deep Q-learning, which is a value based method suitable for problems with a discrete action space. In a typical deep Q-learning method, the Q value of performing the action a in the state s ($Q(s_t, a_t)$) is calculated as follow:

$$Q(s_t, a_t) = r_t + \gamma . max_a Q(s_{t+1}, a) \tag{5.1}$$

In which r_t is the immediate reward, and $max_a Q(s_{t+1})$ is the highest possible Q value

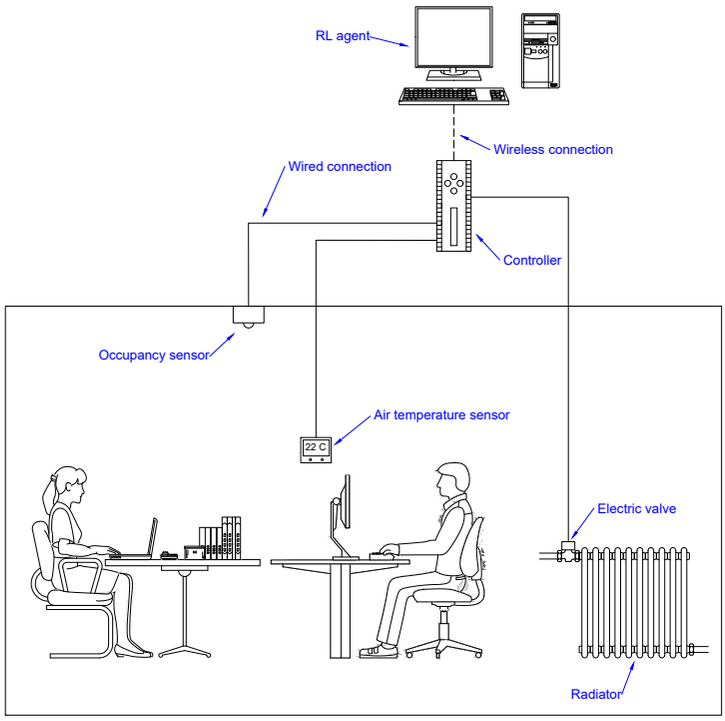


Figure 5.2: Physical layout of the *DeepValve* control framework

in the next state. In this formulation, both the current Q-value ($Q(s_t, a_t)$), and the future Q-value ($max_a Q(s_{t+1})$) are calculated with the same neural network. It is known that it can lead to the overestimation of the Q-value for a specific action, which results in the selection of a non-optimal action. To overcome this issue, a modified version of Deep Q-learning, known as Double Deep Q-learning [206] is used. Double Deep Q-learning involves two neural networks. The first network, called *the online network*, is used to control the environment directly, and its weights are constantly updated. The second network, called *the target network*, is only used to predict the target value, and its weights are updated only after N iterations. Therefore, in a Double Deep-Q learning, the online network calculates the $Q(s_t, a_t)$ term and the target network calculates the $max_a Q(s_{t+1})$ term in Equation 5.1. This solves the overestimation issue and also improves the stability of learning [206]. The RL model takes the current state of the environment and selects between two possible actions: $Disable$ action and $Enable$ action. If $Disable$ action is selected, then the RB model is not called, and the electric valve is turned OFF. If the $Enable$ action is selected, the RB model is activated. The RB model is the typical Rule-Based method commonly used in thermostats. This method follows a simple rule, which turns ON the system when the indoor air temperature is below a lower limit and turns OFF the system once the temperature is above a higher limit. The RB control keeps the air temperature within an interval centered on the setpoint temperature (e.g. $22℃ \pm$

0.5℃). Once the RB model is activated, it observes the current air temperature, considers the setpoint temperature given by the user, and turns ON/OFF the valve to reach and maintain the air temperature within the interval of $T_{setpoint} \pm \Delta T$℃. This study considers a setpoint of 22 ℃ with a dead band of 0.5 ℃. However, both of temperature setpoint and temperature interval can be selected by user with the only limitation that it should be within the thermal comfort zone. This framework integrates the adaptiveness of the RL method with the robustness of the RB method. If a single RL model was used to directly control the electric valve, the agent had to also learn the relationship between the valve opening and the indoor air temperature to properly regulate the air temperature based on the occupancy schedule. But in the proposed framework, the complicated control problem is decoupled into two simpler problems of (1) adapting to occupancy schedule and thermal response that is handled by RL, and (2) tracking the user-input setpoint that is handled by RB. It simplifies the problem for the RL agent, increases the safety and robustness of the controller, and also facilitates the integration of user-input setpoint temperature.

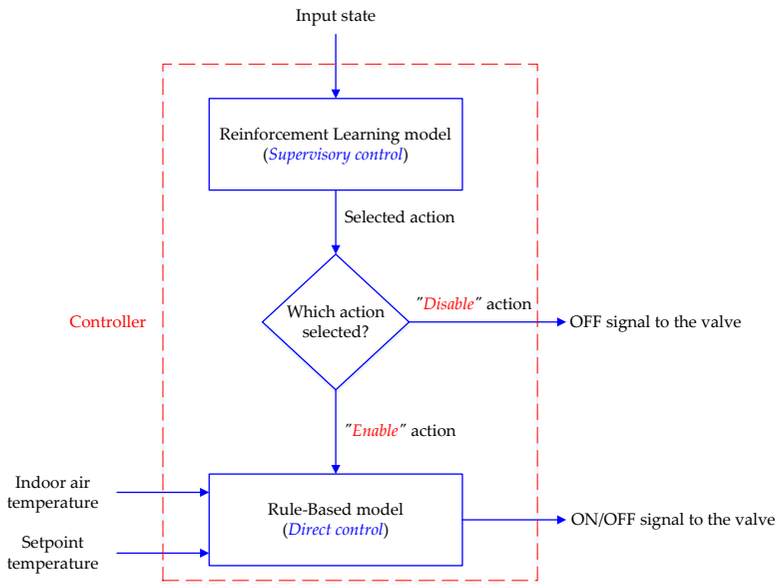


Figure 5.3: Conceptual block diagram of the *DeepValve* control framework

5.3.2.1 Agent and Environment setup

RL is consisted of an agent and an environment. This study uses Tensorforce [131], which is a Python library based on Tensorflow with very customizable classes for modeling the RL agent and environment. Table 5.1 presents the hyperparameters selected for the agent. To balance the exploration and exploitation tradeoff in this study, the $\epsilon - greedy$ method is used. A linear decay is established for exploration, where the ϵ linearly decays from 0.99 to 0.0001 at each time step over the first 30% of the training timesteps. The update

frequency is set to 24 timesteps, which means that the agent updates the policy every 12 hours. The rest of hyperparameters are tuned by running the models multiple times.

Table 5.1: Selected parameters for the agent

Hyperparameter	Value
Agent type	Double Deep-Q network
Learning rate	0.003
Batch size	24
Update frequency	24
Memory	672000
Discount factor	0.9

The environment in this framework is a Python class that receives the selected action by the agent, executes it, and gives back the new state and reward value to the agent. It can execute the action on a simulated model of the system or on an actual system. As will be explained in section *Training and Testing Procedure*, the training and initial testing of the agent is done in a simulated environment, and the final test is done in an actual environment (environmental chamber).

5.3.2.2 States, Actions, and Reward Design

The RL model in this framework uses a timestep of 30 minutes, which means that the RL model receives the states and the reward and selects the next action every 30 minutes. Due the certain response time of the space heating systems, a duration of 30 minutes is considered to ensure that the system has enough time to respond to the selected action, and the agent can observe the consequence of the selected action at the end of the timestep. In practice, a backup controller can be also added to sense the occupancy more frequent (e.g. every 1 minute) and turn ON the system if occupants are present and the temperature is not in comfort zone. Integration of backup controller is not evaluated in this study. Figure 5.4 shows the states and actions used in the RL model (The number of layers and nodes do not represent the actual Neural Network used in this study. The figure is just to show the states and actions). Given the complexity of the occupants' behavior, enough information should be provided to the agent to be able to learn and predict the occupants' behavior properly. The states vector in the RL framework includes 17 parameters; out of them, 16 parameters are focused on the behavior of occupants to provide useful information for predicting the presence of occupants, and 1 parameter is used to provide the current indoor air temperature. The following parameters are included in the state vector:

- $\boldsymbol{OS(t-n)}$: This parameter shows the occupancy status of the office during the timestep starting at $t-n$. For example, if the current time is 16:00, $OS(t-1)$ shows the occupancy status from 15:30 to 16:00, and $OS(t-2)$ shows the occupancy status from 15:00 to 15:30. The status of occupancy is a binary, indicating "1" for the occupied office and "0" for the empty office, regardless of the number of occupants. To be more conservative with the comfort of occupants, if the office was occupied only part of a timestep, it is still considered as 1. The occupancy status during the time steps of $(t-1)$, $(t-2)$, $(t-3)$, $(t-4)$, $(t-5)$ and $(t-6)$ are included to provide the history of occupancy status during the last 3 hours from the current timestep. Considering the repetitive habits and routines in occupancy, to predict the future occupancy status it will be helpful for the agent to know the occupancy status of the same time during the previous day. To include this information, the occupancy status during the timesteps of $(t-43)$, $(t-44)$, $(t-45)$, $(t-46)$, $(t-47)$ and $(t-48)$ are also included. Considering that the duration of time interval is 30 minutes, these parameters show the occupancy status during the next 3 hours from time *t* over the last day. Previous studies on other kinds of occupant behavior, such as hot water use behavior [191, 194], have shown that the future occupant behavior has a strong correlation with the historical behavior. Therefore, providing the mentioned historical occupancy information for the agent is expected to be very useful for the prediction of future occupancy.

- $\boldsymbol{TLO(t)}$: This parameter represents the total number of timesteps passed from the last occupancy. For example, if occupants leave the office at 18:00, *TLO(t)* at 19:30 is equal to 3. This parameter helps the agent to (1) learn the common durations of occupants' absence and (2) distinguish between Monday and other working days.

- $\boldsymbol{Time_{sin}(t)}$ and $\boldsymbol{Time_{cos}(t)}$: Time of day is highly correlated with occupancy [207] and thus should be provided to the agent. To better reflect the cyclic nature of time of day to the agent, each timestep is converted into two coordinates of sin and cos terms as follow:

$$Time_{sin}(t) = sin(\frac{2 \times \pi \times timestep(t)}{48}) \tag{5.2}$$

$$Time_{cos}(t) = cos(\frac{2 \times \pi \times timestep(t)}{48}) \tag{5.3}$$

  Converting to *sin* and *cos* terms, and the use of dummy variables [13] are two common method for encoding the temporal features in Machine Learning to represent their cyclic nature.

- $\boldsymbol{Daytype(t)}$: To help the agent to learn the difference between the occupancy

pattern of working days and weekends, a binary parameter is included to indicate working days as "1" and weekends as "0". Therefore, the agent can learn when this parameter is "0", it is a weekend and no occupancy is expected. Similarly, when it is "1" the normal occupancy of a working day is expected.

- $\boldsymbol{T_{indoor}(t)}$: This parameter shows the current indoor air temperature, which should be known by the agent to properly plan heating cycles.

The values of all these state parameters are normalized into [0-1] by dividing by a constant number. The possible actions of the RL model are as follows:

- $\boldsymbol{Disable}$: Disable the RB control and turn OFF the electric valve.
- $\boldsymbol{Enable}$: Enable the RB control to achieve the desired setpoint air temperature.

The proposed framework aims to reduce energy use while maintaining comfort, with the comfort as a priority to energy saving. Therefore, The reward function is consisted of two competing terms, *energy reward* and *comfort reward*. The ***energy reward*** is calculated as follows:

$$R_{energy} = \begin{cases} -a \times (T_{indoor}(t) - T_{indoor}(t-1)) & If\ T_{indoor}(t) > T_{indoor}(t-1) \\ 0 & If\ T_{indoor}(t) \leq T_{indoor}(t-1) \end{cases} \tag{5.4}$$

where $T_{indoor}(t)$ and the $T_{indoor}(t-1)$ are the indoor air temperatures at the end and the beginning of each timestep, accordingly. The heat given to the indoor space can be calculated as $Q = M_{air} C_{p\,air} \Delta T_{air}$. This formulation of energy reward means that the agent should try to minimize the total temperature increment of the office over the episode, which is proportional to the total heat given to the office air. Although other sources of heat such as internal gains can contribute to the temperature increment, with this formulation the agent learns how to minimize the part of the temperature increment that is dependent on the agent decision. This formulation of the energy reward only requires an air temperature sensor and eliminates the need for an energy meter or water flow and temperature sensors at the inlets and outlets. Also, it is independent of whether the inlet water flow and temperature are fixed or variable. To adjust the importance of the energy term in the total reward function, a scaling factor a is used.

The comfort term punishes the agent if the occupants are present and the indoor air temperature is not within the thermal comfort zone. If the occupants are not present, or if they are present and the temperature is within the comfort zone, the comfort reward

would be zero. The ***comfort reward*** is calculated as follow:

$$R_{comfort} = \begin{cases} -b & If\ occupancy = 1\ and\ (T_{indoor}(t) < 20\ or\ T_{indoor}(t) > 24) \\ 0 & else \end{cases} \tag{5.5}$$

where b is a constant number. Although the comfort reward could be proportional to the distance of the current air temperature from the comfort limits, the use of a constant number facilitates the learning for the agent.

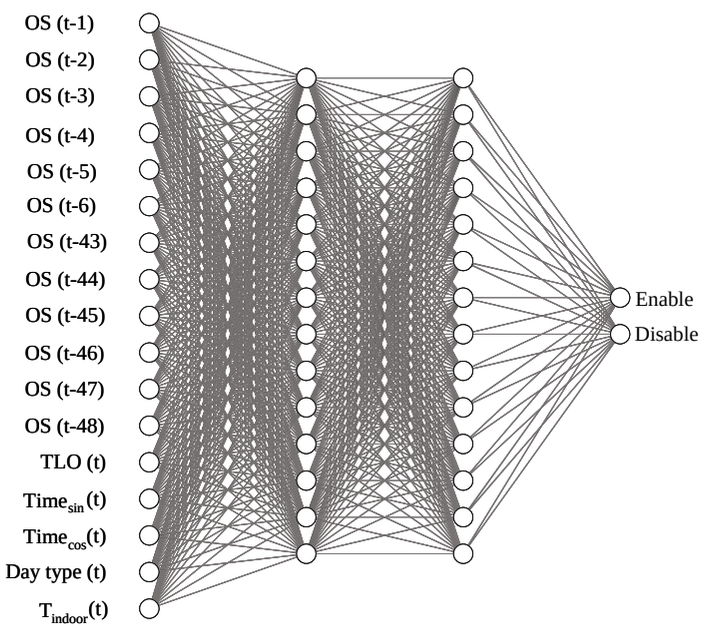


Figure 5.4: States and actions used in the RL model

5.3.2.3 Training and Testing Procedure

Simulation is a safe environment in which the agent can interact with a virtual model of the system and learn the optimal behavior without causing any discomfort to the occupants or any damage to the physical system. Figure 5.5 shows the training and testing steps of the framework. Although the proposed framework can directly start learning on the target office, this work proposes an intensive training process in a simulation environment prior to the implementation on the target office. The first step is the intensive training, which aims to provide a generalizable knowledge to the agent and speed up the learning process when implemented on the target office. To gain generalizable knowledge, in this step the agent is trained on a variation of thermal response time and occupancy profiles. A large dataset of the office occupancy from the literature is collected (e.g., from works [58, 208, 209]). The dataset includes 23 different offices in three different countries as overviewed

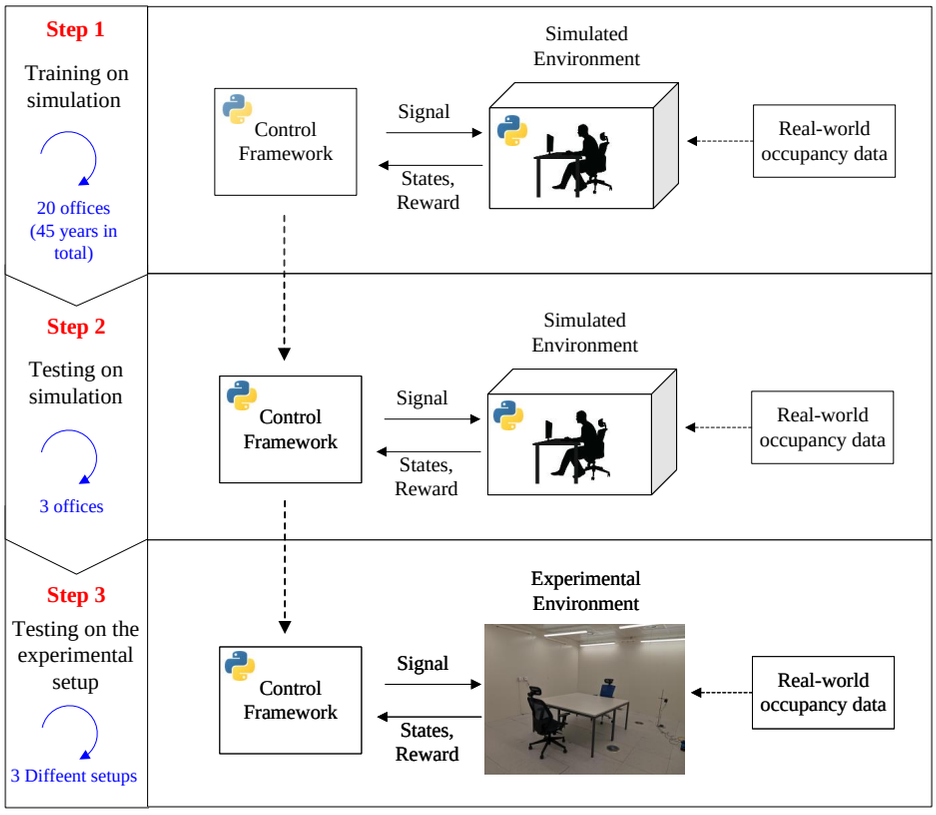


Figure 5.5: Training and testing procedure of the *DeepValve* control framework development

in Table 5.2. Each monitored office included 1 or 2 (maximum) occupants. The variations of thermal response time and occupant behavior prevents the agent to be overfitted to a specific case, and accelerates the adaptation to a new office.

Table 5.2: Overview of different sources used in the occupancy dataset

Offices	Location	Reference
Office 1-17	Frankfurt, Germany	[208]
Office 18	Calabria, Italy	[209]
Offices 19-23	Texas, USA	[58]

Duration of the collected data, as well as boxplots of the occupied hours for each office are shown in Figure 5.6. Offices 1-17 were monitored for an extended period of 2 to 4 years, while office 18 was monitored for around 13 months, and offices 19-23 were monitored for shorter periods of around 2 months. The boxplots show that there is a good variation in the occupied hours between offices. For example, occupied hours has a high

variation in office 10, while it has a low variation in office 20. Out of the collected data, offices 3, 17 and 22, are kept for testing. The data from the rest of the offices are combined to form 45 years of training data.

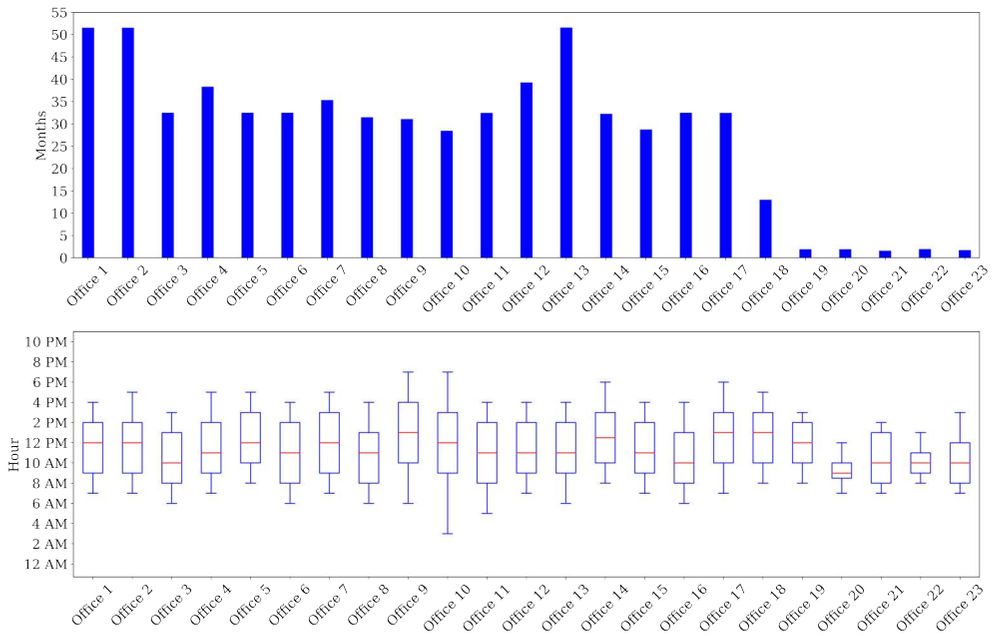


Figure 5.6: Duration of data collection and boxplots of occupancy hours in different offices (Data from [58, 208, 209])

In this study, the extensive training in the simulation environment is more focused on providing experience with potential variations of occupant behavior (as the main stochastic variable), with less emphasis on variations of building characteristics. This is because once the pre-trained agent starts learning on a specific office, the occupant behavior has a higher variation than the building characteristics over the building lifetime. Therefore, to facilitate training on the big dataset of occupancy data, a simple building model written in Python [210] (developed based on the hourly dynamic model of ISO 13790) is used to simulate the office room. The simplicity of the model, and the fact that both agent and environment are in Python, significantly reduces the computational time and makes it feasible to train on a large occupancy dataset. This is more desired than training on a detailed building model (e.g. developed in TRNSYS) but with a smaller dataset of occupancy.

As shown in Figure 5.5, at the beginning of each timestep the environment class executes the control signal (ON/OFF signal of the valve) on the building model and at the end of the timestep it gives back the new state and reward to the agent. The occupancy status is read for each timestep based on the current time of the day from the collected occupancy dataset. The variations in occupancy behavior is based on real-world collected

data, and the variations in thermal response time is implemented by gradually altering the heated area, thermal capacitance, and heat loss at each month. It is equivalent of implementing the agent on a different office every other month. The office area is 30 m^2 in the first month, and increases by 0.1 m^2 every month to 88 m^2 in the last month. The thermal capacitance and heat loss coefficient of the office are altered proportional to the floor area, as indicated in Equations 5.6 and 5.7. It should be noted that these equations use arbitrary numbers and do not reflect the thermal characteristics of any kind of specific office building.

$$Thermal\ capacitance = 400 \times area\ \frac{kJ}{K} \tag{5.6}$$

$$Heat\ loss\ coefficient = 2 \times area\ \frac{W}{K} \tag{5.7}$$

Once the agent is trained, it should be tested and compared to the baseline models before being implemented in practice. Therefore, the second step, as shown in Figure 5.5, aims to evaluate the agent performance and compare with the baseline control methods on 3 different offices. In each office, a different area (thus different thermal characteristics based on Equations 5.6 and 5.7) with a different occupancy behavior that is never seen before by the agent is used. A possible approach for implementing the RL agent on the target offices is static deployment, in which the agent is no longer learning and updating the policy but only controlling the system. Static deployment is computationally efficient and can be implemented using cheap hardware such as Raspberry Pi. But if enough computational power is available, the agent can continue to learn and update control policy while controlling the system. In this case, it can adapt to any future changes in the occupant behavior or system and continuously get smarter over time. In this study, to explore the full potential of the RL agent, the agent continues to learn while controlling the system during the tests. Table 5.3 shows the main specifications of these tests. In each of these tests, two baseline control methods are also modeled, namely Schedule-driven Rule-Based control (SRB) and Occupancy-driven Rule-Based control (ORB). SRB method is the common practice in offices, in which the setpoint air temperature is set at a comfort temperature for specific hours (in this study 22 ℃ from 6 AM to 7 PM), and it is set at a lower temperature (16 ℃ in this study) (called setback) for the rest of the hours and during the weekends. In this study, the setpoint is 22 ℃ from 6 AM to 7 PM over working days and is 16 ℃ for the rest of time. In ORB method, the rule-based controller with the setpoint of 22 ℃ is enabled only when the presence of people is detected by an occupancy detection method and disabled for the rest of the hours. It is not a common approach in practice, but it is modeled to represent

a reactive occupant-centric control where the heat emission control is synchronized with the occupancy schedule.

Table 5.3: Specifications of three simulation tests

Simulation test	Office area (m^2)	Occupancy data	Duration
1	50	Office 3	4 weeks
2	60	Office 22	4 weeks
3	70	Office 17	4 weeks

Once the agent outperforms the baseline models in simulation environment, the third step (shown in Figure 5.5) is to test the RL agent on an actual (physical) system. Due to the simplifying assumptions in simulations, controlling an actual system for the agent is more challenging than controlling a simplified simulation model. For example, in the simulation model it is assumed that no heat is emitted to the room after the valve was closed. But in practice once the water supply valve is closed the heat emission system remains warm and continues to heat the room until its temperature equalizes with the indoor environment. Thus, the experimental tests aim to assess the adaptation potential of the agent to different thermal response time on an actual system. An experimental test facility with a high degree of flexibility is used to assess the agent adaptation to different thermal response time of the office. The thermal response time is altered by using different heat emission systems (e.g., floor heating, ceiling heating, and a combination of both) and by adjusting the air change rate per hour. Altering the type of heating system and air change rate is to represent different offices that the agent can be implemented on. A detailed description of the experimental test setup is provided in the following sections. The objective of the third step in the control framework development is not to compare the RL performance with the baseline models. Instead, the aim is to observe if the agent can properly adapt to different thermal response time to maintain the comfort in an actual system. Due to the limited availability of the experimental facility, long-time tests to compare with baseline models were not possible. While the agent is not expected to outperform the baseline methods only after one day, the SRB method is also tested to better observe the difference of RL with SRB. Table 5.4 summarizes the three experimental tests using RL, as well as the test using the baseline method. The abbreviations in this table reflect the type of the activated heat emission system and the set air change rate per hour (e.g., the label "RL-CF-4" indicates the RL method using both ceiling and floor heating at 4 air changes per hour). For easier comparison, the heat emission system and air change rate per hour of the baseline test are similar to Experimental test 2.

Table 5.4: Specifications of experimental tests

Test	ID	Heat emission system	Air change per hour (ACH)	Supply air temperature (°C)	Duration
Experimental test 1	RL-CF-4	Ceiling and floor panels	4	14	24 Hours
Experimental test 2	RL-F-6	Floor panels	6	14	24 Hours
Experimental test 3	RL-C-8	Ceiling panels	8	14	24 Hours
Baseline test	SRB-F-6	Floor panels	6	14	24 Hours

5.3.3 Testing on the Experimental Setup

5.3.3.1 Experimental Facility

An environmental chamber (EPFL- Smart Living Lab, Switzerland) was used to experimentally test the performance of the RL-based control framework. The chamber is a highly insulated room of 4.3(w) x 5.8(l) x 2.5(h) m^3 volume. It is specially designed as a multi-purpose, versatile, and programmable facility for thermal comfort studies. Figure 5.7 illustrates a drawing of the facility and the interior of the room. The chamber is equipped with radiant panels and air supply diffusers both on the ceiling and the floor. Radiant panels on the ceiling and the floor include 6 independent sections with 8 panels each. Water supply to each section can be controlled independently by an electric valve. All air diffusers could also be controlled independent of each other. Air flow rate and the direction of air supply and extract can be changed by electric dampers. Figure 5.8 details the installation of air dampers and electric valves on the ceiling, floor, and the main water supply pipe. The hot or cold water required for the radiant panels is provided by a water-to-water heat pump (Hidros WZA012HELSRVP2U). The required air supply is also provided by an air handling unit (AHU) located outdoors.

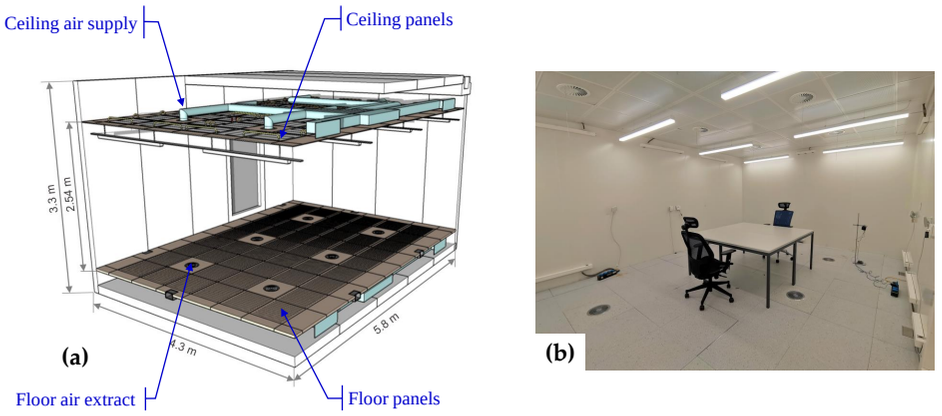


Figure 5.7: Overview of the experimental facility: (a) 3D schematic of the environmental chamber (b) an interior view

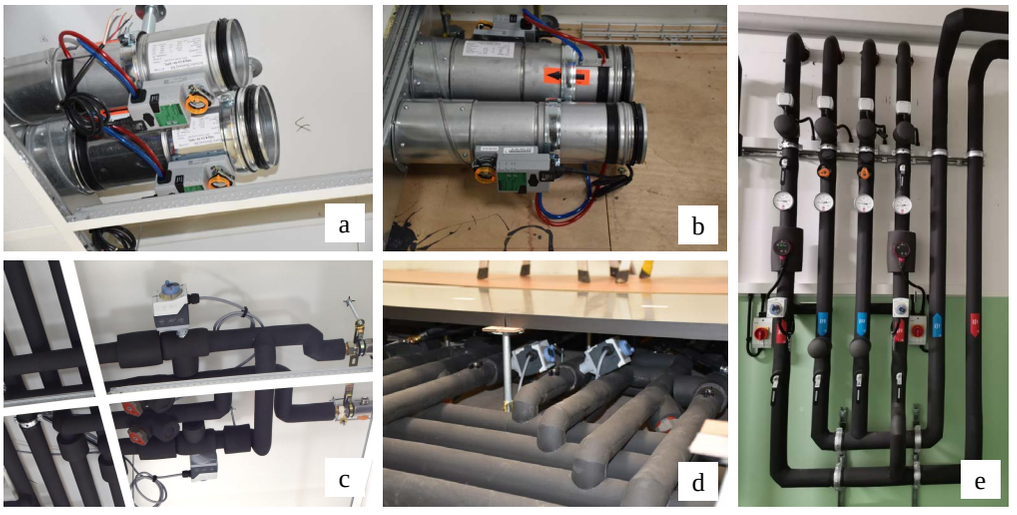


Figure 5.8: Installation of electric valves and dampers for flexible control of air and water systems in the environmental chamber: (a) electric air dampers on the ceiling, (b) electric air dampers on the floor, (c) electric water valves on the ceiling, (d) electric water valves on the floor, (e) electric valves on the main supply pipes (outside the climatic chamber)

Figure 5.9 shows the schematics of the control system in the environmental chamber. The sensors and actuators are wired to the SIEMENS controller, which is connected to the WiFi network. The RL control framework is implemented on a PC connected to the same network. The agent is exactly the same as in the simulation-based tests. The environment class needs to be modified to read state parameters and write control actions to the actual controller rather than the simulation model. The interaction between the RL control framework (on PC) and the controller is based on the BACnet protocol.

To communicate with the controller through this protocol, the environment class uses a Python library called BAC0. BAC0 is developed to process BACnet messages on an IP network. Using this setup, the developed control framework can wirelessly control the environmental chamber and record performance data.

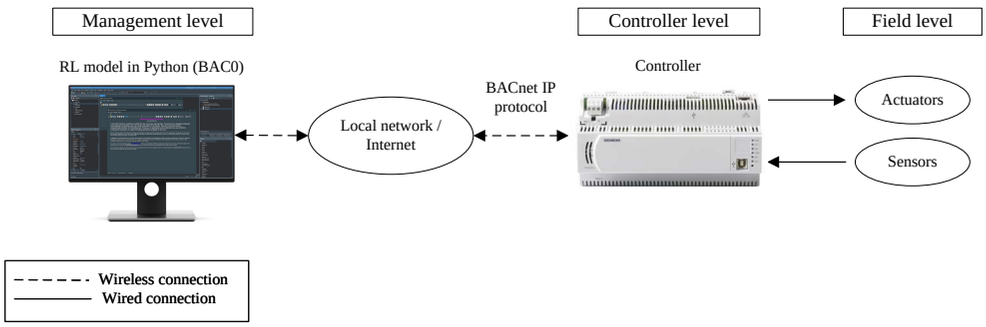


Figure 5.9: Layout of the implemented control system in the environmental chamber

5.3.3.2 Representation of the Occupancy

Typical office work can be characterized as sitting or standing (light activity, typing, filing). Thus, metabolic rate of people would vary in the range of 1.2-1.5 met [176]. The maximum heat emitted by the human body in offices can be considered as 100 W [211]. To mimic the presence of occupants and their impact on the indoor air temperature a programmable electric heater is assembled as shown in Figure 5.10. A mini 100 W electric heater is used to represent the presence of one person in the chamber. This fan-assisted electric heater can heat up and cool down faster than water-based or oil-based systems. It makes the heater a more realistic device for mimicking the presence of occupants. A Raspberry Pi with a relay hat is programmed to turn ON/OFF the heater based on the occupancy data in an excel file.

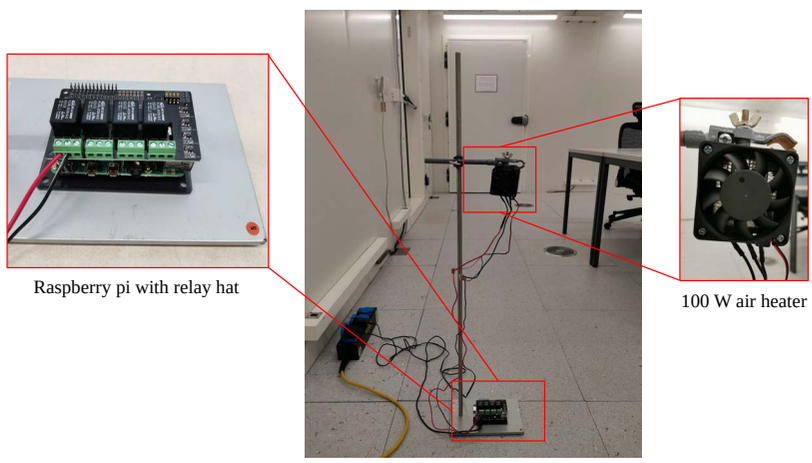


Figure 5.10: Electric heater to represent occupant heat gain

5.4 Results

The results of this study are categorized into the *Training phase*, *Simulation tests*, and *Experimental tests*.

5.4.1 Training Phase

The reward function combines the energy and comfort performance of the agent into a single value. Higher values of the reward indicate better performance of the agent. Lower values or high oscillations during the training phase may indicate that the agent has failed to achieve the optimal control policy. The evolution of the reward value during the training phase is a good indicator to assess whether the agent has converged to the optimal control policy. When evaluating the reward value, the minimum and maximum possible values should be taken into account. The values for the coefficients a and b in the reward function (Equations 5.4 and 5.5) are selected as 1 and 30, respectively. A high value is selected for the comfort reward to ensure that the energy saving is not at the cost of comfort violations. Therefore, the values higher than -30 are only due to the energy use, which is not totally avoidable. The values equal or lower than -30 indicate that the comfort is also violated. Figure 5.11 (a-c) shows the evolution of the total reward, the energy reward, and the comfort reward averaged over each month during the training phase. As seen in Figure 5.11(a), the total reward shows many fluctuations over the time. The evolution of the energy and the comfort rewards indicate that the variations of the total reward are mainly caused by the comfort reward, and the energy reward follows a converging trend. It should be noted that the occupancy data used for the training phase include the data of 20 different offices. Therefore, the agent should deal not only with the daily change of occupancy schedule in each office but also with the change of office (with different occupants). Considering the variations in occupancy data and office thermal characteristics included in the training phase, the agent is not expected to converge the reward value during the first training phase. rather, the aim of the first training phase is to gain a generalizable and transferable knowledge by experiencing different offices. A converging reward value is only expected over the testing steps, in which the agent is applied to one specific office.

5.4.2 Simulation Tests

Each of the three simulation tests includes a different office area (accordingly, a different thermal capacitance and heat losses) and occupancy data that was never observed by the agent before. To better understand and interpret the results of this section, 5 consecutive days of the occupancy data used in each simulation test are shown in Figure 5.12. In conventional schedule-based control methods domain experts assume the occupancy of

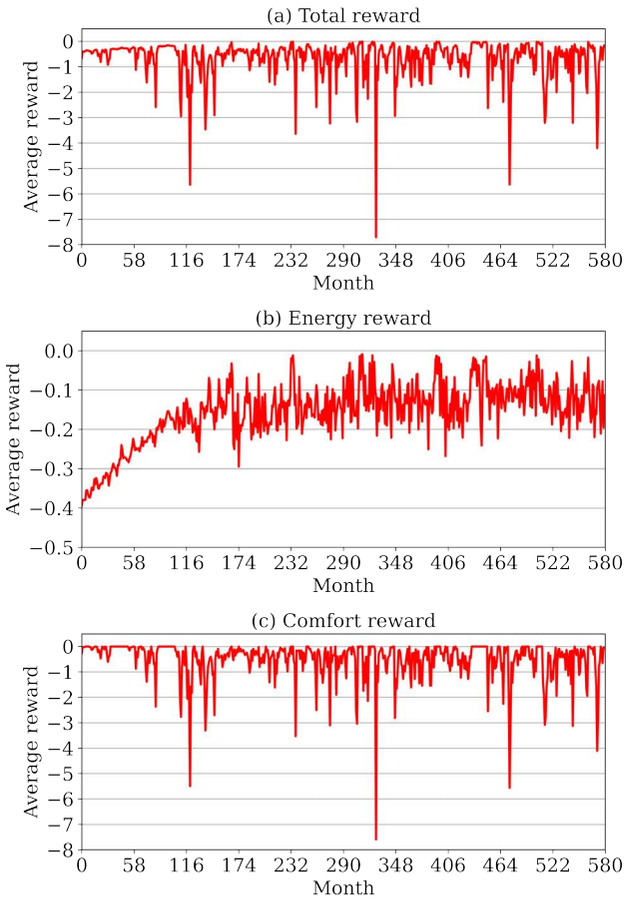


Figure 5.11: Evolution of total reward, energy reward and comfort reward over the training phase

offices are static and similar. However, observing the actual occupancy data indicates that the occupancy schedule can significantly vary from day to day and from office to office. In occupancy data used in *simulation test 2* the occupants usually leave the office quite early, while in that of *simulation test 3* they stay in the office till late afternoon. The occupancy data in *simulation test 2* shows more variations between different days than *simulation test 3*. These insights from the real-world occupancy measurements further highlight the uniqueness of occupant behavior in each office and the importance of integrating actual occupancy into office controls.

Figure 5.13 shows the daily average of the total reward during the three simulation tests. As mentioned, the first training phase included variations in occupancy data and office area between each month. But in this step, each simulation test is performed on a single office without changing the occupancy dataset or office thermal characteristics. This represents the implementation of pre-trained agent on a target office. Therefore, the reward function in this stage is expected to converge or be stable over the days. Figure

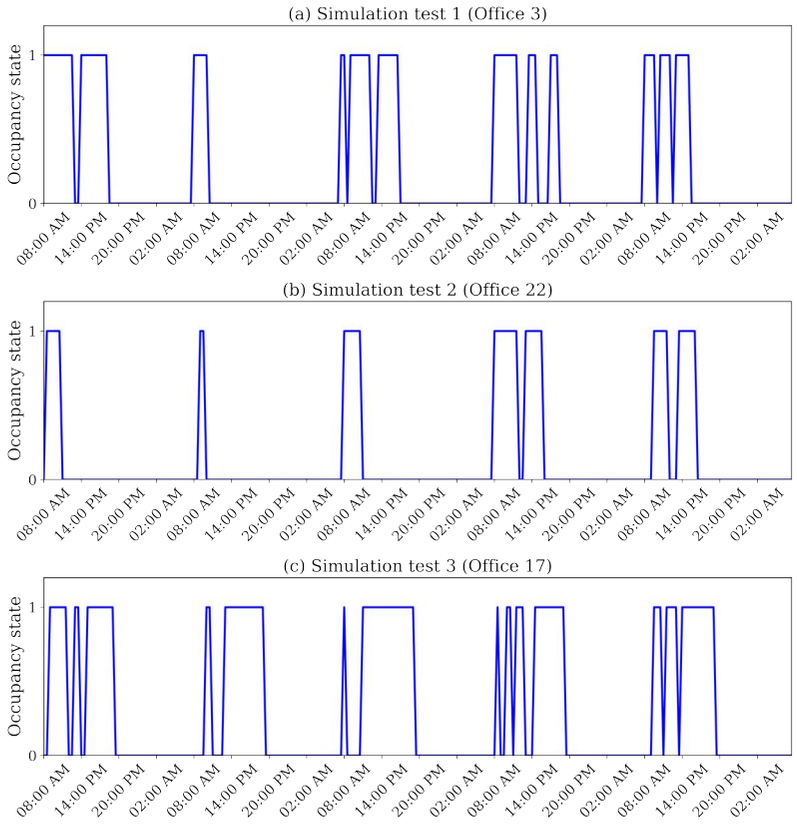


Figure 5.12: One week of occupancy data used in each simulation test

5.13 shows that the reward value in all 3 tests is almost stable, with only few violations of the comfort during the first 3 weeks. The reward value in all offices becomes stable over the last week, which indicates that the agent has adapted the control policy to that specific office.

To observe how the agent has adapted the heating schedule to the occupancy and how it differs from the baseline methods, Figure 5.14 shows office occupancy, indoor air temperature and the temperature setpoints by three control methods. This Figure is only focused on the third week of simulation test 1, where the RL exhibits the most number of comfort violations. The first hour in this Figure is the start of the week (Monday morning). The *Enable* action in RL activates the RB controller with a setpoint of 22, and the *Disable* action deactivates the RB controller and turns OFF the valve. Therefore, the setpoint during the *Enable* action is 22 ℃. For the visualization purpose the *Disable* action is visualized as a setpoint of 0 ℃. As can be seen in Figure 5.14(a), during the working days the agent tries to pre-heat the room from early morning (1 A.M.) to be able to reach to the comfort zone before the occupants arrive. Then it tries to maintain the comfort temperature until the occupants leave the office. It performs the heating in a very cautious manner by sequentially turning ON/OFF the valve. This shows that the agent is

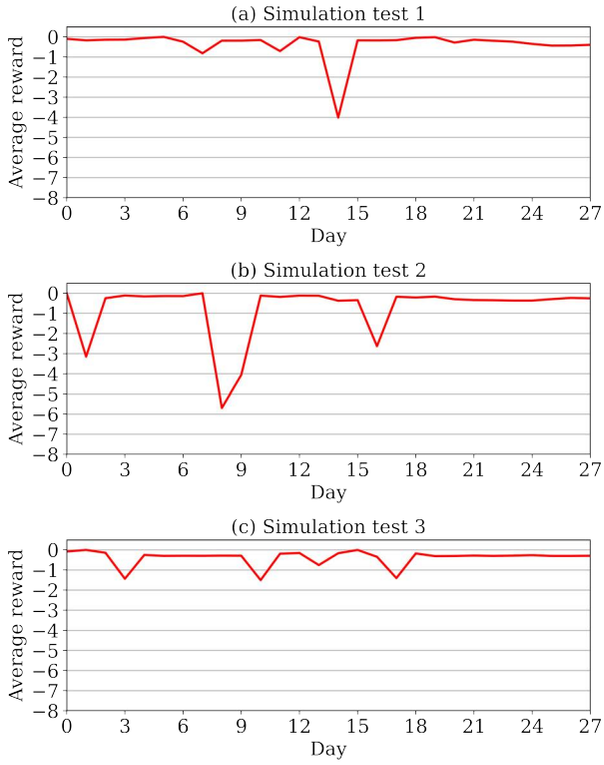


Figure 5.13: Evolution of the total reward over three simulation tests

trying to reach the comfort zone with minimum temperature increment and to stay only slightly above the minimum comfort limit to save energy. An interesting strategy that the agent follows to save energy is turning OFF the heat emission system a few hours before the occupants leave the office and taking advantage of the stored heat in the office for the remaining occupied hours. Most of the days the end of the heating schedule is adjusted very well, such that the air temperature drops below the comfort level only a short time after occupants leave the office. Properly adjusting the end of heating schedule highlights the benefit of learning occupancy behavior. Also, the results indicate that the agent can distinguish between the weekdays and weekends and does not try to reach the comfort level during the weekends. The air temperature at the begining of the week was very low, which indicates that the agent did not heat the office over the preceding weekend days. Consequently, although the agent starts heating in the early morning and continuously heats the office until the arrival of occupants the air temperature did not reach the comfort zone before their arrival. Interestingly, the agent learns from this mistake that the heating rate of the emission system in this office is quite low, and it preheats the office on Sunday to make sure it can reach the comfort level on Monday morning. This example shows that the RL controller becomes smarter over time, which is not the case for baseline methods.

Figure 5.14(b) shows the operation of the conventional SRB control, which is static

and detached from the actual occupancy. In common practice, usually the same heating schedule is set for all the offices in a multi-office building regardless of the fact that the occupancy in each office can be different. To represent this common practice, a constant heating schedule of 6 AM to 7 PM is set in this study without considering the actual occupancy data of each of the three test offices separately. In the office presented in Figure 5.14, the occupants arrive after 6 AM and leave before 7 PM. Therefore, the air temperature is within the comfort zone during occupancy hours. The inefficiency of the SRB method is further highlighted when the occupancy schedule unexpectedly changes on some days. An example of this case can be seen on the last working day in Figure 5.14(b), in which the occupants leave the office a few hours earlier than other working days. In this case, the controller keeps heating a vacant office that will be empty for the next two days. Another drawback of the SRB method is the need for maintaining the indoor temperature at 16 ℃ during the nights and weekends. The setback temperature is considered to make sure a fast temperature rise at the start of working days and if occupants arrive out of the scheduled hours and manually turn ON the system. The setback temperature is actually a conservative approach in which the vacant office is still heated for many hours due to being detahced from the actual occupancy. The RL agent, on the other hand, learns the occupant behavior and only heats the vacant office if it seems necessary based on the expected arrival of occupants. Figure 5.14(c) shows the performance of the ORB method. As expected, it can be seen that this method fails to maintain the occupants' comfort on their arrival. This shows that in the slow-response thermal systems a method for predicting the occupancy is necessary to properly preheat the office. If the occupancy prediction is done seperately, then the experts should properly calculate how advance should the system starts heating at each office. The advantage of the proposed RL model is that the agent learns when to preheat the office and no expert knowledge is needed.

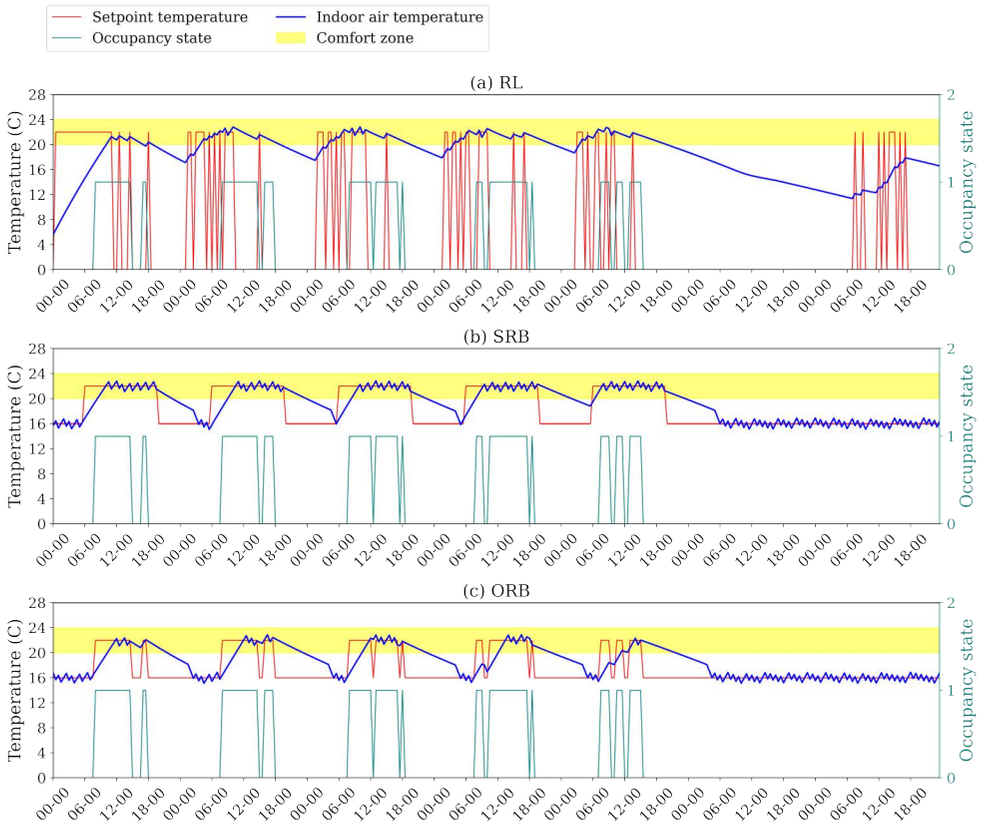


Figure 5.14: Temperature setpoint, indoor air temperature and occupancy during third week in simulation test 1

Comfort of occupants should be considered as the main constraint in developing energy-saving control methods in buildings. Control methods that save energy by violating comfort will result in user dissatisfaction and can not be widely implemented in buildings. This applies to all comfort-related services in buildings, such as lighting [58], space heating [154, 192], and hot water production [191]. For a detailed evaluation of the comfort aspect, Figure 5.15(a) shows the total number of timesteps when indoor air temperature is out of the comfort zone during each day of deployment. Figure 5.15(b) also shows the average temperature of these violations. The RL method has maintained the occupants' comfort very well, with a few violations that have happened in 3 days out of 28 days of testing. On 2 of these days the comfort violation has occurred only for one timestep (30 minutes) with an average temperature very close to 20 ℃, which might not be even noticed by occupants. Another significant day is a Monday, in which the comfort is violated for 6 consecutive timesteps (3 hours in total) with an average temperature of 18.6 ℃. As explained before, this has happened because the agent did not pre-heat the office during the weekend and started to heat the office from a quite low temperature in the early morning of Monday. The agent then learned from this mistake to pre-heat the office

on Sunday, and therefore the comfort is not violated over the next Monday. Although the SRB method is very conservative, there are still a few comfort violations over 4 days of the test period. This is because the occupants arrived a bit earlier than usual and the indoor air temperature has not yet reached the comfort zone. As expected, the ORB method has violated the comfort temperature for 1.5 hours to 4.5 hours after arrival of occupants.

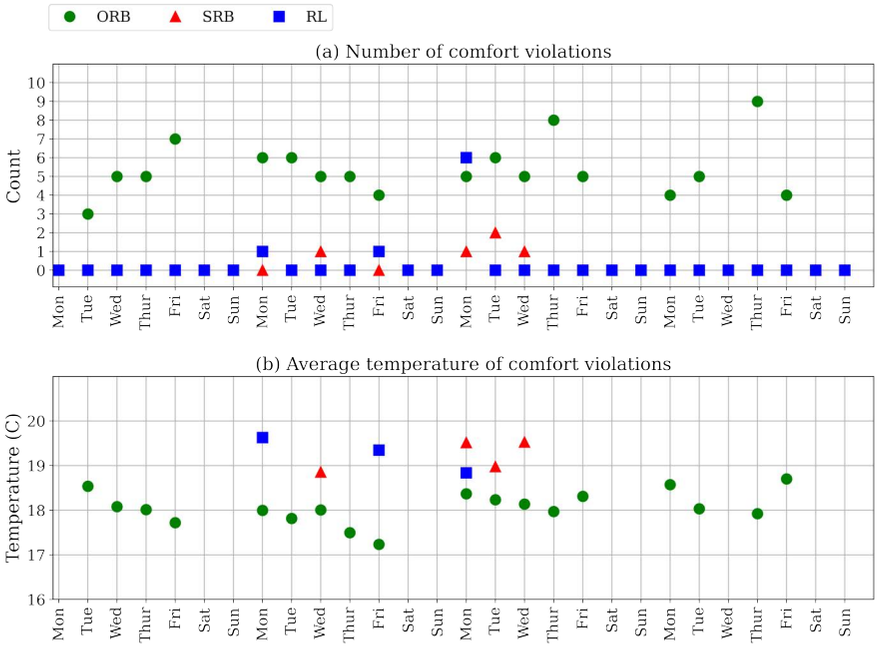


Figure 5.15: Comfort temperature violations by different methods during 28 days of deployment

A set of quantified metrics should be used to compare the control methods and to ensure that the proposed control framework outperforms the baseline methods before implementing on the experimental setup. The total number of timesteps with violated comfort and the average temperature of comfort violations are used to assess comfort aspect. The energy aspect is compared based on the total increment of the indoor air temperature. If the temperature of the building is increased during a timestep, the total heat received by the indoor air can be calculated by Equation 5.8.

$$\begin{cases} Q_{timestep=t} = m_{air} \times Cp_{air} \times (Tindoor_t - Tindoor_{t-1}) \quad \text{if} \quad Tindoor_t > Tindoor_{t-1} \\ Q_{timestep=t} = 0 \quad \text{if} \quad Tindoor_t \leq Tindoor_{t-1} \end{cases} \tag{5.8}$$

The heat loss from building is the same between different methods, and the only internal heat gain is from the heating system. Therefore, it can be assumed that the heating

system that has resulted in a higher total temperature increment has used a higher energy. The total temperature increments, as expressed by Equation 5.9, is the basis for energy use comparison between different methods. This proportion might not be the case in more complicated models with internal heat gains. The total temperature increment is used as a metric because this study is only focused on the zone level without considering the main heating system.

$$\Delta T_{total} = \sum_{t=0}^{n} (Tindoor_t - Tindoor_{t-1})[Tindoor_t > Tindoor_{t-1}] \tag{5.9}$$

Table 5.5 shows the comfort and energy metrics for different methods in three simulation tests. The RL method in all offices has a lower ΔT_{total} and, therefore, a lower energy use than the baseline methods. It shows that the agent has learned to minimize total temperature increment. Regarding the comfort aspect, in simulation tests 1 and 3 the RL method provides the same level of comfort as the SRB method. In both cases, the comfort is violated only in a few timesteps with an average temperature higher than 19 ℃. In the case of simulation test 2, the number of timesteps with a violated comfort in case of the RL method is almost half of the SRB method. However, the average temperature of violations by the RL method is 4℃ lower than the SRB method. This is mainly due to the fact that the occupancy profile used in simulation test 2, as shown in Figure 5.12, is sparse and different from what the RL agent has mostly observed during the training phase. Overall, the comparison of energy and comfort metrics over three different offices indicates that the proposed RL method outperforms the baseline models and can be transferred to the experimental test step for evaluation on a real system. It should be noted that the superior performance of the RL framework is achieved only within a short time of one-month training on the target office. Considering the learning and adaptation potential of the RL agent, it is expected that after a few additional months of gaining experience on the target offices the RL framework will show even better performance and increasingly get better than the baseline models.

5.4.3 Experimental Tests

The results of the simulation tests demonstrated the ability of the agent to adapt to different real-world occupancy schedules. As the following step, the experimental tests are supposed to prove the adaptation potential to the thermal response time of the office. The flexible environmental chamber makes it possible to represent different thermal response times by using different heat emission systems and applying different air change rates. The occupancy data of office 3 is used in all experiments. Figure 5.16 shows the control actions and the indoor air temperature versus occupancy status in the experimental

Table 5.5: Performance metrics of different control methods on three simulation tests over the entire test period (4 weeks)

		RL	SRB	ORB
Simulation test 1	Number of comfort violations	8	5	97
	Average temperature of comfort violations (℃)	19	19.1	18
	ΔT_{total}	248.5	445	426.5
		RL	SRB	ORB
Simulation test 2	Number of comfort violations	22	40	114
	Average of comfort violations (℃)	13.3	17.5	17.8
	ΔT_{total}	270.9	370.7	415.3
		RL	SRB	ORB
Simulation test 3	Number of comfort violations	3	12	161
	Average of comfort violations (℃)	19.6	19.4	17.9
	ΔT_{total}	300.6	307.7	312.7

tests.

Over experimental tests 1 to 3, the air change rate is increased and the heat emission system is changed from the higher heating rate (combination of ceiling and floor panels) to the lower heating rate (ceiling panels). This imposes very different thermal response behaviors as in experimental test 1 the office is heated quickly and reduces in temperature slowly. On the other hand, in experimental test 3 the office is heated slowly and reduces in temperature quickly. Comparison between Figures 5.16(a)-5.16(c) indicates that the agent has always maintained occupants' comfort by properly scheduling the heating cycles despite of the significant difference between thermal response time. Similar to the simulation tests, the agent (i) starts to preheat the office before the occupants arrive to ensure comfort and (ii) maintains the temperature as close as possible to the lower comfort limit and stops heating the office a few hours before the occupants' departure to reduce energy use. The agent can maintain thermal comfort under different conditions, which shows that the training phase has provided generalizable knowledge to the agent and it can adapt to different offices quickly. These results firmly indicate that the proposed framework can be implemented in different offices. Although the agent is trained on the offices with 1 or 2 occupants, considering its learning potential it is expected that it can quickly adapt to the offices with more occupants. In such a case, a higher number of sensors would be needed to capture the occupancy and a longer time might be needed

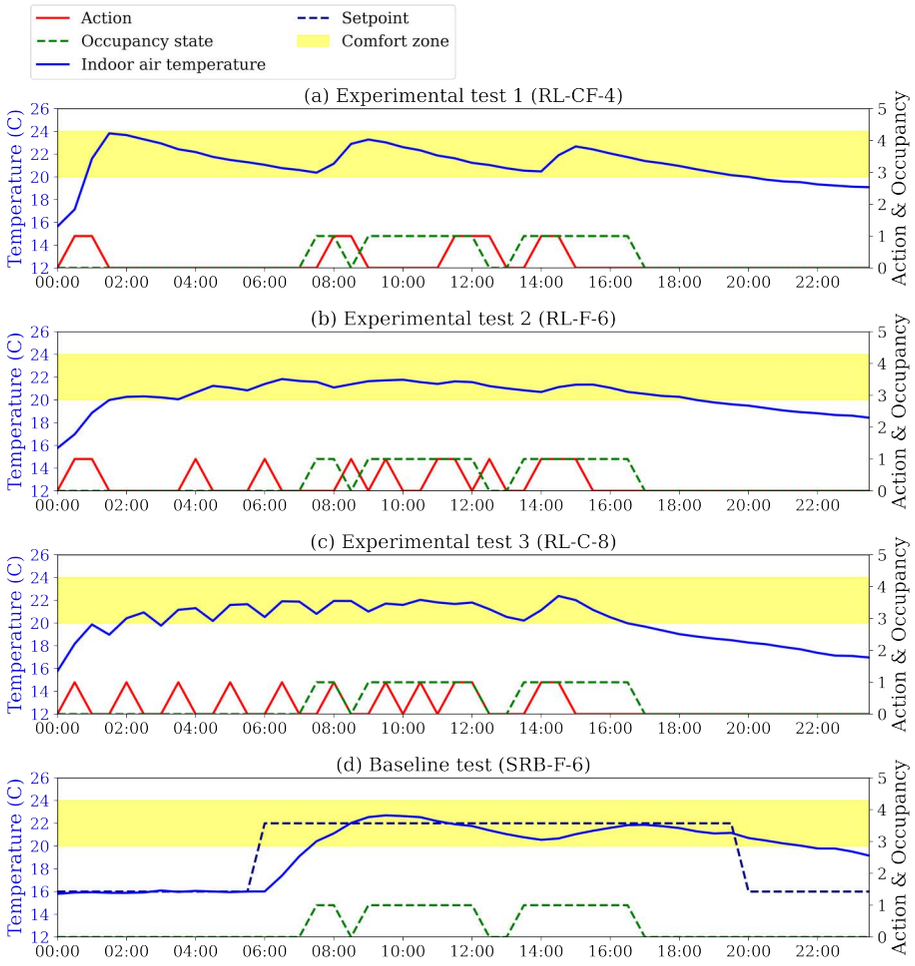


Figure 5.16: Results of experimental tests: variations of control actions and indoor air temperature versus occupancy in 4 tests

to learn the occupancy by multiple people. Figure 5.16(d) shows the performance of the SRB method to provide a better insight into the differences between the proposed method and common practice. The baseline test can be only compared with the experimental test 2 as they are done on a similar setup (floor heating at 6 ACH). In both cases the control system heats the vacant office for few hours. But with a difference that the RL method heats the office before the occupants' arrival to ensure their comfort, while the SRB method heats the vacant office after the occupants' departure. So the heating of the vacant office by the RL method is more rational.

Table 5.6 shows the performance metrics of the experimental tests. Comparison of experimental test 1 with the baseline model shows that the RL model provided the same level of comfort as the baseline model with slightly lower temperature increment. Although the difference is very small and cannot be considered as a superior performance.

As mentioned before, only one day of experimental test cannot be enough for comparison of energy aspects. Furthermore, the RL agent has successfully maintained occupant comfort. The results indicate that the proposed control framework can be transferred to many offices with any type of hydronic heating system to save energy while maintaining the occupants' comfort.

Table 5.6: Performance metrics of experimental tests

Parameter	Experimental test 1 (RL-CF-4)	Experimental test 2 (RL-F-6)	Experimental test 3 (RL-C-8)	Baseline test (SRB-F-6)
ΔT_{total}	13.3	8.3	15	8.6
Number of comfort violations	0	0	0	0

It should be noted that this study only focused on the zone-level control decisions. In practice, the system-level controller should be in harmony with the zone-level controller. For example, the temperature and flow rate of supply water should be selected in accordance with the zone-level actions.

5.5 Conclusion

The following conclusions can be drawn from this study:

- Simulation results indicate that the proposed RL framework can provide a significant energy saving (up to 44% lower total temperature increment) while providing the same or better comfort level compared to the conventional schedule-driven control method (SRB);
- Experimental tests in the environmental chamber by imposing different thermal response behavior indicate that the agent can quickly adapt to different offices and maintain occupants' comfort. While the agent is only trained on a simplified simulation model, the good performance in experiments indicates that it can quickly adapt to an actual system with a higher complexity;
- To maintain thermal comfort of occupants, the agent pre-heats the office before the occupants' arrival. In order to save energy it keeps the temperature as close as possible to the lower comfort limit and stops heating the office a few hours before occupants departure. These strategies indicate that the agent has properly learned the occupancy behavior;

- A simple rule-based occupant-centric control method was simulated that turns ON the heating system immediately when occupancy is detected. Results indicate that this method provides a poor level of comfort, that proves the necessity of using a prediction-based algorithm for the slow-response heating systems;

Theoretical and experimental results prove the generalization and adaptability of the proposed framework and indicate that it can be applied to many offices. Further research steps can be done to facilitate the widespread implementation of *DeepValve* control framework in offices. The dataset, Python codes, and trained agent can be shared with researchers interested in taking next steps.

Chapter 6

Overall conclusions and future outlook

This section discusses:

Overall conclusions from the proposed frameworks
Challenges of real-world implementation
Limitations of this study
Potential future steps

6.1 Overall conclusions

Expert-based building controls rely on the hard-coded knowledge of the experts. They are effective in dealing with problems that can be mathematically described and programmed to the controller by domain experts. Thus, they are limited to expert knowledge. Occupant behavior is a highly stochastic and complex phenomenon and is unique in each building, which makes it hard to deal with for the experts [2]. In programming conventional expert-based controls, occupant behavior is either ignored or over-simplified, which has resulted in a gap between *what is provided by building systems* and *what is actually needed by occupants*. The overall motivation of this study was to *investigate the Learning-based building controls that can perceive, learn, and adapt to the occupant behavior to save energy*. Reinforcement Learning is used to provide the ability of learning and adaptiveness to the controller. Overall, this study contributes to the existing knowledge of occupant-centric controls by:

- Proposing three novel occupant-centric Reinforcement Learning control frameworks that are formulated and trained with the specific consideration about robustness, transferability, fast convergence, and minimum use of sensors and actuators;

- Integration of occupant health (water hygiene) into the commonly used pillars of energy and comfort;
- Integration of Reinforcement Learning with static rule-based method to increase learning speed and robustness;
- Evaluation of intensive virtual trainings (using stochastic or real-world data) to provide a generalizable knowledge and warm-start the agent before controlling the target building;
- Simulation investigations using real-world data;
- Experimental investigations in an environmental chamber;

More specific contributions of each framework are presented in each chapter. Table 6.1 provides a summary of three proposed frameworks.

The following presents some main remarks from this study that can help the future research on Reinforcement Learning for occupant-centric controls.

How to efficiently integrate occupant behavior into a Reinforcement Learning control framework?

This thesis work focused on two forms of occupant behavior, namely hot water use behavior (*DeepHot* and *DeepSolar*), and occupancy behavior (*DeepValve*). In Reinforcement Learning, the agent observes the current condition of the environment through *states*. Similar to the other aspects of the environment such as indoor and outdoor air temperature, occupant behavior can also be monitored through sensors, and collected data can be used to form *occupant behavior-related states*. To follow Markov Decision Process in the Reinforcement Learning framework, the state vector should provide enough information for the agent to decide the next action [212]. As listed in the introduction section, occupant behavior can be influenced by many different parameters (environmental-related, time-related, individual, social, and random parameters). To enable the agent to implicitly predict the future occupant behavior and accordingly plan the next control actions, many of these influencing parameters can be monitored and included in the state vector. A preliminary study by the authors on the prediction of hot water use indicated that the future hot water use is strongly correlated with the previous hot water use, and therefore a vector including historical usage forms a very useful feature for predicting the future usage [13]. Based on this experience, occupant behavior-related state in the Reinforcement Learning framework was formed as a history of previous occupant behavior (hot water use behavior or occupancy behavior). The results indicate

Table 6.1: Summary of main results

	DeepHot	***DeepSolar***	***DeepValve***
System to be controlled	With-tank heat pump water heating system	Solar-assisted heat pump for space heating and hot water production	Zone-level heating of offices using Hydronic heat emission system
Case study building	1 residential building in Switzerland	3 residential buildings in Switzerland	Simulated offices + Environmental chamber
Objectives	Energy use, comfort, hygiene	Energy use, comfort, hygiene	Energy use, comfort
Using real-world data of occupant behavior?	Yes	Yes	Yes
Experimental deployment?	No	No	Yes
Baseline control methods	• Rule-based with conventional setpoints	• Rule-based with conventional setpoints • Rule-based with energy-saving setpoints	• Rule-based with a conventional schedule • Rule-based with a schedule synchronized to occupancy
Energy saving	• 24%	• 22% to 47% compared to Rule-based with conventional setpoints • 7% to 22% energy-saving compared to Rule-based with energy-saving setpoints	• Up to to 44% reduction in total temperature increment

that the agent could properly anticipate and adapt to the occupant behavior. This shows that a low-cost and effective approach for integrating occupant behavior is to include a history of behavior in the state vector. It limits the number of sensors, reduces initial cost, and reduces the risk of failure due to sensor malfunctioning.

The Reinforcement Learning agent has a learning process, including exploration of random actions, that can impose the risk of violating comfort and health of occupants. How to overcome this risk?

To minimize the risk of violating comfort and health during the learning process, this study proposes the following solutions to take into account at the same time:

1. **Pre-training and exploration in the safe simulation environment:** Nowadays, there are several simulation tools available for the energy modeling of buildings. The simulation environment provides a safe environment for the agent to learn initial knowledge and to try random actions (exploration phase). As the first solution, this study proposes to integrate realistic occupant behavior into the simulation environment and pre-train the agent in the simulation environment for a certain duration to reach a stable performance. To include a realistic occupant behavior, in *DeepHot* and *DeepSolar* frameworks, a statistical model of hot water use behavior was integrated to generate hot water use profiles, and in *DeepValve* framework, real-world measured data was used. The exploration phase, when the agent is allowed to perform random actions to better explore the environment, is only allowed during the training phase in simulation. This study demonstrated that the agent pre-trained for a long-time using realistic occupant behavior showed a good performance at the very beginning of the deployment phase on the target house.

2. **Enabling manual interaction:** For implementing the proposed frameworks on real buildings, this study proposes to avoid full automation and enable the possibility of manual interaction for the occupants. For example, if the agent has decided to turn OFF the heat pump, but the occupants need hot water, they should be able to deactivate the agent and manually turn ON the heat pump.

How to ensure the transferability potential to many buildings?

To ensure the proposed frameworks can be easily transferred to many different buildings, this study took the following considerations in all the frameworks:

- **Model-free:** While the integration of a model reduces the learning time and increases the robustness, on the other hand, it will reduce the transferability

potential to the other buildings (similar to the Model Predictive Control) since the model of the system should be updated according to the target building. All the frameworks in this study are model-free to enable an easy transfer to other buildings.

- **Variations in pre-training phase:** To provide a generalizable knowledge to the agent, the pre-training phase includes variations in occupant behavior, system sizes, building area, geographical location, etc. This will prevent the agent from over-fitting to a specific case, and increase the adaptation potential to different buildings.
- **Relying on the minimum number of sensors and actuators:** This will reduce the dependency on the system layout and facilitates implementation in several houses.

Should the agent be continuously trained during the system lifetime?

The central part of the proposed frameworks is occupant behavior, which is characterized by temporal variations and unpredictability. The combination, number, preferences and behavior of occupants can change over time, which can significantly vary their requirements in buildings. The static deployment (deployment without training) indicated a good performance in this study. However, to harness the full adaptation potential of Reinforcement Learning and ensure an optimal operation, the agent should be continuously trained when controlling the system. It might not be economic in some cases since it requires higher computational power. In that case, the pre-training phase should be further enriched by including more occupant behavior-related data. Then the trained agent can be deployed on low-cost hardware such as Raspberry Pi. As a possible improvement, the low-cost hardware can also collect the occupant behavior-related data. Then the agent can be periodically trained in the simulation environment with the new data (for example every 6 months), and the updated agent then be deployed back to the low-cost hardware. It will reduce the cost of hardware but increases the labor work. This process might be automated using a combination of cloud computation and local low-cost hardware.

Can the proposed controls be retrofitted to the existing controls?

The *DeepHot* and *DeepSolar* frameworks are designed to be replaced with the existing controls. But the *DeepValve* can be retrofitted to the end-use level (heat emission unit) of a hydronic heating system to work in parallel with the existing control. If there is a separate supply valve to the zone that is controlled by the main controller, that should be replaced by the *DeepValve* control framework. Otherwise, a valve can be installed on the

supply pipe to the zone and controlled by *DeepValve* to regulate heat emission based on occupancy.

6.2 Challenges for implementing occupant-centric Reinforcement Learning controllers in buildings

While a major contribution of this study has been to demonstrate the potential of Reinforcement Learning for integrating occupant behavior in building controls, implementing RL in real-world settings can be hampered by a few challenges. This section discusses some of these challenges to be considered in real-world implementations.

- **Interpretability of the control policy by the building managers:** A major challenge to the acceptability of occupant-centric Reinforcement Learning controllers is the fact that domain experts and building managers prefer to be able to interpret or to understand the decisions made by the controller. This is while Reinforcement Learning, similar to the other artificial intelligence algorithms, has a black-box nature that is hard to interpret. Advances in explainable artificial intelligence are needed to develop algorithms that are easier to be explained, even with the cost of sub-optimal performance [213].

- **Large state/action space:** In a large building, the control of the energy system might include several pumps, valves, dampers, and plenty of sensors. This will increase the dimension of state and action space which, in turn, significantly increases the complexity of the control problem for the agent. A possible solution is to use multiple agents, that either collaboratively or independently control a sub-system of the problem.

- **Sensitive system constraints that should never be violated:** An agent can perform actions that are non-optimal, for example, due to the sudden changes in the system or occupant behavior. It would be challenging if there are some safety constraints in the system. By including the safety constraints in the reward function and pre-training the agent, the probability of safety violations would be much less. But if the safety aspects are very critical, some supervisory rules can be integrated into the Reinforcement Learning algorithms that post-process the agent actions and revise the risky ones.

- **Sensor failure or malfunctioning:** The agent observes the state of the environment through sensors. In case one of the sensors fails or sends a wrong signal, it can provide wrong information resulting in a non-optimal action. To deal with this

issue, a set of rules can be used to check if the sensor data is correct, for example by comparing the data with the expected range, and raise an alarm if a sensor is malfunctioning. In addition, as taken into account in this study, the RL framework could be designed to rely on minimum number of sensors and actuators.

- **Including multiple objectives:** An occupant-centric control framework can include multiple objectives in addition to energy saving, such as comfort, health, and productivity of occupants. Adjustment of weighting factors in the objective function was found to affect the performance significantly. It is, however, very time-consuming as it requires multiple runs in the simulation environment. Proper adjustment of the weights in a reasonable time might require powerful computation hardware. Automated hyper-parameter optimization methods can be used to adjust the weighting factors properly.

- **Latency in actions:** Reinforcement Learning algorithms, especially if trained continuously, include much more computations than the conventional simple rule-based methods. In addition, they might rely on IoT sensors and data transfer from the cloud. These aspects may result in a delay in executing control action after acquiring the states. To take into account this latency, the timesteps of the control framework should be designed accordingly (e.g., taking actions every 30 minutes). But it would raise a challenge if the timestep cannot be longer than the controller latency. In this case, it might be required to use more powerful computational hardware or to change the hardware layout to reduce the dependency on online data transfer.

6.3 Limitations

With the contributions made, this thesis also have several limitations that are mentioned in this section.

- **Field implementation on residential houses:** The *DeepValve* control framework was experimentally implemented in a real-world setup to show that it can deal with the higher complexity of a real system. However, *DeepHot* and *DeepSolar* frameworks require to be tested on a residential house with real occupants, which is very challenging unless there is an adjustable and programmable setup like NEST building (Empa, Switzerland). So the current study collected real-world occupant behavior with a non-intrusive approach and integrated the data into the simulation environment. With the satisfactory results of this study, future research can be focused on implementing the proposed frameworks in a real house with occupants.

- **Data collection during COVID-19 pandemic:** Data collection from residential houses used for *DeepHot* and *DeepSolar* control frameworks was during the COVID-19 pandemic when all the occupants were mostly working from home. Therefore, the hot water use behavior has been different from the normal (pre-pandemic) period. In addition, for the *DeepSolar* framework, a constant occupancy profile had to be considered because all the occupants have been working from home, and therefore it was assumed that the indoor air temperature should be always within the comfort zone. The *DeepSolar* framework can be further improved by including the occupancy profile, and an additional action of turning OFF the space heating system.

- **Occupancy detection method for *DeepValve* framework:** In the development of the *DeepValve* control framework, it was assumed that there is an occupancy detection method that detects the occupancy with high accuracy. But the method was not discussed as it was out of the research scope. However, the accuracy of the occupancy detection method can also affect the performance of the *DeepValve* control framework. A future study can implement the *DeepValve* control framework with an actual occupancy detection method and evaluate the performance of the whole setup in a real office.

- **Evaluation on a short period:** Given the limited period of available data, all the frameworks have been evaluated for the periods of a few months. However, the climatic conditions and occupant behavior can significantly change over a long period, for example, by the change of season. If the required data are available for a longer period, the frameworks can be evaluated over the long-time to evaluate the adaptiveness to more significant and sudden changes of occupant behavior.

- **Lack of Legionella growth model in the pipe :** Legionella growth is only modeled for the hot water tank, assuming that the length of piping in a residential house is not significant and the pipes will be periodically disinfected manually. Integration of a Legionella growth model in the pipe can further ensure the health of occupants.

- **Investigation on a limited number of houses:** All the frameworks are tested on a few case studies (*DeepHot* on 1 residential house, *DeepSolar* on 3 residential houses, and *DeepValve* on 3 offices and the environmental chamber). Considering the potential variations in occupant behavior, especially between residential houses, these frameworks should be evaluated on a more number of cases with different lifestyles and climatic conditions. However, the cost of implementing IoT monitoring system, and the limited duration of this PhD study, limited the number of possible cases.

- **Evaluation of different sensing layouts:** The sensing approach in *DeepHot* and *DeepSolar* are not exhaustive. There can be alternative state designs (and thus sensing approaches) that could yield different energy saving. For example, the temperature of cold water incoming to the water tank, or temperature of hot water leaving the tank can be included in the state vector.

6.4 Future outlook

Based on the contribution done in this thesis, several research gaps and potential further research topics are discovered. The following list provides a guidance for conducting future studies on this topic:

- **Domain knowledge-assisted Reinforcement Learning:** The proposed frameworks in this study do not rely on domain (expert) knowledge and learn the control strategy from scratch. Future research can be done to evaluate how the domain knowledge can be incorporated into the control framework to increase robustness and data efficiency without limiting the adaptiveness and generalizability potential. Most of the available domain knowledge in building controls is in the form of "*if,then*" rules. Thus, a possible approach to integrate domain knowledge is to develop set of rules based on domain knowledge that provides the expert suggestion to the agent. Based on the suggested action and the current state, the agent then decides the next action. This architecture ensures adaptability since the agent is free to either follow the expert suggestion or to take another action. It also ensures transferability since the expert knowledge is not case-specific. In addition, some pre-checking rules can be used to evaluate current conditions before the agent takes decision, and take an action if the decision of optimal action is obvious. For example, if the occupants are present, and the temperature is too low, the optimal action is obviously turning ON the heating system. These two approaches are expected to speed-up the learning of agent and robustenss of control framework.

- **Integration of a pre-processing method to compress state vector:** To enable the agent to perceive and learn occupant behavior, a history of occupant behavior can be used in the state vector. This will, in turn, increase the dimensionality of the state vector and the learning time for the agent. A possible future research topic is to develop a method that can compress the state vector into less number of parameters while providing the required information to the agent. For example, an auto-encoder network can be used to compress the state parameters into less number of hidden states.

- **Integration of occupancy detection to *DeepHot* and *DeepSolar*:** The history of building occupancy can help the agent to better anticipate future hot water use behavior and accordingly schedule the heating cycles. A low-cost and transferrable occupancy detection method can be integrated into *DeepHot* and *DeepSolar* control frameworks to improve the agent performance without limiting the transferability of the frameworks.

- **Eliminating the necessity of continuous learning for *DeepHot* framework:** Domain experts and current producers of hot water systems can accept a new controller easier if it follows a plug-and-play approach. Since the *DeepHot* control framework is only focused on the hot water systems, it has a good potential to eliminate the need for online training and turn it into a plug-and-play controller. To this aim, the future research can design a very intensive offline training session, that includes variations of water heating system (heat pump, boiler, electric heater, etc.), variations and changes in occupant behavior (a sudden change of occupant behavior, short-term absence, etc.), different climatic locations, and other aspects, to provide a generalizable knowledge to the agent. A backup controller can be also integrated to further assist the acceptability of the occupant-centric controller. After the intensive training, the agent can be deployed on a low-cost hardware to control the hot water system. A local occupant-centric controller (e.g. on a Raspberry pi) is easier to be implemented on the current hot water systems than a cloud-based solution.

- **Field implementation of the *DeepHot* and *DeepValve* control frameworks:** The *DeepHot* and *DeepValve* control frameworks have less complexity and thus are easier to be implemented in actual buildings. Considering the limited field studies on Reinforcement Learning, potential future research is to experimentally implement these frameworks and evaluate their performance.

- **Design and integration of the backup controller:** A backup controller seems to be very important for increasing the robustness of the Reinforcement Learning frameworks. So possible future research is *how to design the backup controller and how to integrate it to the proposed Reinforcement Learning frameworks*.

- **Experts learning from agents:** With proper setup and enough training, the agent can surpass expert knowledge and come up with the solutions that experts are not aware of them. This provides an opportunity for experts, to train the agent with enough data, and then try to convert the agent policy into rules and heuristics that can be used in expert-based controls.

- **Integration of more sophisticated Legionella growth models:** This study, for the first time, integrated a Legionella growth model into the Reinforcement Learning

control framework for hot water systems. This study incorporated a simple model that is only dependent on the water temperature. But more accurate models using other parameters, such as PH of the water, can be also used to further improve the safety and health aspects of the framework.

- **Addressing the privacy aspects:** The proposed frameworks rely on occupant behavior-related data, which always raises privacy concerns. This study was performed after obtaining ethical approval from EPFL-HREC. But the privacy issues can be a barrier to the widespread adoption of the proposed frameworks. A multi-disciplinary study by computer scientists and sociologists can focus on how to improve the privacy and acceptability of the proposed occupant-centric control frameworks, especially for *DeepHot* and *DeepSolar* frameworks that are designed for the residential houses.

- **Automation of hyper-parameter adjustment:** Hyper-parameters found to significantly change the performance of the proposed frameworks. Similar to the Supervised Learning, there is a need to develop automated hyper-parameter optimization methods for Reinforcement Learning.

- **Alternative algorithms:** The focus of this study was more on the application aspect of RL than the algorithm. But with the advances in Reinforcement Learning, alternative novel algorithms can be used in the proposed frameworks to increase stability and data efficiency.

Considering the characteristics of occupant behavior, Reinforcement Learning found to be a promising method for integration of occupant behavior into building controls. Further theoretical and experimental studies can unlock the potentials of Reinforcement Learning and facilitate wide-spread implementation in future buildings.

Chapter 7

Achievements

7.1 Publications

The framework of this Ph.D. thesis allowed for contributing to the current knowledge in literature through a series of journal and international conference papers, including:

- 4 published first author papers in Q1 journals;
- 1 submitted first author paper to a Q1 journal;
- 3 published first author papers in international conferences;

Few more papers were published during the Ph.D. study. But they are not included in the dissertation since their topic have not been directly related to the Ph.D. thesis . Following is a list of publications categorized by the topic.

Initial exploration

Journal papers

Solar Energy	Amirreza Heidari, Dolaana Khovalyg	2020	*Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition*
Sustainable Cities and Society	Amirreza Heidari, Nils Olsen, Paul Mermod, Alexandre Alahi, Dolaana Khovalyg	2021	*Adaptive hot water production based on Supervised Learning*

Conference papers

BEHAVE 2021	Amirreza Heidari, Verena Marie Barthelmes, Dolaana Khovalyg	2021	*Probabilistic Machine Learning for Occupancy Prediction based on Sensor Fusion*
CLIMA 2022	Caroline Risoud, Amirreza Heidari, Dolanaa Khovalyg	2022	*Customized Neural Network training to predict the highly imbalanced data of domestic hot water usage*

DeepHot

Journal papers

Applied Energy	Amirreza Heidari, François Maréchal, Dolaana Khovalyg	2022	*An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach*

Conference papers

CISBAT 2021	Amirreza Heidari, François Maréchal, Dolaana Khovalyg	2021	*An adaptive control framework based on Reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems*

DeepSolar

Journal papers

Applied Energy	Amirreza Heidari, François Maréchal, Dolaana Khovalyg	2022	*Reinforcement Learning for proactive operation of residential energy systems by learning stochastic occupant behavior and fluctuating solar energy: Balancing comfort, hygiene and energy use*

Conference papers

CLIMA 2022	Amirreza Heidari, François Maréchal, Dolaana Khovalyg	2022	*Reinforcement learning for occupant-centric operation of residential energy system: Evaluating the adaptation potential to the unusual occupants ´ behavior during COVID-19 pandemic*

DeepValve

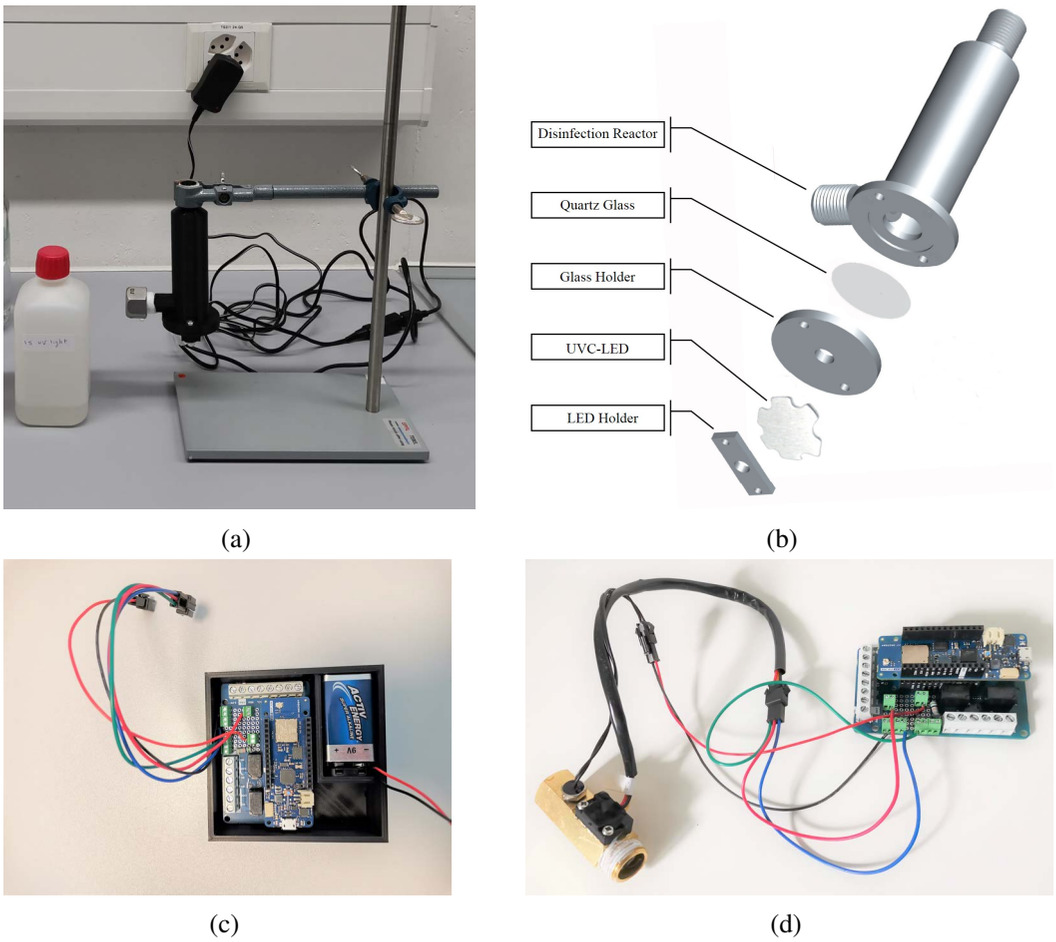
Journal papers

Submitted	Amirreza Heidari, Dolaana Khovalyg	2022	*DeepValve: Development and experimental assessment of a system-independent Reinforcement Learning control framework for occupant-centric space heating in offices*

7.2 Awards

During this Ph.D. study, we realized that Legionella bacteria is an increasing issue in the hot water systems in Switzerland. Legionella is a bacteria that can grow in a water between 20 ℃- 45 ℃ and can be transferred to the humans by breathing in the contaminated water droplets. Constantly increasing number of infections in Switzerland shows that there is a lack of knowledge and technological solutions in practice. Legionella can grow in the stagnant water. Therefore, the risk of Legionella is dependent on the hot water use behavior (frequency of usage) and the temperature variations. Parallel to this Ph.D. work, we developed the prototype of an IoT solution that can (1) predict the Legionella risk based on previous hot water use and temperature variations, (2) communicate the risk to the occupants in advance, and (3) disinfect the water at the point of use (e.g., shower, faucet) using a developed compact disinfection reactor based on UVC-LED. Figure 7.1 shows some pictures of the prototype development and test. The device include a water flow and temperature sensor, an Arduino to predict and communicate the risk and a reactor with UVC-LED to disinfect the water. This prototype has awarded two following grants:

- *Student Incubator Grant*- Smart Living Lab, Baloise Insurance- 2020
- *ENAC Innovation Seed Grants*-EPFL- 2021



(a) (b)

(c) (d)

Figure 7.1: Prototyping of IoT product for prediction and elimination of Legionella risk (a) Testing disinfection reactor (b) Parts of disinfection reactor (c) Arduino-based IoT hardware (d) Hardware connected to water flow and temperature sensor

Bibliography

1. Kim, S., Song, Y., Sung, Y. & Seo, D. Development of a consecutive occupancy estimation framework for improving the energy demand prediction performance of building energy modeling tools. *Energies* **12,** 433 (2019).
2. Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I. & Feng, X. IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings* **156,** 258–270 (2017).
3. Balvedi, B. F., Ghisi, E. & Lamberts, R. A review of occupant behaviour in residential buildings. *Energy and Buildings* **174,** 495–505 (2018).
4. Janda, K. B. Buildings don't use energy: people do. *Architectural science review* **54,** 15–22 (2011).
5. Barthelmes, V. M. Impact of Occupant Behaviour (OB) on building energy use and thermal comfort.
6. Hong, T. & Lin, H.-W. *Occupant behavior: impact on energy use of private offices* techreport (Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2013).
7. Andersen, R. V., Olesen, B. & Toftum, J. *Simulation of the effects of occupant behaviour on indoor climate and energy consumption* **in** *Proceedings of Clima* **2007** (2007), 9th.
8. Chen, J. & Taylor, J. E. Layering residential peer networks and geospatial building networks to model change in energy saving behaviors. *Energy and Buildings* **58,** 151–162 (2013).
9. Mahdavi, A. *The human dimension of building performance simulation. driving better design through simulation* **in** *Proceedings of the 12th Conference of The International Building Performance Simulation Association, Sydney, Australia, eds V. Soebarto, H. Bennetts, P. Bannister, PC Thomas, and D. Leach, K16–K33* (2011).

10. Eguaras-Martınez, M., Vidaurre-Arbizu, M. & Martın-Gómez, C. Simulation and evaluation of building information modeling in a real pilot site. *Applied Energy* **114,** 475–484 (2014).

11. Clevenger, C. M. & Haymaker, J. *The impact of the building occupant on energy modeling simulations* **in** *Joint International Conference on Computing and Decision Making in Civil and Building Engineering, Montreal, Canada* (2006), 1–10.

12. Yohanis, Y. G., Mondol, J. D., Wright, A. & Norton, B. Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and buildings* **40,** 1053–1059 (2008).

13. Heidari, A., Olsen, N., Mermod, P., Alahi, A. & Khovalyg, D. Adaptive hot water production based on Supervised Learning. *Sustainable Cities and Society* **66,** 102625 (2021).

14. Booysen, M., Engelbrecht, J., Ritchie, M., Apperley, M. & Cloete, A. How much energy can optimal control of domestic water heating save? *Energy for Sustainable Development* **51,** 73–85 (2019).

15. Guerra Santin, O. Occupant behaviour in energy efficient dwellings: evidence of a rebound effect. *Journal of Housing and the Built Environment* **28,** 311–327 (2013).

16. Gaetani, I., Hoes, P.-J. & Hensen, J. L. Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy. *Energy and Buildings* **121,** 188–204 (2016).

17. Jeong, B., Kim, J. & de Dear, R. Creating household occupancy and energy behavioural profiles using national time use survey data. *Energy and Buildings* **252,** 111440 (2021).

18. Ouf, M. M., Bowden, E., Park, J. Y. & Gunay, B. A simulation-based approach to test and fine-tune occupant-centric lighting control strategies.

19. Zhou, P., Wang, H., Li, F., Dai, Y. & Huang, C. Development of window opening models for residential building in hot summer and cold winter climate zone of China. *Energy and Built Environment* **3,** 363–372 (2022).

20. Liu, W., Gunay, H. B. & Ouf, M. M. Modeling window and thermostat use behavior to inform sequences of operation in mixed-mode ventilation buildings. *Science and Technology for the Built Environment* **27,** 1204–1220 (2021).

21. Clemente, S., Beauchêne, S. & Nefzaoui, E. Generation of aggregated plug load profiles in office buildings. *Energy and Buildings* **252,** 111398 (2021).

22. Tyagi, R., Vishwakarma, S., Singh, K. K. & Syan, C. Low-cost energy conservation measures and behavioral change for sustainable energy goal. *Affordable and clean energy. Encyclopedia of the UN Sustainable Development Goals. Springer, Cham. https://doi. org/10.1007/978-3-319-71057-0_155-1* (2020).

23. D'Oca, S., Corgnati, S. P. & Buso, T. Smart meters and energy savings in Italy: Determining the effectiveness of persuasive communication in dwellings. *Energy Research & Social Science* **3,** 131–142 (2014).

24. Pothitou, M., Kolios, A. J., Varga, L. & Gu, S. A framework for targeting household energy savings through habitual behavioural change. *International Journal of Sustainable Energy* **35,** 686–700 (2016).

25. Gulbinas, R., Jain, R. K. & Taylor, J. E. BizWatts: A modular socio-technical energy management system for empowering commercial building occupants to conserve energy. *Applied Energy* **136,** 1076–1084 (2014).

26. Orland, B., Ram, N., Lang, D., Houser, K., Kling, N. & Coccia, M. Saving energy in an office environment: A serious game intervention. *Energy and Buildings* **74,** 43–52 (2014).

27. Himeur, Y., Alsalemi, A., Bensaali, F., Amira, A., Varlamis, I., Bravos, G., Sardianos, C. & Dimitrakopoulos, G. Techno-economic assessment of building energy efficiency systems using behavioral change: A case study of an edge-based micro-moments solution. *Journal of Cleaner Production* **331,** 129786 (2022).

28. Fraternali, P., Cellina, F., Herrera, S., Krinidis, S., Pasini, C., Rizzoli, A. E., Rottondi, C. & Tzovaras, D. *A socio-technical system based on gamification towards energy savings* **in** *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (2018), 59–64.

29. Esrafilian-Najafabadi, M. & Haghighat, F. Occupancy-based HVAC control systems in buildings: A state-of-the-art review. *Building and Environment* **197,** 107810 (2021).

30. Wei, S., Jones, R. & De Wilde, P. Driving factors for occupant-controlled space heating in residential buildings. *Energy and Buildings* **70,** 36–44 (2014).

31. Varki, A., Geschwind, D. H. & Eichler, E. E. Human uniqueness: genome interactions with environment, behaviour and culture. *Nature Reviews Genetics* **9,** 749–763 (2008).

32. Fabi, V., Andersen, R. V., Corgnati, S. & Olesen, B. W. Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models. *Building and Environment* **58,** 188–198 (2012).

33. D'Oca, S., Chen, C.-F., Hong, T. & Belafi, Z. Synthesizing building physics with social psychology: An interdisciplinary framework for context and occupant behavior in office buildings. *Energy research & social science* **34,** 240–251 (2017).

34. Hong, T., Taylor-Lange, S. C., D'Oca, S., Yan, D. & Corgnati, S. P. Advances in research and applications of energy-related occupant behavior in buildings. *Energy and buildings* **116,** 694–702 (2016).

35. Wagner, A., O'Brien, W. & Dong, B. Exploring occupant behavior in buildings. *Wagner, A., O'Brien, W., Dong, B., Eds* (2018).

36. Martani, C., Lee, D., Robinson, P., Britter, R. & Ratti, C. ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings* **47,** 584–591 (2012).

37. Masoso, O. & Grobler, L. The dark side of occupants' behaviour on building energy use. *Energy and Buildings* **42,** 173–177. ISSN: 0378-7788. https://www.sciencedirect.com/science/article/pii/S0378778809001893 (2010).

38. Kim, J., Schiavon, S. & Brager, G. Personal comfort models–A new paradigm in thermal comfort for occupant-centric environmental control. *Building and Environment* **132,** 114–124 (2018).

39. Gunay, H. B., O'Brien, W. & Beausoleil-Morrison, I. Development of an occupancy learning algorithm for terminal heating and cooling units. *Building and Environment* **93,** 71–85 (2015).

40. Naylor, S., Gillott, M. & Lau, T. A review of occupant-centric building control strategies to reduce building energy use. *Renewable and Sustainable Energy Reviews* **96,** 1–10 (2018).

41. Park, J. Y., Ouf, M. M., Gunay, B., Peng, Y., O'Brien, W., Kjærgaard, M. B. & Nagy, Z. A critical review of field implementations of occupant-centric building controls. *Building and Environment* **165,** 106351 (2019).

42. Xie, J., Li, H., Li, C., Zhang, J. & Luo, M. Review on occupant-centric thermal comfort sensing, predicting, and controlling. *Energy and Buildings* **226,** 110392 (2020).

43. Stopps, H., Huchuk, B., Touchie, M. F. & O'Brien, W. Is anyone home? A critical review of occupant-centric smart HVAC controls implementations in residential buildings. *Building and Environment* **187,** 107369 (2021).

44. Melfi, R., Rosenblum, B., Nordman, B. & Christensen, K. *Measuring building occupancy using existing network infrastructure* **in** *2011 International Green Computing Conference and Workshops* (2011), 1–8.

45. İçoğlu, O. & Mahdavi, A. VIOLAS: A vision-based sensing system for sentient building models. *Automation in construction* **16,** 685–712 (2007).

46. Lam, K. P., Hoynck, M., Zhang, R., Andrews, B., Chiou, Y.-S., Dong, B., Benitez, D. **andothers**. *Information-theoretic environmental features selection for occupancy detection in open offices* **in** *Eleventh International IBPSA Conference* (2009), 27–30.

47. Ekwevugbe, T., Brown, N., Pakka, V. & Fan, D. Improved occupancy monitoring in non-domestic buildings. *Sustainable cities and society* **30,** 97–107 (2017).

48. Hobson, B. W., Lowcay, D., Gunay, H. B., Ashouri, A. & Newsham, G. R. Opportunistic occupancy-count estimation using sensor fusion: A case study. *Building and environment* **159,** 106154 (2019).

49. Naghiyev, E., Gillott, M. & Wilson, R. Three unobtrusive domestic occupancy measurement technologies under qualitative review. *Energy and Buildings* **69,** 507–514 (2014).

50. Heidari, A., Barthelmes, V. M. & Khovalyg, D. *Probabilistic Machine Learning for Occupancy Prediction based on Sensor Fusion* **in** *"Proc. of 6th European Conference on Behaviour and Energy Efficiency"* ().

51. Curran, K., Furey, E., Lunney, T., Santos, J., Woods, D. & McCaughey, A. An evaluation of indoor location determination technologies. *Journal of Location Based Services* **5,** 61–78 (2011).

52. Misra, A. & Das, S. K. Location estimation (determination and prediction) techniques in smart environments. *Smart Environments: Technologies, Protocols, and Applications,* 193–228 (2004).

53. Li, N., Calis, G. & Becerik-Gerber, B. Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations. *Automation in construction* **24,** 89–99 (2012).

54. Shipman, R. & Gillott, M. Study of the Use of Wireless Behavior Systems to Encourage Energy Efficiency in Domestic Properties. *Hong Kong* (2013).

55. Hay, S. & Harle, R. *Bluetooth tracking without discoverability* **in** *International Symposium on Location-and Context-Awareness* (2009), 120–137.

56. Lee, D. & Oh, S. *Understanding human-place interaction from tracking and identification of many users* **in** *2013 IEEE 1st International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA)* (2013), 112–115.

57. Dodier, R. H., Henze, G. P., Tiller, D. K. & Guo, X. Building occupancy detection through sensor belief networks. *Energy and buildings* **38,** 1033–1043 (2006).

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

59. Heidari, A., Maréchal, F. & Khovalyg, D. Reinforcement Learning for proactive operation of residential energy systems by learning stochastic occupant behavior and fluctuating solar energy: Balancing comfort, hygiene and energy use. *Applied Energy* **318,** 119206 (2022).

60. Heidari, A., Maréchal, F. & Khovalyg, D. An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach. *Applied Energy* **312,** 118833 (2022).

61. Park, J. Y. & Nagy, Z. *HVACLearn: a reinforcement learning based occupant-centric control for thermostat set-points* **in** *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (2020), 434–437.

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

63. Barthelmes, V. M., Heo, Y., Fabi, V. & Corgnati, S. P. Exploration of the Bayesian Network framework for modelling window control behaviour. *Building and Environment* **126,** 318–330 (2017).

64. Tekler, Z., Low, R. & Blessing, L. *Using smart technologies to identify occupancy and plug-in appliance interaction patterns in an office environment* **in** *IOP Conference Series: Materials Science and Engineering* **609** (2019), 062010.

65. Jung, S., Jeoung, J. & Hong, T. Occupant-centered real-time control of indoor temperature using deep learning algorithms. *Building and Environment* **208,** 108633 (2022).

66. Nguyen, T. A. & Aiello, M. Beyond Indoor Presence Monitoring with Simple Sensors. *PECCS* **2012,** 5–14 (2012).

67. Zhang, Z. & Lam, K. P. *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), 148–157.

68. Labeodan, T., De Bakker, C., Rosemann, A. & Zeiler, W. On the application of wireless sensors and actuators network in existing buildings for occupancy detection and occupancy-driven lighting control. *Energy and Buildings* **127,** 75–83 (2016).

69. Xu, Y., Stojanovic, N., Stojanovic, L., Anicic, D. & Studer, R. *An approach for more efficient energy consumption based on real-time situational awareness* **in** *Extended Semantic Web Conference* (2011), 270–284.

70. *The Energy Egg* https://www.thegreenage.co.uk/review/the-energy-egg/ (2022).

71. Erickson, V. L., Carreira-Perpiñán, M. Á. & Cerpa, A. E. *OBSERVE: Occupancy-based system for efficient reduction of HVAC energy* **in** *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks* (2011), 258–269.

72. 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* **144,** 159–168 (2018).

73. Hassabis, D. The Power of Self-learning systems. *Video, Center for Brains Minds+ Machines* **20** (2019).

74. Schreiber, T., Eschweiler, S., Baranski, M. & Müller, D. Application of two promising Reinforcement Learning algorithms for load shifting in a cooling supply system. *Energy and Buildings* **229,** 110490 (2020).

75. bwp. *CO_2 savings through heatpumps* https://www.waermepumpe.de/ (2022).

76. Zhang, D., Li, J., Nan, J. & Wang, L. Thermal performance prediction and analysis on the economized vapor injection air-source heat pump in cold climate region of China. *Sustainable Energy Technologies and Assessments* **18,** 127–133 (2016).

77. Van Kenhove, E., Dinne, K., Janssens, A. & Laverge, J. Overview and comparison of Legionella regulations worldwide. *American journal of infection control* **47,** 968–978 (2019).

78. Rao, R. *Reinforcement Learning: An Introduction; RS Sutton, AG Barto (Eds.); MIT Press, Cambridge, MA, 1998, 380 pages, ISBN 0-262-19398-1* 2000.

79. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95,** 245–258 (2017).

80. Wang, Z. & Hong, T. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy* **269,** 115036 (2020).

81. Otterlo, M. v. & Wiering, M. **in** *Reinforcement learning* 3–42 (Springer, 2012).

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

83. Roveda, L., Maskani, J., Franceschi, P., Abdi, A., Braghin, F., Molinari Tosatti, L. & Pedrocchi, N. Model-based reinforcement learning variable impedance control for human-robot collaboration. *Journal of Intelligent & Robotic Systems* **100,** 417–433 (2020).

84. Chen, B., Cai, Z. & Bergés, M. *Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy* **in** *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation* (2019), 316–325.

85. Hosseinloo, A. H., Ryzhov, A., Bischi, A., Ouerdane, H., Turitsyn, K. & Dahleh, M. A. Data-driven control of micro-climate in buildings: An event-triggered reinforcement learning approach. *Applied Energy* **277,** 115451 (2020).

86. Ryu, M., Chow, Y., Anderson, R., Tjandraatmadja, C. & Boutilier, C. CAQL: Continuous action Q-learning. *arXiv preprint arXiv:1909.12397* (2019).

87. Sabry, M. & Khalifa, A. On the reduction of variance and overestimation of deep Q-learning. *arXiv preprint arXiv:1910.05983* (2019).

88. Van Hasselt, H., Guez, A. & Silver, D. *Deep reinforcement learning with double q-learning* **in** *Proceedings of the AAAI conference on artificial intelligence* **30** (2016).

89. Marszal-Pomianowska, A., Zhang, C., Pomianowski, M., Heiselberg, P., Gram-Hanssen, K. & Hansen, A. R. Simple methodology to estimate the mean hourly and the daily profiles of domestic hot water demand from hourly total heating readings. *Energy and Buildings* **184,** 53–64 (2019).

90. *EFFICIENT PRODUCTION OF DOMESTIC HOT WATER* (Suisse Energie). file:///C:/Users/aheidari/Documents/Downloads/WEB_Effiziente_Warmwasser-Systeme_201706_f.pdf.

91. Aguilar, F., Aledo, S. & Quiles, P. Experimental study of the solar photovoltaic contribution for the domestic hot water production with heat pumps in dwellings. *Applied Thermal Engineering* **101,** 379–389 (2016).

92. Heidari, A. & Khovalyg, D. Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition. *Solar Energy* **207,** 626–639 (2020).

93. Sinha, R., Jensen, B. B., Pillai, J. R., Bojesen, C. & Moller-Jensen, B. *Modelling of hot water storage tank for electric grid integration and demand response control* **in** *2017 52nd International Universities Power Engineering Conference (UPEC)* (2017), 1–6.

94. *Heat Pump Water Heaters as Clean-Energy Batteries* https://www.nrdc.org/experts/pierre-delforge/heat-pump-water-heaters-clean-energy-batteries (2022).

95. George, D., Pearre, N. S. & Swan, L. G. High resolution measured domestic hot water consumption of Canadian homes. *Energy and buildings* **109,** 304–315 (2015).

96. Booysen, M., Engelbrecht, J., Ritchie, M., Apperley, M. & Cloete, A. How much energy can optimal control of domestic water heating save? *Energy for Sustainable Development* **51,** 73–85 (2019).

97. Heidari, A., Olsen, N., Mermod, P., Alahi, A. & Khovalyg, D. Adaptive hot water production based on Supervised Learning. *Sustainable Cities and Society* **66,** 102625 (2021).

98. Booysen, M., Engelbrecht, J., Ritchie, M., Apperley, M. & Cloete, A. How much energy can optimal control of domestic water heating save? *Energy for Sustainable Development* **51,** 73–85 (2019).

99. Stadler, P., Girardin, L., Ashouri, A. & Maréchal, F. Contribution of model predictive control in the integration of renewable energy sources within the built environment. *Frontiers in Energy Research* **6,** 22 (2018).

100. Ritchie, M., Engelbrecht, J. & Booysen, M. A probabilistic hot water usage model and simulator for use in residential energy management. *Energy and Buildings* **235,** 110727 (2021).

101. Hosseinloo, A. H., Ryzhov, A., Bischi, A., Ouerdane, H., Turitsyn, K. & Dahleh, M. A. Data-driven control of micro-climate in buildings: An event-triggered reinforcement learning approach. *Applied Energy* **277,** 115451 (2020).

102. 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* **144,** 159–168 (2018).

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

104. Zou, Z., Yu, X. & Ergan, S. Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. *Building and Environment* **168,** 106535 (2020).

105. Du, Y., Li, F., Munk, J., Kurte, K., Kotevska, O., Amasyali, K. & Zandi, H. Multi-task deep reinforcement learning for intelligent multi-zone residential HVAC control. *Electric Power Systems Research* **192,** 106959 (2021).

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

107. Han, M., May, R., Zhang, X., Wang, X., Pan, S., Da, Y. & Jin, Y. A novel reinforcement learning method for improving occupant comfort via window opening and closing. *Sustainable Cities and Society* **61,** 102247 (2020).

108. Ali, A. & Kazmi, H. *Minimizing grid interaction of solar generation and DHW loads in nZEBs using model-free reinforcement learning* **in** *International workshop on data analytics for renewable energy integration* (2017), 47–58.

109. Brandi, S., Piscitelli, M. S., Martellacci, M. & Capozzoli, A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings* **224,** 110225 (2020).

110. Correa-Jullian, C., Droguett, E. L. & Cardemil, J. M. Operation scheduling in a solar thermal system: A reinforcement learning-based framework. *Applied energy* **268,** 114943 (2020).

111. Kazmi, H., D'Oca, S., Delmastro, C., Lodeweyckx, S. & Corgnati, S. P. Generalizable occupant-driven optimization model for domestic hot water production in NZEB. *Applied Energy* **175,** 1–15 (2016).

112. Ruelens, F., Claessens, B. J., Quaiyum, S., De Schutter, B., Babuška, R. & Belmans, R. Reinforcement learning applied to an electric water heater: From theory to practice. *IEEE Transactions on Smart Grid* **9,** 3792–3800 (2016).

113. Schreiber, T., Eschweiler, S., Baranski, M. & Müller, D. Application of two promising Reinforcement Learning algorithms for load shifting in a cooling supply system. *Energy and Buildings* **229,** 110490 (2020).

114. Soares, A., Geysen, D., Spiessens, F., Ectors, D., De Somer, O. & Vanthournout, K. Using reinforcement learning for maximizing residential self-consumption–Results from a field test. *Energy and Buildings* **207,** 109608 (2020).

115. Lork, C., Li, W.-T., Qin, Y., Zhou, Y., Yuen, C., Tushar, W. & Saha, T. K. An uncertainty-aware deep reinforcement learning framework for residential air conditioning energy management. *Applied Energy* **276,** 115426 (2020).

116. Lissa, P., Deane, C., Schukat, M., Seri, F., Keane, M. & Barrett, E. Deep reinforcement learning for home energy management system control. *Energy and AI* **3,** 100043 (2021).

117. Schreiber, T., Netsch, C., Baranski, M. & Müller, D. Monitoring data-driven Reinforcement Learning controller training: A comparative study of different training strategies for a real-world energy system. *Energy and Buildings* **239,** 110856 (2021).

118. Polydoros, A. S. & Nalpantidis, L. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems* **86,** 153–173 (2017).

119. *Legionella control efforts intensified in Switzerland – new "LeCo" project launched* https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-78327.html (2022).

120. Sabry, M. & Khalifa, A. On the reduction of variance and overestimation of deep Q-learning. *arXiv preprint arXiv:1910.05983* (2019).

121. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

122. Van Hasselt, H., Guez, A. & Silver, D. *Deep reinforcement learning with double q-learning* **in** *Proceedings of the AAAI conference on artificial intelligence* **30** (2016).

123. Gelažanskas, L. & Gamage, K. A. *Forecasting hot water consumption in dwellings using artificial neural networks* **in** *2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG)* (2015), 410–415.

124. Delorme-Costil, A. & Bezian, J.-J. *Forecasting domestic hot water demand in residential house using artificial neural networks* **in** *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2017), 467–472.

125. Armstrong, P. M., Uapipatanakul, M., Thompson, I., Ager, D. & McCulloch, M. Thermal and sanitary performance of domestic hot water cylinders: Conflicting requirements. *Applied Energy* **131,** 171–179 (2014).

126. Jacobs, H., Botha, B. & Blokker, M. *Household hot water temperature–an analysis at end-use level* **in** *WDSA/CCWI Joint Conference Proceedings* **1** (2018).

127. Bertrand, A., Aggoune, R. & Maréchal, F. In-building waste water heat recovery: An urban-scale method for the characterisation of water streams and the assessment of energy savings and costs. *Applied Energy* **192,** 110–125 (2017).

128. Stout, J. E., Best, M. G. & Yu, V. L. Susceptibility of members of the family Legionellaceae to thermal stress: implications for heat eradication methods in water distribution systems. *Applied and environmental microbiology* **52,** 396–399 (1986).

129. Van Kenhove, E., De Backer, L., Delghust, M. & Laverge, J. *Coupling of modelica domestic hot water simulation model with controller* **in** *Building simulation* (2019).

130. *Research SC centre of excellence for agricultural* https://www.agrometeo.ch/ (2022).

131. *TensorForce-modular deep reinforcement learning in TensorFlow* https://tensorforce.readthedocs.io/en/0.2.0/index.html (2022).

132. *Droople company* https://droople.com/ (2022).

133. Majcen, D. Predicting energy consumption and savings in the housing stock. *A+BE| Architecture and the Built Environment,* 1–224 (2016).

134. Sun, K. & Hong, T. A simulation approach to estimate energy savings potential of occupant behavior measures. *Energy and Buildings* **136,** 43–62 (2017).

135. Gill, Z. M., Tierney, M. J., Pegg, I. M. & Allan, N. Low-energy dwellings: the contribution of behaviours to actual performance. *Building Research & Information* **38,** 491–508. eprint: https://doi.org/10.1080/09613218.2010.505371. https://doi.org/10.1080/09613218.2010.505371 (2010).

136. Han, M., Zhao, J., Zhang, X., Shen, J. & Li, Y. The reinforcement learning method for occupant behavior in building control: A review. *Energy and Built Environment* **2,** 137–148 (2021).

137. Li, J., Yu, Z. J., Haghighat, F. & Zhang, G. Development and improvement of occupant behavior models towards realistic building performance simulation: A review. *Sustainable Cities and Society* **50,** 101685 (2019).

138. Hong, T., Chen, Y., Belafi, Z. & D'Oca, S. *Occupant behavior models: A critical review of implementation and representation approaches in building performance simulation programs* **in** *Building Simulation* **11** (2018), 1–14.

139. Harputlugil, T. & de Wilde, P. The interaction between humans and buildings for energy efficiency: A critical review. *Energy Research & Social Science* **71,** 101828 (2021).

140. Yue, T., Long, R. & Chen, H. Factors influencing energy-saving behavior of urban households in Jiangsu Province. *Energy Policy* **62,** 665–675. ISSN: 0301-4215. https://www.sciencedirect.com/science/article/pii/S0301421513006940 (2013).

141. *REmap 2030: A renewable energy roadmap* (International Renewable Energy Agency, 2014).

142. *Grid-integrated distributed solar: addressing challenges for operations and planning* (National Renewable Energy Laboratory, 2016).

143. Dengiz, T., Jochem, P. & Fichtner, W. *Impact of different control strategies on the flexibility of power-to-heat-systems* **in** *Transforming Energy Markets, 41st IAEE International Conference, Jun 10-13, 2018* (2018).

144. Kondziella, H. & Bruckner, T. Flexibility requirements of renewable energy based electricity systems–a review of research results and methodologies. *Renewable and Sustainable Energy Reviews* **53,** 10–22 (2016).

145. Sethi, M., Tripathi, R., Pattnaik, B., Kumar, S., Khargotra, R., Chand, S. & Thakur, A. Recent developments in design of evacuated tube solar collectors integrated with thermal energy storage: A review. *Materials Today: Proceedings* (2021).

146. https://www.waermepumpe.de/presse/zahlen-daten/. Accessed: 2021-10-1.

147. Leppin, L. *Development of operational strategies for a heating pump system with photovoltaic, electrical and thermal storage* 2017.

148. Camacho, E. F. & Alba, C. B. *Model predictive control* (Springer science & business media, 2013).

149. Fiorentini, M., Wall, J., Ma, Z., Braslavsky, J. H. & Cooper, P. Hybrid model predictive control of a residential HVAC system with on-site thermal energy generation and storage. *Applied Energy* **187,** 465–479 (2017).

150. Halvgaard, R., Poulsen, N. K., Madsen, H. & Jørgensen, J. B. *Economic model predictive control for building climate control in a smart grid* **in** *2012 IEEE PES innovative smart grid technologies (ISGT)* (2012), 1–6.

151. Mady, A. E.-D., Provan, G., Ryan, C. & Brown, K. *Stochastic model predictive controller for the integration of building use and temperature regulation* **in** *Proceedings of the AAAI Conference on Artificial Intelligence* **25** (2011).

152. Smarra, F., Jain, A., De Rubeis, T., Ambrosini, D., D'Innocenzo, A. & Mangharam, R. Data-driven model predictive control using random forests for building energy optimization and climate control. *Applied energy* **226,** 1252–1272 (2018).

153. Schreiber, T., Netsch, C., Eschweiler, S., Wang, T., Storek, T., Baranski, M. & Müller, D. Application of data-driven methods for energy system modelling demonstrated on an adaptive cooling supply system. *Energy* **230,** 120894 (2021).

154. Brandi, S., Piscitelli, M. S., Martellacci, M. & Capozzoli, A. Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. *Energy and Buildings* **224,** 110225 (2020).

155. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

156. Liu, S. & Henze, G. P. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis. *Energy and buildings* **38,** 148–161 (2006).

157. Chen, Y., Norford, L. K., Samuelson, H. W. & Malkawi, A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. *Energy and Buildings* **169,** 195–205 (2018).

158. Cheng, Z., Zhao, Q., Wang, F., Jiang, Y., Xia, L. & Ding, J. Satisfaction based Q-learning for integrated lighting and blind control. *Energy and Buildings* **127,** 43–55 (2016).

159. Zou, Z., Yu, X. & Ergan, S. Towards optimal control of air handling units using deep reinforcement learning and recurrent neural network. *Building and Environment* **168,** 106535 (2020).

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

161. Heidari, A., Khovalyg, D. & Marechal, F. *An adaptive control framework based on Reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems* **in** *Vielfalt leben - Offenheit erhalten. Multiperspektivität und Interdisziplinarität in Pflege - Praxis - Wissenschaft.* Forschungswelten 2018St. Gallen, Switzerland, 19–20 **april** 2018 (2021).

162. Correa-Jullian, C., Droguett, E. L. & Cardemil, J. M. Operation scheduling in a solar thermal system: A reinforcement learning-based framework. *Applied Energy* **268,** 114943 (2020).

163. Ali, A. & Kazmi, H. *Minimizing grid interaction of solar generation and DHW loads in nZEBs using model-free reinforcement learning* **in** *International Workshop on Data Analytics for Renewable Energy Integration* (2017), 47–58.

164. Lissa, P., Deane, C., Schukat, M., Seri, F., Keane, M. & Barrett, E. Deep reinforcement learning for home energy management system control. *Energy and AI* **3,** 100043 (2021).

165. Van Amerongen, G., Lee, J. & Suter, J.-M. *Legionella and solar water heaters* (Suter Consulting, Bern, Switzerland, 22 **april** 2013).

166. Van Kenhove, E., De Backer, L., Janssens, A. & Laverge, J. Simulation of Legionella concentration in domestic hot water: comparison of pipe and boiler models. *Journal of Building Performance Simulation* **12,** 595–619 (2019).

167. Mirnaghi, M., Panchabikesan, K. & Haghighat, F. *Application of data mining in understanding the charging patterns of the hot water tank in a residential building: a case study* **in** *IOP Conference Series: Materials Science and Engineering* **609** (2019), 052038.

168. Carlson, K. M., Boczek, L. A., Chae, S. & Ryu, H. Legionellosis and recent advances in technologies for Legionella control in premise plumbing systems: a review. *Water* **12,** 676 (2020).

169. Krawczyk, M., Petruzzelli, M. **andothers**. Legionella 2003: An update and statement by the association of water technologies. *Association of Water Technologies* **26** (2003).

170. Taghdiri, S. *Airborne Dispersion and Plume Modeling of Legionella Bacteria* (Arizona State University, 2014).

171. Sharaby, Y., Rodrıguez-Martınez, S., Oks, O., Pecellin, M., Mizrahi, H., Peretz, A., Brettar, I., Höfle, M. G. & Halpern, M. Temperature-dependent growth modeling of environmental and clinical Legionella pneumophila multilocus variable-number tandem-repeat analysis (MLVA) genotypes. *Applied and environmental microbiology* **83,** e03295–16 (2017).

172. Van Kenhove, E., De Backer, L., Delghust, M. & Laverge, J. *Coupling of Modelica Domestic Hot Water Simulation Model with Controller* **in** *Building Simulation 2019, 16th IBPSA International Conference and Exhibition* **16** (2020), 924–931.

173. Quillen, D., Jang, E., Nachum, O., Finn, C., Ibarz, J. & Levine, S. *Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods* **in** *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), 6284–6291.

174. Gelažanskas, L. & Gamage, K. A. *Forecasting hot water consumption in dwellings using artificial neural networks* **in** *2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG)* (2015), 410–415.

175. Delorme-Costil, A. & Bezian, J.-J. *Forecasting domestic hot water demand in residential house using artificial neural networks* **in** *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2017), 467–472.

176. For Standardization, I. O. *ISO 7730 2005-11-15 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* https://books.google.ch/books?id=p3YcoAEACAAJ (ISO, 2005).

177. Ritchie, M., Engelbrecht, J. & Booysen, M. A probabilistic hot water usage model and simulator for use in residential energy management. *Energy and Buildings* **235,** 110727 (2021).

178. Organization, W. H. **andothers**. WHO housing and health guidelines (2018).

179. Ormandy, D. & Ezratty, V. Health and thermal comfort: From WHO guidance to housing strategies. *Energy Policy* **49,** 116–121 (2012).

180. Quero, S., Párraga-Niño, N., Garcia-Núñez, M., Pedro-Botet, M. L., Gavaldà, L., Mateu, L., Sabrià, M. & Mòdol, J. M. The impact of pipeline changes and temperature increase in a hospital historically colonised with Legionella. *Scientific Reports* **11,** 1–7 (2021).

181. Gooroochurn, M. & Visram, A. Maximization of Solar Hot Water Production Using a Secondary Storage Tank. *Journal of Clean Energy Technologies* **7** (2019).

182. Melius, J., Margolis, R. & Ong, S. Estimating rooftop suitability for PV: a review of methods, patents, and validation techniques (2013).

183. Zhang, Z., Chong, A., Pan, Y., Zhang, C., Lu, S. & Lam, K. P. *A deep reinforcement learning approach to using whole building energy model for hvac optimal control* **in** *2018 Building Performance Analysis Conference and SimBuild* **3** (2018), 22–23.

184. Vanhoudt, D., Geysen, D., Claessens, B., Leemans, F., Jespers, L. & Van Bael, J. An actively controlled residential heat pump: Potential on peak shaving and maximization of self-consumption of renewable energy. *Renewable Energy* **63,** 531–543 (2014).

185. Li, N. & Chen, Q. Experimental study on heat transfer characteristics of interior walls under partial-space heating mode in hot summer and cold winter zone in China. *Applied Thermal Engineering* **162,** 114264 (2019).

186. Stazi, F., Naspi, F. & D'Orazio, M. A literature review on driving factors and contextual events influencing occupants' behaviours in buildings. *Building and Environment* **118,** 40–66 (2017).

187. Stadler, P., Girardin, L., Ashouri, A. & Maréchal, F. Contribution of model predictive control in the integration of renewable energy sources within the built environment. *Frontiers in Energy Research* **6,** 22 (2018).

188. Farrokhifar, M., Bahmani, H., Faridpak, B., Safari, A., Pozo, D. & Aiello, M. Model predictive control for demand side management in buildings: A survey. *Sustainable Cities and Society* **75,** 103381. ISSN: 2210-6707. https://www.sciencedirect.com/science/article/pii/S2210670721006545 (2021).

189. *The power of self-learning systems* https://cbmm.mit.edu/video/power-self-learning-systems. Accessed: 2022-03-15.

190. Qin, Y., Ke, J., Wang, B. & Filaretov, G. F. Energy optimization for regional buildings based on distributed reinforcement learning. *Sustainable Cities and Society* **78,** 103625. ISSN: 2210-6707. https://www.sciencedirect.com/science/article/pii/S2210670721008891 (2022).

191. Heidari, A., Maréchal, F. & Khovalyg, D. An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach. *Applied Energy* **312,** 118833 (2022).

192. Heidari, A., Maréchal, F. & Khovalyg, D. Reinforcement Learning for proactive operation of residential energy systems by learning stochastic occupant behavior and fluctuating solar energy: Balancing comfort, hygiene and energy use. *Applied Energy* **318,** 119206. ISSN: 0306-2619. https://www.sciencedirect.com/science/article/pii/S0306261922005712 (2022).

193. Heidari, A., Olsen, N., Mermod, P., Alahi, A. & Khovalyg, D. Adaptive hot water production based on Supervised Learning. *Sustainable Cities and Society* **66,** 102625. ISSN: 2210-6707. https://www.sciencedirect.com/science/article/pii/S2210670720308428 (2021).

194. Heidari, A., Marechal, F. & Khovalyg, D. An adaptive control framework based on Reinforcement learning to balance energy, comfort and hygiene in heat pump water heating systems. *Journal of Physics: Conference Series* **2042,** 012006. https://doi.org/10.1088/1742-6596/2042/1/012006 (**november** 2021).

195. Soares, A., Geysen, D., Spiessens, F., Ectors, D., De Somer, O. & Vanthournout, K. Using reinforcement learning for maximizing residential self-consumption–Results from a field test. *Energy and Buildings* **207,** 109608 (2020).

196. Marantos, C., Lamprakos, C., Siozios, K. & Soudris, D. **in** *IoT for Smart Grids* 183–207 (Springer, 2019).

197. Lee, D., Lee, S., Karava, P. & Hu, J. *Simulation-based policy gradient and its building control application* **in** *2018 Annual American Control Conference (ACC)* (2018), 5424–5429.

198. Fazenda, P., Veeramachaneni, K., Lima, P. & O'Reilly, U.-M. Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems. *Journal of Ambient Intelligence and Smart Environments* **6,** 675–690 (2014).

199. Yang, T., Zhao, L., Li, W., Wu, J. & Zomaya, A. Y. Towards healthy and cost-effective indoor environment management in smart homes: A deep reinforcement learning approach. *Applied Energy* **300,** 117335 (2021).

200. Lämmle, M., Bongs, C., Wapler, J., Günther, D., Hess, S., Kropp, M. & Herkel, S. Performance of air and ground source heat pumps retrofitted to radiator heating systems and measures to reduce space heating temperatures in existing buildings. *Energy* **242,** 122952 (2022).

201. Park, J. Y., Dougherty, T. & Nagy, Z. *A Bluetooth based occupancy detection for buildings* **in** *Proceedings of of Building Performance Analysis Conference and SimBuild. IBPSA* (2018), 807–814.

202. Shetty, A. D., Shubha, B., Suryanarayana, K. **andothers**. *Detection and tracking of a human using the infrared thermopile array sensor—"Grid-EYE"* **in** *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)* (2017), 1490–1495.

203. Dong, B., Andrews, B., Lam, K. P., Höynck, M., Zhang, R., Chiou, Y.-S. & Benitez, D. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network. *Energy and Buildings* **42,** 1038–1046 (2010).

204. Villariba, K. *A neural network approach to detecting the occupancy state of a home from electricity usage* phdthesis (Mills College, 2014).

205. Yang, Z., Li, N., Becerik-Gerber, B. & Orosz, M. *A non-intrusive occupancy monitoring system for demand driven HVAC operations* **in** *Construction Research Congress 2012: Construction Challenges in a Flat World* (2012), 828–837.

206. Van Hasselt, H., Guez, A. & Silver, D. Deep Reinforcement Learning with Double Q-learning. *CoRR* **abs/1509.06461.** arXiv: 1509.06461. http://arxiv.org/abs/1509.06461 (2015).

207. Zhang, C., Zhao, Y., Zhang, X., Fan, C. & Li, T. An improved cooling load prediction method for buildings with the estimation of prediction intervals. *Procedia Engineering* **205,** 2422–2428 (2017).

208. Schweiker, M., Kleber, M. & Wagner, A. Long-term monitoring data from a naturally ventilated office building. *Scientific data* **6,** 1–6 (2019).

209. Mora, D., Fajilla, G., Austin, M. C. & De Simone, M. Occupancy patterns obtained by heuristic approaches: cluster analysis and logical flowcharts. A case study in a university office. *Energy and Buildings* **186,** 147–168 (2019).

210. *Building Energy Model in Python* https://github.com/timtroendle/simple-simple. Accessed: 2022-06-04.

211. Kumar, R., Aggarwal, R., Sharma, J. & Pathania, S. Predicting energy requirement for cooling the building using artificial neural network. *Journal of Technology Innovations in Renewable Energy* **1,** 113–121 (2012).

212. Sutton, R. S. *On the significance of Markov decision processes* **in** *International Conference on Artificial Neural Networks* (1997), 273–282.

213. Nagy, Z. & Nweye, K. Real-world challenges for reinforcement learning in building control. *arXiv preprint arXiv:2112.06127* (2021).

Amirreza Heidari

Doctoral Assistant

20 May 1993

Passage du Cardinal 13b, CH-1700, Fribourg, Switzerland

+41 762 874 513

amirrez.heidari@epfl.ch

Languages

- Persian ●●●●●
- English ●●●●○

Programming Skills

- Python
- Arduino

Software Skills

- CAD software
  - AutoCAD
- Technical software
  - TRNSYS
  - Thermoflow
  - Node-RED

Hardware Skills

- Raspberry Pi
- Arduino
- 3D printing

Education

Currently — **Doctor of Philosophy**
EPFL, Lausanne, Switzerland

- **Field**: Energy Program

2015-2017 — **Master of Science**
Sharif University of Technology, Tehran, Iran

- **Field**: Energy Systems Engineering
- **G.P.A**: 18.04 out of 20 (4 out of 4)
- **Thesis**: Design of a solar-assisted desiccant system for co-production of water and cooling in hot and humid climates
- *First ranked student among graduates*

2011-2015 — **Bachelor of Science**
Shahid Beheshti University, Tehran, Iran

- **Field**: Mechanical Engineering
- **G.P.A**: 17.94 out of 20 (3.85 out of 4)
- **Thesis**: Efficiency improvement of parabolic trough solar thermal power plants
- *Second ranked student among graduates*

2007-2011 — **High School**
Exceptional talents High School (NODET), Shahrood, Iran

- **Field**: Mathematics and Physics

Awards and Achievements

Nov.2020 — **Student incubator award**, An award and a prototype development grant by Baloise insurance company of Switzerland for development of an IoT-based device for risk assessment and elimination of Legionella bacteria in building water systems

Feb.2017 — **Ranked first between graduates (M.Sc)**, Department of Energy Engineering, Sharif University of Technology, Tehran, Iran

Mar.2017 — **Award for distinguished student in education**, Department of Energy Engineering, Sharif University of Technology, Tehran, Iran

Sept.2015 — **Ranked second between graduates (B.Sc)**, Shahid Beheshti University, Department of Mechanical and Energy Engineering, Tehran, Iran

Sept.2015 — **Exceptional Talents Credit**, Passing from the M.Sc entrance without taking exam using exceptional talents credit

Nov.2015 — **Distinguished graduate award**, Shahid Beheshti University, Department of Mechanical and Energy Engineering

Research Interests

- Machine Learning (Reinforcement Learning, Supervised Learning)
- Internet of Things
- Occupant behavior
- Prototype development with Arduino and Raspberry Pi