User Feedback- Based Reinforcement Learning for Vehicle Comfort Control

Petre, Alexandra

Award date:
2019

Awarding institution:
Coventry University

USER FEEDBACK-BASED REINFORCEMENT LEARNING FOR VEHICLE COMFORT CONTROL

ALEXANDRA-GALINA PETRE

A thesis submitted in partial fulfilment of the University's requirements for the Degree of Doctor of Philosophy

September 2018

Faculty of Engineering, Environment and Computing
Coventry University

Coventry University

ABSTRACT

Occupants adapt to thermal discomfort using three types of thermal behaviours: physiological (e.g., sweating or shivering), changing the environment (e.g., changing heating settings or opening a window), or changing personal elements (e.g., clothing, body position, seating location). Compared to the built environment, the range of thermal behaviours in a vehicle is limited. Modern vehicle climate systems aim to maximise thermal comfort semi-automatically through control modes that allow the occupants to provide feedback via the interface. A novel approach to automatic climate control is the use of Reinforcement Learning. However, past work has ignored the feedback from the user.

The main aim of this thesis is to integrate user-feedback into an Reinforcement Learning (RL)-based vehicle climate control and assess if the system can learn user's preferences within a reasonable time. In order to develop an integrated system that includes the interaction of the user with the climate control interface there is a need for: a) a set of literature-based rules describing the extent of the thermal behaviour; b) a human-agent that mimics the feedback process; c) a method of integrating the simulated feedback in the context of Reinforcement Learning.

For the purpose of modelling the interaction with the climate system, three main rules were identified in the thermal comfort literature related to how likely occupants are to make changes when they are uncomfortable, which setting (temperature, blower or vent) they are likely to select, and which value they are likely to prefer.

The activation likelihood for each rule is found using data from an in-field experiment with 49 trials, monitoring occupant thermal comfort, climate control actions, and the thermal environment parameters. The resulting hybrid model (User-Based Module (UBM)) is validated against a hold-out set of data from the experimental trials. The User-Based Reinforcement Learning (UBRL) climate controller combines the simulated feedback from the UBM with feedback from the thermal environment by means of reward shaping. Three types of reward shaping methods were statistically compared: state shaping, look-back advice, and look-forward advice. Several State-Action-Reward-State-Action (SARSA)-based RL algorithms were used to train the system and their performance was evaluated using a set of test scenarios.

The UBM outperforms simpler models, such as neural network and fuzzy logic, achieving the highest accuracy for estimating setting adjustments. The simulated user feedback from the UBM improves the learning speed of the UBRL controller to 2.9 years of simulated learning. The controller using look-back advice has a statistically higher average reward per trial than alternative methods. Additionally, it requires a lower number of steps to achieve occupant desired equivalent temperature. The UBRL controller using the Double SARSA algorithm achieves on average the occupant's desired comfort in 5.6 minutes, maintaining it 86% of the journey duration and consuming an average power of 1.07 kW.

Therefore the Double SARSA UBRL climate control can significantly improve the comfort of the occupants by learning and maintaining their setting preferences within less than half the life time of their vehicles. Potential avenues for improvement involve a variable exploration rate, further development of the human agent,

and multi-zone climate control, extending its application to a variety of user modelling and control areas.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my Director of Studies, Dr. James Brusey, for giving me the opportunity to undertake this research path. His advice, expertise and guidance were the cornerstones for my development throughout my doctoral journey.

Further, I would like to thank my second supervisor, Professor Elena Gaura, for continuously motivating me through her determination and constructive criticism to progress in my research. Her experience and research perspective have been invaluable towards improving my scientific writing.

Many thanks are extended to Dr. Ross Wilkins, for all the help and continuous encouragement. Ross provided me with valuable insight on how to further improve my work, challenging my potential.

I am extremely grateful for the solid bond that I have forged with my Cogent Labs team: Kojo Sarfo Gyamfi, Gene Palencia, Nicolas Melinge, Gaobo Chen, James Wescott, and Ross Drury, with a mention of Edy Suardiyana and Razvan Ionescu. Thank you to Yevheniia Zhoholieva for her pertinent feedback, to Susan Lasslett for her kindness and reassurance, and to my dear friends who have stayed by my side in crucial moments.

I would like to express my immense love and gratitude to my amazing family, especially my mom and dad, for their everlasting care and support. This research would not have been possible without their sacrifice and belief in me.

Finally, I would like to thank Henry Ho, for his relentless faith in my capabilities, for his never-ending patience and inspiration. To his lovely family, for all the guidance and affection they have shown me. Every step taken towards accomplishing this work was in his presence and for that I am grateful. Thank you for always believing in me!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineers

AUC Area Under Curve

DAgger Dataset Aggregation algorithm

HVAC Heating, Ventilation and Air Conditioning

IEQ Indoor Environmental Quality

KBKR Knowledge-Based Kernel Regression

KL Kullback-Leibler

LCVTP Low Carbon Vehicle Technology Project

MAP Maximum A Posteriori

MDP Markov Decision Process

NV Natural Ventilated

PCT Perceptual Control Theory

PMML Predictive Model Markup Language

PMV Predicted Mean Vote

PPD Percentage People Dissatisfied

Pref-KBKR Preference Knowledge- Based Kernel Regression

RL Reinforcement Learning

ROC Receiver Operating Characteristics

SARSA State-Action-Reward-State-Action

TAMER Training the Agent Manually via Evaluative Reinforcement

TD Temporal Difference

UBRL User-Based Reinforcement Learning

UBM User-Based Module

1

INTRODUCTION

1.1 AIM

The Heating, Ventilation and Air Conditioning (HVAC) system within a car cabin cannot always efficiently satisfy the thermal comfort preferences of the cabin occupants, while maintaining a low level of energy consumption. Car cabin passenger thermal comfort is harder to maintain compared to comfort in homes and offices, given the non-stationary nature of this environment and its thermal asymmetry caused by the use of windows and the HVAC system [NH03].

Several factors related to the thermal environment contribute to the thermal comfort problem in vehicles such as: the speed of the car; the angle at which the sun shines; the solar loading; the outside and inside temperatures; the velocity with which the air is circulated or re-circulated; and the humidity rates in the cabin. When a vehicle is in motion, the external environment changes quickly causing the parameters that impact occupant's thermal comfort to also change rapidly. Alternatively, when passengers re-enter a stationary and unoccupied vehicle, they may feel immediate discomfort. This is due to the fact that the cabin can reach extremely high or low temperatures, given the radiation from the sun and the outside temperature.

The thermal comfort control process can be briefly described as a cycle (figure 1.1). The HVAC system operation is tasked with maintaining the thermal comfort of the passengers. Once they feel uncomfortable they choose from a range of behaviours to regain comfort (e.g. HVAC adjustment, the use of windows, removing or adding

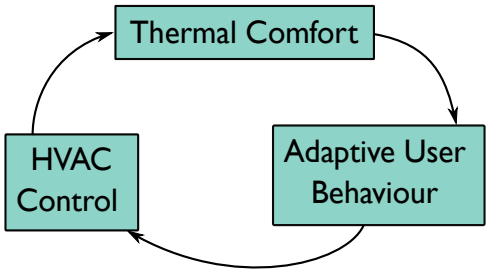

Figure 1.1: The cyclic nature of the thermal comfort control in the car cabin.

clothing items). The occupants' most accessible and comfort-related behavioural option in the car is to interact with the HVAC interface and adjust its settings. The alternative behavioural options are not strictly related to maintaining comfort and are beyond the scope of this thesis. The motivation behind these actions is vague and can be associated with other intentions, such as manoeuvrability for clothing removal, smoking habits and sleep prevention for opening windows.

The occupant's interaction with their HVAC systems has been overlooked, despite researchers' and engineers' efforts to further improve the intelligence of control systems by the use of machine learning algorithms (e.g fuzzy logic, evolutionary, artificial neural networks and reinforcement learning). This interaction is classified as a secondary task because drivers concentrate on the act of driving, which causes them to make sparse changes, only when necessary. Conversely, the changes made to the HVAC system are essential as they can offer vital information about the comfort preferences of the occupants.

Hence, the aim of this thesis is to develop a realistic simulation of occupant adaptive behaviour related to the HVAC system and include this within a machine learning control system. The main focus of the adaptive behaviour is to specifically examine the effect of HVAC setting preferences on the performance of a reinforcement-learning HVAC controller. Namely, if the controller can adjust its behaviour within a reasonable amount of time and offer comfort to passengers according to their desired settings.

Current set-point HVAC systems enable occupants to select a number of settings on the interface in order to better ensure their comfort (e.g. temperature, blower speed, vent orientation). Despite the HVAC interface adjustments, the system may not maintain the desired settings in all circumstances. For example, it only inputs

hot or cold air when the sensed cabin temperature is below or abovethe desired set-point, with maximum air blown into the cabin. Moreover, the system may not maintain the heated or cooled air at the preferred setting configurations throughout the journey duration given the rapidly changing factors that were mentioned above, and also throughout the cabin environment (i.e. temperature stratification between head and foot levels). This can trigger the occupants' discomfort and subsequently a further change in the settings.

The climate control system is programmed by a set of strict manufacturing rules, determining its actions to further accentuate the discomfort of the occupants (e.g. by not achieving the desired set-point temperature or slow vent flow changes). Additionally, in order to achieve the preferred settings a high amount of energy is consumed.

This is why implementing machine learning algorithms within the control process of the HVAC system is an alternative solution to hand-coded procedures. Machine learning has the potential to improve the passengers' comfort and mitigate energy constraints. Even though machine learning based HVAC systems deal with these two major problems, occupant preferences are disregarded. As an integral part of the environment, the occupant prefers specific settings and changes them when feeling uncomfortable, thus providing feedback to the system. This thesis examines the potential of integrating the setting interaction as an adaptive comfort method for the occupant onto a machine learning based HVAC system.

Among climate control systems using machine learning, the Reinforcement Learning (RL) HVAC controller developed by Hintea [Hin14] has the potential to include user feedback in order to improve the time it takes for the system to minimise energy consumption, while maintaining the desired comfort of the passengers. The agent (system) achieves this goal by a set of trial and error steps: observing the states and taking actions. The decision (policy) of which action the agent takes relies on which action brings the maximum reward (numerical value) when chosen. The reward affects both the current and future actions taken, hence it is bounded. The environment represents the car cabin. The state vector is comprised of the cabin, shell, ambient temperatures, and air flow. The action vector

is a combination of the outputs of the HVAC controller (fan speed, temperature, and recirculation).

This thesis proposes to combine a RL based HVAC controller with the modelled occupant interaction in order to explore the impact of user feedback on the learning performance of the control system. Interaction with the HVAC system enables the passengers to manifest their preferences for the thermal environment. As these preferences are reactions to the state of the environment and the current action of the control system, they can serve as corrective or predictive measures for the following state. The user feedback can be used as part of the reward, accelerating the learning process of the controller.

The challenge is to use such preferences in a realistic manner to train an HVAC controller in order for the system to learn in an effective way (to achieve an optimal or nearly optimal policy), within a reasonable amount of time (preferably before the end of a car's lifetime, estimated at 6-8 years).

1.2 RESEARCH QUESTIONS

The fundamental question that this thesis tries to answer is:

Given the limited interaction that users have with the HVAC, can an RL based system learn occupant's desired settings within the expected lifetime of a car?

The three main objectives for answering this research question are:

1. to identify the fundamental aspects that trigger cabin occupants' changes in HVAC settings;
2. to develop and integrate a realistic simulation of how occupants choose their HVAC settings within the framework of RL control;
3. to examine the learning capabilities of the newly developed system (User-Based Reinforcement Learning (UBRL)) in terms of learning time and policy performance.

Three subsequent research questions were developed related to these objectives:

1. What is the set of simple rules that can be drawn from the thermal comfort literature on occupant thermal behaviour related to HVAC control?

2. Can an artificial agent, validated using real-world data, realistically simulate the interaction that humans have with their HVAC system?

3. Can the UBRL HVAC system learn and maintain a nearly optimal policy based on occupant preferences within a reasonable amount of time?

The following sub-sections expand on the main ideas behind the sub-questions.

1.2.1 *Rules of interaction with the climate control*

In a confined space such as a car cabin, the occupants have a limited set of actions. For instance, they cannot change their location in the car or include personal heating or cooling devices (additional fans and heaters). So their opportunity to act is to adjust the settings of the HVAC system. Examining this type of behaviour is essential, given that the motivation behind these actions can be triggered by discomfort. Conversely, drivers are sometimes focused on fuel conservation. Hence they take actions to reduce the car's energy consumption, while sacrificing their personal comfort (e.g. not activating the HVAC system).

Modelling thermal behaviour in a car cabin is a challenging task. It relies on extensive experimental human trials conducted over various seasons, under clearly stated procedures that capture particular aspects of occupant behaviour. Moreover, a method for recording any type of actions is required and the motivation behind each action is also difficult to determine. Realistic driving scenarios are difficult to organise as there are several factors that cannot be controlled (e.g. the weather or traffic conditions). Furthermore, the participants need to be fully focused on the act of driving, which renders subsequent behaviours as infrequent and their motivation as vague (e.g. clothing changes or opening and closing windows). This is why the information related to occupant's adaptive behaviour needs to be readily available to the control system, without using alternative methods of capture. This thesis hypothesises that the first available action for occupants in cars is to make changes to the HVAC settings.

An alternative to capturing adaptive behaviour data is by using surveys to capture participant motivation. These are undertaken either after or during the trial. For the former, the surveys rely on the subject's memories of their past experiences. For the latter, the trials may be biased as the occupants would focus on their actions, thus leading to forced or unnatural behaviour. Instead of using extensive trials or surveys, the testing can initially be done by means of simulation.

In order to examine how to model the setting selections, this research compiles a set of simple rules based on thermal behaviour literature and proposes a model based on these rules, strictly referring to human interaction with the HVAC system. The interaction is separated into three main aspects: the decision to make a change, the selection of the type of setting and the value of the setting. This behaviour is modelled as a set of conditional probabilities linking setting selections with the overall body comfort of the occupants, represented in this thesis by equivalent temperature.

Chapter 2 presents an examination of the potential of improving comfort by using human feedback within the context of the vehicle's environment, with an emphasis on the state of the art in terms of human-RL systems interaction. The set of rules for HVAC setting selection is presented in Chapter 3.

1.2.2 *Deriving a realistic simulation of human interaction with the HVAC system*

Secondary activities to the driving task can be directly related to maintaining the comfort of drivers, since they interact with elements of the car (HVAC settings or heated seats). Hence, a model of how occupants behave when they experience discomfort is needed for a better evaluation and estimation of thermal comfort.

The model resulting from the combination of the three rules, named the User-Based Module, is validated using real-world data and outputs the estimated setting selections of a simulated cabin occupant. Each decision that an occupant makes depends on a conditional probability. The method used to model the conditional probabilities is a set of classifiers. The first rule is modelled as the probability of making an adjustment or not depending on how comfortable the occupant is.

The second rule (selecting a setting) is based on determining the probability of selection for the three types of settings: set-point temperature, blower level or vent distribution. The final rule is related to values selected for each setting (a set of 23 distinct values).

As the rules become more complex, the classification problem for the third rule evolves from estimating the probabilities of two classes into identifying multiple classes. Given the complexity of the model, it is compared with two alternative machine learning models (a neural network, and a fuzzy logic system), in order to test its performance.

While behaviour is important, the key factor is how to integrate the information as feedback to the learning system (the RL based HVAC) in order to improve the system's learning time and adjust the cabin conditions to an occupant's desired comfort. The feedback method that this thesis discusses is based on a task familiar to both drivers and passengers: changing the HVAC settings. The captured interactions are based on data extracted from a set of trails conducted in the car. The processed and analysed data is subsequently used to test and validate the User-Based Module (UBM) described in Chapter 4.

1.2.3 *Performance of an HVAC controller trained by passenger preferences*

The RL-based HVAC system has two major goals: maintaining equivalent temperature-based thermal comfort while using a low amount of energy. This system is designed to be trained offline and achieve a nearly optimal policy (maximising the total reward based on the association of environmental states with greedily selected actions). The policy can be programmed onto the control unit of a car, making use of the existing HVAC capabilities.

Compared to the available HVAC control systems that offer multiple modes, including those that allow cabin passengers to adjust the HVAC settings in terms of set-points [Scho8], the RL control system does not incorporate any form of human participation.

A further problem with software agents based on RL is that the learning takes a considerable amount of time. For example, Dalamakidis [DD08] proposed an RL based system for building comfort control that had an estimated learning time of four years. Even though introducing human feedback has the potential to improve the learning speed of RL systems, drivers are mainly focused on driving in the car cabin and cannot provide constant feedback. Moreover, by the time such an HVAC system makes changes to the environment, the occupants' comfort preferences may have changed and they would no longer be comfortable. Thus it is essential for the system to learn from sparse feedback as fast as possible.

A further problem is how to integrate the preferences of the occupant into the learning environment. A solution is integrating the feedback through the reward function in the form of a penalty when preferences are not met. Alternatively, the reward can be used as motivation and guidance towards future actions. Therefore, setting selections can be used as an avenue to anticipate desired changes to the environment and help the RL-based controller to choose actions closer to the cabin passengers' preferences. Integrating the feedback of a simulated human agent (UBM) within the climate control architecture enables the RL agent to learn a policy by means of potential based, look ahead, and look back shaping.

Additionally, the RL controller proposed by Hintea [Hin14] uses only the State-Action-Reward-State-Action (SARSA) (λ) algorithm. There are alternative algorithms based on SARSA that have a convergence potential, such as Expected SARSA, or do not have an maximisation bias, Double SARSA. The performance of the control system trained with such algorithms is examined. The controllers trained with the SARSA algorithms are compared by means of the highest reward obtained during the training process, how fast the occupant's desired comfort is achieved under cooling or warming of the cabin, the time the occupant spends in comfort and the amount of energy consumed.

Consequently, this research proposes the UBRL HVAC controller that learns from the feedback of a simulated occupant (UBM) and the cabin environment. It examines the behaviour of the climate control system trained with multiple RL algorithms in order to identify the most suitable learning strategy. The UBM represents the

internal process of how the occupant chooses to alter the settings of the HVAC system. This change in settings is registered as feedback by the UBRL HVAC system, which receives an additional penalty or reward for the change.

The description of the system architecture and the exploration of the controller's performance are available in Chapter 5. The combined human and environmental reward enables the system to learn and maintain the occupant's desired comfort within an average of 5.6 minutes, with a training of 2.9 years while providing high comfort performance and energy efficiency.

1.3 CONTRIBUTIONS TO KNOWLEDGE

The overall contribution of the thesis is an integrated simulation of a heating, ventilation and air conditioning controller that learns from the feedback of the vehicle environment as well as the preferences of its occupant. The three main contributions to knowledge are:

1. Identifying a set of simple and understandable rules based on thermal comfort literature about occupant thermal behaviour (detailed in Chapter 2) that represents the changes in HVAC settings (Chapter 3).

2. A method for representing and validating occupant feedback behaviour by the use of empirical data (Chapter 4). The UBM model is based on a set of interconnected conditional probabilities resulting from the combination of a set of seven classifiers that has environmental parameters and equivalent temperature as inputs and the setting adjustment that a human makes to the HVAC system as outputs.

3. The User-Based Reinforcement Learning HVAC is an integrated system that learns from environmental and human feedback. The system uses a shaped reward based on the feedback provided by the UBM. The shaping reward is combined with the comfort and energy usage penalties. This thesis describes the UBRL HVAC controller and examines its performance when trained with

alternative SARSA algorithms in terms of average step reward, the percentage of time in which comfort is provided and energy use (Chapter 5).

1.4 PUBLICATIONS

The work described in this thesis has lead to the following publications:

Journal

- Alexandra Petre, James Brusey, Ross Wilkins: An examination of comfort and sensation for manual and automatic controls of the vehicle HVAC system. Accepted for publishing by SAE International

Conference Proceedings

- Alexandra Petre, James Brusey, Elena Gaura: User Feedback for the Improvement of Thermal Comfort in the Car Cabin. In Proceedings of ICAST (10) September. 2015, pp. 91-92

1.5 THESIS STRUCTURE

- Chapter 2 reviews the literature, detailing the use of human feedback in Reinforcement Learning, HVAC control, and thermal comfort.
- A closer examination of thermal comfort in the car cabin and how people behave in the car when experiencing discomfort is combined into a set of simple rules presented in Chapter 3 (contribution 1).
- Chapter 4 details the UBM and the approach used for developing, testing and validating the agent (contribution 2).

- Chapter 5 describes the probability-based model in the context of the system architecture, which is split into the Reinforcement Learning Agent and Cabin Environment. The capability of the UBRL system to learn from simulated user feedback is examined. This chapter presents an analysis of the learning speed, the quality of learning and reward maximisation of the controller (contribution 3).

- Chapter 6 presents a summary of the answers to the research questions and the conclusions drawn from the presented research, recommendations and directions for future work.

2

LITERATURE REVIEW

This chapter provides the theoretical background and support for the experimental work described in the following chapters. It presents the gaps in knowledge and the current developments of three major topics: i) Reinforcement Learning; ii) Thermal Comfort; and iii) Heating, Ventilation and Air Conditioning control.

This literature review begins with an introduction to the Reinforcement Learning technique. The chapter structure is shown in figure 2.1. It includes details of algorithms and application areas that include user feedback methods (Demonstration, Advice, Shaping, Training the Agent Manually via Evaluative Reinforcement). Subsequently, it highlights the potential of including occupant feedback within the context of the car cabin. The following section examines thermal comfort and the impact it has on the occupants of both buildings and cars, namely their thermal behaviour in terms of perception and preferences. While these aspects represent a platform that has been newly explored in buildings, it is rarely considered for car cabin comfort. The following section explores the state of the art in Heating, Ventilation and Air Conditioning (HVAC) control with an emphasis on Reinforcement Learning (RL) based systems, looking at potential avenues to improve and personalise thermal comfort. Moreover, the ways cabin passengers manifest their preferences related to the HVAC interface is presented. The final section is a summary of the identified avenues for exploration and their relationship with the work accomplished for the scope of this research.

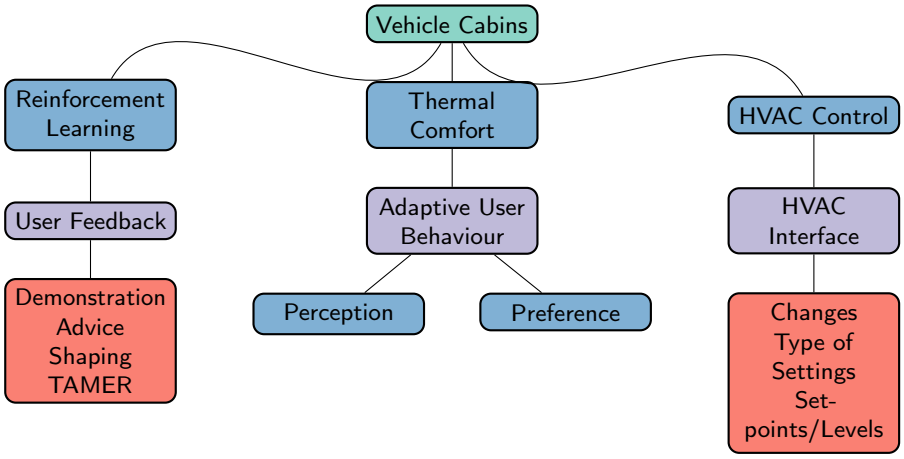

Figure 2.1: Chapter 2 structure with key points, bridging the gap between the user adaptive behaviour and reinforcement learning using the relationship between user feedback and climate interface control.

2.1 REINFORCEMENT LEARNING

RL is a machine learning framework that supports active learning from positive and negative rewards associated with the state of the environment and the actions chosen to alter this state. The environment is assumed to be a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ). The model of the environment is markovian if the transitions of the state are independent from the actions of the agent and the previous states of the environment [KLM96]. Within this framework, a following state s_{t+1} belonging to a set of states, S, is achieved at moment $t+1$ with the probability P, when action a_t of a set of actions, A, is taken at state s_t ($s_t \in S$) at time t. The reward function $R(s_t, a_t)$, represented by $R : S \times A \times S \longrightarrow \mathbb{R}$, relates to an immediate evaluation of the state s_t and action a_t pairs. RL agents aim to maximise the total discounted reward, also known as expected return G_t,

$$G_t = \sum_{k=0}^{T} \gamma^k R(s_{t+k}, a_{t+k}) \tag{2.1}$$

Discounting (the assignment of weights to the rewards) is used to bound the return. The discount factor, $\gamma \in [0\ 1]$, decreases the value of the future reward exponentially

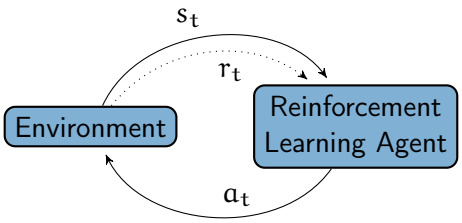

Figure 2.2: General Reinforcement Learning schematic

(the weights are greater for immediate than for distant reward). This determines the influence of future values on current predictions.

An agent acting within a framework of an RL algorithm [SB98] attempts to reach the goal of a specific task by environmental exploration as can be seen in figure 2.2. The behaviour of a learning agent is defined by a policy π ($\pi : S \rightarrow A$) which maps the states of the environment to actions taken by the agent.

The value function $V(s_t)$ determines the total reward (the return) that an agent accumulates starting with the current state and leading to future states [SB98]. It represents the expected return when starting with state s_t and following policy π.

$$V(s_t) = \mathbb{E}_\pi\{G_t | s_0 = s_t\} \tag{2.2}$$

For linear environmental models (model-based RL), the value function aids the agent to choose the action that will determine the following state with the highest value. For environments that do not have an exact model (model-free), state-action values $Q(s_t, a_t)$ are employed in order to determine action selection (equation 2.3). This action-value function is based on the relationship between the total reward and the state-action pairs using policy π.

$$Q(s_t, a_t) = \mathbb{E}_\pi\{G_t | s_0 = s_t, a_0 = a_t\} \tag{2.3}$$

The optimal policy π^* represents a policy that is higher than the alternative policies, which will lead to the maximum action-value function (maximising the

long-term total reward). The maximum action-value function is therefore considered optimal (equation 2.4).

$$Q^*(s_t, a_t) = \max_{\pi} Q_\pi(s_t, a_t) \tag{2.4}$$

Considering RL, two main approaches are available: policy-search and value function optimisation algorithms. Policy search focuses on finding adequate parameters for policy refinement in order to solve the problem of high-dimensional state and action spaces. Value function optimisation relies on modelling the return to obtain a course of action that leads to the desired states of the environment. The agent undertakes an action that not only results in a favourable immediate reward, but also in future states that have a long term effect on the policy, subsequently generating a high value for future rewards. This thesis uses the latter technique. As the problem is based on model-free RL, the agent relies on the exploration of the state-action space to learn a good policy.

There are two categories of learning problems: episodic and continual. Problems that have one or multiple terminal states are episodic. When the states are reached, the episode finishes and the state returns to its initial setting. Continual problems do not have terminal states, they continue indefinitely. The discount factor is useful for these problems as it avoids infinite reward attribution. An episodic problem can be extended to be continuous by introducing the final state as an absorbing one. For this thesis, the agent is trained using episodic scenarios, where the goal is to achieve the desired occupant comfort with the minimum steps taken. It is extended to a continuous scenario when testing if the HVAC controller can achieve and maintain the desired comfort in the case of cooling and warming the cabin.

2.1.1 *Algorithms*

There are two main problems when using RL: temporal credit assignment and generalisation [Lin92]. Temporal credit assignment is directly connected to decision making and performance improvement. This problem relies on how the agent

chooses a method in which credit is distributed to each state-action pair, after the agent has executed a set of actions and has arrived at a certain stage. The generalisation problem stems from an agent's capability to make decisions based on experience in an unexplored part of the state space. This problem appears when the state space is too large for complete exploration, such as in the real-world problems (e.g. riding a bicycle [RA98]). For solving the temporal credit assignment methods, related to dynamic programming, algorithms based on Temporal Difference (TD) are used.

State-Action-Reward-State-Action (SARSA)(λ) is an on-policy algorithm for TD control learning [SB98]. It is based on the agent that experiences the world with starting state s_t, action a_t, reward r_t, the next state s_{t+1}, and the following action a_{t+1}.

The rule for updating the state-action function for SARSA is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[R(+\gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

where α is the learning rate. The difference between this algorithm and the Q-learning algorithm [RN94; SB98] is that the state-action values, Q, are sensitive to the policy that is executed. In other words, the Q-value for the next state and action is chosen in the current state.

SARSA(λ) uses eligibility traces, λ [PMK01]. Eligibility traces are essential in training an RL agent as they are temporary records of the events that register changes in learning. They determine which states and actions are assigned a reward for the TD error [SB98]. The selection of different values of λ determines types of algorithms learnt, from SARSA to Monte Carlo [PMK01]. This work uses SARSA(λ), as it has the potential to solve continuous problems [SSR97] such as HVAC control.

However, SARSA (λ) is known to have slow convergence and its performance can degrade over time. For example, the agent unlearns the policy, or chooses a more complex action that triggers the same amount of rewards over a simpler action that

brings a final maximum reward. Expected SARSA [Van+09] outperforms both SARSA and Q-learning agents and improves the the learning rate of the agent.

Expected SARSA is an algorithm that has lower update variance. It improves on the learning performance by having higher learning rates. It unifies Q-learning [RN94] and SARSA as it approximates the optimal action-value function $Q^*(s_t, a_t)$ independent of the policy [Van+09]. It performs the step update by using the expected value of the action-value function ($\mathbb{E}\{Q(s_{t+1}, a_{t+1})\}$) instead of the function itself (equation 2.6). Therefore the update uses the state-value function (equation 2.5), instead of the action-value function $Q(s_{t+1}, a_{t+1})$.

$$V(s_{t+1}) = \sum_{a} \pi(s_{t+1}, a) Q(s_{t+1}, a) \tag{2.5}$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[R(s_t, a_t) + \gamma \sum_{a} \pi(s_{t+1}, a) Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{2.6}$$

By using the expected value of the following state-value function, the algorithm determines how likely the agent is to choose the next action under the current policy.

The problem with SARSA and Expected SARSA algorithms is that overly-optimistic values for $Q(s_t, a_t)$ are assigned by maximising over the estimates. The maximisation bias [Van10] can be avoided by using double learning. This method entails having two action-value functions $Q_A(s_t, a_t)$ and $Q_B(s_t, a_t)$ that are switched regularly, essentially doubling the memory requirements of the algorithm with no impact on the amount of computation required per step [GDH16]. For Double SARSA, the average of the two functions is used to determine the next action greedily (equation 2.7), with the update rule for the current action-value function $Q_A(s_t, a_t)$ using $Q_B(s_{t+1}, a_{t+1})$ (equation 2.8).

$$\pi = \underset{a \in A}{\operatorname{argmax}}(Q_A(s_t, a_t) + Q_B(s_t, a_t)) \tag{2.7}$$

$$Q_A(s_t, a_t) \leftarrow Q_A(s_t, a_t) + \alpha[R(s_t, a_t) + \gamma Q_B(s_{t+1}, a_{t+1}) - Q_A(s_t, a_t)] \quad (2.8)$$

Similarly, Double Expected SARSA has an update rule depending on the two action-value functions and calculating the state-value function, depending on which element A or B is used (equation 2.9). The difference between Double SARSA and Double Expected SARSA is that the latter has convergence guarantees.

$$Q_A(s_t, a_t) \leftarrow Q_A(s_t, a_t) + \alpha[R(s_t, a_t) + \gamma \sum_a \pi_B(s_{t+1}, a) Q_B(s_{t+1}, a) - Q_A(s_t, a_t)] \quad (2.9)$$

An alternative method for solving control problems by use of actor-critic algorithms [SB98] such as Deep Deterministic Policy Gradient (DDPG) [Lil+15]. This algorithm uses policy gradient to enable an agent to perform continuous actions. Despite this fact, the agent has a reduced frequency of action exploration. An improvement to this algorithm comes from the Trust Region Policy Optimization (TRPO) [Sch+15] that introduces a guarantee that the long-term reward does not decrease. This guarantee is maintained by introducing a Kullback-Leibler (KL) divergence constraint and a surrogate objective function. The problem with this algorithm is that is highly complex. A further optimisation for TRPO is Proximal Policy Optimisation (PPO) [Sch+17] algorithm that alters the surrogate objective function. A disadvantage with these methods is that they are off-policy, meaning that the optimal action is learnt from a policy that is not current.

Advancements in the gaming world use a combination of RL algorithms with deep neural networks (Deep Q-Network- DQN) to improve the learning speed and policy optimisation for a large number of games, achieving higher scores than gaming experts [Lia+16; MKS15]. Hasselt et al. [Van10] proposed a double Q-learning algorithm and showed that it outperformed DQN [Van+16], combating the problem of the latter algorithm of over-estimation. For robustness to randomised rewards, Ganger et al. [GDH16] introduced and examined the performance of

Double SARSA and Double Expected SARSA, which proved to have more stable action-values than SARSA and Expected SARSA.

These alternative SARSA algorithms are used to train the HVAC controller, as they can improve the learning performance of the system, and maintain a stable state-action function.

2.1.2 *Learning from humans*

User feedback is an essential step in improving Reinforcement Learning algorithms, especially for tasks that have an impact on the daily lives of humans. The interaction with an RL agent has two main objectives: improving the learning performance in order to ensure optimisation of the control as well as valuable research material on human behaviour. People, an integral part of the environment, differentiate themselves from their surroundings when interacting with the machine [KFS09]. DiGiovanna et al. [DiG+09] state that human-machine interaction enables the user and the agent to learn from each other. This depends on the quality of the information provided by the user and how this information helps the agent improve its learning process [Gri+13]. The main goal of this section is to explore the state of the art of RL based systems that include human feedback and how they achieve it.

Wang et al. [Wan+03] proposed the integration of user commands at different levels of the RL algorithm in order to ensure a faster learning rate and personalisation. The agent maintains its autonomy to prevent any sparsity of feedback, misinformation or breach of safety. The user feedback acts as a bias function that can enable the system to ignore certain actions, or alter the order in which it executes them. As an outcome, Wang et al. emphasised the utility of user commands for:

1. improvement of the speed of the algorithm by focusing the attention of the agent on specific aspects of the environment;
2. additional information about user preferences;
3. capability of altering the course of a task performed by the agent;

4. change of control strategies without complete control of the policy;
5. setting limits on the reward in order to avoid loops also known as *positive cycles*.

This supports the idea that training an agent with user feedback can produce significant improvement in the overall capabilities of achieving an optimal policy.

Knox et al. [Kno+12] noticed that human awareness in providing feedback can improve the learning performance of agents. Transparency can attract user involvement. Sharing specific information of the task execution can produce tailored feedback from the users, and enables them to adapt to the system requirements, throughout the training process [THB06a]. User requirements and the willingness to participate in the act of providing guidance can affect their contribution and involvement.

The present literature identifies the necessity of adapting the algorithms to take advantage of how users want to share their information in order to avoid preferences towards positive cycles (the agent chooses continually a set of sub-optimal actions that provide high reward) [KFS09; KS09; Kno+12; KS12a; NHR99; Ng03]. This can be achieved with previous knowledge about the available actions, which comes from sensor measurements and also from the user. Zaidenberg et al. [ZRM10] conducted a user preference study for a reinforcement learning based system in order to develop a methodology for the development of a planning assistant for cognitive load reduction.

Different roles come with human involvement that they assume in order to forge a link with the machine (figure 2.3). The three roles assumed by the user are: as a *teacher* (directly intervening in the policy [Mac+05]), as an *evaluator* (having a certain level of expertise [MS96] and predominantly giving negative reward [KFS09]) or as a *guide* (using the reward as motivation and being inclined to give positive rewards [TB06]).

Amershi et al. [Ame+14] present aspects of the teaching role of humans and the criteria for designing algorithms that rely on collaboration with the user, amongst which are several RL agents (software and robot based). When trainers

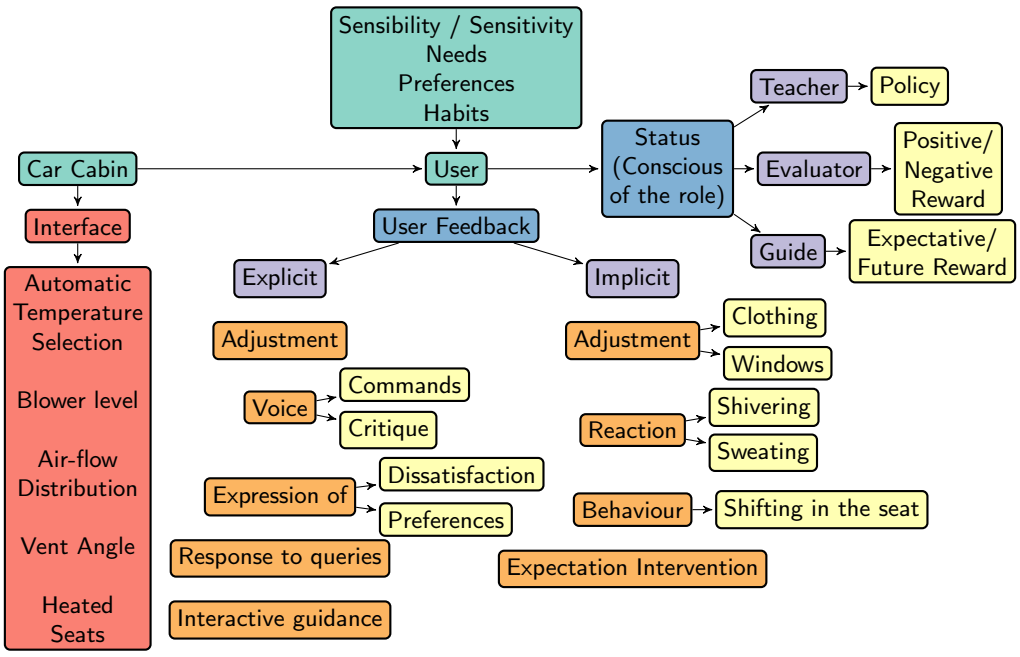

Figure 2.3: Schematic depicting the range of user feedback in the car cabin.

are aware of their role as teachers, they adjust their teaching behaviours in order to respond to the learner's requirements. These requirements are either inferred by the teachers or displayed through information bars or queries. Li et al. [Li+13] designed two interfaces in which the Training the Agent Manually via Evaluative Reinforcement (TAMER) agent's performance and uncertainty were displayed. An outcome was that the trainers became more involved in the training and provided more feedback. A transparent learning mechanism influences the user. The feedback is tailored to respond to the type of information provided by the agent with the goal to maximise performance.

Human perception is limited. When giving priority to the human perspective, the agent will often focus only on a subset problem that is important to the teacher. When assuming the role of teachers, humans tend to favour positive reinforcement [THB06a; THB06b; TB08]. This can affect the agent's performance by the formation of positive cycles for a discount factor of value 1. This causes the agent to stray from the primary goal.

Knox et al. [KS12a] demonstrated that varying the discount factor for episodic tasks influences the amount of feedback given to the agent. For human rewards, discounting needs to be appropriately chosen to user requirements. Knox

et al. [Kno+12] observed that there is no change in learning when the user is aware of evaluating an agent compared to providing comments (evaluation) of a recording. The agent can influence human behaviour in terms of feedback frequency without its performance being affected.

Thomaz et al. [THB06b], [TB08] also highlighted the interaction that happens naturally between humans and how it can be applied to humans and machines (guidance). The researchers discovered that not only can exploration and learning speed be improved, but also efficiency. Guidance has the potential to alter immediate actions and the course of task execution and assign rewards to future expected actions. The reward has a particular meaning to the trainer. If the meaning is not fulfilled by the agent, the guide will alter his or her behaviour and stop giving that specific kind of reward. Alternatively, humans provide information through their actions and gestures, therefore guidance can be manifested unconsciously.

A mixture of guidance and evaluation is considered in this thesis as an appropriate solution for integrating feedback from the occupants of a vehicle. Knowing what the users expect and the modalities in which they provide feedback can lead to further improvements in the performance of control systems.

2.1.3 *User interaction methods*

In research areas such as robotics and gamification, the potential of RL systems that receive knowledge from humans is already acknowledged [Ame+14]. This form of training has an impact on the learning performances of RL agents and allows the trainers to define correct real-time behaviour. What is more, the trainers are not expert programmers, but instead they convey knowledge via natural ways of communication. Similar to Sutton et al. [SB98], Knox [Kno12] also identified in his thesis three ways to convey knowledge: by demonstration, by voice and by shaping. Demonstration, a frequently used technique, relies on human behaviour being captured by sensors and reproduced by a robot [NM03]. Advice relies on voice commands or evaluations of a robot's actions [TMV10]. Shaping refers to the use of human defined rewards (either positive or negative) that help train a system

Table 2.1: Areas of application for reinforcement learning.

Area	Purpose	References
Robotics	Industrial	[PG98]
	Competitions	[MS96][NM03][Tor+10]
	Social	[DiG+09][Ber+06][THB06b]
Simulation	Games	[Isb+06][KSS11][KS12b][KB06][Tay+14]
	Control Systems	[HG01][Mel09]
	Smart Homes	[Dal+07][DD08]
	Vehicles	[Hin14][OHH07]
	Aerospace	[Abb+07][Kim+04][Ng+06]

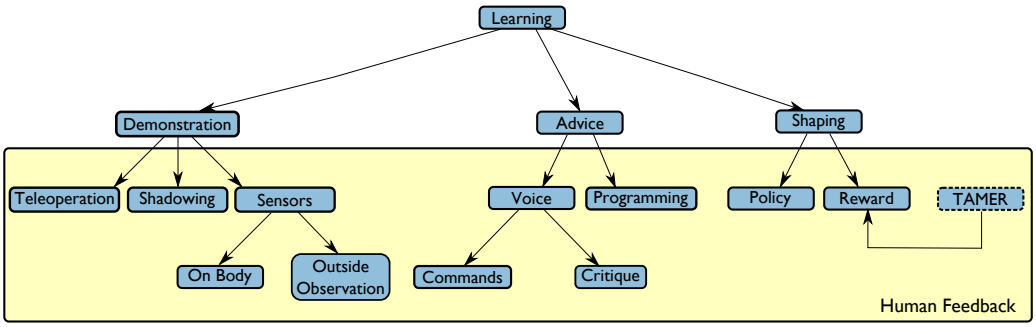

Figure 2.4: Types of learning with human feedback

towards achieving optimal behaviour [KB06]. Table 2.1 highlights the main fields of application that employ a form of RL. The works listed either use a form of user feedback or mention human feedback as an improvement to the existing system. Reinforcement Learning becomes a tool to support user interaction with the control system.

It is worth mentioning the TAMER framework as an interactive shaping method, which is associated with supervised learning through its structure (figure 2.4). It maintains distinct RL traits and is used in combination with the RL algorithm in order to improve the agent's learning ability.

Demonstration

Demonstration is widely applied to mobile robots and involves kinaesthetics for automating tasks in order to ease the workload of humans, especially for operations that require force and strain. Demonstration relies on transmitting the training data to the agent by:

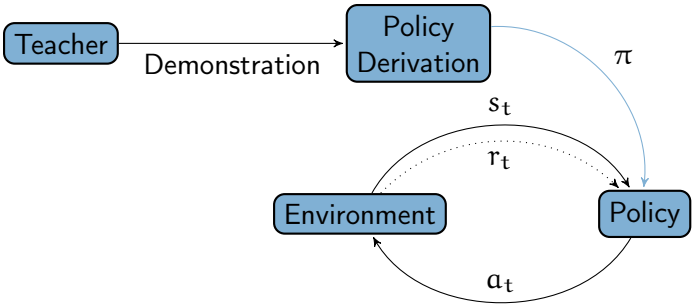

Figure 2.5: General scheme for demonstration.

i) teleoperation (the human directly operates the robot and the movement is recorded via sensors) [Ng+04; NHR99; OHH07];

ii) shadowing (the robot records the motions of the human through its sensors and tries to mimic them at the same time [NM03; Ogi+05];

iii) strictly using sensors to monitor and record the movement of the user (depending on the sensor placement, either directly on the body or in the surrounding environment) [Ber+06; Leo+11; LTM13].

Learning from demonstration is a technique that involves developing a policy from examples that are provided by a teacher and learned by an agent. These examples are represented by pairs of states and their corresponding actions that are recorded during the demonstration of the desired behaviour by the teacher, as captured in figure 2.5. Further work using this feedback technique is examined in appendix A.

Demonstration is targeting physical robots, relying on motion capture. It is not an adequate avenue for implementation in the context of control systems. The cabin occupant would be required to wear on-body sensors or a specific sensor needs to be installed in the car cabin to capture the occupant's motions. This method can be intrusive and inconvenient and can be distracting to drivers if its scope is other than capturing their driving behaviours.

Advice

Advice is the form of guidance through which humans can give recommendations to the RL agent. For this particular type of learning, there are two main categories

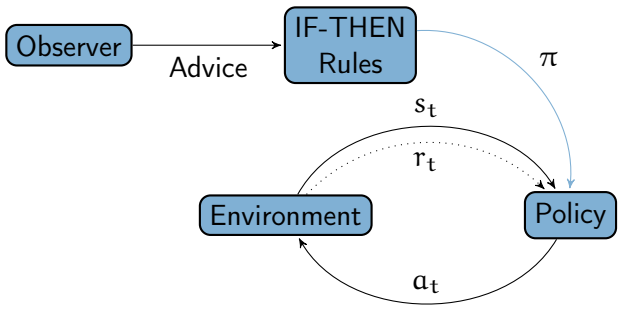

Figure 2.6: General schematic for learning via advice.

of providing guidance either by programming [MS96; Mac+05; Tor+10] (figure 2.6) or by voice recognition [TMP10]. A user can guide an agent by giving recommendations (a natural way of advising), which in turn have to be expressed in a manner that the system can decipher. Even though this type of feedback relies on potentially noisy signals it improves the learning capacity of the agent. Further works employing advice are presented in appendix B.

Advice is a feedback method that has the potential to improve the agent's learning and response time. It is provided in a direct manner (voice cues or programming rules). This means that expertise or instruction as to how the agent operates is required. Additional problems with this method are the fact that humans cannot provide advice continuously and the response to their suggestions can be delayed. This method is therefore not appropriate for capturing occupant feedback within the context of a vehicle.

Shaping

Shaping relies on training using feedback that is specifically oriented towards task performance. It serves to guide the learner towards a specific area of the state space. The incorporation of feedback using shaping is characteristic of dynamic programming. Its implementation into RL is popular, especially for executing goal-oriented tasks [DC98].

The term *shaping* comes from psychology and refers to a form of guidance in the training process of an animal by giving it rewards in order to achieve an intricate task. There are two main meanings to shaping for RL [KB06]: i) the agent learns by executing a simple task that gradually increases in complexity (altering the initial

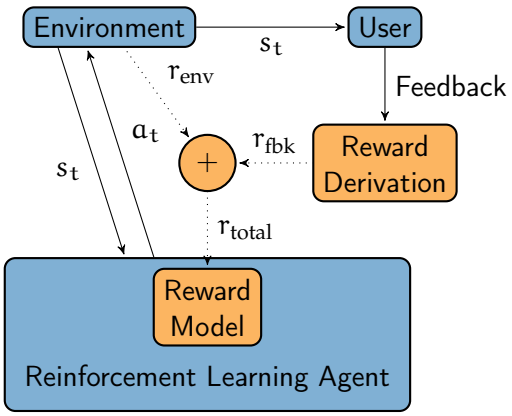

Figure 2.7: General schematic for learning using user feedback as an additional reward.

policy) [Gri+13]; ii) associating rewards to prior knowledge based feedback that guides the agent to a specific part of the state space, ensuring learning optimisation [TB06; THB06a; THB06b; TB08; KB06] (figure 2.7). A disadvantage of the first form of shaping (altering the policy) is that the agent learns a single task but does not generalise to similar tasks.

Reward shaping can take a general form by modifying the reward function ($R(s_t, a_t)$) to combine it with a shaping reward function ($F(s_t, s_{t+1})$). This changes the reward function to an extended problem that directly includes the occupant's feedback (equation 2.10).

$$\bar{R}(s_t, a_t, s_{t+1}) = F(s_t, s_{t+1}) + R(s_t, a_t) \tag{2.10}$$

Griffith et al. [Gri+13] introduced Advise, a Bayesian Q-learning approach that uses human feedback to alter the initial policy. The researchers considered the consistency of the feedback and likelihood that it might happen. Griffith et al. compared the algorithm with other RL techniques such as action biasing, control sharing and even reward shaping. The comparison shows promising results but the experiments were conducted for a static environment via simulation, without using any human feedback model. Reward shaping, as mentioned above, relies on introducing human feedback as an additional reward function, that together with the environmental reward, can reduce the learning time.

Tenorio-Gonzalez et al. [TMV10] explored the introduction of an additional reward based on the feedback given by the user. Tenorio-Gonzalez et al. explored verbal feedback. The cues were converted into rewards added to the environmental ones, forming the reward function. Three types of feedback were used in the on-line training of the agent: continuous, where information is provided constantly and the user has a concrete idea of the goal; occasional, when the user thinks that the agent needs it, generating a sub-goal once emphasis is put on a specific state; or it can be noisy (wrong, delayed or ambiguous feedback). The system improves its learning rate even in the event of noisy feedback. Once the learning process is refined, the user feedback can decrease over time. Even though shaping improves the learning speed, there are vocabulary restrictions. Only certain verbal cues can be used, therefore the user needs to be instructed what these cues are. Moreover, human speech can carry specific intentions and can have different intensities which cannot be interpreted by current sensor systems.

Thomaz et al. [THB06a] used an interactive virtual interface of a kitchen environment, where users could provide positive or negative rewards by clicking the mouse with the goal of teaching the agent how to bake a cake. The authors designed the reward function as a sum of the environmental and human rewards. The teaching behaviours of non-expert users in real-time was observed. It was noted that people have a propensity towards giving positive rewards and further use these rewards as a means for future guidance.

Once the learning pattern of the agent is known, the teachers adapt the way they give feedback in order to improve the learning performance. Thomaz et al. [TB06] developed two separate reward channels: one specifically for feedback and one for future guidance, proving that guidance is an essential step in reducing the agent's failure rate.

Isbell et al. [Isb+06] integrated an agent (Cobot) in the LambdaMOO social environment, enabling the users of the world to directly interact with it. The authors examined the evolution of the agent from the initial steps of gathering data, to adapting it for user interaction by using RL with a shaped reward. The particular reward function was directly affected by feedback, which could either be

directly expressed through written verbs or inferred from the approval remarks of the users. The agent learns a separate user policy for each individual based on separate state features, enabling Cobot to be an integrated part of the virtual world. The implementations of this method have not only shown an increase in the learning performance of the agent, but have also offered insights in human behaviour and the adaptability of the agent to user requirements.

In order to preserve policy optimality, the shaping reward proposed by Ng et al. [NHR99] is constrained by a potential function $\Phi : S \to \mathbb{R}$ (equation 2.11).

$$F(s_t, s_{t+1}) = \gamma\Phi(s_{t+1}) - \Phi(s_t) \tag{2.11}$$

Ng et al. [NHR99] proved that even though subsequent rewards (e.g. feedback rewards) are introduced to an algorithm, the policy remains unaffected using MDP under well-defined conditions. This enables the agent to learn faster and avoids the problem of *positive cycles* [NHR99]. Randlov [Ran00] demonstrated that for a finite MDP with $\gamma < 1$, the learning agent trained with a shaped reward that relies on the physics of the problem changing, will converge to an optimal policy.

Wiewiora et al. [WCE03] extended the potential function to include the action space. Two implementations were developed: look-ahead (the exploration of the advised actions has priority) and look-back (the potential function is a difference between the current and the previous situations that the agent underwent). The first method can be used in connection with the preferred state values, whereas the second is representative for action selection. Look-back (equation 2.12), and look-ahead advice (equation 2.13) are also implemented in this work, in order to examine if an agent that is aware of the past or future actions can maintain the occupant's comfort longer.

$$F(s_t, a_t, s_{t-1}, a_{t-1}) = \Phi(s_t, a_t) - \gamma^{-1}\Phi(s_{t-1}, a_{t-1}) \tag{2.12}$$

$$F(s_t, a_t, s_{t+1}, a_{t+1}) = \gamma\Phi(s_{t+1}, a_{t+1}) - \Phi(s_t, a_t) \tag{2.13}$$

Look-ahead advice has an impact on the policy, as a greedy decision also depends on the shaping function (equation 2.14). A greedy decision refers to the agent selecting the action which would result in the highest estimated reward.

$$\pi = \underset{a}{\operatorname{argmax}}(Q(s_t, a_t) + \Phi(s_t, a_t)) \tag{2.14}$$

Wiewora et al. [Wie03] also showed that for tabular-based temporal difference that uses an advantage-based policy for exploration, initialising the Q-function with the potential function is equivalent to executing reward shaping.

Grzes et al. [GK10] proposed two algorithms. The first one is for model-free RL and uses an abstract level of the value function for the potential function. The first algorithm uses high level generalisation of the states to learn the value function. The second, implemented for a model-based RL, is used to evaluate the value function [GK09]. The learning ofthe potential function and the policy happen at the same time.

Devlin et al. [DK12] generalised the shaping reward to potentials that include a time parameter (also referred as dynamic potentials), with a demonstration that the theoretical properties are preserved (equation 2.15).

$$F(s_t, t, s_{t+1}, t+1) = \gamma\Phi(s_{t+1}, t+1) - \Phi(s_t, t) \tag{2.15}$$

Harutyunyan et al [Har+15] introduced an arbitrary reward function that preserves policy invariance by adapting it to a form of dynamic advice potential. It incorporated external feedback through an additional value function. Laud et al. [LD02] used dynamic shaping to solve a walking problem in the robotics field. The shaping function relied on directly adjusting the parameters by means of an approximate quality function used as input. Laud et al. [LD03] also analysed the impact of shaping on the event horizon of the agent (how far the agent needsto look ahead in order to act according to a near-optimal policy), using an algorithm that explores the visible states of an optimal trajectory horizon.

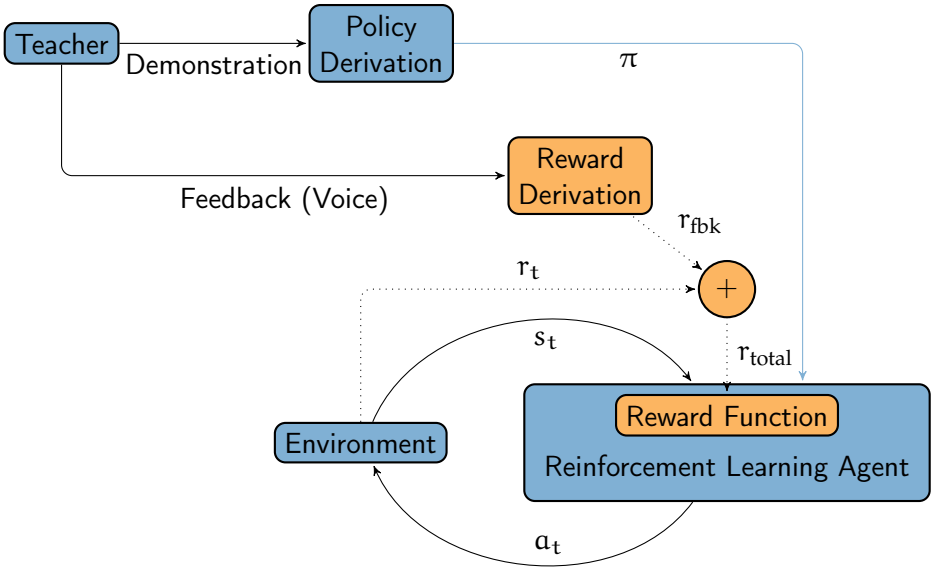

Figure 2.8: Reward shaping combined with demonstration.

Marthi et al. [Mar07] proposed an automatic shaping method that learns the shaped reward at the same time as the learning happens, by means of reward decomposition. A function approximation algorithm identifies a good shaping reward and learns how to perform multiple tasks by potential function adjustment. Additionally, Snel's et al. [SW14] algorithm uses feature selections based on k-relevance to learn a correct potential function for multiple task solutions.

Reward shaping is often coupled with an additional method of providing feedback either by demonstration or advice (figure 2.8). Knox et al. [Kno+12] observed that in 82% of cases, feedback is used in combination with other types of teaching (e.g. demonstration, advice) and is used in 58% of the time after testing. For example, Suay et al. [Sua+16] proposed two approaches using a potential-based function derived from human demonstrations. The first algorithm uses a mixture of Gaussian distributions fitted on the demonstration samples, efficiently capturing local information using heuristics [Bry+15]. The second algorithm is based on Relative Inverse Entropy RL [BKP11; Zie+08] and is robust against bad demonstrations [Sua+16].

Reward shaping can be used in combination with other teaching methods. It offers the opportunity to explore different effects of the direct involvement of the

users. This is a method that can easily be implemented through an alternative reward channel complimentary to the one from the environment and has a high potential to be implemented in real-world applications. The shaping methods used in this thesis combine the feedback from the cabin occupant concerning the states of the environment with the actions of the HVAC controller. Therefore potential shaping (for the state), look-back and look-ahead advice are used in this thesis.

TAMER

The *TAMER* framework [KS08; KS09; KS10; KSS11; Kno+12; KS12a; KS12b; Li+13] is a strictly human feedback based learning method used in combination with RL in order to improve the system's immediate performance. The TAMER framework [KS08] is based on the replacement of the environmental reward with credit directly assigned by the user pointing out good and bad behaviours. Since the human has knowledge about the goal, the agent is taught a policy that is relevant to the user. Knox et al. [KS08; Kno+12] characterise human reinforcement as "a moving target" due to the fact that the task execution is estimated as the agent advances, with humans giving rewards based on their judgement of the execution.

Feedback relies on both current and past states, therefore the transition function needs to be modelled accordingly. Humans consider what the agent intends to do, so the reinforcement relies on inferred current and future actions. Since the feedback channels for environmental and human-given reward are different, combining them can result in missing information. This is why Knox et al. support the idea of direct guidance by people (strictly human-given rewards).

TAMER uses interactive shaping, which essentially means that the agent is trained using positive and negative rewards [KS09]. Knox et al. [Kno+12] implemented this interactive shaping framework on a Tetris^{TM} game, concluding that the environmental reward function does not reflect the guiding behaviour of the participants. The reward function model based on supervised learning techniques, is used to choose actions that would provide the maximum immediate human reward [KS08]. The framework registers a faster learning rate due to the fact that it minimises the number of learning episodes [KS09].

The implications of using only human reinforcement are considered [KS08] in several experiments in which the reward is not completely replaced by the credit assigned by the human. The credit is similar to the gamma probability function [KFS09] and is based on how likely the user responds to the previous environmental state. Knox et al. [KS10] identified the potential of combining the human reward of the TAMER framework that reduces sample complexity and enables fast learning in the initial stages, with the MDP reward that determines a transition function that triggers optimal behaviour. The combination is done through several algorithms, user training happening before applying RL. From this experiment, action biasing and control sharing proved to have the best performance compared to standard SARSA(λ) [SB98] or TAMER [KS12b].

Additionally, Knox et al. [KSS11] examined the use of human intervention at any stage during the learning process (i.e. the agent learning from the environmental and user feedback at the same time) by using an alternative of eligibility traces [SB98] that identify feedback recency [KS12b]. The TAMER framework is suitable for the gaming environment, offering a good perspective on how humans provide feedback. Nevertheless, this method of incorporating human feedback requires the full attention of the participants. Therefore, it is not suitable for application within the cabin environment, where the occupant cannot provide constant feedback. Conversely, this thesis also maintains the idea that the cabin occupant has a separate channel for providing feedback than the physical environment.

2.1.4 *Vehicle domain implementation*

As mentioned above, human feedback is an important means of improving the learning rate and task execution performance of agents designed for interactive environments. Wang et al. [Wan+03] suggested the inclusion of feedback at a low-level (using devices such as joysticks or mouse) as well as a high-level (creating sub-goals or actions either by direct intervention—reward altering—or indirectly by verbal commands). This enables the agent to switch between different operation modes depending on the available feedback.

While gaming and robotics offer invaluable insight about human behaviour and the possibility to train machines using the knowledge of non-expert users, there are additional areas such as control systems (e.g. for comfort, appliances, light and ambient fixtures, or industrial work) that can be equally interesting. Out of all the feedback techniques presented above shaping makes direct use of the feedback within the reward function and can accelerate the learning without compromising the optimality of the learnt policy. In conjunction with the cabin, it is a potential solution for improving the performance of an RL based control system.

Belotti et al. [BE01] and Dey et al. [DDG01] suggested that improving the control efficiency does not rely solely on the involvement of the user but on how easily the final outcome is achieved. A balance needs to be struck between relying too much on the user (via warnings and queries) and a fully automated system. User interaction is only mentioned as one of the solutions that can improve such control systems, without an actual implementation. The fundamental problem is in identifying how the user (the occupant of a car) is providing feedback to an RL based system, what triggers this feedback and how this response can be modelled in the context of thermal comfort. The following sections provide more information on these issues.

2.2 THERMAL COMFORT

According to Haghighat [Hag02], people spend 90% of the time in enclosed environments. These environments can either be static (offices [LWG13], hospitals [Giu+13], gyms [RA14], homes [SMP12]) or dynamic (cars [MSR04], trains [KAA16], aircrafts [GOG15]). Considering the amount of time spent in these environments, an important factor is the level of occupant's satisfaction with their thermal environments, also known as thermal comfort [Sta94; 01]. To avoid discomfort, the core body temperature of 37°C needs to be maintained within a narrow range. When the body exceeds this range, it is exposed to danger from heat, or cold stress [Hag02].

The occupant's thermal comfort is affected by various factors in each environment. It is an extensively researched topic for several fields, such as engineering [SMP12], construction [DHM11], ergonomics[CB07] and design [Cro+15]. Thermal comfort

in static environments is impacted by changes in seasons, the outside environment, the type of construction [Hon+15b] as well as the type of ventilation system used (natural ventilated or air conditioning systems [Hon+15b; CB15; Hon+15a]). A particular environment where thermal comfort needs further attention is the vehicle cabin. Thermal comfort is one of the most significant aspects of car design as it affects the health, concentration, perception, performance and safety of the occupants [Dam+16].

Advancement in vehicle comfort has always relied on research conducted for other industries such as building and clothing [Wal+06]. For example, Ruzic [Ruz11] maintains the idea that cabin comfort needs to be at the standard of building environments. Hence, it is necessary to look at the advancements within the body of research for building interiors and develop an understanding of what can be applied to the cabin environment.

Arens [AZH06b; AZH06a] identified the potential of thermal environments that are transient, asymmetric, and adequately designed to provide occupants with an increased level of comfort. Additionally, Brager et al. [BZA15] suggested a transfer of comfort examination towards environments that are not uniform and dynamic. These key characteristics describe the car cabin environment.

The cabin is characterised as transient and non-uniform [Wal+06] because of the rapid fluctuations of parameters [Vol16] such as air temperature, air flow, humidity, and mean radiant temperature (also known as environmental parameters).

As humans perceive discomfort in terms of how much heat energy is lost from their bodies [AZH06b; FH09], parameters related to the person are also measured (personal parameters). These include the activity levels and clothing insulation that influence the level of comfort experienced by the occupants.

2.2.1 *Modelling*

In order to get a correct estimate of thermal comfort, all these personal and environmental parameters have to be measured [CM07; Dam+16; MSR04; Wal+06]. By grouping these parameters using equations of thermoregulation, various thermal

comfort models have been developed [HHA01; Nil07; Sta94; Zha+05]. Among these, the most notable is Predicted Mean Vote (PMV) [Fan73], developed in 1973 by P.O. Fanger. It essentially represents an index predicting the thermal vote of the occupants based on the equation of the thermal balance of the human body [Cro+15], using all the above-mentioned parameters. The model associates the comfort evaluations of the occupants (a seven point sensation scale [Sta94]) with a thermal vote. While PMV applies well to buildings, it does not measure correctly the comfort in cars [Dam+16]. Fanger also developed an equation estimating the level of dissatisfaction that occupants experience with the conditions of their thermal environment, named Percentage People Dissatisfied (PPD) [Sta94]. For buildings, the equation predicts that at least 5% of the occupants will be dissatisfied at any time. When applied to vehicles, up to 100% of the occupants will be dissatisfied [Dam+16], making PPD unsuitable for this environment.

Other models such as Standard Effective Temperature [GFB86] and Equivalent Temperature [NH03] aim to effectively estimate thermal comfort. For the evaluation of the thermal models, the estimated values are correlated with evaluations of the car cabin occupants. Among the models, the most suitable for car cabin comfort is Nilsson's [Hin+14; Nil07] equivalent temperature model.

Also, the analysis of thermal phenomena is often performed in simulation [Ji+14], in climate chambers with mannequins [MSR04], or using surveys targeting human experience, with the outcomes applied to stationary vehicles. This thesis aims to use a set of data that includes real-world driving scenarios that capture occupant comfort.

2.2.2 *Adaptive Behaviour*

The problem with the afore-mentioned models is that they cover the physical and physiological measurements with a standard evaluation of how comfortable people are. What they are lacking is the fact that thermal comfort is a personal and subjective experience [Sta94]. Furthermore, when given the opportunity, people behave in order to avoid discomfort [Sch+13] which allows them to adapt to en-

vironmental conditions. According to Schlader [Sch14], a powerful mechanism of thermal regulation is thermal behaviour, encouraging further exploration of the processes behind it. The decision to act, when experiencing discomfort at skin level, allows the occupants to maintain a stable core temperature. Zhang [Zha+10b] explained that depending on the thermal state of the body, actions such as immersing the hands in warm water can have a positive or negative impact on the subject's comfort. Haghighat [Hag02] further suggested that the quality of an environment depends on the occupant's response to the surrounding stimuli.

For the building environment, the focus has shifted to understanding and modelling forms of adaptive behaviour. According to de Dear [Dea11], research is currently targeting adaptive methods for local control. Moreover, Luo et al. [Luo+16] identified three levels of adaptation psychological, physiological and behavioural.

The notion of thermal comfort is in itself, vague and individualistic. Zhang et al. [Zha+10a] highlighted that how comfort is actually conceived by the human mind is not explored at all. What is more, Haghigat [Hag02] found that adaptive comfort models support the occupants to execute adaptive behaviours. Zhang et al. [Zha+10a] further elaborated that an action produces a pleasing sensation when it satisfies a need (e.g. when entering a cool environment, on a hot and humid day). This identifies the source of adaptive behaviours as alliesthesia [Dea11]. According to Cabanac [Cab92b], alliesthesia (the sensation of pleasure) is apparent only when the body is in a transient state. This sensation helps the body to return to stable conditions, also known as experiencing a neutral sensation. The subject's experience of thermal neutrality determines the end of any adaptive actions.

Research involving adaptive behaviour [Par02; GOB13; YLL09] is already being explored for the built environment. Behaviour-based models are used in simulation to explore the capabilities of building systems [LWG15]. As thermal comfort was initially an issue addressed in the military clothing industry as well as buildings [Wal+06], it is necessary to look at the applicability of user-trained control for the static environment, what aspects of human behaviour can be taken into account and which human modelling approaches have been proposed. This is because thermal comfort solutions initially start within this area and have been

subsequently applied to dynamic environments, represented in this thesis by the car cabin. As thermal comfort is an issue that affects driving performance [Hod13], there is a need to analyse the adaptive behaviours of the passengers in order to fulfil their requirements and preferences.

There are two aspects related to thermal behaviour present in the literature: perception and preference. Perception relates directly to the occupant's satisfaction with the environment. Preference refers to the actions performed by the humans in order to achieve thermal equilibrium or the desired elements of their thermal environment.

Perception

Factors that influence human perceptions towards thermal comfort in a building environment can equally be applied to a vehicle cabin. Considering human perception of the thermal environment, a series of studies show that for Indoor Environmental Quality (IEQ) thermal comfort is the most important factor that is influenced by time, climate, culture and social environment, and education; as well as the type of buildings and outdoor factors [FW11]. Furthermore, this perception varies with gender, as women have a higher predisposition towards being dissatisfied with their environment due to the fact that they are more susceptible to changing environmental parameters [Kim+13], whereas males are more predisposed to interact with the available control systems [Kar07].

Perceptions of comfort can also be influenced by the purpose in which the environment is used. In hospitals, air quality not set-point temperature, represents a measure of comfort [Giu+13]. The duration of stay in the respective environment plays an important role in the subjective perceptions of patients and doctors. Thermal neutrality (associated with a comfortable state) is also affected by the level of activity or by the type of ventilation systems. For example people in sports facilities prefer their environments to be warmer [RA14]. Conversely, Natural Ventilated (NV) building occupants have more tolerance to seasonal variations and a higher sensitivity to discomfort than those working in HVAC based buildings [LWG13].

This means that the amount of time spent within an environment has an effect on the actions of its occupants. Luo et al. [Luo+16] support this hypothesis by stating that people's perceptions of comfort can be directly linked to their thermal history and are led by their expectations. As occupants get accustomed to poorly air-conditioned environments, it influences their decisions not to make changes to the HVAC systems, which is synonymous with not altering their comfort situation.

Cabin occupants reach a thermally neutral state (steady state) in approximately 30 minutes [Zha+14b]. This is an appropriate window during which more frequent changes to the HVAC system are likely to happen.

Conversely, a closer look needs to be taken to the duration of a car journey. Jeffers et al. [JCR15] explained that an average journey in the United States of America is approximately 20 minutes. On the other hand, Johnson [Joh02] made the point that because of the short nature of car trips, the HVAC system is left on for the entire duration of the journey, since most of the journeys are short (about 92% of journeys last under 40 minutes). This thesis uses the same principle for testing the learning ability of the RL system within a time frame of 20 minutes, and investigating after how many simulated trips the system can learn.

Preference

More recent work considers the adaptability of humans [YLL09] when they experience discomfort and assumes that they change their behaviour [Sch14; Sch+13] in order to achieve thermal comfort either directly (changes made to the environment) or indirectly (e.g.drinking hot or cold drinks). Nevertheless, the adaptive levels of humans vary with the type of environment in which they are situated, hence an adaptive model cannot be standardised for all buildings and cars in a specific region, or for the entire life span of the respective control system [Son+15]. In both cases, humans are required to actively participate in training the control systems so that their preferences and requirements are satisfied in a personalised in an efficient manner.

In the available literature on mechanisms ensuring thermal comfort for the building environment [TC08], human input is often necessary to verify the pre-

dicted results (e.g. NEST and HIVE thermostats). Adjusting a thermostat can be taken as an example. Certain aspects need to be considered in order to enable setting the right temperature for the user: direct access and control of the system; easy interaction and useful information display [Pef+11]. Additionally, a basic knowledge of what the system is trying to achieve (e.g. time duration until the set-point value is reached) can enable the user to understand its workings and detect any problems [TC08]. When given the opportunity to act, individuals can adjust the amount of clothing they are wearing [Par02] or open and close the windows [KK07] instead of adjusting the thermostat. These actions also represent adaptive behaviours [GOB13]. Apart from the need for accessibility and integrability with other devices (e.g. mobile phones) and direct monitoring through the internet, an alternative is to control the system by voice [Pef+11].

Personalisation represents a key objective in the development of smart environments which can improve the occupant's comfort. It relies on customised controls that can prevent energy wastage. The main problem of personalisation is that there are often more than one individual interacting with the controls, leading to a higher amount of data needed in order to construct a preference model.

Existing work done on personalising intelligent systems proposes the use of fuzzy logic. Daum et al. [DHM11] created a model based on occupant preferences in order to develop a personalised system for controlling window blinds. Jazizadeh et al. [JMB13] ran a series of experiments when developing a user interface that revealed that the perception of comfort is mainly associated with set-point temperature. This served as a practical testing method for setting up a personalisation framework using a fuzzy logic based controller [Jaz+13].

Pattern monitoring [YB14] can be used to train a system depending on the users present in the environment. Occupant preferences are related either to the range of acceptable temperatures, which constitute the base for comfort distributions that train a Bayesian network [GTB15]; air flow levels for the summer and winter periods in a climate chamber [CB15]; or have associated selections of lighting and music preferences [KWA09].

Conversely, there is the question of what are the occupant's readily available adaptive methods in a limited space such as a car cabin, where concentration and focus need to be dedicated to the road and the act of driving. An answer comes from Brager et al. [BZA15]; that directly links the improvement of comfort with the occupant's relationship with the architecture and systems present in their surrounding environment, manifested in the form of personal control. This points the thesis in the direction of the HVAC system. This is why perceptions and preferences need to be taken into consideration when designing and implementing HVAC control systems. Frontczak et al. [FW11] not only identified that there is a need to examine the dynamic character of human responses, but also to customise the available controls according to their preferences. This supports the tendency for individual control, to which Ruzic [Ruz11] refers when explaining the functions of the HVAC systems in cabs (i.e. the system should deal with personalised comfort, rather than catering globally for all the occupants of a cabin). The following section explores the HVAC system and the state of the art, with an emphasis on what new avenues can be taken from the building environment and implemented within the car cabin.

2.3 CLIMATE CONTROL

The interior conditions of both built and vehicle environments are changed using HVAC control systems to fulfil the occupant's comfort needs. The term *control* is often encountered within the engineering and mathematical fields. It determines the dynamic behaviour of a device [DFT91], describing the relationship between the inputs and outputs of a system (plant). A controller is designed to capture the error signal between the desired and actual values of the output. This signal is used as feedback for the input of the system in order to enable the convergence of the output with its desired value (reference). Control theory can be used to model simple systems (e.g. thermostats) to complex ones (e.g. vehicle commands – the steering wheel, HVAC system or brakes).

This thesis examines the control of HVAC systems within the car cabin. The car cabin environment is characterised by rapid changes due to different external parameters [CM07] which determine its transient states e.g. solar loading and velocity. Due to these rapid changes, occupant comfort and well-being are affected [Cro+15]. Not only does the HVAC system determine thermal comfort, but also air quality [SC13]. Hence, vehicles incorporate management systems that control air circulation, purification, temperature and humidity.

While the use of actuation is typical for internal HVAC control [Scho8], the HVAC interface or panel deals with climate control (input or set-point control). The main objective of the interface is to maintain a temperature and humidity balance no matter the conditions present outside the car [Scho8], with a focus on temperature control. What is more, current systems have additional settings such as preventing window fogging, and preserving the relative humidity at healthy levels (45-50%) [Wanoo].

Vehicle passengers can set their desired set-points and the system will automatically output the highest levels of air (either heated or cooled) in order to achieve and maintain the desired selections [Scho8]. This causes a non-uniform temperature change from the head to the foot area in the car cabin. The high discrepancy between the thermal sensations experienced at different levels of the body (e.g. cold feet and hands, heated body) impacts the comfort of the occupants [HC04].

The use of a climate control system in this manner also leads to a high amount of energy being consumed [Cro+15], especially for hybrid and electrical vehicles [Che+15; KB14; NW14]. Hence, another constraint for HVAC design is energy efficiency.

Thus, the two constraints of the HVAC system are: maintaining the overall comfort of the passengers, while keeping energy consumption to a minimum. A study conducted in the United Kingdom in 2014 [JDP14] reported that only 31% interviewed drivers were satisfied with their cars' systems. In order to deal with the two constraints and improve the performance of the HVAC system, several machine learning techniques have been proposed.

2.3.1 *State of the art*

Farzaneh et al. [FT08] aimed to optimise the HVAC system in the automotive environment using fuzzy logic. The novelty was that PMV was used for thermal comfort prediction and included in the feedback loop of the climate controls. The PMV based controller was further optimised using a genetic algorithm and compared to a strictly temperature based controller. Its robustness was analysed for uncontrolled environmental variations, namely fluctuations in the outside temperature. Energy consumption was measured by monitoring the evaporator cooling capacity. A mathematical model of the Peugeot 206 control architecture was developed in order to determine blower power and temperature door positions. Their results proved that a PMV based controller outperformed the temperature based one in both comfort and energy consumption.

Kranz et al. [KNG12] used a black box approach to automate HVAC control. Data was gathered by sensor measurements, comfort evaluations and HVAC interface settings for blower, temperature, and flap positions on a Volkswagen Polo in South Africa [Kra11]. An artificial neural network was trained using this data in order to control the blower level and flap positions. Research on comfort preferences and draft evaluations were captured by means of a joystick based interface. The training algorithm used was conjugate gradient descent achieving a classification performance of 87%. Kranz's work was more oriented towards how to improve HVAC interfaces to better incorporate comfort evaluations, demonstrating that intelligent systems can be trained directly by the use of these evaluations. A linear function was used for modelling temperature settings depending on the ambient temperature, instead of including its control as for the blower and flap positions.

Conversely, RL can be used for the optimal or near-optimal control of comfort and the reduction of energy consumption [Bru+17; Hin14] for vehicle HVAC systems. A one-dimensional simulation of a car cabin was developed, using vectors for the cabin state, and the HVAC actions. The bounded reward, associated with the following state that the cabin achieves and the current state-action pair, relies on two controlled parameters: Nilsson's Equivalent Temperature for measuring

comfort [NH03; Nil07] and energy consumption. The system outperformed a fuzzy logic and a standard controller in terms of policy performance. Even though the system learns the controls and consequently provides a near-optimal policy, it does not account for the preferences of the individual. What is more, the RL system needs 6.8 years of training trials until a near-optimal policy is achieved, which is not feasible in a real-world implementation.

These machine learning based controllers ([Bru+17; FT08; Hin14; KNG12]) aim to refine comfort metrics and energy consumption without considering the fact that humans have specific habits and preferences concerning the control of HVAC systems. The current design of control systems has a limited capability to sense or infer certain aspects of human behaviour. This behaviour is contextual, individual, and cannot be accurately modelled. It is difficult to quantify and measure all the factors that drive the actions of the occupants. Therefore, the targetted aspects of the occupants' behaviours need to be precisely defined in order to be monitored. One of the main objectives of this thesis is to develop a set of definitions or rules that identify explicit thermal behaviour actions related to thermal comfort.

A control system can deal with basic aspects of context that are either repetitive, pattern-related or strictly defined and conditioned [BE01]. It can benefit from refinement by receiving feedback from the user, preventing it from making false assumptions (e.g. rapid heat reduction by blowing cold air into the cabin, a preferred interval of set-point temperatures). The most suitable system for including the occupant's feedback is an RL based HVAC system.

Human feedback can be defined as additional information or opinions provided by a user (human) about the agent's (machine) performance of a specific task. A series of questions emerge when considering human feedback: Which parts of human behaviour can be considered as feedback? What is missing from comfort modelling? To answer these questions, further research on adaptive models for the building environment is presented in the following subsection.

2.3.2 *Adaptive behaviour models for HVAC control in the built environment*

According to Perceptual Control Theory (PCT) [Pow+11], *control* refers to the capability to act upon the surrounding environment with the goal of maintaining a specific experience. This theory supports the modelling of a person's internal mechanisms as a negative feedback control loop [GWP11]. PCT depicts how an external stimulus perceived by a person (e.g. sensation, cognition, perception) is compared to a desired experience triggering a particular behaviour towards the external environment. PCT establishes the link between perception and action, outlining the fact that perception has an impact on the human-environment interaction [VP00].

Powers et al. [Pow+11] took the example of thermoregulation when the skin or body comes into contact with cold air, triggering shivers. Core temperature is the variable that the individual tries to control. The neural control system acts as a comparator and the action to counteract the error for example is the act of shivering. If the body is not capable of dealing with the amount of heat loss, a person will employ a different action such as using a heater.

Fabi et al. [Fab+12] examined occupant behaviour in order to improve comfort and energy efficiency. Karjalainen et al. [KL11] identified the need for users to control their environment and proposed a modular user interface for joint control over building or home systems, having both comfort and energy efficiency as motivation. Gunay et al. [GOB13] encouraged the use of occupant control to improve the robustness of control systems in offices, giving recommendations on how to build a human behaviour model depending on one or multiple variables.

The research of Fabi et al. [FAC15; Fab+15] consists of the verification and validation of behavioural models, analysing the difference between stochastic and probabilistic behaviour for set-point and window adjustment. Logistic regression is used to develop predictive models for opening windows and validating them against real-world data. Zhao et al. [Zha+14b] proposed a model that takes account of complaint behaviour related to the IEQ depending on the states of the individual (transient or steady state).

Langevin et al. [LWG13; LWG15] developed a model of occupant adaptive behaviour in the office environment using climatic, building interior and occupant preferences data.. Bayesian estimation, using Gibbs sampling, was applied in order to project the satisfaction, thermal acceptability and preference for controls. Four versions of the developed agent-based model (with clothing variation, standard acceptability, reordered behaviour, and with realistic ranges) were compared to random guess, logistic regression, Humpreys algorithm, and Haldi regression models. The full version of the agent-based model outperformed the alternative methods by having the highest balanced accuracy (70-76%). The alternative models overly-predicted window use, and under-predicted use of fans. Wong et al. [WMC14] also used the Bayesian estimation model in order to examine the level of adaptation to the environment, registering less dissatisfaction among the elderly, and in student classrooms. Hong et al. [Hon+15a; Hon+15b] identified the need for a joint framework that can define and plan for any experimentation on occupant behaviour providing a unified and standardised view.

Kim et al. [Kim+18] developed personal comfort models for seats using a set of machine learning algorithms: classification tree, Gaussian process classification, gradient boosting, kernel support vector machine, random forest and regularised logistic regression. For training and testing the models k-fold cross validation was used on combined data from surveys and trails with 38 participants. The model with the highest prediction accuracy (Area Under Curve (AUC) of 0.71) was random forest surpassing PMV models (0.5), the minimum amount of surveys necessary for model convergence was 64. The research demonstrated the potential of personal comfort adjustments as an alternative method for comfort prediction to using surveys. Kim et al. [KSB18] also proposed a framework for personal comfort modelling recommending the use of regression, decision tree, Bayesian and kernel algorithms.

Among the proposed intelligent methods of providing comfort that are strictly related to HVAC control, Dalamagkidis et al. [DD08; Dal+07] modelled a RL based controller for the home environment. It is based on a radial basis function for the feature vector, which is then multiplied with a weight vector. The reward function

is a combination of comfort, energy consumption, and air quality. The authors proposed a user simulator module based on Fanger's PPD [Sta94]. The performance of the controller was proven to be comparable to fuzzy logic controllers. Despite this fact, the agent's learning is slow (4 years). Moreover the algorithm (Q-learning) does not converge to an optimal solution in simulation.

Fazenda et al. [Faz+14] used a reinforcement learning HVAC controller in conjunction with a simulated human, based on an α-fuzzy logic. The simulated behaviour represents occupant's interaction with a thermostat. It is based on prior devised schedules for when the occupant is working or is out of the building and on the adjustment of the thermostat when the occupant is uncomfortable. Fuzzy logic was used for modelling set-point temperature selection. The controller was based on Q-learning with neural network adjustment of the weights and was used to examine two cases: a standard on-off control (Bang-bang control) and a set-point adjustment control. The HVAC controller could pre-heat the room given that the simulated occupant's behaviour did not change (i.e. the same schedule was maintained). While it has the potential of integrating occupant's behaviour to train the system, the caveat is that the learning is slow (80 days of training) and does not converge under any circumstances.

These models were developed in order to improve comfort control in both home and working environments, enabling designers to create more realistic scenarios in simulations involving the occupants. Human behaviour has a high degree of uncertainty, especially in the car cabin. Changes to the HVAC system are seldom, but are targeted to the occupants' comfort needs. Using probability distributions to model such behaviour has the convenience of varied certainty, while maintaining the bivalence of the response (a statement is either true or false). Modelling adaptive behaviour via probability distributions ([Fab+15; GOB13; LWG16; YLL09]) computes the likelihood of whether or not a specific behaviour happens, such as changing the settings.

Occupants' decisions to act are influenced by their perceptions, that vary from person to person. Nevertheless, certain preferences are exhibited throughout the literature and need to be compiled in the form of a set of rules, as people convey

knowledge of their perceptions via natural language. The rules can be expressed as conditional probabilities linking comfort and preference. A solution is a combination of Bayes inference with other machine learning techniques to identify appropriate posterior probabilities. Alternative machine learning algorithms can be used for modelling the occupants' actions such as neural networks, support vector machines, conditional inference trees, logistic regression and random forest.

2.3.3 *Adaptive behaviour in the cabin environment*

Numerous trials have been conducted in enclosed environments [Kim+13] using surveys [FW11] and sensor measurements [Hin+11], but fewer driving or in the car [KNG12]. An integral aspect of how people perceive their thermal environment is what actions they choose in order to adapt to it [Sch+13]. This aspect of thermal comfort is just emerging within the building environment [Fab+12; LWG16] and is not yet observed within vehicular research. It is mainly overlooked as it is considered a secondary task, priority being given to the task of driving [PBR07].

Models of human behaviour that go beyond physical and psychological characteristics, are vital for improving the experience of cabin occupants. The HVAC systems mentioned in subsection 2.3.1 rely only on the evaluation of comfort from estimated values (equivalent temperature, PMV) and ignore the fact that humans act when they experience discomfort. The occupants' adaptive actions can alter the estimated comfort values. For example, removing an item of clothing can change the value of clothing insulation. Human evaluations commonly rely only on the scales [Sta94] of thermal comfort and sensation. These are completed either after experimental trials take place, or during the trials by means of queries. In the car cabin, localised comfort is a significant problem, as a difference in temperature at the head and foot level produces more than 10% dissatisfaction [Hod13].

When experiencing thermal discomfort in buildings, people tend to use actions [Sch14] in order to achieve their thermal equilibrium [Hag02] (e.g. removing or putting on more clothes [Sch+13], or getting close to heating or cooling sources). In enclosed environments such as car cabins, there is a limited number of action pos-

sibilities [FF86], hence the subjective and preferential responses of the passengers are also limited.

Fugiglando et al. [Fug+18] aimed to develop a personal thermal comfort model for HVAC control in cars by using data collected from 10 cars through the CAN-bus. The data includes information on window opening, wind shield wiper activation, HVAC and air conditioning compressor turning on and off, heated seat activation, as well as external and internal temperatures, however the thermal comfort of the occupants was not measured. The results showed that the activation of the heated seats is correlated to the HVAC heating mode being selected (Pearson's R of 0.42), to using the air-conditioning in heated mode (Pearson's R of 0.46), and to turning on the the wipers(Pearson's R 0.36) . The method used for developing the personal comfort model was a regression tree model, that registered poor estimation performance. However alternative models are recommended for use such as support vector machines, artificial neural networks and random forest.

When cabin passengers experience discomfort they use adaptive methods to improve their state. The range of behaviour when experiencing thermal discomfort can be limited to:

1. Removing or adding more items of clothing.
2. Use of windows and additional heated surfaces (e.g. heated seats or steering wheel).
3. Changing the HVAC interface settings.

The motivation behind the first two points is vague as it cannot be solely related to thermal comfort, but also to alternative aspects such as ease of driving (for clothing) or smoking (for opening the windows). These aspects are detailed in appendix C.

HVAC Interface

The HVAC interface is conveniently placed as part of the instrument panel [PBR07; Sal01; Hor01] (figure 2.9). Specific legislation is designed for the placement of the control systems so as to enable drivers to easily reach for them [Foc06; NHT12; OIC15]. Symbols on the control panels are standardised in order for passengers to

Figure 2.9: Heating, Ventilation and Air Conditioning system panel for a Ford Escape, with set-point temperature, blower and vent control [Coo17].

easily recognise their use [Rie+13]. The types of panels can include push-buttons, knobs, or LCD screens. The interfaces offer the operator a range of features that can include outside air temperature measurements as well as input controls for heating and air-conditioning. Alternative interfaces employ either speech, handwriting and gesture recognition [Bla+10; HN98; Ich04; Rie+13]. A problem with these types of interfaces is that they contribute to driver distraction [PBR07; Ran+00; Sal09].

Three types of HVAC system are identified: automatic, semi-automatic or manual [Hod13]. The control interfaces have a head-up or down display [Rie+13]. All systems allow the direct adjustment of temperature, blower level or air distribution. A problem with the current HVAC systems is that even though passengers know what settings are available, they do not understand how the HVAC system operates internally once they make adjustments to the settings [Com16]. Hence people experience impatience and dissatisfaction [DeM15].

The HVAC system has multiple modes. Amongst these modes, *AUTOMATIC* has already built-in rules programmed in the electronic control unit. These rules enable the HVAC to blow in cold or hot air, depending on the state of the cabin environment. In their climate control descriptions, car brands detail the *AUTOMATIC* mode [Com16; Vol16], where a set-point temperature is the only additional setting that can be adjusted. This specific mode enables the climate control system to blow air until a pre-defined set-point temperature is reached, at which point it reduces the blower speed to a minimum (level 1 or 0 depending on HVAC sys-

tem). When an increase or decrease of the specific temperature is registered by a temperature sensor, the system resumes blowing air into the car cabin until the set-point is subsequently reached. For *FULL AUTOMATIC* selection the passenger is required to switch on the *AUTO* button and select the knob for fan control (also on *AUTO*) [Scho8]. Once a change in settings is registered (i.e. a change in the speed of the fan or selection of recirculation) the interface activates its semi-automatic function and the *FULL AUTO* mode is no longer available.

Cabin occupants are not always content with the *AUTOMATIC* settings and make their own settings selections through *MANUAL* mode. This enables them to: i) select a specific temperature; ii) select a blower level; iii) select airflow mode. For *MANUAL* settings, a set point is selected by increasing or decreasing the default temperature on a screen or turning a knob to the left or right. A general range for temperature selection is 15 °C to 32 °C with increments of one degree [Scho8]. Temperature regulation happens only when the set-point or knob is placed between the minimum and maximum points. Once either of these points is selected, the system outputs the maximum heated or cooled air. The blower level controls how much air is blown into the cabin (the levels control the speed at which the fan moves). Depending on the model of the controls, there are 5 to 7 levels of fan speed that represent the percentage of the duty cycle [NXP15]. For air flow or air distribution control, the actuator motor driver is positioned towards a body part towards which the air is blown (*Head*, *Foot*, *Both*, *Ambient*). It extends to 4 to 5 positions, or levels [NXP15]. The term *Both* defines the direction that the air is blown towards the feet and the head, whereas *Ambient* identifies a neutral position.

Knox [Kno12] proposed modelling the feedback delay as a negative uniform distribution between 0.2 and 0.8 seconds, the point of origin being the feedback to the learning system, and having a backward view of the event that is approximated. This thesis uses a forward view, with the event (in this case the human feeling discomfort) triggering the process of a change in settings. The following chapter will examine interactions with the HVAC interface concerning the adjustment of temperature, blower, and vent angle settings as a set of rules based on the literature.

2.4 SUMMARY

In this chapter, the main problem with state of the art HVAC controllers is presented. They do not offer the occupants the opportunity for personal control. Among the machine learning techniques for HVAC control, Hintea's [Bru+17; Hin14] RL controller could benefit from integrating interactions with the HVAC interface as feedback to improve its learning speed and the occupant's comfort. Exploring the feedback methods already encountered in robotics and gaming, shaping has a high potential of generalisability. Furthermore, it offers the opportunity for simulated interaction with the settings to directly train an RL controller by using the reward function.

As thermal comfort plays a significant role in how occupants act within their environment, a closer look at modelling adaptive behaviour in the context of the building environment was necessary. The main problem with modelling thermal behaviour is that it is individual and situational. Therefore, a clear identification is needed of which behaviour aspects are examined. A solution is to identify simple rules based on how occupants react in a car cabin that can be modelled as probability distributions. Classifiers can be used in order to estimate the probabilities of actions. Moreover, examining the opportunities of action within the car cabin, complementary to the driving task, gives insight into how to tailor the HVAC system to the individual. This chapter establishes that only changes in climate control can represent a first step towards providing direct occupant feedback to an HVAC system, as the motivation behind the use of other methods is ambiguous.

That is why the aim of this thesis is to model occupant preferences related to climate control and use them as feedback for training an RL based HVAC system to improve its learning performance and personalise comfort within the car. The following chapter details the literature-based rules that are considered in this thesis, with the purpose of developing a model of HVAC setting adjustments to be executed by the occupant.

3

OCCUPANT HVAC INTERACTION RULES

Chapter 2 examined the importance of a user feedback-trained Reinforcement Learning (RL) system and identified which method is suitable in the case of vehicle thermal comfort. The chapter established the link between comfort and the user's thermal behaviour, distinguishing that the most suitable form of feedback to the climate control is the adjustment of the Heating, Ventilation and Air Conditioning (HVAC) settings. The main focus of comfort literature is on which factors impact thermal comfort, and how it can be effectively modelled. Little is known about how occupants act in order to improve their comfort, or how to model this thermal behaviour in the context of a car cabin.

Karjalanien et al. [Kar07] stated that there are three main aspects that need to be considered for tailored control: the HVAC system; the available controls for the occupant and the strategy used. Moreover Zhang et al. [Zha+10a] recommended developing user models based on variable rules through a trial-and-error method. Based on this recommendation, this chapter describes a set of three simple rules based on the available literature. The rules define how occupants use their climate controls.

The research question that this chapter answers is : "*What is the set of simple rules that can be drawn from the thermal comfort literature on occupant thermal behaviour related to HVAC control?*".

The choice of rules is motivated by the fact that they need to be understandable and establish a probabilistic relationship between thermal comfort and HVAC control.

The author does not deem the set of rules to be exhaustive but a start towards a closer look at thermal comfort behaviour and its modelling potential.

The rules are as follows:

1. R1: When people are uncomfortable they are more likely to make changes to the HVAC interface than when they are comfortable (section 3.1).
2. R2: People are more likely to make changes to the temperature settings, than the blower and vents (section 3.2).
3. R3: Occupants prefer specific settings depending on the type of environment (either hot, cold or neutral, section 3.3).

3.1 R1: MAKING CHANGES TO THE CLIMATE CONTROLS

Schlader [Sch14] states that when people experience thermal discomfort, they choose to act. For example, when occupants experience discomfort with their thermal environments, they change the HVAC settings to ensure that a stable core temperature of 37 °C is maintained [Hag02].

There are several external factors mentioned in the thermal comfort literature that determine how users interact with control settings. Cabin occupants experience discomfort due to external environmental changes and solar radiation, that cause them to make changes to the climate controls [HC04]. When going from an extreme environment (e.g. cold or hot) to a neutral environment people tend to feel uncomfortable. People act in order to change their surrounding [KK07] until thermal equilibrium is reached, depending on the type of transition from one environment to another [Du+14; Liu+14]. Zhao et al. [Zha+14a] discovered that the behaviour and complaints of occupants are different when they are in transient state, than when they are in a thermally steady state (i.e. reaching a stable skin temperature). Moreover, according to de Dear [Dea11], occupant's adaptive mechanisms are activated when they are in non-steady states. Therefore, occupants are more likely to use HVAC settings in transient states.

Part of the decision to make changes to the controls is the fact that occupants' health and driving performance are impacted by a rapidly-changing environment, combined with the air quality in the cabin [Ruz11; Wal+06]. According to Luo et al. [Luo+16], their thermal satisfaction depends on their immediate experience and their thermal comfort expectations. In the building literature, personal control over the environment increases occupant's psychological comfort satisfaction [BCB98; BPD04; Hal+15; Hui+06; LWG12; PBZ18]. Moreover, Singh et al. [SBS15] stated that the thermal satisfaction of the vehicle occupants is related to their use of an available system (in this case, the HVAC interface).

The purpose of the changes is two-fold, according to Cabanac [Cab92a]:

1. to reduce displeasure;

2. to indicate desire for comfort improvement.

Ruzic [Ruz11] stated that it is not always guaranteed that the occupants will experience a higher level of thermal comfort by performing changes to the settings. This needs to be considered in the first rule.

Rule R1, states that occupants are likely to make changes to the HVAC interface when they experience discomfort. R1 is based on Johnson's [Joh02] hypothesis that when people experience dissatisfaction with their thermal environment, they activate the air conditioning. To further examine this, Johnson [Joh02] assumed that the Percentage People Dissatisfied (PPD) [ASH10] model determines the percentage of time for which the occupant uses the HVAC. Similar to Johnson, this thesis avoids examining the use of the interface for other purposes such as automatic mode, defrost mode, noise responses or opening and closing windows. Additionally, Zhang et al. [Zha+10a] incorporated a similar rule to R1: for transient environments or when occupants interact with the control system, their overall comfort is higher than the body-part comfort of the two most uncomfortable regions. The rule is used to model thermal comfort responses for the building environment. Fugiglando et al. [Fug+18] also infer the thermal comfort of the occupants from their actions assuming that drivers' changes to the climate controls are triggered by their discomfort.

R1 refers to the overall changes made to the HVAC interface when in MANUAL mode. It refers to the action of whether or not to make a change to the HVAC, therefore it can take a binary value (0 for no change, 1 for change). It also depends on the state of the occupant, i.e. if the occupant is comfortable or not.

The measure of comfort used in this thesis is equivalent temperature [NH03] (T_{eq}). According to Hintea [Hin14], the equivalent temperature model is the most accurate model for the car cabin environment. The overall body equivalent temperature is calculated using Bedford's equation [Nil07]:

$$T_{eq} = 0.522 \times T_a + 0.478 \times T_w - 0.01474 \times \sqrt{v} \times (100 - T_a) \tag{3.1}$$

where, T_a, is the air temperature, T_w is the average temperature of the surroundings (both expressed in degrees Fahrenheit), and v is the air speed (measured in feet/minute). In this work the following environmental parameters are used: for air temperature, cabin air temperature ($T_a = T_C$); the surrounding surfaces temperature ($T_w = T_{Int}$) and air speed v (calculated using the air flow $\dot{V}$). By surrounding surfaces temperature, this work refers to the average temperature of the cabin surfaces surrounding the occupant. The surrounding surfaces can be the instrument panel, the dashboard, the trim or the car seats.

The relationship between comfort and changes can be described as a conditional probability, represented as $P(Change = 1|T_{eq})$ and $P(Change = 0|T_{eq})$. Given the sparsity of the interactions with the HVAC interface, it is expected that the probability of change would be lower than the probability of no change.

According to Nilsson and Madsen [Nil07], the average overall body equivalent temperature at which occupants are comfortable is 22°C. Equivalent temperature is also used by Brusey et al. [Bru+17] for the comfort associated reward for the RL HVAC control system with a target temperature of 24°C. Conversely, several studies in comfort performance indicate that there is no exact temperature for comfort, as each passenger experiences it at different ranges [BZA15; Ruz11; Lim+12].

R1 identifies when changes are made to the HVAC system, therefore it is a global action similar to that proposed by Fazenda et al. [Faz+14] for estimating when an office worker is uncomfortable, or is present in the building. R1 represents the start of the interaction with the system, it is followed by the decision of what setting to select, presented in the following section. R2 builds on R1 in order to include more information about the type of changes made to the HVAC.

3.2 R2: SELECTING A SETTING OF THE CLIMATE CONTROL

Singh et al. [SBS15] identified the main control parameters of the HVAC system. These parameters are the set-point temperature, air speed, the vent numbers, the temperature in the cabin and the relative humidity. Temperature selection is most frequently used, as occupants are likely to associate their comfort strictly with the adjustment of this setting. Supporting this fact, Karjalainen et al. [Kar07] suggested localised temperature controls for buildings. Fazenda et al. [Faz+14] modelled the decision for a preferred set-point temperature of a building occupant as an α-level fuzzy set. Moreover, Ruzic [Ruz11] proposes that in taxis an increase in air temperature has the potential to increase the amount of energy saved. Singh et al. [SBS15] added that it is difficult to identify a single temperature that provides uniform comfort for all occupants of the cabin. Even for the automatic mode, set-point temperature adjustment is most commonly used. On the other hand the discomfort of the occupants can be accentuated by dual-zone systems catering to different desired temperatures [DeM15].

Conversely, Ruzic [Ruz11] recommended combining existent settings by selecting a range of values (set-points or levels) preferred by the occupants to increase their perception of comfort. Even though temperature selection is associated with thermal comfort, the blower level is subsequently adjusted by the passengers to control the amount of hot or cold air entering the cabin [CB15]. Knowing that the system blows in hot air, the cabin occupant selects the blower level, with no other interventions after this action [Pir76]. Additionally, high blower speeds are

recommended by Singh et al. [SBS15] in cool down conditions, with a vent angle of 30° for 4 minutes. A decrease in speed is expected after this duration.

This research is in line with the arguments of Singh et al. [SBS15]. While vehicle systems are advancing scientifically, including new technologies onto the HVAC can increase the weight of the cabin, impacting the engine. Thus the system becomes inefficient and costly. This is why the second rule, R2, averts attention to the three main adjustment options readily available to the cabin occupant: set-point temperature (T_{set}), blower level (B_{set}), and vent distribution (V_{set}). These three settings are available to the occupant on the instrument board, with recognisable symbols to differentiate between their purposes (equation 3.2).

$$Setting = \{T_{set}, B_{set}, V_{set}\} \tag{3.2}$$

R2 is related to R1 as the decision to select a specific setting depends on the decision to make a change. It also depends on the comfort level of the occupant (equivalent temperature). If the person does not make a change then no setting selection happens $P(Setting|Change = 1, T_{eq})$.

R2 outlines that the most frequently used setting is temperature (T_{set}) as comfort is often associated with temperature adjustment. Blower speed (B_{set}) follows, as high levels of speed promote the circulation of the air in the cabin and rapid changes to the thermal environment. Vent distribution (V_{set}) is the least favoured setting, as it depends on passenger's preferences and health. By using binary values for the setting selection and having a model for each setting selection, multiple selections within a single time step can be predicted. One of the points that this research will examine is if the equivalent temperature can be the only input parameter for modelling occupant behaviour or if elements of the thermal environment (e.g. ambient, cabin, surrounding surfaces temperatures, air flow) need to be considered.

R2 only deals with preferences for the types of settings, therefore the information contained is still insufficient, as each setting is based on a number of set-points or levels. These values serve as comparison measures to the sensed data captured

from the cabin environment. To further improve on this rule, once the decision to change at least one setting is made, the final step is to adjust the value of the setting.

3.3 R3: IMPACT OF THE ENVIRONMENT ON SELECTING A SETTING VALUE

Thermal comfort is directly associated with thermal alliesthesia (a term that is used to differentiate pleasure from a neutral thermal level) [Dea11]. Depending on the type of environmental stimuli, the extremities (e.g. hands, feet, ears) experience discomfort first. People who are sensitive to cold will try to adjust the HVAC settings towards those body parts [AZH06a; FH09]. Other body parts, such as the head for hot and abdomen for cold environments [AZH06a] are targeted when passengers try to achieve body part comfort by changing the orientation of the air distribution [Ruz11; SBS15]. Alternatively, Ruzic [Ruz11] recommended inputting air (either heated, cooled or dry) into the cabin by means of vents placed on the instrument panel and at the foot region.

In extreme cold or hot environmental conditions (e.g. tropical or seasonal e.g. winter and summer), the cabin temperatures reach extremely high or low levels [Ruz11]. Passengers use their HVAC system to cool down or heat up the car to improve their driving performance [JCR15; Joh02]. Nilsson et al. [NH03] found that there is a shift in set-point temperature preference when the seasons change. Zhang et al. [Zha+05] identified that specific body parts can be targetted for cooling (the head) or warming (the feet) depending on the types of environment (cold, hot or neutral).

One of the main complaints that people have when entering a car is that, when the environment is cold, the HVAC blows cold air until the engine is heated, hence they are uncomfortable [Com16]. When occupants are in a hot environment, they do not choose settings that will make their surroundings even hotter. Alternatively, for a cold environment, they will not select settings that will make it colder. R3 incorporates the following types of environment: extreme, such as hot or cold,

including warm (higher than 24 °C) and cool temperatures (lower than 20 °C); and neutral (between 20 °C and 24 °C).

For each type of environment, there is a specific set of selections that cabin passengers choose, or conversely avoid, related to set-point temperature, blower level, and vent orientation. When the selection of a set-point temperature is in place, the aim is to bring the cabin to the selected set-point. Due to the speed of the blower and the desired vent angles, there is a high discrepancy between the air temperature at different regions of the body, such as the head and feet, which accentuates occupant discomfort. Zhang et al. [Zha+05] established a link between the operative temperature of the environment and the acceptable limit between the head and ankles of the occupant. The limit is different than the one recommended by the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Standard 55 (2004) of 3 °C. An average temperature for a neutral environment is between 25.3 °C and 25.8 °C, with an acceptable variation of 7 °C between the head and foot regions. For cool or warm surroundings, stratification is not acceptable because it causes body-part discomfort (i.e. cold feet, warm head).

According to Zhang et al. [Zha+05] the asymmetry of the environment is not the main cause of discomfort but the actual local discomfort experienced at the extremities. Therefore an occupant's first target is a local comfort when aiming to obtain overall body comfort. According to Zhang et al. [Zha+10a], localised comfort is impacted by the comfort perceived at other regions of the body, and the whole body comfort. Furthermore a cool breathing zone is favoured in all types of environment, whereas a warm one is not tolerated. A warm foot region impacts the overall comfort of the subjects. Localised cooling of the trunk is a problem, as excessive cooling of the chest and lower back can further cause discomfort.

The sensations at specific regions such as the face, head, chest, lower back, hands, and feet have clear associations with discomfort. Nakamura [Nak+08] stated that cooling of the face is favoured for warm environments. For cool environments, heating the chest and abdomen produced a comfort response, whereas warming or cooling the face did not have any impact. Therefore,there is a clear preference for specific areas of the body to be cooled or warmed, depending on the type of

environment. The occupant can target specific regions of the body by adjusting the angle of the vent. Walgama et al. [Wal+06], also observed that the regions of the body that are in contact with the seat are not directly impacted by the air flow.

Concerning sensation reports, there is a phenomenon called overshoot that occurs when the occupant passes from an extreme into a neutral environment and vice-versa, or transitions into another extreme one. Thermal sensation vote overshoot means that the reported sensation scores exceed the final steady-state values. For example, occupants can report a sensation vote of "+3", associated on the ASHRAE 7 point sensation scale with the term "Hot", before reporting a steady state value of "0", which corresponds to a "Neutral" sensation [ASH04]. Liu et al. [Liu+14] examined the effects of step changes in transient environments, namely from warm to neutral and back to warm. By looking at mean skin temperature and overall sensation, Liu et al. [Liu+14] discovered that a thermal sensation vote overshoot is present both when going from warm to neutral, and from neutral to warm environments. Skin temperature reaches a steady state after 10 minutes in the step down case and after 20 minutes in step up. Additionally, skin temperature is not a direct determinant of sensation and comfort, but heat loss at skin level is. Du et al. [Du+14] share similar findings for the transition from cool to neutral, and neutral to cool environments in terms of heat loss. The thermal sensation vote overshoot is present only for transitions from cool to neutral environmentswith the skin temperature reaching steady state after 10 minutes. Whereas in the neutral-cool transition steady state is still not reached even after 20 minutes. Du et al. [Du+14] recommend that the set-point temperature difference between the cold environment and the air-conditioned environment is less than 5°C. Moreover, alliesthesia can be a factor that alleviates the sensations of the participants for the cool-neutral transitions.

Horikoshi [HF93] examined the step change transition from cold to hot environments and vice-versa. Horikoshi [HF93] found that the comfort level rapidly deteriorated when moving from a hot to a cold environment. According to Horikoshi et al. [HF93] human response is non-linear, with a delay in response when transitioning from cold to hot environments. An important factor influencing both comfort

and sensation votes is represented by sweating, which started immediately as the participants entered the hot environment, and stopped when transitioning to the cold one. The measured skin temperature reached steady state after 10 minutes in both cases, with an increase in comfort when transitioning from the cold to the hot environment.

Since vehicle cabins represent transitional environments, it is rarely possible for the occupants to experience steady state thermal sensations. Often passengers perceive the environment as being extreme, when the car is left under the sun in summer or left outside in winter. For warm up and cool down processes, occupants experience transience not only due to the movement of the car, but also due to the activation of the HVAC system, that aims to rapidly establish a neutral environment. The type of behaviour exhibited in such cases is likely to trigger overshoots in the selections of set-point temperatures. To be noted that according to Luo et al. [Luo+16] occupants can easily adapt to solutions of improved comfort.

Related to climate controls, there are specific actions that take place when extreme environments have an effect on the car and rapid cool down or warm up takes place. Jeffers et al. [JCR15] maintain the idea that occupants activate their HVAC systems because of the impact of the environmental conditions, the driving schedule, the settings of the interface and the cool-down time. Once the temperatures in the cabin reach steady state, the impact is reduced. Singh et al. [SBS15] examined the cool down process of a car left in the sun using different types of vents. They recommend an initial maximum speed and high air flow for approximately 3-4 minutes, after which a reduction is necessary. Currle et al. [CM00] recommended the use ofside and centre vents, with a 30° direction. Moreover, Jeffers et al. [JCR15] provided a description of the settings they used in their cool-down experiments: blower speed at level 7; maximum recirculation for the panel vents (alternative vents being turned off); a set-point temperature lower than 60 degrees Fahrenheit (approximately 16°C). The process is similar at warm-up with maximum blower speed and high-temperature set-point selection.

R3 is related to the occupant's decision to select a specific value for a desired setting. The three types of settings (temperature, blower, and vent) have different

Table 3.1: Blower speed levels reflecting the a percentage of maximum blower speed , according to [AMC].

Blower Level	Percentage of Max. Blower Speed
1	0.4%
2	12.5%
3	25.0%
4	37.5%
5	50.0%
6	62.5%
7	100.0%

values depending on their purpose within the cabin context (equation 3.3). The temperature varies from 16°C to 28°C in 1 degree increments (T_{val}). Blower speed (B_{val}) has 7 levels corresponding to a percentage of the maximum blower speed (table 3.1). Vent distribution (V_{val}) has four levels: Ambient, Head, Both, Foot. Instead of these classes the corresponding angle at which the vents are oriented is used: 0°, 30°, 60°, 90°.

The probability of a value being selected depends on the occupant's choice of setting (activating R1 and R2) and the level of discomfort felt (equivalent temperature).

$$Value = \{T_{val}, B_{val}, V_{val}\} \tag{3.3}$$

Even when the overall body equivalent temperature is within the comfort range (22-27°C [NH03]), there is a probability occupants will still make a change depending on the preferred setting and the rate of discomfort $P(Value|Setting, Change = 1, T_{eq})$.

This rule also introduces the direct impact that the thermal environment has on the decision making process. Depending on the type of environment (either

hot: $T_{Env} > 24$, neutral: $T_{Env} \in [20, 24]$ or cold: $T_{Env} < 20$) the occupant is likely to select from a reduced range of settings.

$$T_{val} = \begin{cases} a \in \{24, ..., 28\} & T_{Env} < 20 \\ a \in \{20, ..., 24\} & T_{Env} \in [20, 24] \\ a \in \{16, ..., 20\} & T_{Env} > 24 \end{cases} \tag{3.4}$$

$$B_{val} = \begin{cases} b \in \{5, 6, 7\} & T_{Env} < 20 \\ b \in \{1, 2, 3, 4, 5\} & T_{Env} \in [20, 24] \\ b \in \{5, 6, 7\} & T_{Env} > 24 \end{cases} \tag{3.5}$$

$$V_{val} = \begin{cases} c \in \{60°, 90°\} & T_{Env} < 20 \\ c \in \{0°, 30°, 60°, 90°\} & T_{Env} \in [20, 24] \\ c \in \{30°, 60°\} & T_{Env} > 24 \end{cases} \tag{3.6}$$

R3 is the final rule in the context of the occupant's interaction with the HVAC system. To reiterate, the scope of this thesis is narrowed to a set of simple rules related to the available settings of the HVAC interface.

3.4 LIMITATIONS OF THE RULES

This chapter presents the three main rules based on the available thermal comfort literature and seeks to further use their combination towards a first attempt at modelling occupant interaction with HVAC controls. The rules represent a start for the examination of what are the available thermal adaptive actions of vehicle occupants.

The rules can also be expanded to cover more aspects of the cabin environment than the HVAC interface. The three rules are essential but not all-encompassing, as the rule-base can be extended to incorporate other control features (e.g. heated seats or wheel) and other features of the cabin environment (e.g. humidity or air quality). Given the fact that one of the main objectives of this thesis is to develop a rule-based model for a simulated occupant that changes the HVAC settings, the rules derive conditional probabilities that specific actions are selected by an individual.

3.5 SUMMARY

Examining the climate control selections of occupants can shed light onto their preferences. As thermal behaviour is restricted in vehicles, this thesis hypothesises that there are three rules that answer the research question: "*What is the set of simple rules that can be drawn from the thermal comfort literature on occupant thermal behaviour related to HVAC control?*". The first relates to the opportunity to make changes to the HVAC system interface in order to control the environment (section 3.1). The changes made are related to the three main HVAC settings: set-point temperature, blower level, and vent distribution (section 3.2). The third and final rule covers the value selection for the desired setting (section 3.3). The three rules offer a clear and simple trajectory for how people interact with the HVAC control interface.

As the rules derive conditional probabilities, chapter 4 presents the available opportunities for modelling interaction with the HVAC system by means of the probabilities based on three rules.

4

THE USER-BASED MODULE

Automatic climate control systems are not fully autonomous as they enable the user to change their control behaviour via i) target temperature adjustment; ii) turning on and off the vents, or adjusting their angles; iii) the air distribution selections; iv) fan speed selection. When designing a system that allows user feedback, it is essential to define the nature of the feedback and then design algorithms based on this definition. In chapter 3 three main rules for occupant interaction were identified from the available literature. This chapter aims to construct a model that mimics the occupant's interaction with the Heating, Ventilation and Air Conditioning (HVAC) interface and examine if the model can outperform alternative simple models such as a neural network or fuzzy logic.

The main contribution of this chapter is a hybrid model, named the User-Based Module, combining a set of seven classifiers that mimics the occupant's decision making process of selecting HVAC settings. The model is tested and validated against real-world data captured under a set of driving trials. Each step of the process is based on a probabilistic rule that is then modelled using classification methods. Each model is then compared with alternative classifiers by means of error metrics. This chapter provides the following:

1. the method behind the mathematical development, experimental procedure, testing, and validation (section 4.1);

2. results and discussion for the set of rules (section 4.2);

3. the hybrid model architecture and comparison to a fuzzy-logic rule based model and a simple neural network (section 4.3).

4.1 MODELLING APPROACH

4.1.1 *Bayesian Probabilistic Method for the Classification Problem*

Chapter 3 identified three rules of interaction of occupants with their HVAC interfaces. The purpose of defining these rules was to identify the probability distribution of the possible actions that the occupant can take when feeling uncomfortable. The way that people act is not defined by the initial conditions and parameters of the environment, there are aspects of their behaviour that are difficult to determine, sense, or model (e.g. an occupant can change their mind as to which setting to select). Given the lack of further information, the actions are modelled stochastically. Therefore R1 determines the probability of change or no change given equivalent temperature. R2 determines the probability of a setting being selected given the decision to make a change and the equivalent temperature. R3 is represented by the probability of value selection given the preferred setting and the equivalent temperature. For R3, additional environmental features affect the probability distribution over the possible values for the respective setting.

Within this framework, modelling the desired settings selected by the passenger becomes a classification problem. For an already known distribution of the data, Vapnik [Vap98] recommends using a Bayesian classifier. The classifier relies on the latent distribution of the inputs to outputs being modelled as a conditional density function [NJ02] (often Gaussian). The Bayesian classifier essentially attributes the input to a specific class using the posterior probability derived from the relationship between the outputs and inputs based on the Maximum A Posteriori (MAP) decision rule [GL94].

The framework of the hybrid model is based on n feature vectors of dimension m, $X = [x_1, ...x_n]$, where $x_i \in R^m, \forall i \in \{1, ..., n\}$ as input. The features include: the cabin temperature, the internal mass temperature of the cabin, the ambient

exterior temperature, the overall-body equivalent temperature, and the cabin air velocity. A subset of the set of features D, which represents equivalent temperature, $D \in \{d_1, ..., d_n\}$ is used as initial input for the models. It is then expanded to the entire feature set depending on the performance of the models.

Each rule was initially modelled as a binary classification problem. The data is appropriately labelled, indicating if each rule is activated or not, then separated into two classes $C = \{C_1, C_2\}$.

Given the set of two classes C, let there be a label for each class, using the MAP decision rule the most likely class to be selected $C^*(D)$ is given by:

$$C^*(D) = \underset{C_k}{\operatorname{argmax}}\, P(C_k|D),\ k \in \{1, 2\} \tag{4.1}$$

In order to compute the posterior probabilities $P(C_1|D), P(C_2|D)$ the classifier relies on the Bayes rule:

$$P(C_1|D) = \frac{P(D|C_1)P(C_1)}{P(D)} \tag{4.2}$$

similarly for C_2. The following sections examine the applicability of the Bayesian classifier on the first rule.

4.1.2 *Modelling Changes to the HVAC Interface (R1)*

The labels are binary for R1, which means that $C_1(D) = \text{no change}, C_2(D) = \text{change}$, hence $y = \{0, 1\}$ is the label for changes and no changes corresponding to $\{C_1(D), C_2(D)\}$.

By definition:

$$P(y|D) = \frac{P(y \cap D)}{P(D)} \tag{4.3}$$

The joint probability $P(y \cap D)$ can also be written as the product of the probability of change $P(y)$ and the conditional probability $P(D|y)$:

$$P(y \cap D) = P(D|y) \times P(y) \tag{4.4}$$

The probability of equivalent temperature can be calculated using the marginal rule:

$$P(D) = \sum_{y} P(D|y) \times P(y) \tag{4.5}$$

Using the Bayesian Rule, the posterior probability for estimating change given the equivalent temperature can be calculated using 4.4, 4.5 in equation 4.3:

$$P(y|D) = \frac{P(D|y)P(y)}{\sum_{y} P(D|y) \times P(y)} \tag{4.6}$$

From equation 4.6 the marginal probability can be calculated using equation .

$$\sum_{y} P(D|y) \times P(y) = P(D|y = 1) \times P(y = 1) + P(D|y = 0) \times P(y = 0) \tag{4.7}$$

Hence, obtaining the posterior probability $P(y|D)$ of a change occurring when the occupant is uncomfortable relies on the conditional probability $P(D|y)$. Given that there are two class labels, there are also two conditional distributions $P_y(D)$, the dataset being separated into two depending on the labels (equation 4.8).

$$P_y(D) = \begin{cases} P_0(D) & P(D|y = 0) \\ P_1(D) & P(D|y = 1) \end{cases} \tag{4.8}$$

The general assumption is that the conditional probability is already known, and can be used to calculate the posterior. In the case of the User-Based Module (UBM), the set of conditional probabilities $P_y(D)$ is modelled as a mixture of Gaussian distributions [FLJ99] based on a real-world data set presented in section 4.1.6. The prior distributions are calculated by the relative frequency of the data.

Each probability is defined as a mixture model of K components:

$$P_y(D) = \sum_{k=1}^{K} \alpha_k P_y(D)_k \tag{4.9}$$

where α_k represents the k-th component mixing weight, with the property $\sum_{k=1}^{K} \alpha_k = 1$. When substituting each distribution by a Gaussian probability density function of form $N(\mu, \sigma^2)$, formula 4.9 becomes:

$$P_y(D) = \sum_{k=1}^{K} \alpha_k N(D|\mu_k, \sigma_k^2) \tag{4.10}$$

where μ_k is the mean and σ_k^2 is the variance of component k.

For the component fitting, the Expectation-Maximization Algorithm [GH+96] is used from the Mixtools package (R-library) [Ben+09]. The iterative algorithm consists of three steps: initialisation, expectation, and maximisation. It starts by using an initial estimate of the hidden parameters $\alpha_k, \mu_k, \sigma_k^2$ for the selected K components. The maximum number of components was $K = 5$ determined using bootstrapping [You08]. Then it iteratively updates the parameters according to the expectation (E) and maximisation (M) steps, until log-likelihood convergence. The E-step calculates the expected value of the log-likelihood function, while the M-step generates up-dated values of the hidden parameters, which maximise the expected values of the log-likelihood.

The likelihood denotes the goodness of fit of the model $l(\alpha, \mu, \sigma|D) = P_y(D)$, and it is generally modified to calculate the log-likelihood (equation 4.11).

$$\log(l(\alpha, \mu, \sigma|D)) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \alpha_k N(D|\mu_k, \sigma_k^2)) \tag{4.11}$$

The higher the value of the log-likelihood the better the model will fit.

In order to test how many Gaussian components are needed to estimate the conditional distribution of equivalent temperature given changes, and no changes, Kullback-Leibler (KL) divergence is used 4.12:

$$KL(P||S) = \sum_{D} P_y(D) \log \frac{P_y(D)}{S_y(D)} dx \tag{4.12}$$

where $S_y(D)$ is the conditional distribution for the actual data points, whereas $P_y(D)$ the conditional distribution of predicted values from the model. Between 2-5 component models were developed, the maximum component number is determined by bootstrapping. The final model uses Bayes inference to determine the prior distributions of change and no change using the Gaussian mixture with the lowest KL divergence fitted to the input.

A set of binary classifiers were selected for comparison with the Bayesian inference model available in the caret package in R [KJ13]. The selected classifiers are: naive Bayes (nb); support vector machine with radial basis function kernel (svmRadial); neural network (nnet); penalised logistic regression (plr); the Bayesian generalised linear model (Bayesglm); conditional inference trees (ctree) [KJ13]. The classifiers are mentioned in the building literature for designing human behaviour models (chapter 2).

The four main objectives for R1 were:

1. To determine how many Gaussian components are needed to estimate correctly the equivalent temperature distribution using KL divergence.

2. To determine if the probability of change and no change given a specific value of equivalent temperature can be correctly estimated using the 3-Gaussian Bayes classifier compared to alternative classifiers.

4.1.3 *Modelling Setting Selection (R2)*

Setting selection is estimated by introducing another variable to the probability set for R2. Let there be the setting variable z, defining one of the available HVAC

interface settings (temperature, blower, vent). The solution for this problem is still binary (whether a setting is selected or not) $z = \{0, 1\}$. Each setting can be modelled using a Bayesian classifier, the outputs being combined in a vector $z_{\mathtt{setting}} = \left(z_{\text{temperature}}\ z_{\text{blower}}\ z_{\text{vent}}\right)^{\mathsf{T}}$.

A single input (equivalent temperature) is not sufficient to correctly estimate the probability of setting selection. The input set was expanded to include the sensed elements within the thermal environment (the cabin, and surrounding surfaces, ambient temperatures, and the air velocity). The classifiers used for the comparison are the same as for R1.

The objective for R2 is :

1. To determine if the conditional probability (that a setting is or is not selected at the activation of R1 given a multiple feature input) can be correctly estimated using the set of binary classifiers mentioned in section 4.1.2.

4.1.4 *Modelling Value Selection (R3)*

R3 represents the probability of value selection depending on the comfort experienced by the occupant and their decision to make a change to a selected setting. Let there be the variable for value selection w, with labels $w = \{0, 1\}$ depending on the type of setting selected (temperature, blower, vent), the value is within the ranges presented in chapter 3.

The multi-class problem emerges when an instance has more than two classes. For example, the equivalent temperature suddenly drops and the passenger decides to make a change and selects a temperature. Therefore, the occupant is left with the decision to select a value from 12 set-points or, alternatively, 7 levels for blower speed or 4 directions for vent distribution.

Similar to the other rules, this distribution can be modelled as a Gaussian probability function. The problem is that such a model increases the complexity as there would be a total of 46 prior distributions, and the same amount of posterior distributions for two class label- models. A subsequent problem for R3, also encountered in R2, is that the input parameter used (equivalent temperature) is not enough

for an accurate estimation of the setting values. This is why introducing the other elements of the cabin environment (specified in section 4.1.4) and including setting selections (R2) is necessary.

The advantage of using classification models is that they estimate the probability of an event happening in the form of continuous values, and a predicted class in discrete form. While the discrete prediction is useful in terms of estimating a specific decision, or an event happening, the probabilities determine the confidence in the particular classification.

Given the problem of multiple classes, one of the most common solutions is applying algorithm adaptation techniques. The algorithms used for modelling R3 are: support vector machines with radial basis function kernel (svmRadial); with linear kernel (svmLinear); and dual linear kernel (svmLinear2); neural network (nnet); k-nearest neighbours (knn); stochastic gradient boosting (gbm); conditional inference trees (ctree); recursive partitioning and regression trees (rpart); random forest (rf); rule-based classifier (PART) [KJ13].

The main objective for R3 is:

1. To determine if the conditional probability (that a value for temperature, blower or vent is selected given the features of the environment, the comfort of the occupant, and the activation of R2) can be correctly estimated using a set of multi-class classifiers.

4.1.5 *The Final Hybrid Model*

The final goal is to combine the classifiers for each rule into a hybrid model that predicts the selections made by the user. The hypothesis is that a single-model used for classification or a strictly rule-based model cannot efficiently predict patterns of thermal behaviour. For the single model a neural network was chosen, whereas for the rule-based one fuzzy logic was used. The selected models were previously used in modelling feedback of the occupants or estimating setting selections (described in chapter 2).

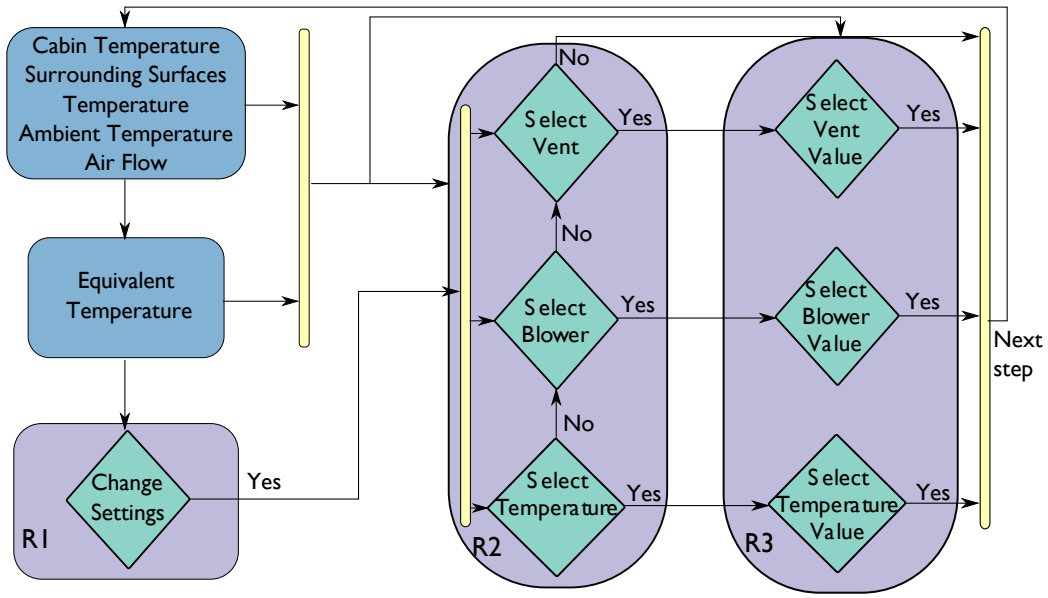

Figure 4.1: The UBM architecture each rule being activated when the change, selection and value adjustment are predicted depending on the comfort of the occupant and the state of the environment.

Therefore, a more complex model is developed, UBM, that combines the highest performing classifiers for each of the three rules (figure 4.1). The interaction with the HVAC system is essentially a three level classification problem, each level depending on the activation of the previous rule. The problem gradually escalates from a binary output for the first two rules, to a set of multi-class outputs for the final one.

4.1.6 *Experimental method*

This section details the data gathering and the pre-processing procedures, followed by the training, testing and validation method in section 4.1.7 (figure 4.2).

For the data gathering, the Comfort Studies Data Bank was obtained by Cogent Labs in collaboration with Jaguar Land Rover and MIRA during the Low Carbon Vehicle Technology Project (LCVTP). For LCVTP, the series of comfort studies aimed to obtain a catalogue of quantitative and qualitative data including cabin, physiology and perception of thermal comfort over a variety of conditions (hot, cold and neutral environments).

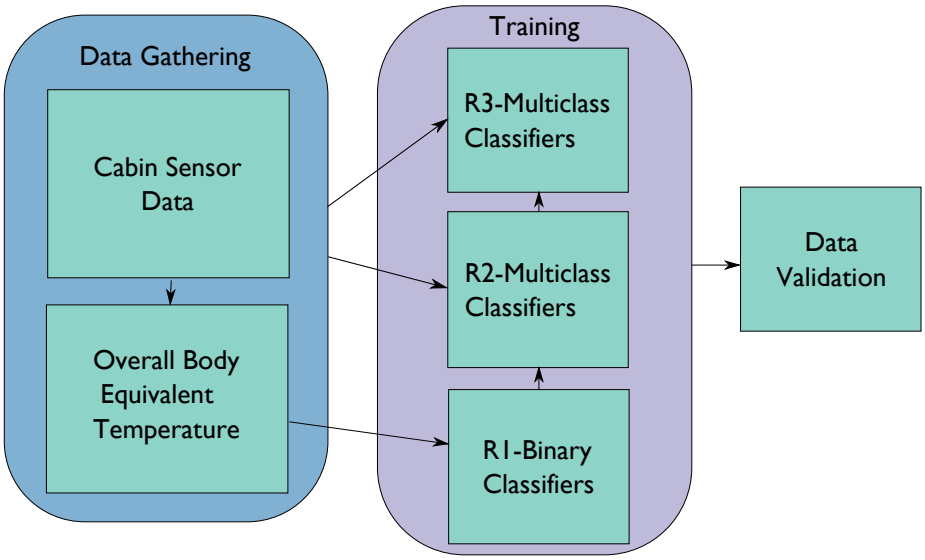

Figure 4.2: Method for testing and validation of the combined hybrid model.

Figure 4.3: Jaguar X3 used for the LCVTP trials, courtesy of Jaguar Land Rover.

Table 4.1: Subject details for the LCVTP experiments.

Subject	Sex	Age (years)	Height (cm)	Weight (kg)	Nationality
1	Male	46	173	78	English
2	Female	37	157.5	73	English
3	Male	56	166	70.5	English
4	Male	49	178	75	English
5	Female	24	162	48	Romanian
6	Male	26	176.5	77	English
7	Female	34	160	55	Mexican
Average		38.9	167.6	68.1	
SD		11.1	7.66	10.9	
Min		24	157.5	48	
Max		56	178	78	

The sets of trials involved 7 subjects (details included in table 4.1). Each subject was allocated a time slot for each day. Clothing was standardised for all subjects and tests. It consisted of long trousers and a short-sleeved shirt or blouse (approximately 0.7 Clo clothing insulation). A dry test run was performed to ensure that the procedure was correctly followed. An observer sitting at the right side of the rear passenger seat monitored the actions of the subjects. The start of the trial was marked by the subjects entering the vehicle and providing initial reports on their subjective thermal sensation and comfort.

The trials included gathering information about the control adjustments made by the cabin occupants in various conditions according to the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Standard 55 [ASH04]. Part of the experiments took place in the wind tunnel, and the other on the road under real-driving conditions. The procedure involved pre-conditioning the subjects to cold (16°C), neutral (22°C), and hot (28°C) temperatures. The HVAC system was set to condition the car at the same target temperatures: 16°C, 22°C , 28°C. The subjects remained in the pre-conditioning room for 20 minutes, then they entered the car and remained inside for 15 minutes. During this time they were permitted to adjust the air conditioning in order to make themselves more comfortable. The control adjustments were manually logged by the observer. Additionally, thermal comfort and sensations were reported every 2 minutes. The first report

was at the start of the test. For each subject, a total of six tests were performed: for neutral, hot and cold temperatures; with and without solar loading.

For the driving trials, the same pre-conditioning procedure was used. The difference between the two sets of trials was that the subjects had to drive the car on the road with no additional solar loading. The duration of the trials was determined by the observer running the experiments. Each subject was required to make turns and change speeds at frequent intervals in order to simulate the level of concentration normally required during driving. The tests for hot and cold temperatures were performed twice per subject, but only once for the neutral temperatures, as the subjects made no significant control changes during them.

For the thermal comfort and temperature monitoring the data was extracted from:

1. sensors placed on the body;
2. sensors placed on clothing;
3. cabin surface mounted temperature, humidity, and air velocity sensors;
4. solar loading sensors;
5. a Flatman manikin;
6. subjective reports by the occupants;
7. annotations of manual control adjustments of the HVAC system.

For each occupant positioned in the driver seat, the surface temperature was measured at different locations throughout the car (figure 4.4). Other parameters such as air temperature, relative humidity, air velocity, and CO_2 concentration were monitored using data loggers (figure 4.5) from Cogent Labs embedded onto the mbed platform (NXP LPC1768 microcontroller, I2C, analogue and serial interfaces, digital inputs and outputs for buttons and LCD screen, Ethernet support for time synchronisation and direct data download, microSD flash support for data logging).

The Flatman manikin was used for the automated calculation of the Predicted Mean Vote (PMV) (figure 4.6) using Dry Heat Loss sensors to determine the effects of

Figure 4.4: Sensors measuring surface temperature at 30 locations including: top, front and bottom of the instrument panel, front seat, windscreen, back of the driver and passenger seats, inside of the door and window, roof, seat locations: back and underneath.

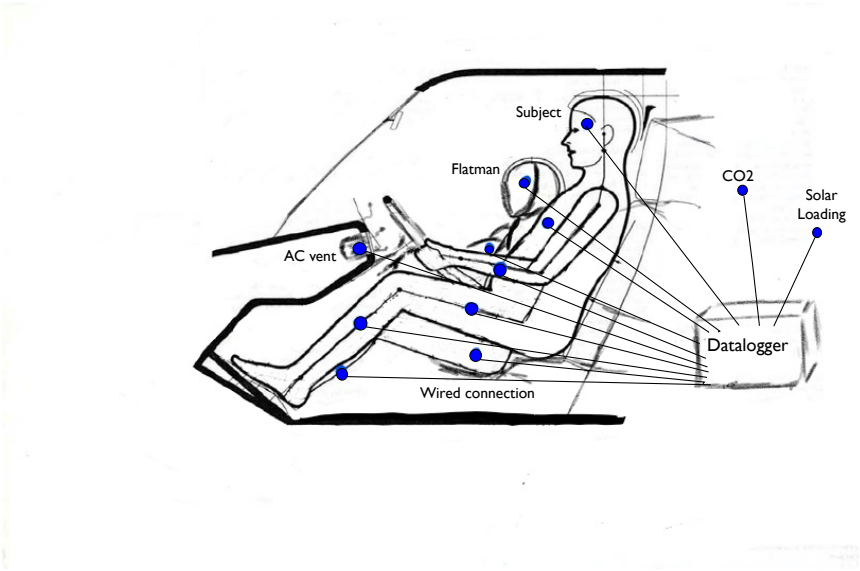

Figure 4.5: Overview of the sensor and data logger connections (wired). To the AC vents air temperature, relative humidity, and velocity sensors were installed. For the subject air temperature, relative humidity, air velocity, and skin temperature sensors were attached on the face, hands, chest, tighs, and calfs. The flatman (positioned in the front passenger seat) used dry heat loss sensors. Additionally the data logger was connected to CO_2 and solar loading measuring sensors.

Figure 4.6: Flatman manikin with data logger.

air temperature, mean radiant temperature and air velocity on a simulated person. The measurements were carried out using Analogue Digital ADT75a temperature sensors. The Flatman manikin was configured with a metabolic rate of 1.2 Met, an overall clothing value of 0.7 Clo based on the weighting of each body location. The body locations measuring equivalent temperature were: head; upper arms (left and right); lower arms (left and right); chest; thigh and calf. The Flatman manikin was positioned on the front passenger seat.

The overall body equivalent temperature measured by the manikin was modelled as a weighted sum of all body parts according to the information sheet [SVG14]:

$$\begin{aligned} T_{eq} = [&10.3 \times T_{eq-head} + 31.1 \times T_{eq-chest} \\ &+ 6.4 \times (T_{eq-left-up-arm} + T_{eq-left-low-arm}) \\ &+ 6.4 \times (T_{eq-right-up-arm} + T_{eq-right-low-arm}) \\ &+ 22.8 \times T_{eq-thigh} + 23 \times T_{eq-calf}]/100 \end{aligned} \quad (4.13)$$

Thermal sensation values reported by the occupants were performed using the ASHRAE thermal sensation scale included in the ISO7730 and ASHRAE Standard 55 [ASH10], modified to include ratings of -4 to $+4$ ([AZH06b]). For the sensation

Table 4.2: Annotations and ranges for the changes in HVAC settings available to the participants of the trials.

Data	Range	Annotation
Set Point Temperature	16-28	
	Climate	(C)
Blower Level	1-7	
	Automatic	(Auto)
	Head	(H)
	Feet	(F)
Vent Selection	Both (head and feet)	(B)
	Ambient	(A)
	Automatic	(Auto)

scale, 0 is the desired value representing thermal neutrality. Similarly a thermal comfort scale of -3 to 3 was used. Rating of 3 indicates a high level of comfort for the passengers, and -3 extreme discomfort. The difference between thermal sensation and thermal comfort is that the former is a result of the impact of the thermal environment on the occupants, whereas the latter is a measure of how satisfied the occupants are with their surrounding environment.

The data used for the UBM contains the time in seconds since the start of the test, the control changes made by the subject separated by target temperature setting (T), blower speed (B), and vent selection mode (F) (table 4.2). The ranges for each setting are identical to those described in chapter 3. Blower speed is on a scale from 1 to 7. Vent selection depends on orientation towards the head (H), feet (F), both head and feet (B), and ambient (A). Also, there are notations for the automatic selection of both the blower speed and vent (Auto) and climate control for temperature (C). The reasons for changing the control settings were written down by the observers, when specified by the subject. This data was merged with the Flatman data, the reports of sensations and comfort (overall and per body part), solar loading, pre-conditioning, and the subject's gender.

Prior to using the data for modelling purposes the following pre-processing steps were executed:

1. Sensor measurements that were outside the ranges were removed.

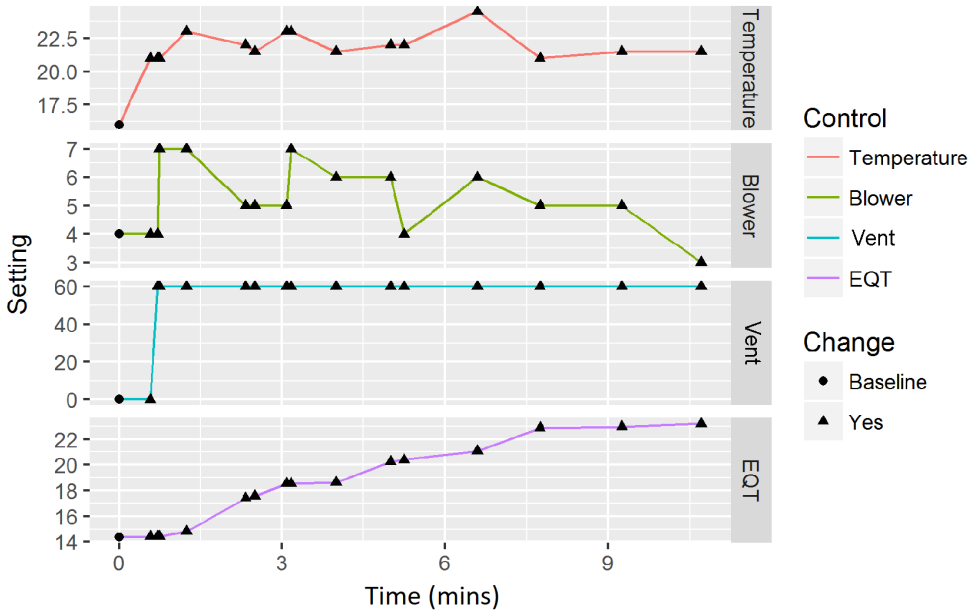

Figure 4.7: Overall equivalent temperature, and corresponding HVAC setting selections for a cold environment trial.

2. Sensor data and observer reports were combined by day, time and type of trial.

3. Overall equivalent temperature was calculated using equation 4.13.

4. The data was linearly interpolated to correspond to the selection intervals.

5. Setting labels were changed to correspond to the classification labels mentioned in the previous sections, in order to correspond to each of the rules.

Figure 4.7 displays the setting selections and the measured body equivalent temperature under cold conditions. The temperature and blower selections vary from higher to lower values over the trial duration. While equivalent temperature is gradually rising, the vent orientation stays the same throughout the duration of the trial, registering only one change within the first minutes of the trial.

From a total of 49 recorded trials under hot, cold and neutral environmental conditions, a total of 306 individual changes were recorded with 125 for set-point temperature, 112 for blower level, and 69 for vent direction.

Modelling of the numerical and categorical data was executed in R-Studio [Tee11] using the caret package [KJ13] for training, validating and testing the individual classifiers, and the mixtools package [Ben+09] for fitting the mixture of Gaussians.

The highest performing classifiers were converted into Predictive Model Markup Language (PMML) format using the pmml package in R [Wil+19]. The PMML classifiers were then opened in Java using the jpmml library [Ruu18] and coded into the UBM. The UBM was included into the cabin simulation, which is a simple lumped capacitance model [Bru+17] and used to train the Reinforcement Learning (RL) HVAC system. Further details can be found in Appendix E. The following section presents the training and validation methods for the classifiers used for each of the three rules.

4.1.7 *Testing and Validation Methods*

The dataset is split into 90% for training and development of the classification models and 10% for testing the final model. The classification model dataset used a training and validation set split using k-fold cross-validation (90%) [Koh95], and a hold-out test set for performance evaluation (10%). The hold-out method prevents over-fitting on the data, the final test scenario evaluation compares the hybrid model against more simplistic models.

As occupants are focused on driving, interacting with the HVAC interface becomes a secondary task. Therefore the instances of change or selection are expected to be less than the instances of non change or non-selection. This means that one outcome significantly outweighs the other, resulting in a class imbalance. The solution for class imbalance used in this thesis, similar to Kim et al. [Kim+18] is over-sampling of the minority class. This method of balancing the classes relies on randomly reproducing (with replacement) the minority class (e.g. the class for changes). This method is used within the cross-validation step in the caret package [KJ13]. On the one hand, the advantage of this method is that there is no loss of information, which can happen when using the under-sampling technique. On the other hand, there is a risk of over-fitting, as the minority class observations are replicated. This problem is prevented using of the hold-out method.

In determining which method was more suitable to use, alternative sampling methods were compared when training the Bayesian model (R1). The alternative

methods used are under-sampling, both, ROSE (Random Over-Sampling Examples). Under-sampling is a method that reduces the samples of the majority class, more suitable for large data sets. Both uses under-sampling with replacement of the majority class and over-sampling with replacement the minority class. Finally ROSE is a method of synthetic generation of data from the conditional density of the classes using bootstrapping [LMT14].

For evaluating the classification performance of a model, the most common used method is the confusion matrix. It displays the differences between the observed and predicted classes by means of a table. The diagonal cells of the table indicate the correctly predicted classes (TP – samples that are events, TN – samples that are non-events, and are predicted as such) while the alternative numbers indicate the classification errors (FP – non-events, classified as events, and FN – events misclassified as non-events). In the case of a binary problem (R1 and R2), the evaluation methods are used for determining event occurrences. Given that the data is imbalanced, the changes or selections are minority classes, and therefore will be classified as non-events. Two measures based on the confusion matrix are essential for model evaluation: sensitivity and specificity. Sensitivity (equation 4.14) is known as the true positive rate at which an event is predicted correctly.

$$Sensitivity\,(TPR) = \frac{TP}{TP + FN} \tag{4.14}$$

Specificity (true negative rate) is the rate at which samples are accurately predicted as non-events (equation 4.15). Related to the dataset, the setting selection is less frequent than non-selection, which determines the classifiers to treat the changes as non-events.

$$Specificity\,(TNR) = \frac{TN}{TN + FP} \tag{4.15}$$

Generally, there is a trade-off between sensitivity and specificity. Given the nature of the data and the purpose of the model, specificity has a higher performance

impact than sensitivity, as the cost for misclassifying a non-event as an event is higher than the cost of estimating an event where there is none (equation 4.16),

$$C_{01} > C_{10} \tag{4.16}$$

An alternative to evaluate the trade-off between the two rates is the Receiver Operating Characteristics (ROC) curve. The curve evaluates the classification probabilities by means of a continuous threshold (often 50%) by plotting the true positive against the false-positive rate. A model that is not effective has the curve following the diagonal between the two axes (with an Area Under Curve (AUC) of 0.5). The Area Under Curve can be used as a comparison method between various binary classification models. Among the advantages of using the AUC is that it is not sensitive to class disparities, while a disadvantage is that it leads to obscured information. The model that has a large AUC is considered as the most effective. Despite this, since the area of interest is within the lower end of the curve, this measure on its own might not be indicative of the best model.

Additionally, a metric for evaluating the error rate is accuracy (equation 4.17). This metric is easy to interpret as it is an indicator of how the predicted data agrees with the observed data.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{4.17}$$

One of the problems of this metric is that it does not offer information about the error type. The accuracy is used for R1, R2 (as an additional metric), R3 (in conjunction with Cohen's kappa) and the final model (on its own).

Cohen's kappa [KJ13] is a complimentary evaluation method to accuracy, initially designed as a method of assessing agreement between two classes. The advantage of kappa is that it includes the accuracy of a model in the case of the events are being chosen by chance. In equation 4.18, the two accuracies are related to the estimated and observed data and are based on the marginal totals of the confusion matrix. Kappa can take values from -1 to 1, with 0 indicating no agreement between the estimated and observed data. Negative values indicate an opposite direction of the

truth. Depending on the scales (Landis et al. [LK77], Fleiss [Fle71], Altman [Alt90]) the values of kappa can indicate agreement (from 0.30-0.50- according to Kuhn et al. [KJ13]). The advantage of kappa is that it can evaluate the models that are predicting multiple classes, such as those used for R3.

$$\text{Kappa} = \frac{\text{Accuracy}_{\text{observed}} - \text{Accuracy}_{\text{estimated}}}{1 - \text{Accuracy}_{\text{estimated}}} \tag{4.18}$$

From the above presented metrics, the following are used for:

1. R1: KL divergence to evaluate the best model based on Gaussian mixtures.
2. R1 and R2: specificity, sensitivity, AUC and accuracy.
3. R3: accuracy, and Cohen's kappa for multi-class model comparison.
4. UBM: accuracy.

4.2 RESULTS AND DISCUSSION

4.2.1 *Modelling the Change in Settings (R1)*

To model R1, all the instances of change and non-change depending on the estimated overall body equivalent temperature were grouped to examine the overall response frequency (figure 4.8). The frequency of non-changes was considerably higher than the changes made to the HVAC interface, which highlights the data imbalance. The probability of making changes increased when occupants feel uncomfortable, registered as an increase in the frequency of changes for equivalent temperature values higher and lower than 20°C indicating that once the occupants feel discomfort they will make changes to the settings. Rare instances where equivalent temperature is higher than 40°C were due to the increase in solar loading and reduced cabin air speed. Only a couple of trials regsitered these instances, the number of changes being minimal due to the small interval in which the occupants experienced these equivalent temperatures. Additionally the adaptation to the

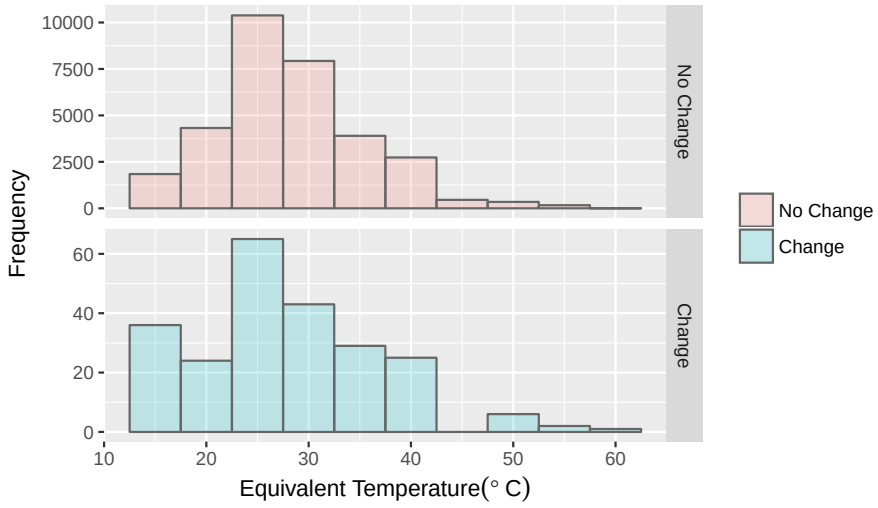

Figure 4.8: Frequency of change and no change data depending on the estimated overall-body equivalent temperature.

warm environment lowers occupant comfort expectancy and reduces the frequency of change.

The distribution of equivalent temperature given the change and no change is unknown. Therefore multiple Gaussians were fitted to the density distribution. In order to determine the number of components for the Gaussian mixture, boot-strapping was used for hypothesis testing. The maximum number of Gaussian mixtures that could be fitted is 5, the size of the mixture increasing when the difference between the current component number and the next is significant (95% confidence level). The method used 100 boot-strap replicates. The null hypothesis (for 4 components being fitted instead of 3) was rejected in the case of change ($p = 0.04 < 0.05$) and no-change ($p = 0.01 < 0.05$). Therefore for estimating the conditional probability for each of the two classes a 3-component Gaussian mixture provides the best fit (at 95% confidence level).

K-fold cross validation was used to further test the hypothesis that a 3 component Gaussian mixture Bayesian model produces the best fit for the data against alternative models that use 2, 4, and 5 components. Using the Expectation-Maximization algorithm, the Gaussian mixtures for 2 to 5 components were fitted against the equivalent temperature training data (figure 4.10). When using more than 5 com-

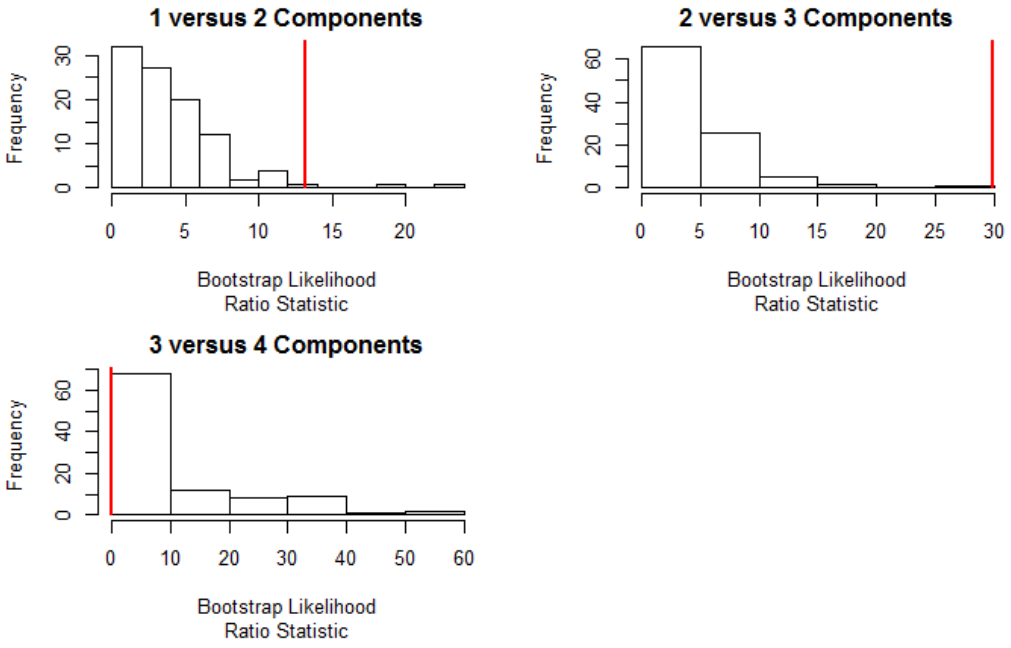

Figure 4.9: Boot strapping method to examine the number of Gaussian components that can be fitted to the changes data.

ponents for both distributions the curves have overlapping regions, as the means of the components are close in values.

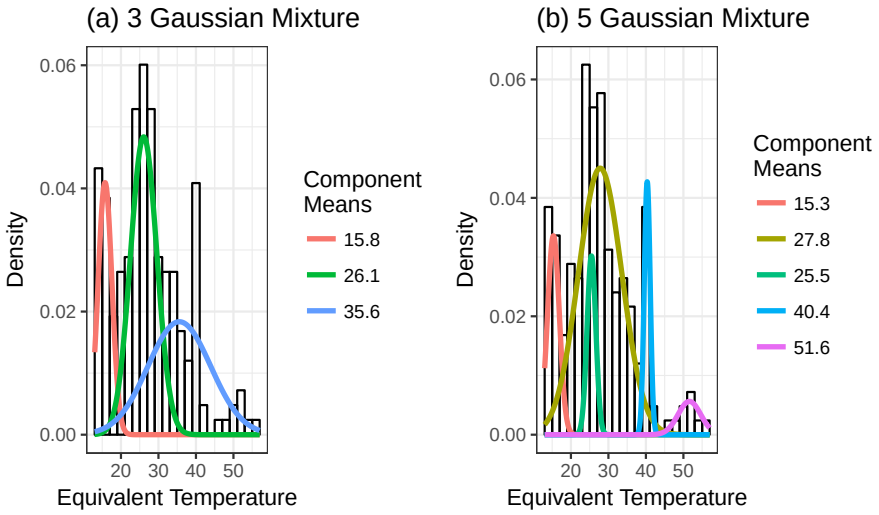

Figure 4.10: Fitting a 3-Gaussian mixture (left) and 5-Gaussian mixture (right) on the changes data. The 5-component mixture is over-fitting the data, as the Gaussian distributions are overlapping.

Examining the estimated values for equivalent temperature against the original values from the testing set and given the fact that there is a reduced number of data points for changes, the density area is lumped and more difficult to estimate.

Despite this, the 3-component model estimations correspond to the original data for both change and no change estimations (figures 4.11).

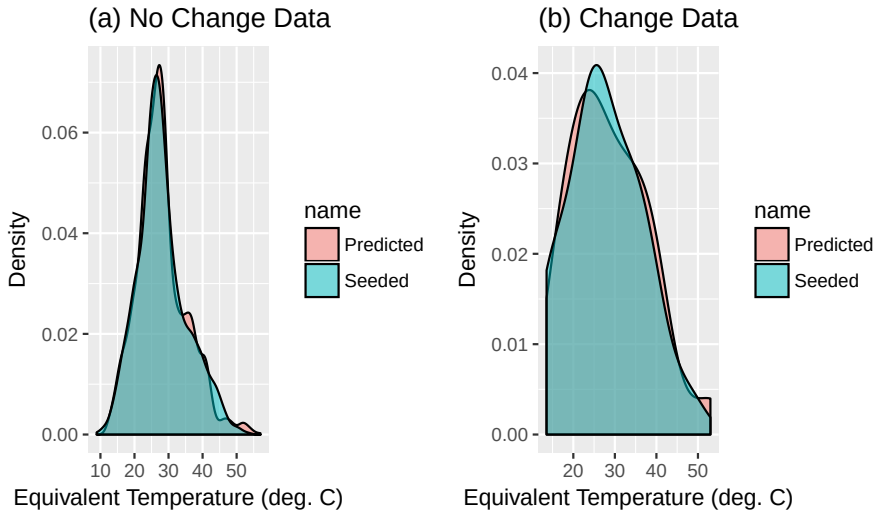

Figure 4.11: The original testing data against the estimated output of the 3-component Gaussian model for change and no change, as the density plots overlap this indicates that the model correctly estimates both changes and no changes depending on the overall-body equivalent temperature.

To strengthen the hypothesis that the 3 component model fits the data better for both change and no change, KL divergence was calculated for the 4 types of Gaussian mixture models (kl2, kl3, kl4, kl5) (figures 4.12,4.13). The 3- component model (kl3) has a lower difference between the estimated and the real data, and less variation between the quartiles than the other Gaussian mixture models. Therefore the 3-component mixture Gaussian Bayes model is used for comparison with alternative classifiers. It is highly likely that as more changes are recorded, the model will be closer to the original response, leading to improvements in model prediction with the increase in training and testing data points.

Testing the performance of the model using various sampling techniques, over-sampling method proved to have the highest AUC (0.67) and maintaining an accuracy of 84%. Compared to the model trained with unbalanced data, and the alternative methods (down, both, and ROSE), the model manages to predict the minority class with a 50% rate compared to 42% for the alternative methods. It registers a slight reduction in estimating the majority class 84% compared to 87% for the unbalanced model, and 86% for both. An interesting aspect is that the model

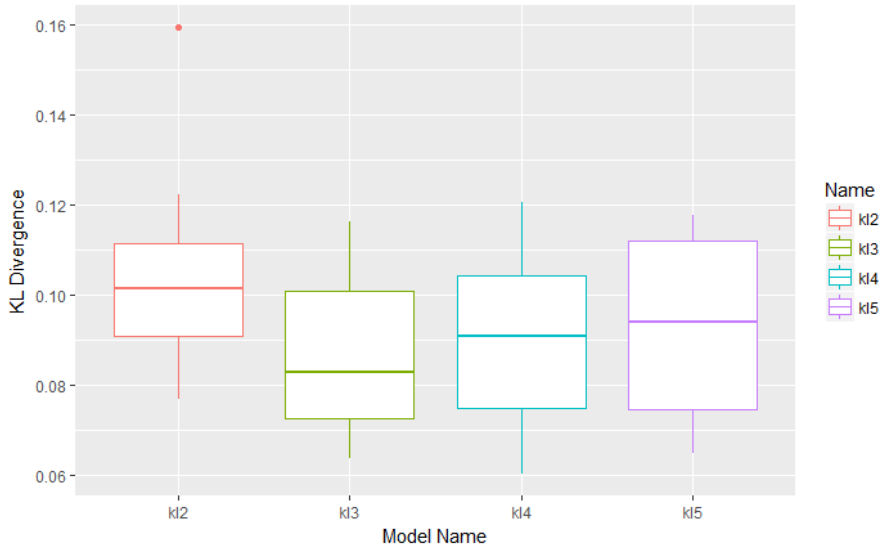

Figure 4.12: Box plots for the KL divergence for the 2 to 5 component models for estimating change, the 3-Gaussians mixture model (kl3) estimating the data better than the alternative models.

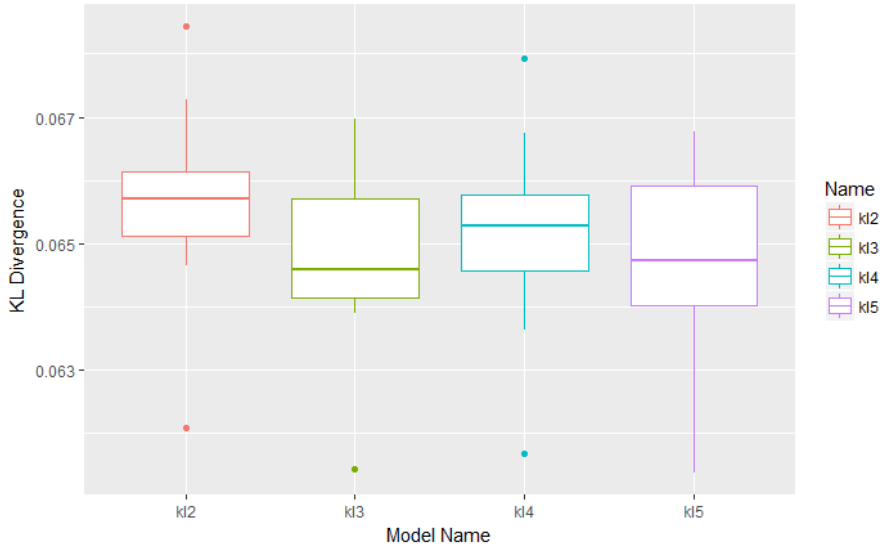

Figure 4.13: Box plots for the KL divergence for estimating no change, the 3-Gaussians mixture model (kl3) estimating the data better than the alternative models.

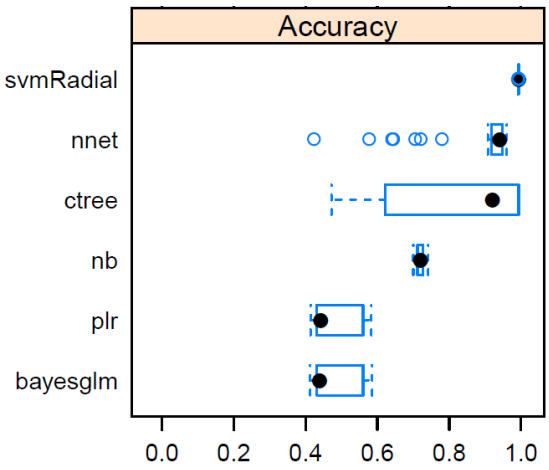

Figure 4.14: Box plot of accuracy margins for the classification models trained with original and over-sampled data.

trained with samples generated from synthetic data (ROSE) had the lowest AUC (0.6) and a significant decrease in sensitivity (78%).

Additionally, alternative binary classification models trained with up-sampled data were used for comparison. While the median accuracies (figure 4.14) for the support vector machine and neural network are the highest, both have poor performance (high sensitivity, low specificity). The models predict only the majority class of no changes to the HVAC interface. The conditional inference tree model has high variance, whereas the naive Bayes model has a median accuracy of 0.7.

Comparing the results on the testing set (table 4.3), the proposed Bayesian model has the highest performance compared to the alternative classifiers. It estimates correctly the change at the rate of 50%, and non-changes at 84%, with an AUC of 0.67. The alternative models have low estimations for the minority class (less than 1-2% true negative rates), despite a high accuracy (0.95 for the neural network and 0.99 for the support vector machine). The Bayesian model using 3-component mixture is therefore used as the R1 component of the UBM as it has a higher estimation rate for changes to the climate control.

Table 4.3: Evaluation of R1 classifiers, the proposed Bayesian model outperforming the other classification models.

Model	TPR	TNR	Accuracy	AUC
bayes	0.84	0.5	0.84	0.67
bayesglm	0.99	0.01	0.44	0.53
ctree	0.99	0.01	0.43	0.59
nnet	0.99	0.02	0.95	0.61
svmRadial	0.99	-	0.99	0.58
nb	0.99	0.01	0.71	0.59
plr	0.99	0.01	0.44	0.53

4.2.2 *Modelling Setting Selection (R2)*

R2 refers to the occupant's decision to select a specific setting from the HVAC interface options. The most selected setting in the experimental trials was temperature (figure 4.15). It was followed by blower speed selection, and finally vent distribution. For equivalent temperatures such as 35°C and 55°C, blower selection was preferred. Vent distribution was the least selected setting (the maximum frequency of non-selection). The sample size was reduced to overall 231 instances for each step second corresponding to the recorded instances of change. The objectives for modelling R2 were similar to those of R1. The exception is that only bootstrapping was used to determine the number of Gaussian components fitted to the conditional distribution.

The use of bootstrapping determined that a single Gaussian component was needed for fitting the prior distribution. The classifiers trained with equivalent temperature as a single input, including the Bayesian model, displayed poor performance for specificity, sensitivity, and AUC, estimating only the majority class. To improve the performance of the classifiers additional features of the environment were included: cabin, average surrounding surfaces, ambient temperatures, and air velocity. Table 4.4 displays the model comparison for all types of selections (temperature, blower, vent) using the hold-out data. Six models were used for the comparison (neural network, conditional inference trees, Bayesian generalised linear model, penalised logistic regression, support vector machine, and naive

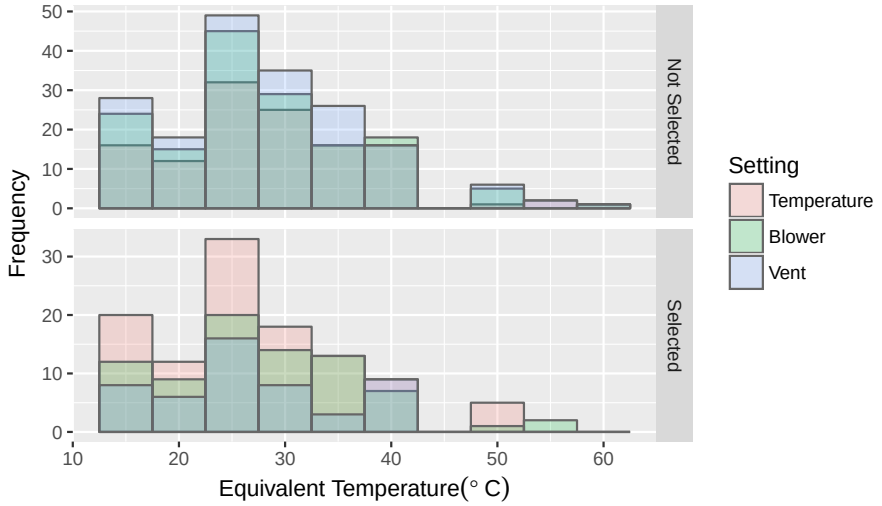

Figure 4.15: HVAC setting selections for the LCVTP data set. Temperature (pink) was the most selected setting, followed by blower (green). The vent (purple) has the highest frequency of non-selection.

Bayes). The performance metrics used for comparing the models for estimating each setting selection are specificity (TNR), sensitivity (TPR), accuracy, and AUC.

Comparing the median accuracy for temperature selections, the models had similar performance, with Bayesian generalised linear model having a slightly higher median than the rest (figure 4.16). Estimating temperature selection the model with the highest AUC (0.66) was the naive Bayesian model fitting Gaussian density functions to the input data (table 4.4). The true negative rate (58%) and true positive rate (64%) were the highest among the selected models, with an accuracy of (65%). The naive Bayes model was tuned by varying the Laplace function and the adjustments (weights). The model used has 1 adjustment and no Laplace corrections. Given that the highest AUC and specificity are achieved by the naive Bayes model, it is used for prediction of temperature setting selection in the UBM.

The highest performing model for estimating the blower selections was a neural network. When examining the validation data, the model had the highest median accuracy (figure 4.17). On the hold-out data (table 4.4), the model registered the highest specificity (50%), a 70% sensitivity, and an accuracy of 0.68. Compared to the alternative classifiers the neural network had the highest AUC (0.69). The neural network was tuned with 6 neurons for the input, 1 neuron for the output, and 4 neurons for the hidden layer with a size of 4 and a weight decay of 0.001.

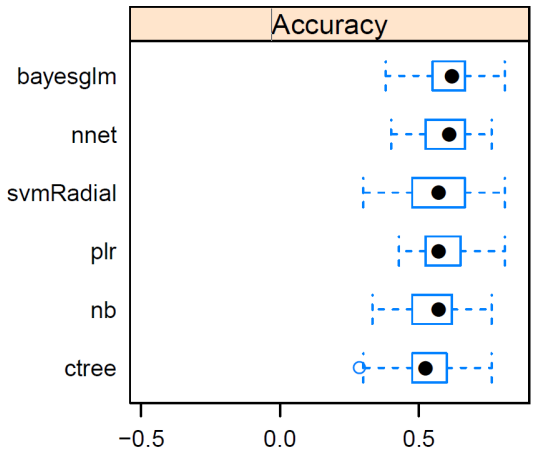

Figure 4.16: Box plot of accuracy margins for the classification models estimating temperature selection, performance is similar with bayeglm model having a slightly higher median accuracy.

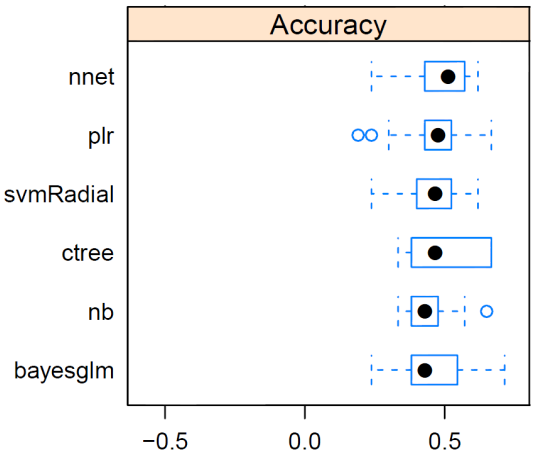

Figure 4.17: Box plot of accuracy margins for the classification models estimating blower selection, the neural network model has the highest median accuracy.

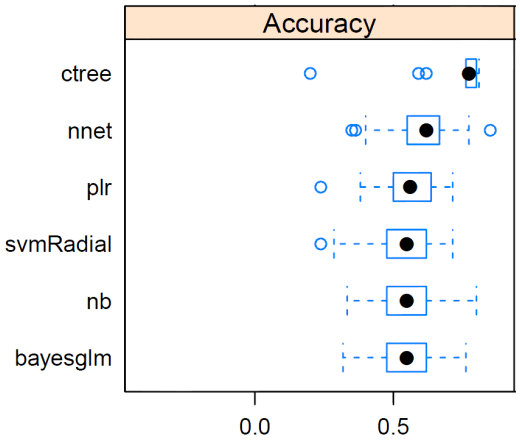

Figure 4.18: Box plot of accuracy margins for the classification models for estimating vent selection, the highest median accuracy is registered for conditional inference trees, however there are multiple outliers indicating a high discrepancy across the validation data.

Table 4.4: Model performance using original and over-sampled data, for estimating binary selection of temperature, blower, and vent. The highest specificity, accuracy are in bold.

Setting	Temperature				Blower				Vent			
Model	TPR	TNR	Accuracy	AUC	TPR	TNR	Accuracy	AUC	TPR	TNR	Accuracy	AUC
nnet	0.5	0.46	0.48	0.53	**0.7**	**0.5**	**0.68**	**0.69**	0.78	-	0.78	0.77
ctree	0.54	0.5	0.52	0.42	0.5	0.28	0.31	0.51	0.82	0.33	0.7	0.8
bayesglm	0.5	0.46	0.48	0.59	0.7	0.33	0.5	0.65	0.86	0.33	0.65	0.84
plr	0.54	0.5	0.52	0.5	0.71	0.33	0.45	0.61	0.86	0.33	0.65	0.8
svmRadial	0.58	0.55	0.57	0.47	0.69	0.33	0.55	0.63	0.82	0.33	0.7	0.85
nb	**0.64**	**0.58**	**0.61**	**0.65**	0.5	0.28	0.31	0.63	**0.88**	**0.5**	**0.78**	**0.86**

Estimating vent selections, the model with the highest median accuracy (figure 4.18) was the conditional inference trees, however the model had significant outliers. In this case, the accuracy measure was not a good indicator of model performance, as the tree-based model estimates only the majority class (non-selection) on the testing set (table 4.4). Examining the performance of all the models, the naive Bayesian model with fitted kernel density functions had highest rate of estimating of vent selections of 50%. The naive Bayesian model outperformed the alternative classifier as it had the highest trade-off between sensitivity and specificity (an AUC of 0.86). The final model had no Laplace correction, one adjustment and used kernel density functions fitted to the inputs. The naive Bayesian classifier will be used for estimating the selection of the vent setting for the behavioural model.

4.2.3 *Modelling Value Selection (R3)*

The following classifiers were used for modelling R3: support vector machine with radial basis function kernel (svmRadial); with linear kernel (svmLinear2); and dual linear kernel (svmLinear3); neural network (nnet); k-nearest neighbours (knn); stochastic gradient boosting (gbm); conditional inference trees (ctree); recursive partitioning and regression trees (rpart); random forest (rf); and a rule-based classifier (PART). Cabin, surrounding surfaces, equivalent temperatures, air velocity, type of environment (depending on the ambient temperature) were used as inputs, with the output being the selection of a value for the desired setting (temperature, blower, vent).

Temperature

The equivalent temperature is indicative of the impact that the type of environment has on the comfort of the occupant (as it is directly connected to the exterior and personal parameters that impact the cabin environment, and influences the selections of the set-point temperature). The most frequently selected set-point is 22°C corresponding to equivalent temperatures between 23-26°C. The higher the equivalent temperature was (figure 4.19) the lower the set-point selections and vice-versa.

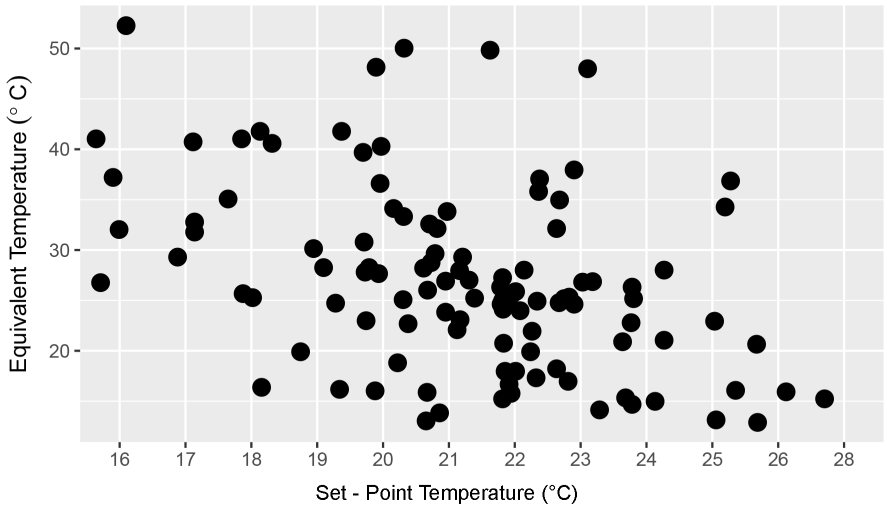

Figure 4.19: Set-point temperature selections depending on the equivalent temperature measures.

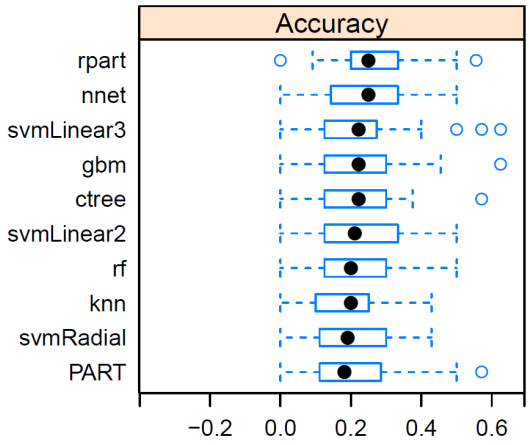

Figure 4.20: Accuracy box plot for the cross-validated data, the median accuracies of the nnet and rpart models are the highest and close in values compared to the alternative classifiers.

The median accuracy of the models was between 0.2 and 0.4 without tuning (figure 4.20), the neural network and regression trees model having the highest median accuracies among the used classifiers. Compared to the regression trees model, the neural network had higher variability. Despite this, the former model had outliers on both sides of the box plot. This indicates that there was less difference in model comparison if these outliers were included, the range of accuracy for the the regression trees being higher than the neural network.

In order to validate the performance of the models the hold-out testing set was used. The model with the best performance on the test set was the neural network (table 4.5). Among the multi-class models, it had a 0.706 kappa measure and a high accuracy of 0.8. According to Fleiss [Fle71] the kappa rate is between fair to good, whereas according to Landis and Koch [LK77] the model was in substantial agreement with the data. The neural network was tuned via resampled training data (figure 4.21) using a total number of weights of 35, with a weight decay of 0.5, 10 neurons for the inputs, 12 neurons for the outputs and 1 hidden layer. The neural network is used for determining set-point selections in the hybrid model.

Table 4.5: Performance metrics for estimating set-point temperature value selections, the neural network model has the highest Cohen's kappa and accuracy.

Model	Accuracy	Kappa
nnet	0.8	0.706
ctree	0.6	0.375
rpart	0.4	0.211
rf	0.4	0.167
PART	0.2	-0.035
svmLinear2	0.2	-0.177
svmLinear3	0.2	-0.177
gbm	0.0	-0.19
svmRadial	0.2	-0.25
knn	0.2	-0.25

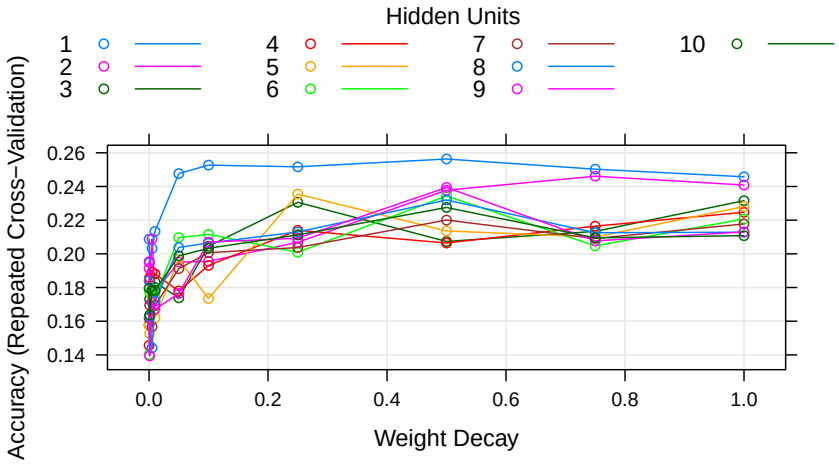

Figure 4.21: Tuning process for the neural network by varying the weight decay and the number of hidden units, when using 1 hidden layes the accuracy for the weight decay is higher than when using multiple hidden units. The highest accuracy of 26% is registered for a weight decay of 0.5.

Blower

The highest frequency of blower speed selection is between equivalent temperatures of 21-25°C for level 5, followed by level 7 between 13-18°C. Level 7 was selected for both high and low equivalent temperatures, with a gap between 20-30°C, indicating that it is the least preferred selection for passengers that are in the comfortable band of 21-24°C equivalent temperature [NH03] (figure 4.22). Lower blower speeds were also selected for equivalent temperatures of 15-35°C.

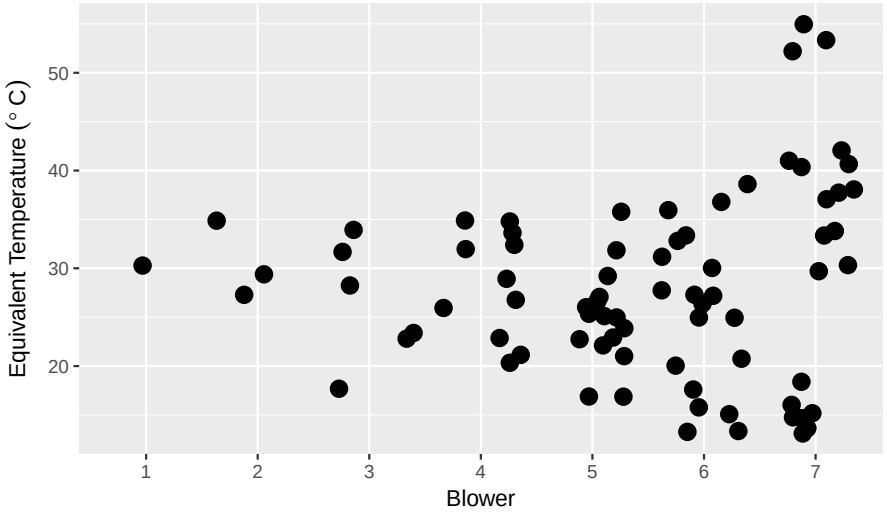

Figure 4.22: Blower speed selections depending on the equivalente temperature.

Similar mean accuracy performance was registered for the following models: support vector machine with dual linear and radial kernels, random forest, stochastic gradient boosting, conditional inference trees, and rule-based (figure 4.23). This indicates that multiple models have similar estimation capabilities, with a higher variability for random forest, rule-based and conditional inference trees.

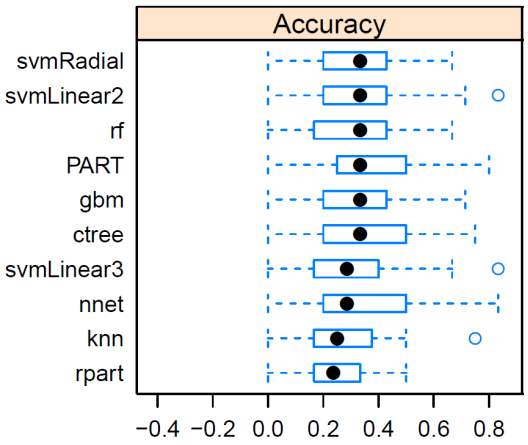

Figure 4.23: Model performance for blower level estimations on the cross-validated data is similar for all models, the generalised bayesian model, random forest, PART and ctree having higher variability. The svmLinear3, nnet, knn and rpart models have lower median accuracies than the other models indicating a decrease in performance.

In order to select the best performing model the hold-out set was used (table 4.6). The tuned conditional inference trees model had the highest accuracy of 0.65 and a kappa of 0.55 (according to Landis et al. [LK77] the model is moderate and for

Altman [Alt90], it is good). The conditional inference tree was tuned by varying the minimum criterion of 0.875, with 3 terminal nodes (figure 4.24). Having the highest performance out of the compared classification models, the conditional inference tree model is going to be used for estimating blower level values in the hybrid model.

Table 4.6: Performance metrics for estimating blower level value selections, the conditional inference trees model has the highest Cohen's kappa and accuracy.

Model	Accuracy	Kappa
ctree	0.65	0.55
rpart	0.55	0.41
svmRadial	0.55	0.41
gbm	0.5	0.34
knn	0.5	0.33
svmLinear2	0.5	0.31
nnet	0.45	0.29
svmLinear3	0.4	0.22
knn	0.5	0.333
rf	0.33	0.172
PART	0.33	0.078

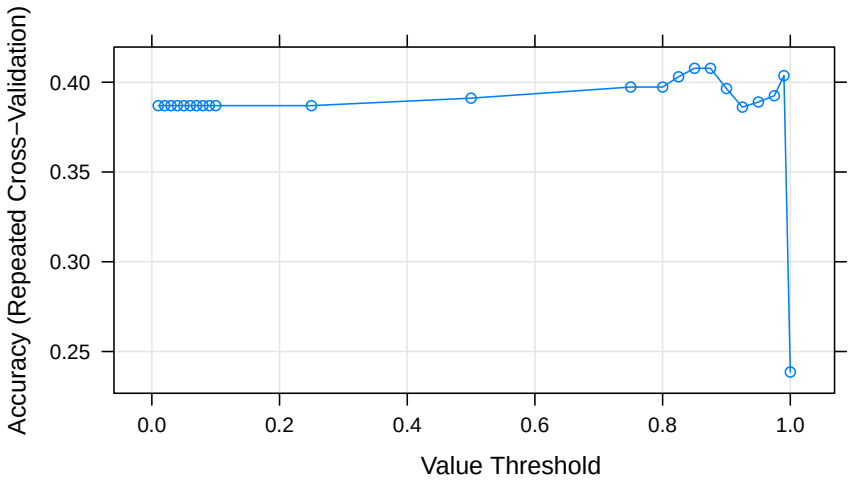

Figure 4.24: Conditional inference trees tuning with a variation of the minimum criterion, the highest accuracy of 44% is achieved using a threshold value of 0.875.

Vent

For vent distribution selections there are four main classes depending on the body part favoured by the occupant. The most selected distribution was towards both head and feet for equivalent temperatures between 20-30°C. The lowest frequency of selection was for the ambient setting (figure 4.25) for temperatures of 20°C and 35°C, indicating that for front occupants this setting is not as relevant as the alternative options. Out of the 50 independent selections, the head region was preferred for high equivalent temperatures (impact of a hot environment), whereas foot region was preferred for lower temperatures (cold environment).

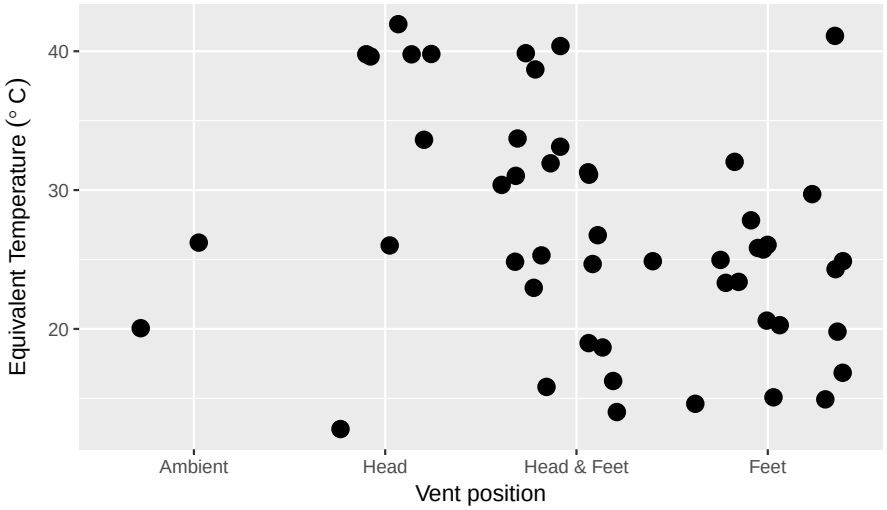

Figure 4.25: Vent distribution selections depending on the equivalent temperature.

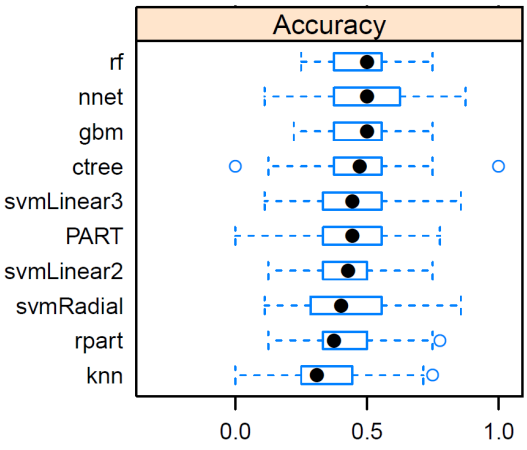

Figure 4.26: Model performance for vent distribution estimations on the cross-validated data is similar of the most models, with gbm, nnet and rf having the highest median accuracy. The neural network model has high variation in accuracy, whereas gbm and rf have comparable performances.

The models with similar accuracy performance on the cross-validated data were the stochastic gradient boosting and random forest. While the neural network model had a similar median accuracy, although it displayed high variability. For the validation procedure (figure 4.27), the model with the highest performance was random forest, with 2 random predictors. While most models had an accuracy 0.5 (table 4.7), the factor that influenced the prediction power of the random forest model was a Cohen's kappa of 0.78 (according to Fleiss [Fle71], kappa indicates good model performance and for Landis-Koch [LK77], the model is in agreement with the data). This can be linked to the lower number of classes for vent direction compared to the temperature and blower classes. It is also related to the 10% testing margin which only has a small number of selection instances. Nevertheless, the random forest identifies the classes correctly. It does not estimate only a single class, which can be the case for an over-fitting model. The random forest model is going to be used as a predictor for R3 vent distribution selection in the final model.

Table 4.7: Performance metrics for estimating vent distribution value selections, the random forest model has the highest Cohen's kappa and accuracy.

Model	Accuracy	Kappa
rf	0.86	0.78
nnet	0.67	0.5
svmLinear3	0.67	0.4
gbm	0.67	0.4
knn	0.67	0
rpart	0.33	0
svmLinear2	0.67	0
svmRadial	0.67	0
ctree	0.67	0
PART	0.33	-0.5

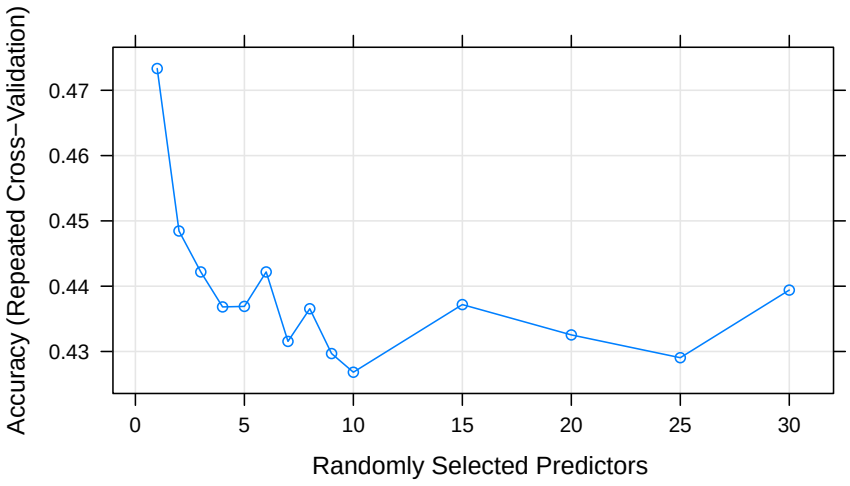

Figure 4.27: Tuning process of the number of randomly selected predictors for the random forest model, the highest accuracy of 48% is registered when the model is using 2 predictors.

4.3 THE USER-BASED MODULE

The User-Based Module is a hybrid model combining the seven best performing classifiers presented above. The model has the overall body equivalent temperature experienced by the passenger as input (figure 4.28). Additional input parameters are included in the modelling stage of R2 and R3. The parameters are related to the state of the thermal environment: the type of environment (hot, neutral, cold) depending on the ambient temperature; cabin and surrounding surfaces temperatures; and air velocity in the car cabin. These are essentially elements of a state vector used in the RL agent simulation for the thermal environment.

Equivalent temperature is used as input for the activation of the R1 Bayesian model, which predicts the probability of the occupant making change to the HVAC settings. R2 is activated only when a change is predicted. For predicting individual setting selections a naive Bayes model with fitted Gaussian distributions (temperature), neural network model (blower), and a naive Bayesian model with fitted kernel density functions (vent) are used. R3 is activated once a type of setting is selected. Similar to R2, for R3 each type of setting uses a distinct classification model. Each model outputs the desired value of the occupant from the available

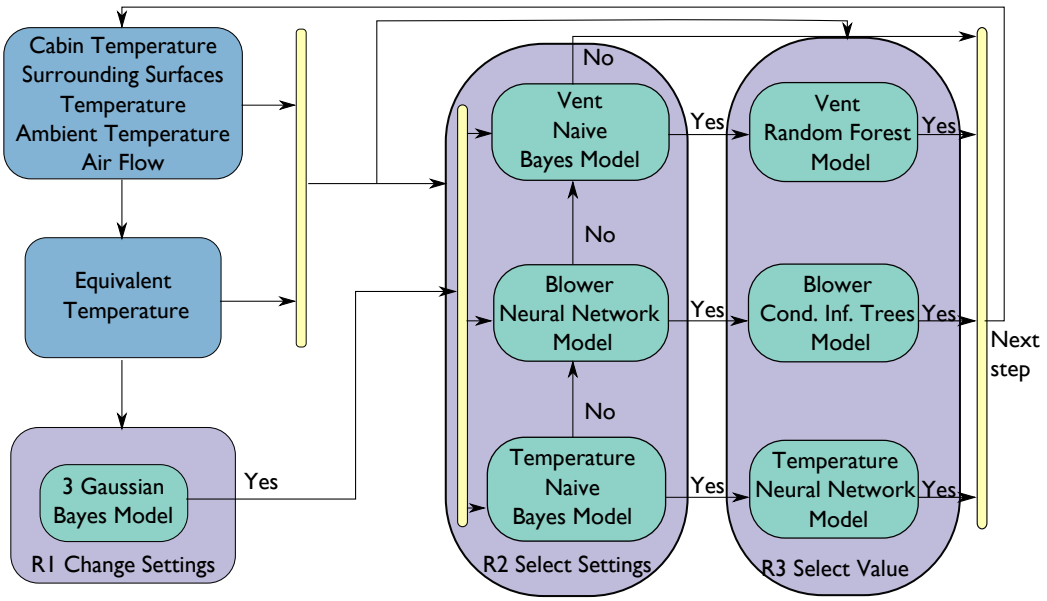

Figure 4.28: Combination of the seven classifiers with rule activation, basic structure of the hybrid UBM.

range of the HVAC interface settings in section 4.1.6. A neural network is used for estimating the set-point temperature, for blower levels conditional inference trees, and for vent distribution random forest.

The purpose of using a hybrid model is to mimic the occupants' decision making system as a sequential process (figure 4.29), using essential information from the cabin to predict their response. The main reason for using this process is to imitate the step by step rationale of the occupant and incorporate it into the context of cabin comfort modelling.

4.3.1 *Model comparison*

The UBM is based on the combination of multiple classifiers, therefore, can be considered complicated. Its performance was compared to a simple neural network model, and a fuzzy logic model using a final hold-out set (10% of the data). The trials were randomly selected out of the dataset, representing each type of environment (cold, neutral, warm).

The neural network model (NNET) has 5 input neurons, 23 output neurons (12 for temperature, 7 for blower level, and 4 for vent selections), with one hidden layer.

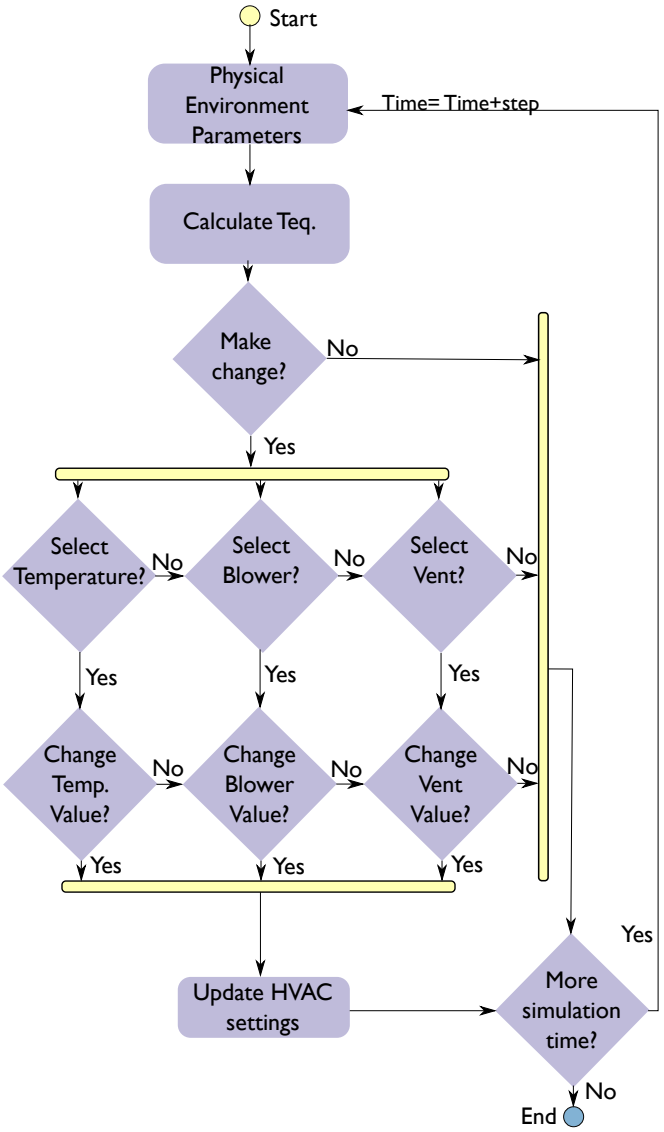

Figure 4.29: UBM model architecture diagram detailing the time step activation of each rule.

The network was trained with the same data as the UBM. The fuzzy logic model (FL) has 5 inputs and is based on 44 rules combining the inputs and outputs using the centre of gravity method. The inputs for both models are the same as the ones used for UBM, with the outputs being value selections for the HVAC settings.

The hybrid model has the highest accuracy (100%) for estimating temperature adjustment for the neutral environment scenario (table 4.8) compared to the neural network (72%) and fuzzy logic model (88%). Blower and vent selections are estimated poorly for the neutral environment by the neural network and fuzzy logic models. The simple models (NNET, FL) do not estimate correctly the values for the settings, and, furthermore, cannot determine adequately setting selections. Conversely, UBM has very good accuracy for vent distribution estimations in all types of environment. The hybrid model has the lowest accuracy for blower level in the neutral scenario (60%) and the highest for the cold environment (100%).

Given the small test set the results can be over-optimistic. There needs to be a consideration that the accuracy of the model will increase with more data being available, improving the estimations for the temperature and blower settings. Using an alternative dataset for model validation can be a solution. However, the problem with existent datasets is they have different features, especially, that they do not capture the setting selections of the occupants.

The neural network can be fine-tuned by varying the input parameters, and number of hidden layers. Moreover the fuzzy logic rules can be optimised using different activation methods. Nevertheless, a degree of uncertainty is necessary when modelling occupant behaviour. The UBM model displays good accuracy rates for estimating value selections for temperature, blower, and vent thus supporting the hypothesis.

4.3.2 *Limitations of the UBM*

For developing the hybrid model, each aspect of the interaction rules was defined as a classification problem. This approach is one of the many alternatives that can be used for modelling the changes made to the HVAC control by an occupant. The

Table 4.8: Accuracy of neural network (NNET), fuzzy logic (FL), and hybrid model (UBM) using the held-out test set.

		Setting Accuracy		
Model	Environment	Temp.	Blower	Vent
UBM	Cold	0.75	1	1
	Neutral	1	0.6	1
	Hot	0.71	0.75	1
FL	Cold	0.33	0.17	0.5
	Neutral	0.88	0.11	0.33
	Hot	0.33	0.17	1
NNET	Cold	0.09	0.17	1
	Neutral	0.72	0.37	0.46
	Hot	0.4	0.17	1

advantage is that each of the classifiers were trained and validated with a subset of a real-world dataset, and the hybrid model's performance was evaluated with a final hold-out set.

One of the problems with the validation of the final model is that there is insufficient data for an extensive evaluation, the performance of the final model can be improved by using an alternative data set as a hold-out set provided that the recorded parameters correspond to the training data set. Alternatively, the number of trial participants might not be sufficient to represent the entire population. However, the general trends depicted in the literature are observed through the occupant responses.

Moreover, the choice of splitting the data in multiple sub-sets impacts the performance of the classifiers, as the AUC is rarely above 0.6, indicating a high degree of randomness when estimating binary classes. This can be due to insufficient samples for training and validating the models. Alternatively, sample size can be adequate but the feature set is limited to elements of the environment and the equivalent temperature. A solution is to have a more extensive feature set and better model constraints. Furthermore only a limited set of classifiers is examined, which were identified in the literature and are available in the caret package (chapter 2).

Even though the risk of over-fitting is eliminated by the data split, there are alternative techniques for improving performance, for instance varying the threshold

for the Receiver Operating Characteristics curve and tuning the models. For the multi-class models, the one-vs-one method [Gal+11] can be used to obtain the average AUC. However this technique reduces the multi-class models to binary models, having an extensive number of combinations and increasing the complexity of the analysis.

Furthermore, by using binary classifiers for estimating each setting selection individually, there is a probability that even though a change is estimated, there will be no selection. This can be viewed as an error of the simulated human but it is also similar to how occupants behave (e.g. accidentally touch the interface, get distracted, forget or change their minds).

The hybrid model described in this chapter serves as an example of modelling occupant behaviour. As the main goal of this thesis is to examine the impact that the feedback of a simulate agent has on a machine learning based control system (presented in the following chapter), UBM fulfils this purpose.

4.4 SUMMARY

In this chapter, the User-Based Module is presented as an answer to the research question *"Can an artificial agent, validated using real-world data, realistically simulate the interaction that humans have with their HVAC system?"*. The model is a hybrid of seven distinct classifiers based on three inter-connected rules that are literature-based. The first rule estimates the probability of change that an occupant can make to the HVAC interface. The best performing model (AUC of 0.67, specificity of 0.5) is a Bayesian model using a 3 Gaussian mixture fitted to the input data. This model has a single comfort parameter as input, which is the equivalent temperature.

Once a change is estimated, the second rule becomes active determining the selection of a specific setting (either temperature, blower, or vent). The best performing models for estimating the setting selection were: a naive Bayesian model using a Gaussian distribution function for estimating temperature selection (AUC 0.65, specificity 0.58); a neural network model for estimating blower level selection (AUC 0.69, specificity 0.5); and a naive Bayesian model using a kernel density function for

estimating vent selection (AUC 0.86, specificity 0.5). A single input parameter was not sufficient for estimating selections, therefore the input feature set was extended to elements of the cabin environment.

Once a setting is selected the third rule is activated which estimates the value selection of the setting. For this classification problem a set of ten classifiers were trained for each type of setting selection. The problem was a multi-class one, the models being evaluated in terms of accuracy, and Cohen's kappa. The classifiers with the highest performance were a neural network model for set-point temperature, conditional inference trees model for blower level, and random forest for vent selections.

The seven presented classifiers are combined into the final model coded in Java. Its response was compared with a rule-based fuzzy logic model, a neural network, and the real-world occupant responses for three types of environment (hot, cold, and neutral) on a final hold-out set. While the responses were not identical, the performance of the UBM was similar to the real-world occupant adjustments (having the highest accuracy), therefore answering the research question. The UBM is a first step towards modelling the thermal actions of occupants within the context of a vehicle cabin. It can be further optimised using additional data and simplified by the combination of the rules.

The following chapter will examine the integration of the UBM within the Reinforcement Learning architecture, with the purpose of identifying which shaping technique can enable the agent to learn faster and maintain the comfort of the occupant longer.

5

USER-BASED REINFORCEMENT LEARNING CLIMATE CONTROLLER

Changes made to the Heating, Ventilation and Air Conditioning (HVAC) interface in the car cabin serve as feedback to the system, in the sense that people make changes to the settings in order to achieve their preferred comfort level. The learning performance of the Reinforcement Learning (RL) based HVAC system proposed by Hintea [Hin14] can be improved by using the occupant's feedback in the form of additional rewards by means of the shaping method. This new system, User-Based Reinforcement Learning (UBRL) HVAC, combines the feedback from the environment with the feedback from the occupant in order to efficiently learn a variable comfort temperature target.

The problem with including an additional reward is that the agent can learn a sub-optimal policy by choosing actions that ensure its immediate gain (e.g. riding a bike in circles, instead of in a line). A solution to this issue is to use a potential-based shaping reward [NHR99]. Potential-based shaping relies on the difference between potential functions associated with the state transitions. It does not change the objective of the task and is policy invariant.

Shaping advice methods are an extension of state shaping that use a potential function connected to the states and actions. Given the fact that the reward is a function of the state of the environment (a desired target temperature for the cabin selected by the occupant) and the control actions of the system (the air flow input to the cabin by means of blower level), this chapter investigates what is the most suitable method of shaping that enables the proposed UBRL system to learn.

This chapter aims to answer the following question: *"Can the UBRL HVAC system learn and maintain a nearly optimal policy based on occupant preferences within a reasonable amount of time?"*. The main contribution of this chapter is a UBRL HVAC controller, which efficiently learns from the cabin environment and occupant feedback, to maintain variable occupant thermal comfort control.

In order to examine the performance of the UBRL system, the controller is trained and tested under simulation conditions using the User-Based Module (UBM) (chapter 4) as the simulated response of the occupant. This chapter presents the proposed architecture of the system, detailing the methods used for integrating the feedback, the choice of the potential function, and shaping methods section 5.1.

The chapter further examines the performance of the UBRL system compared to a standard controller and determines how long it takes for the agent to learn an optimal policy when being trained by a simulated occupant (Section 5.2). Moreover, the UBRL HVAC agent is trained with alternative algorithms to State-Action-Reward-State-Action (SARSA) (λ), presented in chapter 2 in order to reduce the maximisation bias and further improve control system's response time of the control system.

5.1 METHOD FOR INTEGRATING OCCUPANT FEEDBACK FOR AN RL CONTROLLER

5.1.1 *UBRL HVAC System Architecture*

The HVAC control in the vehicle can be formulated as a Markov Decision Process problem that is defined by the tuple (S, A, T, R, γ), with a set of drawn states S and actions A available for the controller to take. The transitions $T : S \times A \rightarrow S$ are determined by mapping the state and action pairs to the following states by the deterministic model of the environment (equation 5.1). The reward function $R : S \times A \rightarrow \Re$ reflects the subsequent reward $R(s_t, a_t)$ resulting from the agent

taking an action for a specific state. The discount factor $0 \leqslant \gamma \leqslant 1$, determines the impact that future rewards have on the learning performance of the agent.

$$T(s_t, a_t, s_{t+1}) = P(s_{t+1} = s' | s_t = s, a_t = a) \tag{5.1}$$

The policy π maps the states to actions $\pi : S \rightarrow A$ and is the method for solving a Markov Decision Process (MDP). The optimal policy π^* represents the maximum long-term discounted reward.

Figure 5.1 shows the overall architecture of the UBRL HVAC control, as used in this thesis. The system has two parts: the cabin environment and the RL agent. The cabin represents the environment that is explored by the agent. The state of the car environment, also known as *sensation*, serves as input to the RL agent. The agent chooses an *action*, that maximises the *cabin reward*. The reward is bounded to satisfy all the following constraints: estimated comfort, occupant preference, and energy efficiency.

The parameters for the *cabin state* and *human feedback state* are stored in the state vector of the cabin environment (*sensation*). Due to the large state space for the cabin environment, tile coding is used. Tile coding is a function approximation method that approximates the state-action function $Q(s, a)$ by means of a smoothing function. Similar to Brusey et al. [Bru+17], the total number of tiles used was 30 (10 for states and actions, 20 for actions).

5.1.2 *The Cabin Environment*

Compared to the system proposed by Hintea [Hin14], the UBRL system has the cabin environment split into the physical environment (the car cabin model) and the UBM (the occupant that makes setting changes to the HVAC). The cabin environment has as its input the actions of the UBRL controller (internal temperature, fan air flow, and recirculation control). Its outputs are the updated sensed states of the cabin and interface (cabin, interior and ambient temperatures, air flow of the cabin, occupant's desired set-point temperature, blower level, and vent angle).

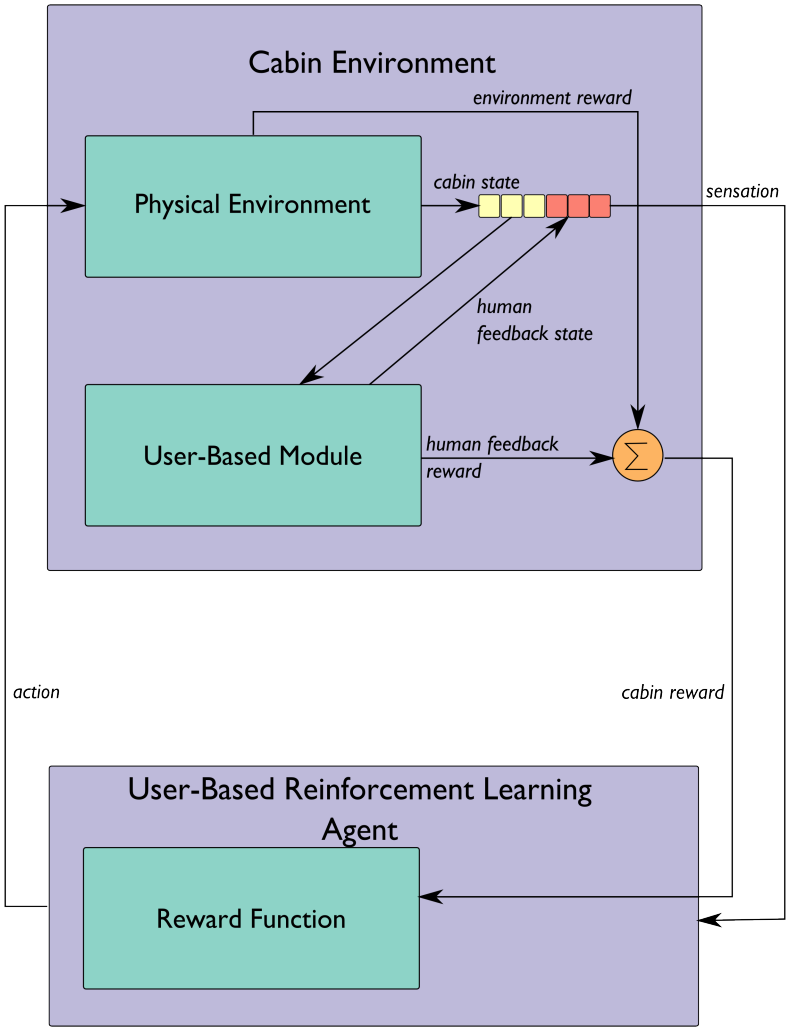

Figure 5.1: Integrated system with the occupant (UBM) as part of the Cabin Environment, the UBRL Agent learns from a combined occupant and environment reward.

Physical Environment

The physical environment is based on a mathematical model of the thermodynamic processes that a car cabin undergoes when in transit. It is a simplified model of the conduction, convection and radiation processes based on the following equation:

$$Q_O + Q_S = Q_I \tag{5.2}$$

meaning that the input heat, Q_I, to the system is equal to the sum of output, Q_O, and absorbed heat, Q_S. It is based on the lumped capacity model proposed by Lee et al. [Lee+15] and further developed and compared to empirical data by Brusey et al. [Bru+17]. The model relies on the following equations that preserve the heat balance:

$$\triangle Q_{heat} + \alpha I_{bl}\,(T_A - T_C) = I_{bl}\,(T_{mix} - T_C) \tag{5.3}$$

$$I_{bl}\,(T_{mix} - T_C) + \triangle Q_s + \triangle Q_{occ} + \frac{T_{Int} - T_C}{R_{Int}} + \frac{T_A - T_C}{R_C} = C_C \frac{dT_C}{dt} \tag{5.4}$$

$$\frac{T_C - T_{Int}}{R_{Int}} = C_{int} \frac{dT_{Int}}{dt} \tag{5.5}$$

The temperatures included are ambient air (T_A), cabin air (T_C), interior surrounding surfaces (T_{Int}), mixed air (T_{mix}), which depending on the recirculation factor (α), is either cabin or ambient air. The recirculation factor is given by the percentage of heated or cooled air recirculated from the cabin (I_{bl}) or input from outside (I_{in}):

$$\alpha = \frac{I_{in}}{I_{bl}} \tag{5.6}$$

The heat balance equation 5.2 is preserved through equation 5.3. The stored and the output heat are replaced by the input heat in equation 5.4 that represents the step update for the cabin temperature. Equation 5.5 depicts the step update for the average temperature of the surfaces surrounding the occupant.

The solar load ($\triangle Q_s$) was maintained at 150W and the occupant load ($\triangle Q_{occ}$) at 120W (corresponding to a single occupant). The change in heat pump energy is depicted as $\triangle Q_{heat}$, the absolute value of which is also considered as the energy

Table 5.1: Table of constant cabin and interior resistivity and capacitance [Bru+17].

Elements	Resistivity	Capacitance
Cabin	$1/(5.741626794 \times 4)\,\mathrm{K}.W^{-1}$	Variable
Interior	$1/75 \times 1.08\,\mathrm{K}.W^{-1}$	$450 \times 0.02 \times 7850\,\mathrm{J}.\mathrm{K}^{-1}$

consumed by the HVAC system with the blower energy costs being negligible (equation 5.7).

$$W_{HVAC} = |(\triangle Q_{heat})| \tag{5.7}$$

The thermal resistances and capacities related to the cabin and surrounding surfaces are constants (table 5.1), with the exception of cabin capacitance, which is calculated using cabin capacitance factor ($k = 8$), air mass (m_C) and specific heat (c_p). The mass is calculated using the volume of the cabin ($V_C = 2.5\,m^3$) and the density of the air (ρ_C).

$$C_C = k \times m_C \times c_p = k \times V_C \times \rho_C \times c_p \tag{5.8}$$

The state of the physical environment of the *cabin state* includes the air temperatures for cabin (T_C) , interior (T_{int}), exterior (T_A) , and air flow ($\dot{V}$). Additionally, the overall body equivalent temperature (T_{eq}) obtained using the Bedford equation (equation 3.1) is not explicitly integrated in the state. The comfort temperature is then compared to a desired variable target (T_{target}), detailed in the following section.

The air velocity (equation 5.9) for the physical environment was calculated by using the cross-sectional area (S) of the vents (estimating 2 blowers were used on the dashboard) and the volumetric mass air flow ($\dot{V}$). The cross-sectional area is $5.04 \times 10^{-3}\,m^2 (50.4\,cm^2)$, corresponding to Fojtlin et al. [Foj+16].

$$v = \frac{\dot{V}}{2 \times S} \tag{5.9}$$

User-Based Module

The UBM represents a novel approach of simulating cabin occupant interaction with the HVAC interface, based on three thermal behaviour rules present in the literature. Each rule is based on a set of classifiers that estimate the probabilities of making a change (R1), selecting a HVAC setting (R2), and selecting a value for the desired setting (R3). The *cabin state* parameters, together with equivalent temperature, act as external stimuli influencing occupant's behaviour and are used as inputs to the UBM. The *human feedback state* is a vector of the output parameters of each of the rules of the UBM. The vector includes the change (C), settings selection (T_{set}, B_{set}, or V_{set}) and the values for the settings (T_{val}, B_{val}, or V_{val}). The vent value is passed as a next step state for the cabin environment (V_{vent}), corresponding to static vent actuation.

The state vector is updated with the human feedback parameters, in order to incorporate the elements of the HVAC interface controlled by the occupant. These parameters are not explicitly mapped in order to maintain unaltered the tile-coding function. Moreover, the desired equivalent temperature (T_{target}) is calculated using the Bedford equation by replacing the values of cabin temperature with the desired set-point temperature (T_{val}) and blower speed (B_{val}). The inclusion of the target temperature is based on the supported hypothesis that there is no single comfort temperature that corresponds to all occupants of a thermal environment. This means that there are different comfort temperatures that improve the occupant's comfort satisfaction, depending on the type of environment and their preferences [BCB98; BPD04; Hal+15; Hui+06; LWG12; PBZ18].

5.1.3 *UBRL Agent*

The HVAC controller is an agent that has a set of available actions: control of fan air flow ($\dot{V}_{fan}$); air temperature of the vent (T_{mix}); recirculation flap positions (A_r) for determining if cabin air or external air is used. The total number of actions is 60, with 4 possible actions for air flow, 5 for inlet temperature, and 3 for recirculation.

In order to avoid abrupt changes in air flow and to better represent the gradual change in HVAC control, hysteresis is performed on the air flow connecting the state with the action (equation 5.10).

$$\dot{V}_{t+1} = 0.9 \times \dot{V}_{fan} + 0.1 \times \dot{V}_t \tag{5.10}$$

At the beginning of an episode, an initial state of the cabin environment is randomly selected $s_0 \in S$. The UBRL agent chooses an action based on the policy depending on the exploration parameter ε ($\varepsilon-$greedy action selection, equation 5.11).

$$\pi = \begin{cases} random_{a \in A} & with\, probability\, \varepsilon \\ argmax_{a \in A}\, Q(s,a) & with\, probability\, 1-\varepsilon \end{cases} \tag{5.11}$$

The agent explores the associated actions as long as $\varepsilon \neq 0$, once it is 0, the agent greedily selects the action and exploits the nearly optimal policy.

Given the fact that driving scenarios are limited in time (the average driving session lasts approximately 20 minutes [JCR15]), this thesis expands the episode length to an average driving trial, compared to Brusey et al. [Bru+17]. As each time-step is equivalent to 1 second, the maximum number of steps to the end of the episode is increased to 1200.

As the learning becomes episodic, an examination of the number of steps an agent is required to take in order to achieve the target is essential in determining how fast the agent learns a correct policy. The episode ends either when the maximum number of steps is achieved, or when a terminal state is reached (equation 5.12).

$$s = \begin{cases} terminal & |T_{eq} - T_{target}| \leqslant 0.5 \\ s_{t+1} & |T_{eq} - T_{target}| > 0.5 \end{cases} \tag{5.12}$$

As the desired comfort target (T_{target}) depends on the selected HVAC settings, it can be argued that the agent receives a form of feedback from the occupant.

According to Sutton et al. [SB98], the notation for an episodic task can be extended to a continuous scenario by having the terminal state transition to an absorbing state with an expected reward $R(s_t, a_t) = 0$, which is the case for this work. The episodic problem is extended to a continuous one for testing the agent. The agent can use and maintain a learnt policy, given sufficient time and exploration of the state-action pairs in a setting where it needs to maintain control.

According to Abbeel et al. [AN], the reward function is a suitable task definition as it is transferable and robust. Therefore, in order to enable the agent to efficiently learn a good policy, the reward function presented in the following section incorporates the occupant's feedback.

Reward Function

For UBRL HVAC, the agent preserves the objectives to maintain comfort (achieve maximum thermal comfort duration) and reduce energy consumption (minimum use of energy).

As the agent learns by interacting with the environment (model-free reinforcement), the reward function becomes a tool for achieving these objectives as a weighted sum of all rewards (equation 5.13).

$$R(s_t, a_t) = w_C R_C(s_t) + w_E R_E(s_t, a_t) + w_G R_G(s_{t+1}) \tag{5.13}$$

In order to differentiate the terminal state from all the other states, an additional reward for achieving the goal is integrated (equation 5.14).

$$R_G(s_{t+1}) = \begin{cases} +1 & \text{if } s_{t+1} = \text{terminal} \\ 0 & \text{otherwise} \end{cases} \tag{5.14}$$

The weight of the reward for reaching a terminal state is $w_G = 100$ (as positive reinforcement) [Grz17; MR18]. The weight for the comfort reward is $w_C = 1$ based on Brusey et al. [Bru+17], the comfort reward is equivalent to the agent receiving a penalty equal to the range between the actual equivalent temperature (T_{eq}),

and the desired equivalent temperature target (T_{target}). The difference is that the comfort target does not depend on a fixed temperature (e.g. Brusey et al. [Bru+17] using 24 ± 1 °C) but on various target temperatures (equation 5.15) based on the occupant's feedback.

$$R_C(s_t) = -|T_{eq} - T_{target}| \tag{5.15}$$

The energy cost (equation 5.16) is represented by the energy reward $R_E(s_t, a_t)$, with a weight equivalent to $w_E = 300W$ for a 1% thermal comfort improvement derived by Brusey et al. [Bru+17].

$$R_E(s_t, a_t) = -(|W_{HVAC}| + 2\dot{V}_{fan}) \tag{5.16}$$

The RL agent takes a long time to learn a good policy by means of the reward function, therefore additional shaping rewards can be used to serve as guidance.

Reward Shaping

In order to preserve the optimality of the original policy and directly use the occupant's feedback, reward shaping can be utilised. This method preserves the optimal policy of the original problem [NHR99] and gives the agent an easier opportunity to solve the task using additional rewards. Finding an appropriate potential function impacts on how fast and smoothly the RL agent is guided. The potential function used in this work is similar to that proposed by Ng et al. [NHR99] by using elements of the state pertaining to the cabin (cabin temperature, blower speed, vent distribution) and to the occupant (set-point temperature, velocity derived from the blower level, and desired vent distribution). The occupant's desired values are considered as sub-goals pointing the agent towards an optimal solution (equation 5.17), where $1 - \varepsilon$ represents the percentage that the selected action is the desired one.

$$\Phi(s_t) = \frac{-(|T_C - T_{val}| + |v_{\dot{V}} - v_{B_{val}}| + |V_{vent} - V_{val}|)}{1 - \varepsilon} \tag{5.17}$$

A problem with this type of function is that the agent is left to discover what actions are appropriate to choose from the rewards associated with the states. Wiewiora et al. [WCE03] proposed two distinct methods of extending the shaping function to the actions taken by the agent, offering a complete view of the state-action space to the agent. Look-back and look-ahead advice (equations 2.12,2.13) are implemented, as the user feedback information is likely to target the actions of the climate control.

The extended potential function for the advice methods includes the fan air flow (part of the action vector) by deriving the velocity of the blower and comparing it to the desired velocity (equation 5.18). The difference between the two methods is that look-ahead advice requires the potential function to be further included within the greedy policy, becoming biased (equation 2.14).

$$\Phi(s_t) = \frac{-(|T_C - T_{val}| + |\nu_{\dot{V}_{fan}} - \nu_{B_{val}}| + |D_{vent} - D_{V_{val}}|)}{1-\varepsilon} \tag{5.18}$$

The UBRL agent is trained using the SARSA (λ), which is an on-policy algorithm that selects the next action a_{t+1} depending on the policy π and the current reward $R(s_t, a_t)$. By using shaping, the reward function is changed to the compound reward and the action value function update includes the new reward (equation 5.19), where α is the learning rate parameter controlling the step-size used to process the current reward $\bar{R}(s_t, a_t, s_{t+1})$ for action a_t. In this work the α parameter is constant.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[\bar{R}(s_t, a_t, s_{t+1}) + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{5.19}$$

Additionally, the agent is trained using Expected, Double, and Double Expected SARSA with a similar modification of the reward for the action value function (equations 2.6, 2.8, 2.9).

The controllers based on these alternative algorithms are compared in order to establish if the originally proposed or the alternative controllers have an increased performance. Details of the evaluation methods are presented in the following section.

5.1.4 *Evaluation methods*

An initial comparison of the three shaping methods was drawn by means of the average trial reward, number of steps and step reward, for 20 runs with randomised seeds. To test that the results for the three shaping methods were statistically different, the Kruskal-Wallis rank sum test was used as an alternative to one-way ANOVA, in case the variance (Lavene test) and the normality (Shapiro-Wilk test) requirements for the data were not fulfilled. For testing the statistical difference between groups of two methods, the Wilkoxon Rank Test was used for the not-normal data. The performance of the UBRL controllers trained with the variations of the SARSA algorithm was evaluated in terms of the numbers of steps taken to reach the goal, average reward per episode, as well as with a test set scenario.

The test scenario set [Bru+17] included 200 pre-selected start states that were randomised after each 1000 episodes of learning. This test provides a good comparison between alternative controllers, while maintaining a standard evaluation of their learning capabilities. The UBRL -based controllers were compared to a standard air-conditioning or bang-bang controller that blows air into the cabin at the maximum speed in order to bring the cabin temperature to a 1°C range of the target. The target in this case was the desired cabin temperature set by the occupant (T_{val}). The performance metrics used to evaluate the systems were the mean reward achieved during the test scenarios, the average percentage of time spent in comfort, the average time to achieve the target, the average power used by the controller, as well as the average changes made by the occupant per trial. The following equation was used to calculate the HVAC power.

$$P_{HVAC} = \frac{W_{HVAC}}{\triangle t} \tag{5.20}$$

The power consumed is obtained by dividing the energy used by the HVAC system by the time step (which in this case is 1 second). As the energy used by the blower is assumed negligible, the energy consumed by the system is equivalent to the heat

pump energy. Using equations 5.3 and 5.7, the power can be derived using the heat from elements of the state and action (equation 5.21).

$$P_{HVAC} = \frac{|\triangle Q_{heat}|}{\triangle t} = \frac{I_{bl}\left[(T_{mix} - T_C) - \alpha\,(T_A - T_C)\right]}{1} \tag{5.21}$$

The performance of the UBRL HVAC controller compared to the bang-bang controller was also examined under two processes: cool-down and warm-up. The cool-down process relies on the HVAC introducing cool air into the cabin with the purpose of lowering the temperatures of the environment in hot weather. Conversely, the warm-up process introduces warm air into the cabin in cold weather. For the cool-down process, the starting cabin, surrounding surfaces temperatures were 35°C, and the exterior temperature was 25°C. The warm-up process started with the cabin at 15°C, surrounding surfaces temperature at 5°C and the exterior temperature at 15°C.

5.2 RESULTS AND DISCUSSION

The problem with the RL HVAC car controller is that it takes approximately 6.3 years of simulated learning time [Bru+17] (equivalent to 200000 episodes). This duration, when implementing the controller in the car cabin, is equivalent to the estimated lifetime of a car (6-8 years). This is undesirable as it would take the entire expected use of the car, with the feedback from the occupant, for the controller to learn a nearly optimal policy. Even with a pre-learnt policy, the agent needs to further explore and potentially unlearn parts of the state-action space, as the only comfort target is 24°C, whereas in the car there can be multiple comfort temperature targets depending on the occupant's preferences.

The RL HVAC controller proposed by Hintea [Hin14] can effectively learn how to maintain a target equivalent temperature of 24°C when a car cabin is in a cool-down process (introducing cold air into the cabin compartment in order to lower the temperatures of the environment in hot weather). Alternatively, even when

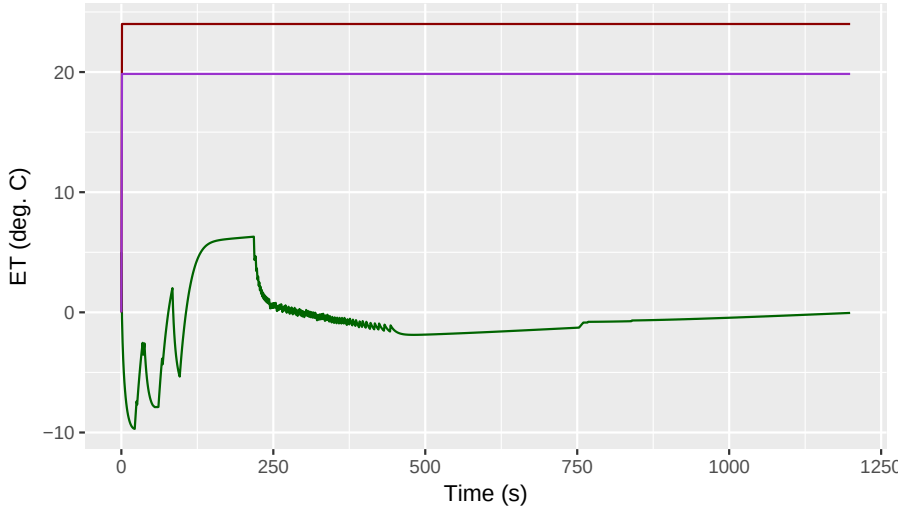

Figure 5.2: Equivalent temperature (green line) of the occupant for the original RL HVAC system proposed by Brusey et al. [Bru+17] under the warm-up process. The equivalent temperature does not reach the target equivalent temperature of 24°C (red line), or the occupant desired target temperature of 20°C (purple line).

increasing the episode duration to 20 minutes, leading to a total training time of 7.6 years (calculated using equation 5.22) for the controller.

$$t_{training} = \frac{N_{episodes} \times N_{steps} \times N_{seconds\,per\,step}}{3600 \times 24 \times 365} \tag{5.22}$$

The number of episodes ($N_{episodes}$) is 200000, the number of steps (N_{steps}) is 1200 and the number of seconds per step ($N_{seconds\,per\,step}$) is 1, in order to get the estimated training time in years they are divided by the the total number of seconds in a year. To be noted that this value refers to the simulated time for training the RL agent rather than elapsed real time for running the simulation. The RL HVAC does not achieve the target equivalent temperature under warm-up conditions (introducing hot air into the cabin in order to increase the temperatures of the environment in cold weather). The equivalent temperature is maintained at 0°C (figure 5.2). Moreover, not even the occupant's desired equivalent temperature is achieved, based on the changes made to the HVAC settings, which is lower than the fixed target of 24°C.

Introducing a variable target equivalent temperature within the comfort reward (equation 5.15) improves the performance of the RL HVAC controller, as it reaches the cool-down target, and manages to get close to the warm-up target (figure 5.3).

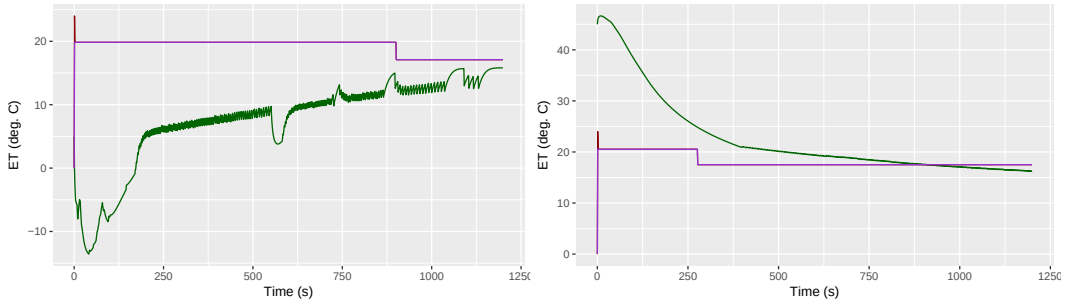

Figure 5.3: Warm-up (left) and cool-down (right) processes of the RL HVAC controller trained for 200000 episodes with an occupant's desired equivalent temperature instead of a fixed target temperature.

The controller was trained for the same number of trials as the original. The desired target was achieved much slower than the controller trained with a fixed target (after 14.17 minutes of a 20 minutes trial duration for cool-down). It is to be expected, as the agent requires more time to explore the state-action space when the comfort target is variable. Compared to the fixed target of 24°C the desired equivalent temperature shifts from 20°C to 18°C for the cool-down process and from 20°C to 17°C for the warm-up. For a controller trained with only a comfort and energy reward, it takes longer to achieve occupant comfort for the cool-down process despite the fact that the desired temperature is much lower than the fixed target. Conversely, it is easier to achieve the desired comfort target for the warm-up process for the same reason.

There are two problems with the RL HVAC controller trained with variable targets: the training process takes longer than the lifetime of the car and even with the learnt policy, the occupant can achieve comfort only at the end of the driving session (for both the warm-up and cool-down processes). The following section presents the comparison between the three shaping methods proposed, examining how well the UBRL controller learns by means of shaping rewards.

5.2.1 *Shaping methods comparison*

The three main methods of shaping explored in this thesis are: potential state shaping; look-back advice; look-ahead advice. The difference between these methods is

Table 5.2: Average steps per trial, average reward per trial, and average reward per step for the three different shaping methods.

Shaping Method	Mean Steps	Mean Reward/Trial	Mean Reward/Step
State	504.8 ± 68.31	-4.11 ± 0.75	-0.59 ± 0.72
Look-back Advice	496.25 ± 61.02	-3.99 ± 0.71	-0.54 ± 0.75
Look-ahead Advice	531.1 ± 37.96	-5.24 ± 0.61	-1.23 ± 0.66

that the first is strictly related to the environmental state (equation 2.11). The second method includes the actions (in this case the air velocity) and compares the states and actions taken in the last step to the current states and actions (equation 2.12). The final method compares the predicted states and actions for the next steps with the current states and actions (equation 2.13).

While the RL HVAC controller was trained for 200000 episodes, the UBRL HVAC controller using all shaping methods required only 75000 episodes of training, equivalent to 2.9 years (using equation 5.22), which is less than half the life-time of a car. All agents were trained with $\varepsilon = 0.16$ for 70000 episodes, after which $\varepsilon = 0$ for the remaining 5000, the step size used was $\alpha = 0.01$, with a discount factor of $\gamma = 1$, meaning the agent is far sighted (future rewards having a higher impact on the learning).

As the problem is episodic in nature, the number of steps taken until target comfort is reached was analysed (figure 5.4). The look-back advice agent took the lowest average number of steps until reaching the goal (the response was averaged over 20 runs with random seeds), with the mean of 496.25 (table 5.2). State shaping and look-back advice had comparable total rewards per trial (figure 5.5), the later method having the lowest mean rewards per trial (-3.99) and per step (-0.54). This means that the look-back advice agent managed to achieve the desired comfort goal faster than the agents trained with the alternative shaping method, while incurring the lowest penalty.

Due to the fact that the mean rewards for look-back advice and state shaping agents are close in value, the Krukscal-Wallis rank test was used to test the difference in means for the three agents. The statistical test is an alternative to one-way ANOVA, as the data does not satisfy the Shapiro-Wilk normality test for the

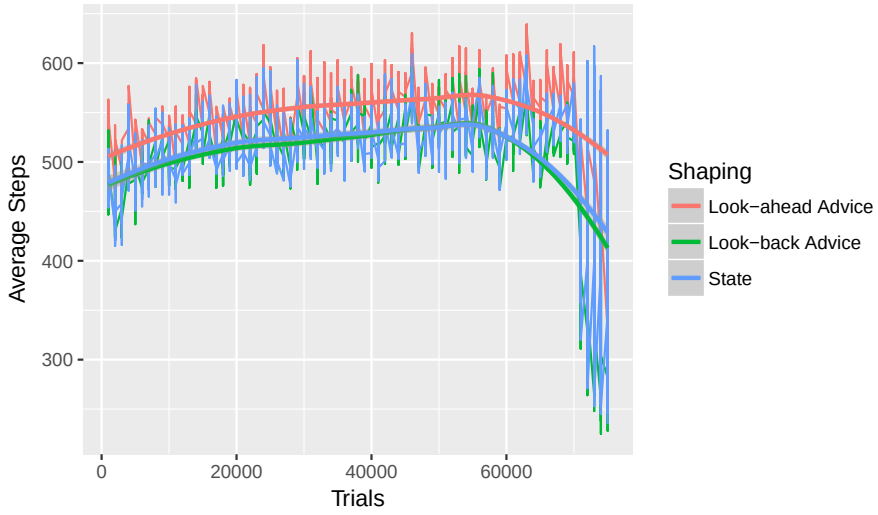

Figure 5.4: Average steps taken until reaching the occupant's desired equivalent temperature using the shaping methods for 20 runs with different seeds (including error bars and Loess fit with 95% confidence level shaded).The agent trained using previous actions achieves a comfort target in less steps than the state-shaping and future-actions trained agents.

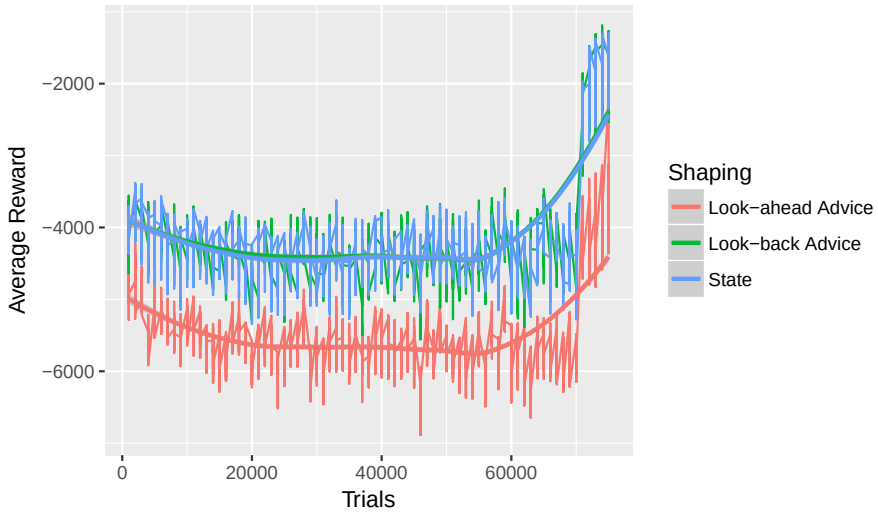

Figure 5.5: Average cumulative reward per trial for 20 runs with different seeds (including error bars and Loess fit with 95% confidence level shaded). State shaping and look-back advice have comparable performance.

Table 5.3: Significance values for the Wilcoxon Rank Test, all values are lower than the 0.05 threshold indicating statistically significant difference between the groups.

Groups	Steps (p-value)	Reward (p-value)
State & Look-back Advice	0.01888	0.021
State & Look-ahead Advice	0.00047	$< 2e^{-16}$
Look-back & Look-ahead Advice	$1.8e^{-8}$	$< 2e^{-16}$

number of steps and for the reward per trial ($p < 2.2e^{-16}$). The null hypothesis was rejected for both the number of steps ($p = 2.47e^{-8}$) and the reward per trial ($p < 2.2e^{-16}$), meaning that there is a statistically significant difference between the three shaping techniques. Additionally, the Wilcoxon rank test was used for paired comparisons between the three methods as it includes corrections for multiple testing. There is a statistically significant difference between each shaping method at a 95% level (table 5.3).

5.2.2 *SARSA-based controllers*

There are alternative algorithms that can further increase the learning speed and reduce the maximisation bias that SARSA (λ) -based agents display. The proposed UBRL controller trained with SARSA (λ) using these shaping methods , namely state shaping (sarsa-fs), look-back advice (sarsa-lba), and look-ahead advice (sarsa-laa) is compared with:

- a set of standard controllers that measure cabin temperature (bang-bang-air), equivalent temperature (bang-bang-et), or the average between the cabin and the surrounding surfaces temperatures (bang-bang-avg);
- the controller trained with Expected SARSA (exp-fs, exp-lba, exp-laa);
- the controller trained with Double SARSA (dsarsa-fs, dsarsa-lba, dsarsa-laa) ;
- and the controller trained with Double Expected SARSA (dexp-fs, dexp-lba, dexp-laa).

These algorithms have not previously been implemented for the HVAC field and combined with the shaping methods.

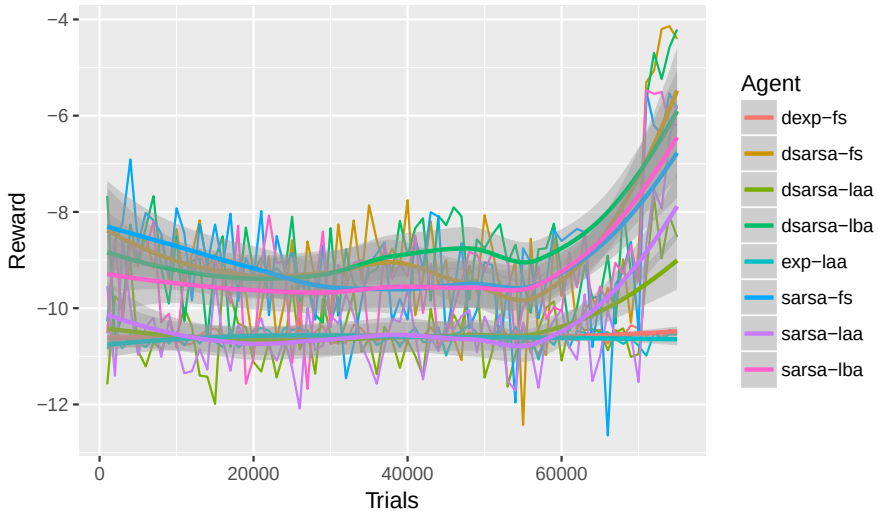

Figure 5.6: Policy performance during learning for the SARSA algorithms, for 70000 episodes the agent is in exploration ($\varepsilon = 0.16$), the rest in exploitation (500 episodes). The Double SARSA agent learning from look-back advice has the highest test scenario reward (Loess fit with 95% confidence band).

The performance of the controllers was tested using the 200 episode scenario, with the Expected SARSA and Double Expected SARSA algorithms having the lowest reward under policy invariance (figure 5.6). While these algorithms generally have a good short term performance, they behave as their Q-learning counterparts for the car cabin environment, given the discount factor ($\gamma = 1$) and the use of a greedy-policy. This means that their on-line behaviour is poor compared to SARSA and Double SARSA, as the agents learn an optimal policy without the impact of action selection (in this case, the controller learns to maintain the actions at a minimum in order to minimise the energy consumption, while increasing the occupant's discomfort).

The highest reward under the test scenario, as well as the average reward per step (figure 5.7) is registered for the Double SARSA UBRL controller using look-back advice. The performance is maintained when using alternative randomised training seeds. Compared to the Double SARSA and SARSA controllers using state shaping (that register a higher reward at the start of the trials but their learning degrades as the number of episodes increases), the look-back advice agents maintain a steady reward through exploration and have the highest increase when exploitation is enabled. Moreover, the Double SARSA agent steadily increases the test scenario

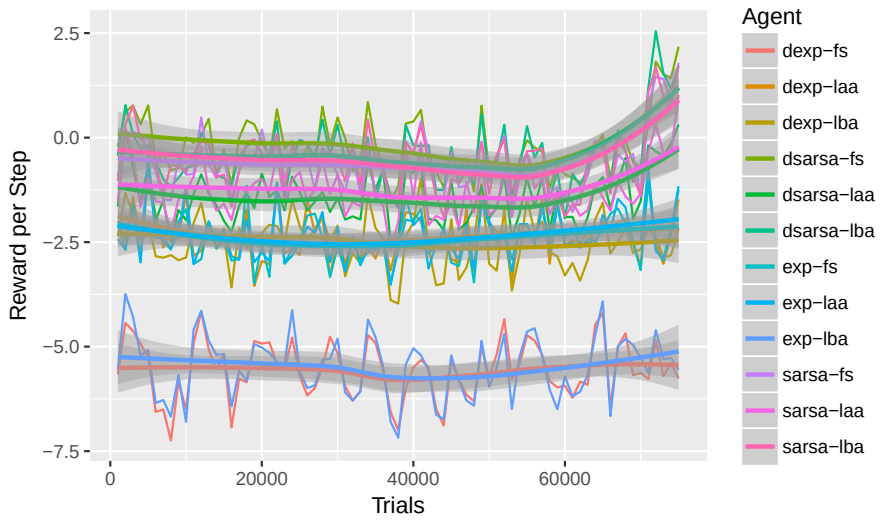

Figure 5.7: The look-back-advice Double SARSA agent has the highest reward per step than the alternative algorithms.

reward due to the use of the look-back shaping method, that helps the agent keep track of the past action choices.

The trade-off between learning and execution can be observed in figure 5.8, as the SARSA algorithms with look-back advice and state shaping achieve the target equivalent temperature in less steps than the alternative algorithms. This can be due to the overly-optimistic maximisation over the expected return as the greedy-selection of the action is not based on a true value, but an estimate of that value.

The controller that has the highest performance in terms of average reward (-4.46) and percentage of time in which the occupant is comfortable (85.87%) is the Double SARSA controller with look-back advice (table 5.4). It surpasses the standard and the original SARSA-based controllers. The caveat for maintaining comfort for a longer period of time is the amount of power used, the controller having the highest average consumption of power of 1.07 kW, but also dealing with one of the highest rates of changes made by the occupants, and having an estimated 5.55 minute response rate.

Analysing the performance of the agents trained with reward shaping, the UBRL Double SARSA HVAC controller trained with look-back advice manages to achieve and maintain the occupant's desired comfort target under warm-up and

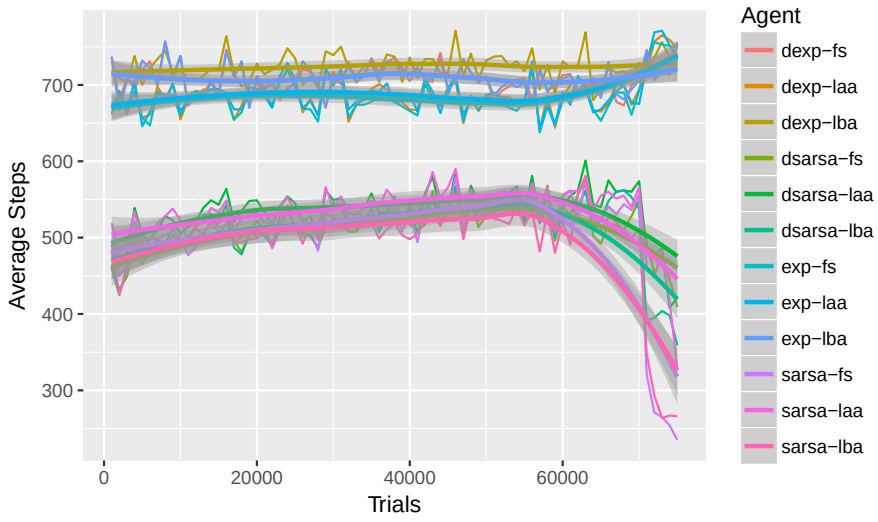

Figure 5.8: Average steps per trial, the SARSA agent achieves the target goal in less steps than the other algorithms (with state shaping and look-back advice).

Table 5.4: Performance of the various controllers for the test set scenario.

Agent	Reward	Avg.Time Target (mins)	Time Comfort (%)	Avg. Power (kW)	No.Changes
exp-laa	-10.17	10.61	8.67	0.02	12.43
dexp-fs	-10.05	10.55	8.05	0.02	12.57
exp-fs	-10.03	11.13	7.98	0.02	12.08
exp-laa	-10.17	10.61	8.67	0.02	12.43
exp-lba	-10.01	10.43	8.3	0.02	12.45
dexp-lba	-9.86	10.65	7.96	0.02	12.95
dexp-laa	-9.79	10.95	7.93	0.02	12.48
dsarsa-laa	-8.077	7.38	61.42	0.69	12.44
bang-bang-et	-6.68	5.82	72.56	0.84	8.44
sarsa-laa	-6.57	5.98	72.91	0.95	16.24
bang-bang-avg	-6.53	5.81	71.19	0.9	13.91
sarsa-fs	-6.25	4.33	80.91	0.84	11.88
bang-bang-air	-6.19	6.23	74.28	0.97	16.7
sarsa-lba	-6.19	4.5	82.92	0.93	12.99
dsarsa-fs	-4.81	7.63	55.85	0.69	22.66
dsarsa-lba	-4.46	5.55	85.87	1.07	14.7

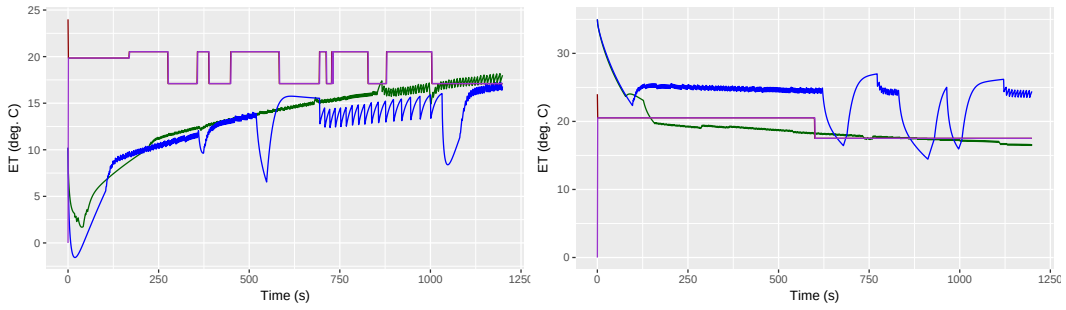

Figure 5.9: Warm-up (left) and cool-down (right) processes of the Double SARSA UBRL HVAC controller trained with look-back advice (green line) and a variable equivalent temperature target (purple line), compared to the bang-bang controller achieved equivalent temperature(blue line).

cool-down conditions, surpassing the alternative algorithms using the shaping methods. Moreover, compared to a standard set-point controller (also known as bang-bang), the UBRL HVAC controller achieves the desired equivalent temperature faster (figure 5.9). It manages to maintain, and alternatively achieve a smoother transition when a change in target is registered, because the desired settings for set-point temperature (5.10), air flow depending on the blower level (5.11), and vent distribution are reached and preserved. The bang-bang controller on the other hand, abruptly inputs hot or cold air into the environment in order to hit the target, struggling to maintain the passenger's comfort.

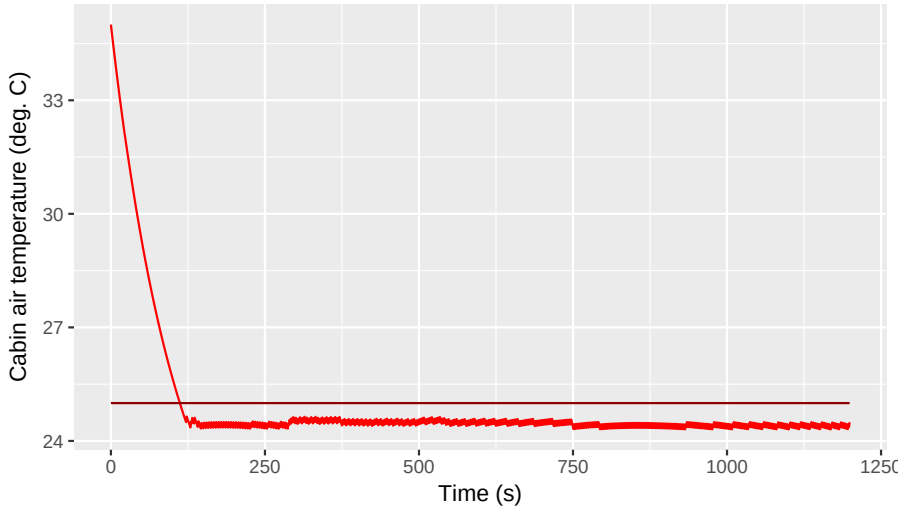

Figure 5.10: Cabin temperature (red) is maintained close to the desired set-point temperature (dark red) by the UBRL controller for the cool-down condition.

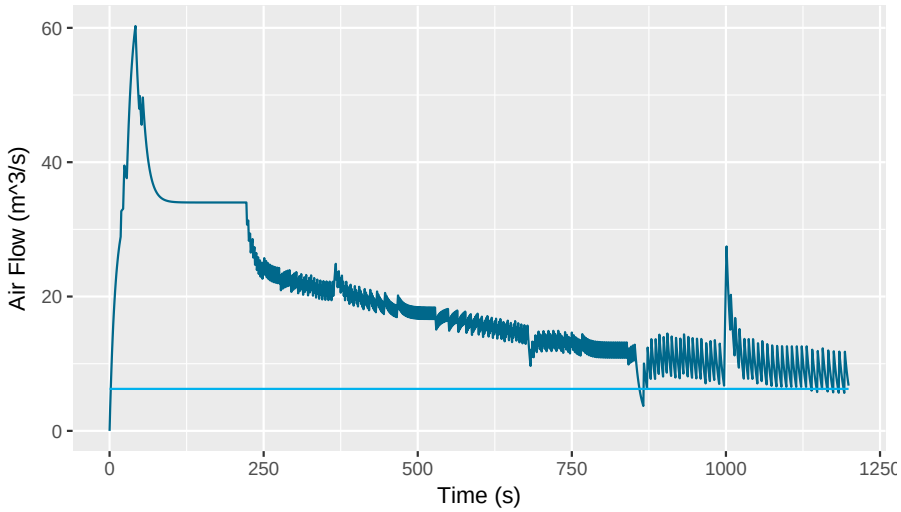

Figure 5.11: Cabin air flow (dark blue) is achieved and maintained to the desired level (blue line) by the UBRL controller for the warm-up condition.

5.3 SUMMARY

This chapter proposes an integrated system, named UBRL HVAC, that includes the cabin environment, capturing both the physical parameters from the physical model and the selected settings from the user model (UBM). The reinforcement learning agent learns a variable target temperature by means of reward shaping. The shaped reward combines the the feedback from the physical model, and the feedback of the UBM in terms of HVAC setting selections.

This chapter aimed to answer the question *"Can the UBRL HVAC system learn and maintain a nearly optimal policy based on occupant preferences within a reasonable amount of time?"*. The Double SARSA UBRL HVAC controller using look-back advice as a shaping method, managed to learn a policy within 2.9 years of training (section 5.2.1) which is less than half the life-time of a car. Furthermore, it achieves and maintains an occupant's desired equivalent temperature within an average of 5.6 minutes (section 5.2.2), ensuring the comfort of the passenger for 86% of the time, surpassing the standard and alternative SARSA-based controllers and avoiding maximisation bias.

The training time is thus improved compared to a system that does not benefit from shaping rewards. However the training duration has room for improvement,

potentially using dynamic shaping methods that take into account a time-based potential function. Depending on the environmental conditions, the agent can provide and maintain comfort for an extended time using the learnt policy.

6

CONCLUSIONS

Human feedback (from experts and non-experts) accelerates Reinforcement Learning (RL) for robotic and gaming platforms. It has not been used for Heating, Ventilation and Air Conditioning (HVAC) RL control within a vehicle cabin environment.

In order to mimic the interaction with the HVAC system as a form of personal comfort control, this thesis proposed the User-Based Module (UBM) that simulates a human agent. The model, based on the combination of three rules, is validated against real-world data. Each rule is based and motivated from existing thermal comfort and thermal behaviour literature.

This research examines the implementation of user feedback using reward shaping, a technique specifically targeting the reward function of RL algorithms. The most suitable shaping method for HVAC control is look-back advice (section 5.2.1), that extends the potential function to the states and actions of the thermal environment. The potential function includes the feedback from the occupant represented by desired changes to HVAC interface as goals for the RL controller. The resulting User-Based Reinforcement Learning (UBRL) system using Double State-Action-Reward-State-Action (SARSA) algorithm outperforms the SARSA trained HVAC controller by maintaining comfort for a longer time but consumes a higher amount of power due to the increased number of occupant setting selections (section 5.2.2). This method combined with the look-back advice (section 5.2.1), eliminates any maximisation bias for action selection and improves the learning speed of the HVAC controller.

6.1 RESEARCH QUESTIONS

This thesis aimed to answer the following over-arching question:

Given the limited interaction that users have with the HVAC, can an RL based system learn occupant's desired settings within the expected lifetime of a car?

The research question, was subsequently split into three sub-questions:

1. *What is the set of simple rules that can be drawn from the thermal comfort literature on occupant thermal behaviour related to HVAC control?*
2. *Can an artificial agent, validated using real-world data, realistically simulate the interaction that humans have with their HVAC system?*
3. *Can the UBRL HVAC system learn and maintain a nearly optimal policy based on occupant preferences within a reasonable amount of time?*

6.2 RESEARCH QUESTION 1

What is the set of simple rules that can be drawn from the thermal comfort literature on occupant thermal behaviour related to HVAC control?

The emerging focus of building thermal comfort literature is modelling the adaptive thermal behaviour of the occupants. There is little investigation in how vehicle cabin occupants behave in order to maintain their desired thermal comfort. Chapter 3 identified the main aspects related to occupant interaction with HVAC controls and combined them in three simple rules:

R1: When people are uncomfortable they are more likely to make changes to the HVAC interface than when they are comfortable.

R2: People are more likely to make changes to the temperature settings, than the blower and vents.

R3: Occupants prefer specific settings depending on the type of environment (either hot, cold or neutral).

The purpose of the rules is to identify what aspects of comfort can be linked to occupant control of the HVAC system, which can be easily represented as conditional probabilities. R1 identifies the relationship between the changes made and the occupant's comfort, which in this thesis is estimated by equivalent temperature (section 3.1). R2 concerns setting selections (the occupant having a choice between temperature, blower, and vent). R2 depends on whether a change is made (R1) and on the overall-body equivalent temperature (section 3.2). Once a setting is selected (R2), R3 identifies how likely the occupants are to select a value depending on their comfort and the thermal environment (section 3.3).

The main objective of answering the first research question is to highlight that occupant thermal behaviour is a key element for thermal comfort modelling especially for the vehicle environment. The literature rules represent the ground truth of how occupants control their HVAC systems, and can be extended to incorporate alternative thermal behaviours such as the activation of heated surfaces.

6.3 RESEARCH QUESTION 2

Can an artificial agent, validated using real-world data, realistically simulate the interaction that humans have with their HVAC system?

Yes. Based on the set of three rules a human agent model is presented in chapter 4. A real-world data set that monitored the occupant's thermal comfort and HVAC interaction is used to validate the rules and their combination.

The agent, named the UBM, is the result of combining a set of classifiers, each estimating an aspect of the occupant's response. A Bayesian model was proposed for estimating R1. The model outperformed alternative classifiers by means of true negative rate, accuracy, and Area Under Curve (AUC). From R2 three models determining each setting selection (temperature, blower, vent) were identified. Two naive-Bayes classifiers were used for estimating temperature selection and vent distribution and a neural network for blower level selection. The three multi-class models (R3) with the highest accuracy and Cohen's kappa were: a neural network

for estimating temperature set-points, conditional inference trees for blower levels, and random forest for vent distribution.

The hybrid model was validated on a final with-held dataset (the testing set). The UBM was compared and outperformed a simple neural network, and a fuzzy logic model by means of accuracy. Therefore the UBM human agent can realistically model the occupant control of the HVAC system and be used in simulation for training a machine learning climate controller.

6.4 RESEARCH QUESTION 3

Can the UBRL HVAC system learn and maintain a nearly optimal policy based on occupant preferences within a reasonable amount of time?

The answer is yes. The UBM model was implemented within the UBRL framework as part of the cabin environment in chapter 5. The occupant feedback (output of the UBM) was combined with the elements of the state and action space by means of a potential function. The resulting shaping reward was combined with the environmental reward in order to train the RL agent. Three shaping techniques were used: standard potential-based shaping (also known as state shaping); look-back advice and look-forward advice. Among the three methods, look-back advice has statistically significant higher reward and number of steps performance than state shaping (section 5.2.1). Both methods surpass look-forward advice by means of number of steps and average reward per trial.

To answer the research question the UBRL HVAC controller trained with the Double SARSA algorithm can learn and maintain a nearly optimal policy under a set of scenarios. For warm-up and cool-down of the cabin the algorithm achieves and maintains comfort within 14 minutes and, respectively, 2 minutes. Additionally, the controller outperforms the standard bang-bang controllers, as well as alternative SARSA trained controllers, ensuring occupant comfort 86% of the duration of the journey. In approximately 5.6 minutes the controller reaches a desired occupant equivalent temperature. Conversely, in order to achieve and maintain occupant comfort, the controller uses more power than alternative controllers (1.07 kW). The

reduction of power consumption will be the aim of future research, by concentrating on finding the appropriate weight for the energy reward, and examining the power usage for a more complex cabin model.

The overall learning performance is reduced to less than half the lifetime of a car. This is reasonable as it surpasses the original SARSA controller that does not even manage to achieve a target temperature for the warm-up process. Despite this fact, the learning is not immediate, therefore a solution is to train the controller in simulation and then implement it on the electronic control unit with variable exploration in order to learn the occupant's desired settings. To improve the training time of 2.9 years, alternative methods that use less samples will be explored for training the controller. Future work will concentrate on examining the effect of dynamic potentials [DK12] (including a time parameter within the shaping function) as well as combining shaping with alternative feedback methods such as advice (section 2.1.3) and demonstration (section 2.1.3).

6.5 OVER-ARCHING QUESTION

This thesis aimed to answer the over-arching question:

Given the limited interaction that users have with the HVAC, can an RL based system learn occupant's desired settings within the expected lifetime of a car?

The answer is yes, it can learn an occupant's desired settings and furthermore maintain them within less than the lifetime of a car (approximately 2.9 years simulation time). This thesis further brings to the attention of its readers the importance of taking a closer look at the adaptive thermal behaviour exhibited by occupants in the context of vehicle cabins and how little is known in this area. It establishes a link between thermal comfort and the decisions to make changes to the climate control by means of three rules.

These rules are expressed as conditional probabilities for the decision to make a change, select the type of setting and the value for that setting. Given the discrete and binary, solution for the rules, the performance of various classifiers

was evaluated and validated using a real world data set. The final model, the UBM, validated against a hold-out set is an approach to modelling a human agent.

Capturing and realistically modelling the thermal actions of cabin occupants is in itself a challenge due to the sparsity of the responses, which can determine class imbalance and randomness. Nevertheless, it is achievable and despite its complexity, is essential for improving the occupants' comfort.

The outputs of the model (setting value selections) are used as user-feedback. The feedback is included within the RL shaping function as occupant desired goals. These are compared with the state of the environment and the actions of the controller, using the look-back advice method. This type of shaping function is included as an additional reward within a Double SARSA -based controller. The Double SARSA HVAC learns to achieve the occupant's desired comfort, maintaining it 86% of the time and surpassing alternative controllers. A drawback of the controller is that it uses on average more power than alternative controllers. This is a result of the weighting system that prioritises a higher level of comfort for the occupants than the power usage. However, it achieves the lowest reward per test scenario and the desired comfort temperature within an average of 5.6 minutes of a journey.

This thesis shows that the feedback of the occupants is valuable for air-conditioning system control in vehicles as it improves the performance of machine learning systems and bridges the gap between thermal comfort and control. The original contributions to the software code include firstly the User-Based Module (Chapter 4), which represents a simulated occupant that changes the settings of the HVAC interface (based on classifiers trained in R using the Caret package, and coded in Java). Secondly the UBRL HVAC controller (Chapter 5) that combines alternative SARSA algorithms with 3 shaping methods (Java coded) in order to identify the best performing system that learns from occupant feedback.

6.6 FUTURE WORK

6.6.1 *Thermal behaviour avenues*

Motivation plays an important role when it comes to the internal process of decision making. For changes in clothing and interactions with additional elements of the vehicle cabin, motivation for choosing these actions is often ambiguous [Ruz11]. The additional adaptive behaviours identified in this thesis (appendix C) require further investigation on the nature of their relationship with thermal comfort. The actions should be monitored and the reasoning behind their use needs to be clarified in further trials and surveys.

Moreover additional factors that can influence occupants' thermal behaviour can be included as input features to the human agent and part of the environmental state. Firstly, body-part equivalent temperature should be considered, namely the difference in temperature between the head and feet, as any difference larger than 3°C [ASH04] can cause thermal discomfort. Moreover skin sensitivity influences occupants' decision to act, and impacts the sensations and comfort felt at the various regions of the body.

Additionally, a detailed analysis concerning the impact of gender, age, country of residence, health is necessary to build a case by case training method for the human agent. The model can be further developed and implemented using occupant profiles (determining how many seats are occupied and their position) for personalised comfort controls. Among these factors the most significant one is gender, as women are more sensitive to cold than men in the context of the built environment [Kar07], therefore an investigation if this hypothesis is valid for the vehicle environment is necessary.

An important aspect of the HVAC system that can restrict the occupants' use of the controls is the noise that the blower produces. This factor has an impact on both ambiance and thermal comfort of the occupants and should be included within the parameters of the environment.

Finally, humidity should be considered as it can influence the passengers' decision to increase or decrease the set-point temperatures and can reduce the fluctuations in the percentages (30-70%) [SBS15] towards the recommended band. It is also important when considering the demisting functions of the controller, using window fogging as an additional constraint in the reward function.

6.6.2 *Simulation related improvements*

There are several aspects that can be improved in the context of the UBRL HVAC system. As exploration was abruptly cut after 70000 episodes, for the remainder 5000 episodes only exploitation is available to the agent. A solution would be variable exploration using either an epsilon with exponential decay within each episode, or an adaptive greedy exploration, as proposed by Tokic [Tok10].

A simple lumped capacitance vehicle cabin model was used. This could be extended to a more complex model as proposed by Lee et al. [Lee+15] or a dual zone model as proposed by Torregrosa et al. [Tor+15].

The system was trained with the feedback of a single person, as the car cabin model included one occupant. The simulation can be expanded to include at least two occupants with similar or completely different behaviours. The advantage of the human agent model is that the nature of the responses is probability based, which means that the frequency of climate control changes will vary for each occupant.

The combination of classifiers is quite high, in relation to the UBM model (further description of the limitations can be found in section 4.3.2). A solution would be to combine the response of the R2 classifiers into one multi-class classifier (example available in appendix D). Moreover, vent distribution control is static. According to Ruzic [Ruz11] dynamic vents can improve the comfort of the occupant. Therefore, adapting the control to a dynamic vent system would extend the values to a range. By using continuous instead of discrete values, R3 becomes a regression problem, for which alternative models could be tested.

Furthermore, the UBRL architecture can be implemented using alternative programming tools. A more complex thermal system, cabin model, and engine model are available in programs such as Theseus-FE, with GT-SUITE, and integrating the Double SARSA and the UBM agents using Java scripts.

Alternatively, the UBRL architecture can be adapted to alternative environments such as smart homes for occupancy monitoring expanding the setting preferences to use of windows, blinds, music, and lighting. Additionally, the system can be used for remote monitoring of patient conditions and uses of medical equipment at home. The system can also be implemented on alternative means of transportation such as trains or planes, combining multiple occupancy profiles.

6.7 CONCLUDING REMARKS

This thesis showcased the benefit of training RL climate controllers with occupant feedback by use of shaping methods. The proposed UBRL HVAC system ensures the comfort of the occupants by achieving their desired (variable) targets. The controller learns a nearly optimal policy using the Double SARSA algorithm in less than half of the lifetime of a car. There are multiple avenues for development and exploration of the UBRL HVAC system from including additional factors to the use of more complex simulation tools. This thesis represents the start of a new journey towards improving occupants' thermal comfort in vehicles based on their preferences and feedback.

REFERENCES

[Abb+07] Pieter Abbeel, Adam Coates, Morgan Quigley and Andrew Y. Ng. 'An application of reinforcement learning to aerobatic helicopter flight'. In: *In Advances in Neural Information Processing Systems 19*. MIT Press, 2007, p. 2007 (cit. on p. 24).

[AN] Pieter Abbeel and Andrew Y. Ng. 'Apprenticeship Learning via Inverse Reinforcement Learning'. In: ACM Press (cit. on pp. 119, 182).

[AB14] Fahad Almutawa and Hanan Buabbas. 'Photoprotection: Clothing and Glass'. In: *Dermatologic Clinics* 32.3 (2014). Photodermatology, pp. 439–448. DOI: http://dx.doi.org/10.1016/j.det.2014.03.016. URL: http://www.sciencedirect.com/science/article/pii/S073386351400031X (cit. on p. 190).

[Alt90] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990 (cit. on pp. 86, 100).

[AMC] AMC. *Intelligent Temperature Monitor and PWM Fan Controller*. URL: http://www.ti.com/lit/ds/symlink/amc6821-q1.pdf (cit. on p. 63).

[Ame+14] Saleema Amershi, Maya Cakmak, W. Bradley Knox and Todd Kulesza. 'Power to the People: The Role of Humans in Interactive Machine Learning'. In: *AI Magazine* (Dec. 2014). URL: http://research.microsoft.com/apps/pubs/default.aspx?id=238110 (cit. on pp. 21, 23).

[AZH06a] Edward Arens, Hui Zhang and Charlie Huizenga. 'Partial- and whole-body thermal sensation and comfort Part II: Non-uniform environmental conditions'. In: *Journal of Thermal Biology* 31 (2006), pp. 60–66. DOI: http://dx.doi.org/10.1016/j.jtherbio.2005.11.027. URL: http://www.sciencedirect.com/science/article/pii/S0306456505001233 (cit. on pp. 35, 59).

[AZH06b] Edward Arens, Hui Zhang and Charlie Huizenga. 'Partial- and whole-body thermal sensation and comfort, Part I: uniform environmental conditions'. In: *Journal of Thermal Biology* 1.31 (2006), pp. 53–59 (cit. on pp. 35, 80).

[Arg+09] Brenna Argall, Sonia Chernova, Manuela Veloso and Brett Browning. 'A Survey of Robot Learning from Demonstration'. In: *Robotics and Autonomous Systems* 67 (2009), pp. 469–483 (cit. on p. 181).

[Asa+12] Shoichi Asaoka, Takashi Abe, Yoko Komada and Yuichi Inoue. 'The factors associated with preferences for napping and drinking coffee as countermeasures for sleepiness at the wheel among Japanese drivers'. In: *Sleep Medicine* 13.4 (2012), pp. 354–361. DOI: http://dx.doi.org/10.1016/j.sleep.2011.07.020. URL: http://www.sciencedirect.com/science/article/pii/S1389945711003911 (cit. on p. 190).

[ASH04] ASHRAE. *Standard 55-2004: Thermal Comfort Conditions for Human Occupancy*. 2004 (cit. on pp. 61, 77, 141).

[ASH10] ASHRAE. *Standard 55-2010:Thermal Environmental Conditions for Human Occupancy*. 2010 (cit. on pp. 55, 80).

[01] 'ASHRAE Handbook-Fundamentals'. In: ASHRAE, 2001. Chap. 8, Thermal Comfort (cit. on p. 34).

[BCB98] Fred Bauman, T Carter and A Baughman. 'Field study of the impact of a desktop task/ambient conditioning system in office buildings'. In: (1998) (cit. on pp. 55, 117).

[BE01] Victoria Bellotti and Keith Eduards. 'Intelligibility and Accountability: Human Considerations in Context Aware Systems'. In: *Human-Computer Interaction* 16.2 (Dec. 2001), pp. 193–212 (cit. on pp. 34, 44).

[Ben+09] Tatiana Benaglia, Didier Chauveau, David Hunter and Derek Young. 'Mixtools: An R package for analyzing finite mixture models'. In: *Journal of Statistical Software* 32.6 (2009), pp. 1–29 (cit. on pp. 71, 82, 197).

[Ber+06] Matt Berlin, Jesse Gray, Andrea L. Thomaz and Cynthia Breazeal. 'Perspective Taking: An Organizing Principle for Learning in Human-Robot Interaction'. In: *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2006, pp. 1444–14450 (cit. on pp. 24, 25, 182).

[Bla+10] Tyler Blake, Freeman Thomas, Matthew Edwards and Andrei Markevich. 'Mobile device interface for use in a vehicle'. US Patent App. 12/765,185. 2010 (cit. on p. 50).

[BKP11] Abdeslam Boularias, Jens Kober and Jan Peters. 'Relative entropy inverse reinforcement learning'. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 182–189 (cit. on p. 31).

[BZA15] Gail S Brager, Hui Zhang and Edward Arens. 'Evolving opportunities for providing thermal comfort'. In: *Building Research & Information* 43.3 (2015), pp. 274–287. DOI: 10.1080/09613218.2015.993536. URL: c%20http://dx.doi.org/10.1080/09613218.2015.993536 (cit. on pp. 35, 41, 56).

[BPD04] Gail Brager, Gwelen Paliaga and Richard De Dear. 'Operable windows, personal control and occupant comfort.' In: *ASHRAE Transactions* 110.2 (2004), pp. 17–35 (cit. on pp. 55, 117).

[BP99] JE Brooks and KC Parsons. 'An ergonomics investigation into human thermal comfort using an automobile seat heated with encapsulated carbonized fabric (ECF)'. In: *Ergonomics* 42.5 (1999), pp. 661–673 (cit. on p. 190).

[Bru+17] James Brusey, Diana Hintea, Elena Gaura and Neil Beloe. 'Reinforcement learning-based thermal comfort control for vehicle cabins'. In: *Mechatronics* (2017) (cit. on pp. 43, 44, 52, 56, 83, 113, 115, 116, 118–120, 122–124).

[Bry+15] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor and Ann Nowé. 'Reinforcement Learning from

Demonstration through Shaping.' In: *IJCAI*. 2015, pp. 3352–3358 (cit. on p. 31).

[Cab92a] M. Cabanac. 'Pleasure: the common currency'. In: *Journal of Theoretical Biology* 155 (1992), pp. 173–200 (cit. on p. 55).

[Cab92b] M. Cabanac. 'What is sensation?' In: *Biological Perspectives on Motivated Activities* (1992) (cit. on p. 37).

[CM07] F. Cascetta and M. Musto. 'Assessment of thermal comfort in a car cabin with sky-roof'. In: *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 221.10 (Oct. 2007), pp. 1251–1258 (cit. on pp. 35, 42).

[CB07] Tulin Gunduz Cengiz and Fatih C. Babalik. 'An on-the-road experiment into the thermal comfort of car seats'. In: *Applied Ergonomics* 38 (2007), pp. 337–347 (cit. on pp. 34, 190).

[Cha+11] Tong-Bou Chang, Deng-Maw Lu, Wen-Yu Yeh and Siou-Ci Syu. 'Air-conditioning clothes used in car'. US Patent App. 12/458,404. 2011 (cit. on p. 189).

[Che+15] Kuo-Huey Chen, Jeffrey Bozeman, Mingyu Wang, Debashis Ghosh, Edward Wolfe and Sourav Chowdhury. *Energy Efficiency Impact of Localized Cooling/Heating for Electric Vehicle*. Tech. rep. SAE Technical Paper, 2015 (cit. on pp. 42, 190).

[CB15] Marta Chludzinska and Anna Bogdan. 'The effect of temperature and direction of airflow from the personalised ventilation on occupants' thermal sensations in office areas'. In: *Building and Environment* 85 (2015), pp. 277–286. DOI: http://dx.doi.org/10.1016/j.buildenv.2014.11.023. URL: http://www.sciencedirect.com/science/article/pii/S0360132314003953 (cit. on pp. 35, 40, 57).

[Com16] Ford Motor Company. *Setting and adjusting climate controls*. 2016. URL: http://owner.ford.com/how-tos/sync-technology/myford-touch/settings/setting-and-adjusting-climate-controls.html (cit. on pp. 50, 59).

[Coo17] Brian Cooley. *How to understand your car's climate controls*. 2017. URL: https://www.cnet.com/how-to/how-to-understand-your-cars-climate-controls/ (cit. on p. 50).

[Cro+15] Cristiana Croitoru, Ilinca Nastase, Florin Bode, Amina Meslem and Angel Dogeanu. 'Thermal comfort models for indoor spaces and vehicles.Current capabilities and future perspectives'. In: *Renewable and Sustainable Energy Reviews* 44 (2015), pp. 304–318. DOI: http://dx.doi.org/10.1016/j.rser.2014.10.105. URL: http://www.sciencedirect.com/science/article/pii/S1364032114009332 (cit. on pp. 34, 36, 42).

[CMoo] J. Currle and J. Maue. *Numerical study of the influence of air vent area and air mass flux on the thermal comfort of car occupants*. Tech. rep. SAE Technical Paper, 2000 (cit. on p. 62).

[DDo8] K. Dalamagkidis and Kolokotsa D. 'Reinforcement Learning for Building Environmental Control'. In: *Reinforcement Learning*. Ed. by Cornelius Weber, Mark Elshaw and Norbert Michael Mayer. Rijeka: IntechOpen, 2008. Chap. 15. DOI: 10.5772/5286. URL: https://doi.org/10.5772/5286 (cit. on pp. 8, 24, 46).

[Dal+07] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis and G.S. Stavrakakis. 'Reinforcement learning for energy conservation and comfort in buildings'. English. In: *Building and Environment* 42.7 (2007), pp. 2686–2698. DOI: 10.1016/j.buildenv.2006.07.010 (cit. on pp. 24, 46).

[Dam+16] Radu Mircea Damian, Mihaela Simion, Lavinia Socaciu and Paula Unguresan. 'Factors which Influence the Thermal Comfort Inside of Vehicles'. In: *Energy Procedia* 85 (2016), pp. 472–480. DOI: http://dx.doi.org/10.1016/j.egypro.2015.12.229. URL: http://www.sciencedirect.com/science/article/pii/S1876610215028945 (cit. on pp. 35, 36, 189).

[DSH14] Mayank Daswani, Peter Sunehag and Marcus Hutter. 'Reinforcement Learning with Value Advice'. In: *JMLR: Workshop and Conference Proceedings*. Vol. 39. 2014, pp. 299–314 (cit. on p. 186).

[DHM11] David Daum, Frederic Haldi and Nicolas Morel. 'A personalized measure of thermal comfort for building controls'. In: *Building and Environment* 46.1 (2011), pp. 3–11. DOI: http://dx.doi.org/10.1016/j.buildenv.2010.06.011. URL: http://www.sciencedirect.com/science/article/pii/S0360132310001915 (cit. on pp. 34, 40).

[Dea11] Richard de Dear. 'Revisiting an old hypothesis of human thermal perception: alliesthesia'. In: *Building Research & Information* 39.2 (2011), pp. 108–117. DOI: 10.1080/09613218.2011.552269. eprint: http://dx.doi.org/10.1080/09613218.2011.552269. URL: http://dx.doi.org/10.1080/09613218.2011.552269 (cit. on pp. 37, 54, 59).

[DeM15] Doug DeMuro. *QOTD: Why Do You Hate Automatic Climate Control?* 2015. (Visited on 19/12/2016) (cit. on pp. 50, 57).

[DK12] Sam Devlin and Daniel Kudenko. 'Dynamic potential-based reward shaping'. In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems. 2012, pp. 433–440 (cit. on pp. 30, 139).

[DDG01] A. K. Dey, Salber D. and Abowd G.D. 'A conceptual framework and a toolkit for supporting the rapid prototyping of context- aware applications'. In: *Human Computer Interaction* (2001). Special Issue (cit. on p. 34).

[DiG+09] Jack DiGiovanna, Babak Mahmoudi, Jose Fortes and Justin C. Sanchez. 'Co-adaptive brain-machine interface via Reinforcement Learning'. In: *IEEE Trans Biomed England* 56.1 (Jan. 2009), pp. 54–64 (cit. on pp. 20, 24).

[DC98] Marco Dorigo and Marco Colombetti. *Robot shaping: an experiment on behavior engineering*. MIT Press, 1998 (cit. on p. 26).

[DFT91] John Comstock Doyle, Bruce A. Francis and Allen R. Tannenbaum. *Feedback Control Theory*. Prentice Hall Professional Technical Reference, 1991 (cit. on p. 41).

[Du+14] Xiuyuan Du, Baizhan Li, Hong Liu, Dong Yang, Wei Yu, Jianke Liao, Zhichao Huang and Kechao Xia. 'The Response of Human Thermal Sensation and Its Prediction to Temperature Step-Change (Cool-Neutral-Cool)'. In: *PLoS ONE* 9.8 (Aug. 2014), e104320. DOI: 10.1371/journal.pone.0104320. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0104320 (cit. on pp. 54, 61).

[Eil99] Andreas Eilemann. *Practical Noise and Vibration Optimization of HVAC Systems*. Tech. rep. SAE Technical Paper, 1999 (cit. on p. 191).

[FAC15] Valentina Fabi, Rune Korsholm Andersen and Stefano Corgnati. 'Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings'. In: *Building and Environment* 94, Part 1 (2015), pp. 371–383. DOI: http://dx.doi.org/10.1016/j.buildenv.2015.08.016. URL: http://www.sciencedirect.com/science/article/pii/S0360132315300974 (cit. on p. 45).

[Fab+12] Valentina Fabi, Rune Vinther Andersen, Stefano Corgnati and Bjarne W. Olesen. 'Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models'. In: *Building and Environment* 58 (2012), pp. 188–198. DOI: http://dx.doi.org/10.1016/j.buildenv.2012.07.009. URL: http://www.sciencedirect.com/science/article/pii/S0360132312001977 (cit. on pp. 45, 48).

[Fab+15] Valentina Fabi, Martina Sugliano, Rune Korsholm Andersen and Stefano Paolo Corgnati. 'Validation of Occupants' Behaviour Models for Indoor Quality Parameter and Energy Consumption Prediction'. In: *Procedia Engineering* 121 (2015). The 9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) joint with the 3rd International Conference on Building Energy and Environment (COBEE), 12-15 July 2015, Tianjin, China, pp. 1805–1811. DOI:

http://dx.doi.org/10.1016/j.proeng.2015.09.160. URL: http://www.sciencedirect.com/science/article/pii/S1877705815029884 (cit. on pp. 45, 47).

[Fan73] Povl Ove Fanger. 'Assessment of man's thermal comfort in practice'. In: *Occupational and Environmental Medicine* 30.4 (1973), pp. 313–324 (cit. on p. 36).

[FT08] Yadollah Farzaneh and Ali A. Tootoonchi. 'Controlling automobile thermal comfort using optimized fuzzy controller'. In: *Science Direct* 28 (2008), pp. 1906–1917 (cit. on pp. 43, 44).

[Faz+14] Pedro Fazenda, Kalyan Veeramachaneni, Pedro Lima and Una-May O'Reilly. 'Using Reinforcement Learning to Optimize Occupant Comfort and Energy Usage in HVAC Systems'. In: *J. Ambient Intell. Smart Environ.* 6.6 (Nov. 2014), pp. 675–690. URL: http://dl.acm.org/citation.cfm?id=2693820.2693826 (cit. on pp. 47, 57).

[FLJ99] Málrio A.T. Figueiredo, José MN Leitão and Anil K Jain. 'On fitting mixture models'. In: *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer. 1999, pp. 54–69 (cit. on p. 70).

[Fle71] Joseph L Fleiss. 'Measuring nominal scale agreement among many raters.' In: *Psychological bulletin* 76.5 (1971), p. 378 (cit. on pp. 86, 97, 102).

[Foc06] Driver Focus-telematics. *Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems*. June 2006 (cit. on p. 49).

[Foj+16] Milos Fojtlin, Michal Planka, Jan Fiser, Jan Pokorny and Miroslav Jicha. 'Airflow Measurement of the Car HVAC Unit Using Hot-wire Anemometry'. In: *EPJ Web of Conferences*. Vol. 114. EDP Sciences. 2016, p. 02023 (cit. on p. 116).

[FF86] Bruno S Frey and Klaus Foppa. 'Human behavior: possibilities explain action'. In: *Journal of Economic Psychology* 7.2 (1986), pp. 137–160 (cit. on p. 49).

[FW11] Monika Frontczak and Pawel Wargocki. 'Literature survey on how different factors influence human comfort in indoor environments'. In: *Building and Environment* 46.4 (2011), pp. 922–937. DOI: 10.1016/j.buildenv.2010.10.021. URL: http://dx.doi.org/10.1016/j.buildenv.2010.10.021 (cit. on pp. 38, 41, 48, 191).

[Fug+18] U. Fugiglando, D Santucci, I. Bojic, T. Chin To Cheoung, S. Schiavon and C. Ratti. 'Developing Personal Thermal Comfort Models for Control of HVAC in Cars Using Field Data'. In: *Windsor Conference: Rethinking Comfort*. Apr. 2018 (cit. on pp. 49, 55).

[FH09] T. Fukazawa and G. Havenith. 'Differences in comfort perception in relation to local and whole body skin wettedness'. In: *European JournaL of Applied Physiology* 106.1 (2009), pp. 15–24 (cit. on pp. 35, 59).

[GFB86] A.P. Gagge, A.P. Fobelets and L.G. Berglund. *A standard predictive index of human response to the thermal environment*. Vol. 92:2B. Jan. 1986 (cit. on p. 36).

[Gal+11] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince and Francisco Herrera. 'An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes'. In: *Pattern Recognition* 44.8 (2011), pp. 1761–1776 (cit. on pp. 108, 193).

[GDH16] Michael Ganger, Ethan Duryea and Wei Hu. 'Double Sarsa and Double Expected Sarsa with Shallow and Deep Learning'. In: *Journal of Data Analysis and Information Processing* 4 (2016), pp. 159–176. DOI: 10.4236/jdaip.2016.44014. URL: http://www.scirp.org/journal/jdaip (cit. on pp. 18, 19).

[Gar08] Jose J Garcia. 'Heated and cooled steering wheel'. US Patent D559,158. Jan. 2008 (cit. on p. 190).

[GL94] Jean-Luc Gauvain and Chin-Hui Lee. 'Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains'. In: *IEEE transactions on speech and audio processing* 2.2 (1994), pp. 291–298 (cit. on p. 68).

[GTB15] Ali Ghahramani, Chao Tang and Burcin Becerik-Gerber. 'An online learning approach for quantifying personalized thermal comfort via adaptive stochastic modeling'. In: *Building and Environment* 92 (2015), pp. 86–96 (cit. on p. 40).

[GH+96] Zoubin Ghahramani, Geoffrey E Hinton et al. *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto, 1996 (cit. on p. 71).

[GOG15] Carlo Giaconia, Aldo Orioli and Alessandra Di Gangi. 'A correlation linking the predicted mean vote and the mean thermal vote based on an investigation on the human thermal comfort in short-haul domestic flights'. In: *Applied Ergonomics* 48 (2015), pp. 202–213. DOI: http://dx.doi.org/10.1016/j.apergo.2014.12.003. URL: http://www.sciencedirect.com/science/article/pii/S0003687014002932 (cit. on p. 34).

[Giu+13] Valeria De Giuli, Roberto Zecchin, Luigi Salmaso, Livio Corain and Michele De Carli. 'Measured and perceived indoor environmental quality: Padua Hospital case study'. In: *Building and Environment* 59 (2013), pp. 211–226. DOI: http://dx.doi.org/10.1016/j.buildenv.2012.08.021. URL: http://www.sciencedirect.com/science/article/pii/S0360132312002235 (cit. on pp. 34, 38).

[GWP11] David M. Goldstein, John White and William T. Powers. *Perceptual Control Theory (PCT) Applied to Personality, Psychotherapy, and Psychopathology*. Dec. 2011. URL: http://www.pctweb.org/Goldsteinetal2011.pdf (cit. on p. 45).

[Gri+13] S. Griffith, K. Subramanian, J. Scholz, C.L. Isbell and A. Thomaz. 'Policy Shaping: Integrating Human Feedback with Reinforcement

Learning'. In: *Advances in Neural Information Processing Systems (NIPS)* (2013), pp. 2625–2633 (cit. on pp. 20, 27).

[Grz17] Marek Grześ. 'Reward Shaping in Episodic Reinforcement Learning'. In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '17. São Paulo, Brazil: International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 565–573. URL: http://dl.acm.org/citation.cfm?id=3091125.3091208 (cit. on p. 119).

[GK09] Marek Grzes and Daniel Kudenko. 'Learning shaping rewards in model-based reinforcement learning'. In: *Proc. AAMAS 2009 Workshop on Adaptive Learning Agents*. Vol. 115. 2009 (cit. on p. 30).

[GK10] Marek Grzes and Daniel Kudenko. 'Online learning of shaping rewards in reinforcement learning'. In: *Neural Networks* 23.4 (2010), pp. 541–550 (cit. on p. 30).

[GOB13] H Burak Gunay, William O'Brien and Ian Beausoleil-Morrison. 'A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices'. In: *Building and Environment* 70 (2013), pp. 31–47 (cit. on pp. 37, 40, 45, 47).

[Hag02] Fariborz Haghighat. 'Thermal comfort in housing and thermal environments'. In: *Sustainable Built Envrionment* I (2002) (cit. on pp. 34, 37, 48, 54).

[Hal+15] John Halloran, Setiadi Yazid, Dan Goldsmith, Ross Wilkins and Elena Gaura. 'Cool to Warm Up? Understanding Student Energy Behaviour In Indonesian University Buildings'. In: *Conference2015 TAU Conference* (2015) (cit. on pp. 55, 117).

[HN98] Hollister A Hartman and Jerome Go Ng. 'Graphical user interface with electronic feature access'. US Patent 5,821,935. 1998 (cit. on p. 50).

[Har+15] Anna Harutyunyan, Sam Devlin, Peter Vrancx and Ann Nowé. 'Expressing Arbitrary Reward Functions as Potential-Based Advice.' In: *AAAI*. 2015, pp. 2652–2658 (cit. on p. 30).

[HC04] Victor Albert Walter Hillier and Peter Coombes. *Hillier's fundamentals of motor vehicle technology*. Nelson Thornes, 2004 (cit. on pp. 42, 54).

[Hin14] Diana Hintea. 'Reinforcement Learning-based Thermal Comfort for Vehicle Cabins'. PhD thesis. Coventry University, 2014 (cit. on pp. 3, 8, 24, 43, 44, 52, 56, 111, 113, 123).

[Hin+11] Diana Hintea, James Brusey, Elena Gaura, Neil Beloe and David Bridge. 'Mutual information-based sensor positioning for car cabin comfort control'. In: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Vol. 6883. Lecture Notes in Computer Science. Springer, 2011, pp. 483–492 (cit. on p. 48).

[Hin+14] Diana Hintea, John Kemp, James Brusey, Elena Gaura and Neil Beloe. 'Applicability of thermal comfort models to car cabin environments'. In: *Informatics in Control, Automation and Robotics (ICINCO), 2014 11th International Conference on*. Vol. 1. IEEE. 2014, pp. 769–776 (cit. on p. 36).

[Hit+12] Sara C Hitchman et al. 'Predictors of car smoking rules among smokers in France, Germany and the Netherlands'. In: *The European Journal of Public Health* 22.suppl 1 (2012), pp. 17–22 (cit. on p. 190).

[Hod13] Simon Hodder. 'Automotive Ergonomics'. In: ed. by Nikolaos Gkikas. CRC Press, 2013. Chap. 7-Thermal Environments and Vehicles, pp. 97–122 (cit. on pp. 38, 48, 50, 189).

[Hoh+14] Silke Hohls, Thomas Biermeier, Ralf Balschke and Stefan Becker. 'Psychoacoustic analysis of HVAC noise with equal loudness'. In: *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*. Vol. 249. 6. Institute of Noise Control Engineering. 2014, pp. 1800–1806 (cit. on p. 191).

[Hon+15a] Tianzhen Hong, Simona D'Oca, Sarah C. Taylor-Lange, William J.N. Turner, Yixing Chen and Stefano P. Corgnati. 'An ontology to represent energy-related occupant behavior in buildings. Part II: Implementation of the DNAS framework using an XML schema'. In: *Building and Environment* 94, Part 1 (2015), pp. 196–205. DOI: http://dx.doi.org/

10.1016/j.buildenv.2015.08.006. URL: http://www.sciencedirect.com/science/article/pii/S0360132315300871 (cit. on pp. 35, 46).

[Hon+15b] Tianzhen Hong, Simona D'Oca, William J.N. Turner and Sarah C. Taylor-Lange. 'An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework'. In: *Building and Environment* 92 (2015), pp. 764–777. DOI: http://dx.doi.org/10.1016/j.buildenv.2015.02.019. URL: http://www.sciencedirect.com/science/article/pii/S0360132315000761 (cit. on pp. 35, 46).

[Hor01] Michael F Hordeski. *HVAC control in the new millennium*. CRC Press, 2001 (cit. on p. 49).

[HF93] Tetsumi Horikoshi and Yoshimaru Fukaya. 'Responses of human skin temperature and thermal sensation to step change of air temperature'. In: *Journal of Thermal Biology* 18.5 (1993), pp. 377–380. DOI: http://dx.doi.org/10.1016/0306-4565(93)90061-W. URL: http://www.sciencedirect.com/science/article/pii/030645659390061W (cit. on p. 61).

[HR99] Jim Horne and Louise Reyner. 'Vehicle accidents related to sleep: a review.' In: *Occupational and environmental medicine* 56.5 (1999), pp. 289–294 (cit. on p. 190).

[HG01] M.N. Howell and T.J. Gordon. 'Continuous action reinforcement learning automata and their application to adaptive digital filter design'. In: *Engineering Applications of Artificial Intelligence* 14.5 (2001), pp. 549–561. DOI: http://dx.doi.org/10.1016/S0952-1976(01)00034-3. URL: http://www.sciencedirect.com/science/article/pii/S0952197601000343 (cit. on p. 24).

[Hui+06] Charlie Huizenga, Sahar Abbaszadeh, Leah Zagreus and Edward A Arens. 'Air quality and thermal comfort in office buildings: results of a large indoor environmental quality survey'. In: *Healthy Buildings 2006* 3 (June 2006), pp. 393–397 (cit. on pp. 55, 117).

[HHA01] Charlie Huizenga, Zhang Hui and Edward Arens. 'A model of human physiology and comfort for assessing complex thermal environments'. In: *Building and Environment* 36.6 (2001), pp. 691–699 (cit. on p. 36).

[Ich04] Toshihiko Ichinose. 'Touch panel input for automotive devices'. US Patent 6,819,990. 2004 (cit. on p. 50).

[Isb+06] Charles Lee Isbell, Michael Kearns, Satinder Singh, Christian R. Shelton, Peter Stone and Dave Kormann. 'Cobot in LambdaMOO: An adaptive social statistics agent'. In: *Autonomous Agents and Multi-Agent Systems* 13.3 (2006), pp. 327–354. DOI: 10.1007/s10458-006-0005-z (cit. on pp. 24, 28).

[JDP14] J.D.Power. *2014 UK Vehicle Ownership Satisfaction Study*. 2014. URL: http://www.jdpower.com/de/press-releases/2014-uk-vehicle-ownership-satisfaction-study-voss (cit. on p. 42).

[Jag+08] Anke Jager et al. 'Numerical and Experimental Investigations of the Noise Generated by a Flap in a Simplified HVAC Duct'. In: *14th AIAA/CEAS Aeroacoustics Conference*. May. 2008, pp. 5–7 (cit. on p. 191).

[Jaz+13] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo and M. Orosz. 'Personalized Thermal Comfort-Driven Control in HVAC-Operated Office Buildings'. In: *Computing in Civil Engineering (2013)*. Vol. 23-25. American Society of Civil Engineers, June 2013. Chap. 28, pp. 218–225. DOI: 10.1061/9780784413029.028. eprint: http://ascelibrary.org/doi/pdf/10.1061/9780784413029.028. URL: http://ascelibrary.org/doi/abs/10.1061/9780784413029.028 (cit. on p. 40).

[JMB13] Farrokh Jazizadeh, Franco Moiso Marin and Burcin Becerik-Gerber. 'A thermal preference scale for personalized comfort profile identification via participatory sensing'. In: *Building and Environment* 68 (2013), pp. 140–149. DOI: http://dx.doi.org/10.1016/j.buildenv.2013.06.011. URL: http://www.sciencedirect.com/science/article/pii/S0360132313001893 (cit. on p. 40).

[JCR15] Matthew A Jeffers, Larry Chaney and John P Rugh. 'Climate control load reduction strategies for electric drive vehicles in warm weather'. In: *SAE International* April (2015), pp. 21–23. DOI: 10.4271/2015-01-0355.Copyright (cit. on pp. 39, 59, 62, 118).

[Ji+14] HoSeong Ji, YoonKee Kim, JangSik Yang and KyungChun Kim. 'Study of thermal phenomena in the cabin of a passenger vehicle using finite element analysis: human comfort and system performance'. In: *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 228.12 (Oct. 2014), pp. 1468–1479 (cit. on p. 36).

[Joh02] Valerie H. Johnson. 'Fuel Used for Vehicle Air Conditioning: A State-by-State Thermal Comfort-Based Approach'. In: *Society of Automotive Engineers, Inc.* 01.724 (2002), pp. 1957–1970. DOI: 10.4271/2002-01-1957 (cit. on pp. 39, 55, 59).

[Jon+09] Miranda R Jones, Ana Navas-Acien, Jie Yuan and Patrick N Breysse. 'Secondhand tobacco smoke concentrations in motor vehicles: a pilot study'. In: *Tobacco control* 18.5 (2009), pp. 399–404 (cit. on p. 190).

[KLM96] Leslie Pack Kaelbling, Michael L Littman and Andrew W Moore. 'Reinforcement learning: A survey'. In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285 (cit. on p. 14).

[KB14] Kiran R. Kambly and Thomas H. Bradley. 'Estimating the HVAC energy consumption of plug-in electric vehicles'. In: *Journal of Power Sources* 259 (2014), pp. 117–124. DOI: http://dx.doi.org/10.1016/j.jpowsour.2014.02.033. URL: http://www.sciencedirect.com/science/article/pii/S037877531400216X (cit. on p. 42).

[Kar07] Sami Karjalainen. 'Gender differences in thermal comfort and use of thermostats in everyday thermal environments'. In: *Building and Environment* 42.4 (Apr. 2007), pp. 1594–1603. DOI: 10.1016/j.buildenv.2006.01.009. URL: http://dx.doi.org/10.1016/j.buildenv.2006.01.009 (cit. on pp. 38, 53, 57, 141, 192).

[KK07] Sami Karjalainen and Olavi Koistinen. 'User problems with individual temperature control in offices'. In: *Building and Environment* 42.8 (2007), pp. 2880–2887. DOI: http://dx.doi.org/10.1016/j.buildenv.2006.10.031. URL: http://www.sciencedirect.com/science/article/pii/S0360132306003349 (cit. on pp. 40, 54, 191, 192).

[KL11] Sami Karjalainen and Veijo Lappalainen. 'Integrated control and user interfaces for a space'. In: *Building and Environment* 46.4 (2011), pp. 938–944 (cit. on p. 45).

[KAA16] George Katavoutas, Margarita N. Assimakopoulos and Dimosthenis N. Asimakopoulos. 'On the determination of the thermal comfort conditions of a metropolitan city underground railway'. In: *Science of The Total Environment* 566-567 (2016), pp. 877–887. DOI: http://dx.doi.org/10.1016/j.scitotenv.2016.05.047. URL: http://www.sciencedirect.com/science/article/pii/S0048969716309780 (cit. on p. 34).

[KWA09] A. Khalili, C. Wu and H. Aghajan. 'Autonomous Learning of Users Preference of Music and Light Services in Smart Home Applications'. In: *Proceedings of the German AI Conference*. 2009 (cit. on p. 40).

[Kim+04] H. J. Kim, Michael I. Jordan, Shankar Sastry and Andrew Y. Ng. 'Autonomous Helicopter Flight via Reinforcement Learning'. In: *Advances in Neural Information Processing Systems 16*. Ed. by S. Thrun, L.K. Saul and B. Schölkopf. MIT Press, 2004, pp. 799–806. URL: http://papers.nips.cc/paper/2455-autonomous-helicopter-flight-via-reinforcement-learning.pdf (cit. on p. 24).

[KSB18] J. Kim, S. Schiavon and G. Brager. 'Personal comfort models - new paradigm in thermal comfort for occupant-centric environmental control'. In: *Windsor Conference: Rethinking Comfort*. Apr. 2018 (cit. on p. 46).

[Kim+18] Joyce Kim, Yuxun Zhou, Stefano Schiavon, Paul Raftery and Gail Brager. 'Personal comfort models: predicting individuals' thermal

preference using occupant heating and cooling behavior and machine learning'. In: *Building and Environment* 129 (2018), pp. 96–106 (cit. on pp. 46, 83).

[Kim+13] Jungsoo Kim, Richard de Dear, Christhina Candido, Hui Zhang and Edward Arens. 'Gender differences in office occupant perception of indoor environmental quality (IEQ)'. In: *Building and Environment* 70 (2013), pp. 245–256. DOI: http://dx.doi.org/10.1016/j.buildenv.2013.08.022. URL: http://www.sciencedirect.com/science/article/pii/S0360132313002485 (cit. on pp. 38, 48, 191).

[KC98] MM Klarin and JM Cvijanovic. 'The optimization of the interior of the passenger car'. In: *International journal of vehicle design* 19.4 (1998), pp. 448–453 (cit. on p. 190).

[Kno12] W. Bradley Knox. 'Learning from Human-Generated Reward'. PhD thesis. Massachusetts Institute of Technology, 2012 (cit. on pp. 23, 51).

[KFS09] W. Bradley Knox, Ian Fasel and Peter Stone. 'Design Principles for Creating Human-Shapable Agents'. In: *AAAI Spring 2009 Symposium on Agents that Learn from Human Teachers*. Mar. 2009 (cit. on pp. 20, 21, 33).

[Kno+12] W. Bradley Knox, Brian D. Glass, Bradley C. Love, W. Todd Maddox and Peter Stone. 'How Humans Teach Agents: A New Experimental Perspective'. In: *International Journal of Social Robotics* 4 (4 2012), pp. 409–421 (cit. on pp. 21, 23, 31, 32).

[KSS11] W. Bradley Knox, Adam Setapen and Peter Stone. 'Reinforcement Learning with Human Feedback in Mountain Car'. In: *Association for the Advancement of Artificial Intelligence* 05 (Spring 2011), pp. 36–41 (cit. on pp. 24, 32, 33).

[KS08] W. Bradley Knox and Peter Stone. 'TAMER: Training an Agent Manually via Evaluative Reinforcement'. In: *IEEE 7th International Conference on Development and Learning*. Aug. 2008 (cit. on pp. 32, 33).

[KS09] W. Bradley Knox and Peter Stone. 'Interactively Shaping Agents via Human Reinforcement: The TAMER Framework'. In: *The Fifth International Conference on Knowledge Capture*. Sept. 2009. URL: http://www.cs.utexas.edu/users/ai-lab/?KCAP09-knox (cit. on pp. 21, 32).

[KS12a] W. Bradley Knox and Peter Stone. 'Reinforcement Learning from Human Reward: Discounting in Episodic Tasks'. In: *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man)*. Sept. 2012 (cit. on pp. 21, 22, 32).

[KS12b] W. Bradley Knox and Peter Stone. 'Reinforcement Learning from Simultaneous Human and MDP Reward'. In: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Valencia, Spain, June 2012 (cit. on pp. 24, 32, 33).

[KS10] W.Bradley Knox and Peter Stone. 'Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning'. In: *Proceedings of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*. May 2010 (cit. on pp. 32, 33).

[Koh95] Ron Kohavi. 'A study of cross-validation and bootstrap for accuracy estimation and model selection'. In: *IJCAI*. Vol. 14. 2. Stanford, CA. 1995, pp. 1137–1145 (cit. on p. 83).

[KST04] M. Kolich, N. Seal and S. Taboun. 'Automobile seat comfort prediction: statistical model vs. artificial neural network'. In: *Applied Ergonomics* 35 (Jan. 2004), pp. 275–284 (cit. on p. 190).

[KB06] George Konidaris and Andrew G. Barto. 'Autonomous shaping: knowledge transfer in reinforcement learning'. In: *Proceedings of the 23rd Internation Conference on Machine Learning*. 2006, pp. 489–496 (cit. on pp. 24, 26, 27).

[KNG12] J. Kranz, T.I. van Niekerk an H.F.G Holdack-Janssen and G. Gruhler. 'Automotive thermal comfort control-A black box approach'. In: *SAIEE*

Africa Research Journal 103.2 (June 2012), pp. 66–76 (cit. on pp. 43, 44, 48).

[Kra11] Jurgen Kranz. 'Intelligent automotive thermal comfort control'. PhD thesis. 2011 (cit. on p. 43).

[KJ13] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013 (cit. on pp. 72, 74, 82, 83, 85, 86, 197).

[LK77] J. R. Landis and G. G. Koch. 'The measurement of observer agreement for categorical data'. In: *Biometrics* (1977), pp. 159–174 (cit. on pp. 86, 97, 99, 102).

[LWG12] Jared Langevin, Jin Wen and Patrick L Gurian. 'Relating occupant perceived control and thermal comfort: statistical analysis on the ASHRAE RP-884 database'. In: *HVAC & Research* 18.1-2 (2012), pp. 179–194 (cit. on pp. 55, 117).

[LWG13] Jared Langevin, Jin Wen and Patrick L. Gurian. 'Modeling thermal comfort holistically: Bayesian estimation of thermal sensation, acceptability, and preference distributions for office building occupants'. In: *Building and Environment* 69 (2013), pp. 206–226. DOI: http://dx.doi.org/10.1016/j.buildenv.2013.07.017. URL: http://www.sciencedirect.com/science/article/pii/S0360132313002151 (cit. on pp. 34, 38, 46).

[LWG15] Jared Langevin, Jin Wen and Patrick L. Gurian. 'Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors'. In: *Building and Environment* 88 (2015). Interactions between human and building environment, pp. 27–45. DOI: http://dx.doi.org/10.1016/j.buildenv.2014.11.037. URL: http://www.sciencedirect.com/science/article/pii/S0360132314004090 (cit. on pp. 37, 46, 189).

[LWG16] Jared Langevin, Jin Wen and Patrick L. Gurian. 'Quantifying the human-building interaction: Considering the active, adaptive occupant in building performance simulation'. In: *Energy and Buildings* 117

(2016), pp. 372–386. DOI: http://dx.doi.org/10.1016/j.enbuild.2015.09.026. URL: http://www.sciencedirect.com/science/article/pii/S037877881530267X (cit. on pp. 47, 48, 189).

[LD02] Adam Laud and Gerald DeJong. 'Reinforcement learning and shaping: Encouraging intended behaviors'. In: *ICML*. 2002, pp. 355–362 (cit. on p. 30).

[LD03] Adam Laud and Gerald DeJong. 'The influence of reward on the speed of reinforcement learning: An analysis of shaping'. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 440–447 (cit. on p. 30).

[Lee+15] Hoseong Lee, Yunho Hwang, Ilguk Song and Kilsang Jang. 'Transient thermal model of passenger car's cabin and implementation to saturation cycle with alternative working fluids'. In: *Energy* 90 (2015), pp. 1859–1868. DOI: https://doi.org/10.1016/j.energy.2015.07.016. URL: http://www.sciencedirect.com/science/article/pii/S0360544215009111 (cit. on pp. 115, 142).

[LTM13] Adrain Leon, Ana C. Tenorio-Gonzalez and Eduardo F. Morales. 'Human Interaction for Effective Reinforcement Learning'. In: *Machine Learning and Knowledge Discovery in Databases*. Vol. 8188. Lecture Notes in Computer Science 23-27. European Conference, ECML PKDD 2013. Springer, Sept. 2013 (cit. on pp. 25, 182, 183, 187).

[Leo+11] Adrian Leon, Eduardo F. Morales, Leopoldo Altamarino and Jamie R. Ruiz. 'Teaching a Robot to Perform Task through Imitation and On-line Feedback'. In: *Springer-Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* 7042 (2011), pp. 549–556 (cit. on pp. 25, 182, 183).

[Li+13] Guangliang Li, Hayley Hung, Shimon Whiteson and W. Bradley Knox. 'Using Informative Behavior to Increase Engagement in the TAMER Framework'. In: *AAMAS 2013: Proceedings of the Twelfth International*

Joint Conference on Autonomous Agents and Multi-Agent Systems. May 2013, pp. 909–916 (cit. on pp. 22, 32).

[Lia+16] Yitao Liang, Marlos C Machado, Erik Talvitie and Michael Bowling Franklin. 'State of the Art Control of Atari Games Using Shallow Reinforcement Learning'. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2016, pp. 485–493. URL: https://www.fandm.edu/uploads/files/617813975725918530-aamas2016-shallow-rl.pdf (cit. on p. 19).

[Lil+15] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver and Daan Wierstra. 'Continuous control with deep reinforcement learning'. In: *arXiv preprint arXiv:1509.02971* (2015) (cit. on p. 19).

[Lim+12] Varad M Limaye, MD Deshpande, M Sivapragasam and Vivek Kumar. 'Design of dynamic airvents and airflow analysis in a passenger car cabin'. In: *SASTECH* 11.1 (2012), pp. 41–48 (cit. on p. 56).

[Lin92] Long-Ji Lin. 'Self-improving reactive agents based on reinforcement learning, planning and teaching'. In: *Machine learning* 8.3-4 (1992), pp. 293–321 (cit. on pp. 16, 185).

[Liu+14] Hong Liu, Jianke Liao, Dong Yang, Xiuyuan Du, Pengchao Hu, Yu Yang and Baizhan Li. 'The response of human thermal perception and skin temperature to step-change transient thermal environments'. In: *Building and Environment* 73.2014 (2014), pp. 232–238. DOI: 10.1016/j.buildenv.2013.12.007. URL: http://dx.doi.org/10.1016/j.buildenv.2013.12.007 (cit. on pp. 54, 61, 191).

[LMT14] Nicola Lunardon, Giovanna Menardi and Nicola Torelli. 'ROSE: A Package for Binary Imbalanced Learning'. In: *R Journal* 6.1 (2014) (cit. on p. 84).

[Luo+16] Maohui Luo, Richard de Dear, Wenjie Ji, Cao Bin, Borong Lin, Qin Ouyang and Yingxin Zhu. 'The dynamics of thermal comfort ex-

pectations: The problem, challenge and impication'. In: *Building and Environment* 95.2015 (2016), pp. 322–329. DOI: 10.1016/j.buildenv.2015.07.015. arXiv: 0360-1323. URL: http://dx.doi.org/10.1016/j.buildenv.2015.07.015 (cit. on pp. 37, 39, 55, 62).

[MS96] Richard Maclin and Jude W. Shavlik. 'Creating Advice-Taking Reinforcement Learners'. In: *Machine Learning* (1996), pp. 251–282 (cit. on pp. 21, 24, 26, 186).

[Mac+05] Richard Maclin, Jude W. Shavlik, Lisa Torrey, Trevor Walker and Edward W. Wild. 'Giving Advice about Preferred Actions to Reinforcement Learners Via Knowledge-Based Kernel Regression'. In: *AAAI Press / The MIT Press* (2005) (cit. on pp. 21, 26, 185).

[MR18] Ofir Marom and Benjamin Rosman. 'Belief Reward Shaping in Reinforcement Learning'. In: (2018). URL: https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16912 (cit. on p. 119).

[Mar07] Bhaskara Marthi. 'Automatic shaping and decomposition of reward functions'. In: *Proceedings of the 24th International Conference on Machine learning*. ACM. 2007, pp. 601–608 (cit. on p. 31).

[MSR04] N. A. G. Martinho, M. C. G. Silva and J. A. E. Ramos. 'Evaluation of thermal comfort in a vehicle cabin'. In: *In Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 218.2 (Feb. 2004), pp. 159–166 (cit. on pp. 34–36).

[Mel09] Roderick V.N. Melnik. 'Coupling control and human factors in mathematical models of complex systems'. In: *Engineering Applications of Artificial Intelligence* 22.3 (2009), pp. 351–362. DOI: http://dx.doi.org/10.1016/j.engappai.2008.10.015. URL: http://www.sciencedirect.com/science/article/pii/S0952197608001759 (cit. on p. 24).

[MKS15] Volodymyr Mnih, Koray Kavukcuoglu and David Silver. 'Human-level control through deep reinforcement learning'. In: *Nature* 518.14236 (2015), pp. 529–539. DOI: 10.1038/nature14236. arXiv: 1604.03986. URL: http://dx.doi.org/10.1038/nature14236 (cit. on p. 19).

[MH09] Radu Musat and Elena Helerea. 'Parameters and models of the vehicle thermal comfort'. In: *Acta Universitatis Sapientiae, Electrical and Mechanical Engineering* 1 (2009), pp. 215–226 (cit. on p. 189).

[Mye04] Donald G Myers. 'Heated steering wheel and method of making same'. US Patent 6,707,006. Mar. 2004 (cit. on p. 190).

[Nak+08] M. Nakamura, T. Yoda, L.I. Crawshaw, S. Yasuhara, Y. Saito, M. Kasuga, K. Nagashima and K. Kanosue. 'Regional differences in temperature sensation and thermal comfort in humans'. In: *Journal of Applied Physiology* 105 (2008), pp. 1897–1906 (cit. on p. 60).

[NW14] Jeremy Neubauer and Eric Wood. 'Thru-life impacts of driver aggression, climate, cabin thermal management, and battery thermal management on battery electric vehicle utility'. In: *Journal of Power Sources* 259 (2014), pp. 262–275. DOI: http://dx.doi.org/10.1016/j.jpowsour.2014.02.083. URL: http://www.sciencedirect.com/science/article/pii/S0378775314002766 (cit. on p. 42).

[Ng03] Andrew Y. Ng. 'Shaping and policy search in Reinforcement Learning'. PhD thesis. University of Californa, Berkeley, 2003 (cit. on p. 21).

[Ng+06] Andrew Y. Ng, Adam Coates, Mark Diel, Varun Ganapathi, Jamie Schulte, Ben Tse, Eric Berger and Eric Liang. 'Autonomous Inverted Helicopter Flight via Reinforcement Learning'. In: *Experimental Robotics IX*. Ed. by Marcelo H. Ang and Oussama Khatib. Vol. 21. Springer Tracts in Advanced Robotics. Springer Berlin Heidelberg, 2006, pp. 363–372 (cit. on p. 24).

[NHR99] Andrew Y. Ng, Daishi Harada and Stuart Russell. 'Policy invariance under reward transformations: Theory and application to reward shaping'. In: *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann, 1999, pp. 278–287 (cit. on pp. 21, 25, 29, 111, 120).

[Ng+04] Andrew Y. Ng, H. Jin Kim, Michael I. Jordan and Shankar Sastry. 'Inverted autonomous helicopter flight via reinforcement learning'.

In: *International Symposium on Experimental Robotics*. MIT Press, 2004 (cit. on p. 25).

[NJ02] Andrew Y Ng and Michael I Jordan. 'On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes'. In: *Advances in neural information processing systems*. 2002, pp. 841–848 (cit. on p. 68).

[NHT12] NHTSA. *Visual-Manual NHTSA Driver Distraction Guidelines*. Feb. 2012. URL: https://www.nhtsa.gov/staticfiles/rulemaking/pdf/Distraction_NPFG-02162012.pdf (cit. on p. 49).

[NM03] Monica N. Nicolescu and Maja J. Mataric. 'Natural Methods for Robot Task Learning: Instructive Demonstrations, Generalization and Practice'. In: *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*. 2003, pp. 241–248 (cit. on pp. 23–25, 182, 183).

[Nil07] H. O. Nilsson. 'Thermal comfort evaluation with virtual manikin methods'. In: *Building and Environment* 42.12 (2007). Indoor Air 2005 Conference, pp. 4000–4005. DOI: http://dx.doi.org/10.1016/j.buildenv.2006.04.027. URL: http://www.sciencedirect.com/science/article/pii/S0360132306003702 (cit. on pp. 36, 44, 56, 189).

[NH03] Hakan O Nilsson and I Holmer. 'Comfort climate evaluation with thermal manikin methods and computer simulation models'. In: *Indoor air* 13.1 (2003), pp. 28–37 (cit. on pp. 1, 36, 44, 56, 59, 63, 98, 189).

[NWS08] Minoru Niwa, Danielle L White and Stephanie A Schneider. 'Heated or cooled steering wheel'. US Patent App. 12/262,246. Oct. 2008 (cit. on p. 190).

[NKM87] Kazushi Noda, Moriyuki Komatsu and Hiroshi Mitsunaga. 'Heated or cooled steering wheel'. US Patent 4,640,340. Feb. 1987 (cit. on p. 190).

[NS07] S. Nordbakke and F. Sagberg. 'Sleepy at the wheel: Knowledge, symptoms and behaviour among car drivers'. In: *Transportation Research Part F: Traffic Psychology and Behaviour* 10.1 (2007), pp. 1–10. DOI: http:

//dx.doi.org/10.1016/j.trf.2006.03.003. URL: http://www.sciencedirect.com/science/article/pii/S1369847806000246 (cit. on p. 190).

[NXP15] NXP. *Automotive HVAC Control System with LCD Interface for S12ZVH Family Devices*. Mar. 2015 (cit. on p. 51).

[Ogi+05] M. Ogino, H. Toichi, M. Asada and Y. Yoshikawa. 'Imitation faculty based on a simple visuo-motor mapping towards interaction rule learning with a human partner'. In: *Proceedings of the 4th International Conference on Development and Learning*. 2005, pp. 148–148 (cit. on p. 25).

[Oi+12] Hajime Oi, Koji Tabata, Yasuhito Naka, Akira Takeda and Yutaka Tochihara. 'Effects of heated seats in vehicles on thermal comfort during the initial warm-up period'. In: *Applied Ergonomics* 43.2 (2012). Special Section on Product Comfort, pp. 360–367. DOI: http://dx.doi.org/10.1016/j.apergo.2011.05.013. URL: http://www.sciencedirect.com/science/article/pii/S0003687011000718 (cit. on p. 190).

[OIC15] OICA. *OICA-Position-Paper-Driver-Distraction*. Mar. 2015. URL: http://www.oica.net/wp-content/uploads/OICA-Position-Paper-Driver-Distraction-Final-2015-03-03.pdf (cit. on p. 49).

[OHH07] M. Ollis, W.H. Huang and M. Happold. 'A Bayesian approach to imitation learning for robot navigation'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2007*. Oct. 2007, pp. 709–714 (cit. on pp. 24, 25, 183).

[Par02] K.C. Parsons. 'The effects of gender, acclimation state, the opportunity to adjust clothing and physical disability on requirements for thermal comfort'. In: *Energy & Buildings* 34 (2002), pp. 593–599 (cit. on pp. 37, 40, 189).

[Pef+11] Therese Peffer, Marco Pritoni, Alan Meier, Cecilia Aragon and Daniel Perry. 'How people use thermostats in homes: A review'. In: *Building and Environment* 46.12 (Dec. 2011), pp. 2529–2541. DOI: 10.1016/j.

buildenv.2011.06.002. URL: http://dx.doi.org/10.1016/j.buildenv.2011.06.002 (cit. on p. 40).

[PMK01] Leonid Peshkin, Nicolas Meuleau and Leslie Kaelbling. 'Learning policies with external memory'. In: *CoRR* cs.LG/0103003 (2001) (cit. on p. 17).

[PBR07] C. A. Pickering, K. J. Bumnham and M. J. Richardson. 'A Review of Automotive Human Machine Interface Technologies and Techniques to Reduce Driver Distraction'. In: *System Safety, 2007 2nd Institution of Engineering and Technology International Conference on*. Oct. 2007, pp. 223–228. DOI: 10.1049/cp:20070468 (cit. on pp. 48–50).

[PBZ18] Margaret Pigman, Gail Brager and Hui Zhang. 'Personal control: windows, fans, and occupant satisfaction'. In: *Windsor Conference: Rethinking Comfort*. Apr. 2018 (cit. on pp. 55, 117).

[PLD02] Charles S Pillsbury IV, George E Lancaster and Attila K Dalkilic. 'Steering wheel with self-regulating heating element'. US Patent 6,495,799. Dec. 2002 (cit. on p. 190).

[Pir76] G H Pirie. 'Thoughts on Revealed Preference and Spatial Behaviour'. In: *Environment and Planning A* 8.8 (1976), pp. 947–955. URL: http://EconPapers.repec.org/RePEc:sae:envira:v:8:y:1976:i:8:p:947-955 (cit. on p. 57).

[Pow+11] William T. Powers, Bruce Abbott, Timothy A. Carey, David M. Goldstein, Warren Mansell, Richard S. Marken, Bruce Nevin, Richard Robertson and Martin Taylor. *Perceptual Control Theory A Model for Understanding the Mechanisms and Phenomena of Control*. Aug. 2011. URL: http://www.pctweb.org/PCTUnderstanding.pdf (cit. on p. 45).

[PG98] Sameer M. Prabhu and Devendra P. Garg. 'Fuzzy-logic-based Reinforcement Learning of Admittance Control for Automated Robotic Manufacturing'. In: *Engineering Applications of Artificial Intelligence* 11.1 (1998), pp. 7–23. DOI: http://dx.doi.org/10.1016/S0952-

1976(97)00057-2. URL: http://www.sciencedirect.com/science/article/pii/S0952197697000572 (cit. on p. 24).

[Pyl+15] M. Pylkkanen, M. Sihvola, H.K. Hyvarinen, S. Puttonen, C. Hublin and M. Sallinen. 'Sleepiness, sleep, and use of sleepiness countermeasures in shift-working long-haul truck drivers'. In: *Accident Analysis & Prevention* 80 (2015), pp. 201–210. DOI: http://dx.doi.org/10.1016/j.aap.2015.03.031. URL: http://www.sciencedirect.com/science/article/pii/S000145751500113X (cit. on p. 190).

[RBN95] T H Rammsayer, E Bahner and P Netter. 'Effects of cold on human information processing: application of a reaction time paradigm.' In: *Integrative physiological and behavioral science : the official journal of the Pavlovian Society* 30.1 (1995), pp. 34–45. URL: http://www.ncbi.nlm.nih.gov/pubmed/7794784 (cit. on p. 192).

[Ranoo] Jette Randlov. 'Shaping in Reinforcement Learning by Changing the Physics of the Problem.' In: *ICML*. 2000, pp. 767–774 (cit. on p. 29).

[RA98] Jette Randlov and Preben Alstrom. 'Learning to Drive a Bicycle Using Reinforcement Learning and Shaping.' In: *ICML*. Vol. 98. Citeseer. 1998, pp. 463–471 (cit. on p. 17).

[Ran+oo] Thomas A. Ranney, Elizabeth Mazzae, Riley Garrott and Michael J. Goodman. *NHTSA driver distraction research: Past, present, and future*. Tech. rep. SAE Technical Paper, 2000 (cit. on p. 50).

[RA14] Gian Marco Revel and Marco Arnesano. 'Perception of the thermal environment in sports facilities through subjective approach'. In: *Building and Environment* 77.2014 (2014), pp. 12–19. DOI: 10.1016/j.buildenv.2014.03.017. URL: http://dx.doi.org/10.1016/j.buildenv.2014.03.017 (cit. on pp. 34, 38).

[Rie+13] A. Riener et al. 'Standardization of the In-car Gesture Interaction Space'. In: *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI '13. Eindhoven, Netherlands: ACM, 2013, pp. 14–21. DOI: 10.1145/

2516540.2516544. URL: http://doi.acm.org/10.1145/2516540.2516544 (cit. on p. 50).

[RB10] Stephane Ross and J. Andrew (Drew) Bagnell. 'Efficient Reductions for Imitation Learning'. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. May 2010 (cit. on p. 186).

[RGB11] Stéphane Ross, Geoffrey J. Gordon and J. Andrew Bagnell. 'No-Regret Reductions for Imitation Learning and Structured Prediction'. In: *AISTATS*. 2011 (cit. on p. 186).

[RN94] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*. University of Cambridge, Department of Engineering, 1994 (cit. on pp. 17, 18).

[Ruu18] Villu Ruusmann. *Java PMML API*. 2018. URL: https://github.com/jpmml (cit. on pp. 83, 197).

[Ruz11] Dragan Ruzic. 'Improvement of Thermal Comfort in a Passenger Car By Localized Air Distribution'. In: *Acta Technika Corviniensis - bulletin of Engineering* (2011), pp. 63–67 (cit. on pp. 35, 41, 55–57, 59, 141, 142, 189, 190, 192).

[Sal09] Dario D Salvucci. 'Rapid prototyping and evaluation of in-vehicle interfaces'. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 16.2 (2009), p. 9 (cit. on p. 50).

[Sal01] Dario D. Salvucci. 'Predicting the Effects of In-car Interfaces on Driver Behavior Using a Cognitive Architecture'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. Seattle, Washington, USA: ACM, 2001, pp. 120–127. DOI: 10.1145/365024.365064. URL: http://doi.acm.org/10.1145/365024.365064 (cit. on pp. 49, 190).

[SSR97] J. C. Santamaria, Richard S. Sutton and Ashwin Ram. 'Experiments with reinforcement learning in problems with continuous state and

action spaces'. In: *Adaptive behavior* 6.2 (1997), pp. 163–217 (cit. on p. 17).

[Sch14] Z Schlader. 'The relative overlooking of human behavioral temperature regulation: An issue worth resolving'. In: *Temperature* 1.1 (2014), pp. 20–21. DOI: 10.4161/temp.29235. URL: http://www.tandfonline.com/doi/abs/10.4161/temp.29235 (cit. on pp. 37, 39, 48, 54).

[Sch+13] Zachary J Schlader, Blake G Perry, M Rahimi Che Jusoh, Lynette D Hodges, Stephen R Stannard and Toby Mundel. 'Human temperature regulation when given the opportunity to behave'. In: *Springer* 113 (2013), pp. 1291–1301 (cit. on pp. 36, 39, 48).

[Scho8] M. Schnubel. *Today's Technician: Automotive Heating & Air Conditioning*. Cengage Learning, 2008. URL: https://books.google.com.pe/books?id=KTDqmgEACAAJ (cit. on pp. 7, 42, 51).

[Sch+15] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan and Philipp Moritz. 'Trust region policy optimization'. In: *International Conference on Machine Learning*. 2015, pp. 1889–1897 (cit. on p. 19).

[Sch+17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford and Oleg Klimov. 'Proximal policy optimization algorithms'. In: *CoRR* abs/1707.06347 (2017) (cit. on p. 19).

[SC13] Jinwon Seo and Yunho Choi. 'Estimation of the air quality of a vehicle interior: The effect of the ratio of fresh air to recirculated air from a heating, ventilation and air-conditioning system'. In: *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 227.8 (Aug. 2013). first published in December 2012, pp. 1162–1172 (cit. on p. 42).

[SMP12] Liyanage C. De Silva, Chamin Morikawa and Iskandar M. Petra. 'State of the art of smart homes'. In: *Engineering Applications of Artificial Intelligence* 25.7 (2012). Advanced issues in Artificial Intelligence and Pattern Recognition for Intelligent Surveillance System in Smart Home Environment, pp. 1313–1321. DOI: http://dx.doi.org/10.1016/j.

engappai.2012.05.002. URL: http://www.sciencedirect.com/science/article/pii/S095219761200098X (cit. on p. 34).

[SBS15] O P Singh, Mrinmoy Biswas and Ramji Singh. 'Effect of Dynamic Vent on Thermal Comfort of a Passenger'. In: *Journal of Mechanical Engineering* 61.2015 (2015), pp. 1–12. DOI: 10.5545/sv-jme.2015.2469 (cit. on pp. 55, 57–59, 62, 142).

[SW14] Matthijs Snel and Shimon Whiteson. 'Learning potential functions and their representations for multi-task reinforcement learning'. In: *Autonomous agents and multi-agent systems* 28.4 (2014), pp. 637–681 (cit. on p. 31).

[Son+15] Xiaoji Song, Liu Yang, Wuxing Zheng, Yimei Ren and Yufan Lin. 'Analysis on Human Adaptive Levels in Different Kinds of Indoor Thermal Environment'. In: *Procedia Engineering* 121 (2015). The 9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) joint with the 3rd International Conference on Building Energy and Environment (COBEE), 12-15 July 2015, Tianjin, China, pp. 151–157. DOI: http://dx.doi.org/10.1016/j.proeng.2015.08.1042. URL: http://www.sciencedirect.com/science/article/pii/S1877705815027708 (cit. on p. 39).

[Sta94] International Organization for Standardization. *ISO 7730: Moderate Thermal Environments - Determination of the PMV and PPD Indices and Specification of the Conditions for Thermal Comfort*. 1994. URL: https://books.google.co.uk/books?id=0cclPwAACAAJ (cit. on pp. 34, 36, 47, 48, 189).

[Sua+16] Halit Bener Suay, Tim Brys, Matthew E Taylor and Sonia Chernova. 'Learning from demonstration for shaping through inverse reinforcement learning'. In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2016, pp. 429–437 (cit. on p. 31).

[SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning-An Introduction*. Ed. by Thomas Dietterich. 3rd. MIT Press, 1998 (cit. on pp. 15, 17, 19, 23, 33, 119).

[SVG14] SVG. *Thermal Manikin "Flatman" & Thermal Comfort Data Logger*. SVG, Sept. 2014. URL: http://site.svg-tech.com/clients/svg-tech/Downloads/1221_Thermal_Flatman97201424250PM2.pdf (cit. on p. 80).

[Tay+14] Matthew E. Taylor, Nicholas Carboni, Anestis Fachantidis, Ioannis Vlachavas and Lisa Torrey. 'Reinforcement learning agents providing advice in complex video games'. In: *Connection Science* 14 (Mar. 2014), pp. 1–20 (cit. on pp. 24, 185).

[Tee11] Paul Teetor. *R cookbook*. Beijing: O'Reilly, 2011. URL: http://www.amazon.com/Cookbook-OReilly-Cookbooks-Paul-Teetor/dp/0596809158 (cit. on p. 82).

[TMV10] Ana C. Tenorio-Gonzalez, Eduardo F. Morales and Luis Villasenor-Pineda. 'Dynamic Reward Shaping: Training a Robot by Voice'. In: *Springer* 6433 (2010), pp. 483–492 (cit. on pp. 23, 28).

[TMP10] Ana Cecilia Tenorio-Gonzalez, Eduardo F. Morales and Luis Villasenor Pineda. 'Teaching a Robot to Perform Tasks with Voice Commands'. In: *Lecture Notes in Computer Science, Springer* 6437 (2010), pp. 105–116 (cit. on p. 26).

[THB06a] A.L. Thomaz, Guy Hoffman and Cynthia Breazeal. 'Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots'. In: *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 206. ROMAN 2006.* Sept. 2006, pp. 352–357 (cit. on pp. 21, 22, 27, 28).

[TB06] Andrea L. Thomaz and Cynthia Breazeal. 'Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance'. In: *American Association for Artificial Intelligence* (2006) (cit. on pp. 21, 27, 28).

[TB08] Andrea L. Thomaz and Cynthia Breazeal. 'Teachable robots: Understanding human teaching behavior to build more effective robot learners'. In: *Artificial Intelligence* 172.6-7 (2008), pp. 716–737 (cit. on pp. 22, 23, 27).

[THB06b] Andrea L. Thomaz, Guy Hoffman and Cynthia Breazeal. 'Experiments in Socially Guided Machine Learning: Understanding How Humans Teach'. In: *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*. HRI '06. Salt Lake City, Utah, USA: ACM, 2006, pp. 359–360. DOI: 10.1145/1121241.1121315. URL: http://doi.acm.org/10.1145/1121241.1121315 (cit. on pp. 22–24, 27).

[Tok10] Michel Tokic. 'Adaptive epsilon-greedy Exploration in Reinforcement Learning Based on Value Differences'. In: *Annual Conference on Artificial Intelligence*. Springer. 2010, pp. 203–210. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.464%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf (cit. on p. 142).

[Tor+15] Barbara Torregrosa-Jaime, Filip Bjurling, Jose M Corberan, Fausto Di Sciullo and Jorge Paya. 'Transient thermal model of a vehicle's cabin validated under variable ambient conditions'. In: *Applied Thermal Engineering* 75 (2015), pp. 45–53 (cit. on p. 142).

[Tor+10] Lisa Torrey, Jude W. Shavlik, Trevor Walker and Richard Maclin. 'Transfer Learning via Advice'. In: *Advances in Machine Learning I - Studies in Computational Intelligence* 262 (2010), pp. 147–170 (cit. on pp. 24, 26, 182, 186).

[TC08] Wai Leung Tse and Wai Lok Chan. 'A distributed sensor network for measurement of human thermal comfort feelings'. In: *Sensors and Actuators, A: Physical* 144 (2008), pp. 394–402 (cit. on pp. 39, 40).

[UC91] P. Utgoff and J. Clouse. 'Two kinds of training information for evaluation function learning'. In: *Procedeeings for the Ninth National Conference of Arificial Intelligence*. 1991, pp. 596–600 (cit. on p. 185).

[Van10] Hado Van Hasselt. 'Double Q-learning'. In: *Advances in Neural Information Processing Systems*. 2010, pp. 2613–2621. URL: https://papers.nips.cc/paper/3964-double-q-learning.pdf (cit. on pp. 18, 19).

[Van+16] Hado Van Hasselt, Arthur Guez, David Silver and Google Deepmind. 'Deep Reinforcement Learning with Double Q-learning'. In: *AAAI*. Vol. 2. Phoenix, AZ. 2016, p. 5. URL: http://www0.cs.ucl.ac.uk/staff/D.Silver/web/Applications%7B%5C_%7Dfiles/doubledqn.pdf (cit. on p. 19).

[Van+09] Harm Van Seijen, Hado van Hasselt, Shimon Whiteson and Marco Wiering. 'A Theoretical and Empirical Analysis of Expected Sarsa'. In: *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*. 2009, pp. 177–184. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.4144%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf (cit. on p. 18).

[VP00] Jeffrey B. Vancouver and Dan J. Putka. 'Analyzing Goal-Striving Processes and a Test of the Generalizability of Perceptual Control Theory'. In: *Organizational Behavior and Human Decision Processes* 82.2 (2000), pp. 334–362. DOI: http://dx.doi.org/10.1006/obhd.2000.2901. URL: http://www.sciencedirect.com/science/article/pii/S0749597800929017 (cit. on p. 45).

[Vap98] Vladimir Naumovich Vapnik. *Statistical learning theory, Vol. 1*. 1998 (cit. on p. 68).

[Vol16] Volkswagen. *Air-conditioning and climate control Creating the perfect temperature in your car*. 2016. URL: http://www.volkswagen.co.uk/technology/comfort-and-convenience/air-conditioning-and-climate-control (cit. on pp. 35, 50).

[Wal+06] C Walgama, S Fackrell, M Karimi, A Fartaj and GW Rankin. 'Passenger thermal comfort in vehicles - A Review'. In: *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering* 220.5 (2006), pp. 543–562 (cit. on pp. 35, 37, 55, 61, 189).

[Wanoo] S. Wang. *Handbook of Air Conditioning and Refrigeration*. McGraw-Hill Education, 2000. URL: https://books.google.co.uk/books?id=0tVSAAAAMAAJ (cit. on p. 42).

[Wan+03] Y. Wang, M. Huber, V.N. Papudesi and D.J. Cook. 'User-Guided Reinforcement Learning of Robot Assistive Task for an Intelligent Environment'. In: *IEEE International Conference on Intelligent Robots and Systems* (Oct. 2003), pp. 424–429 (cit. on pp. 20, 33).

[Whi91] S. Whitehead. 'A complexity analysis of cooperative mechanisms in reinforcement learning'. In: *Proceedings of the Ninth National Conference on Artificial Intelligence*. Anheim, 1991, pp. 607–613 (cit. on p. 185).

[Wie03] Eric Wiewiora. 'Potential-based shaping and Q-value initialization are equivalent'. In: *Journal of Artificial Intelligence Research* 19 (2003), pp. 205–208 (cit. on p. 30).

[WCE03] Eric Wiewiora, Garrison W Cottrell and Charles Elkan. 'Principled methods for advising reinforcement learning agents'. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 792–799 (cit. on pp. 29, 121).

[Wil+19] Graham Williams, Tridivesh Jena, Wen Ching Lin and Michael Hahsler. *Generate PMML for Various Models*. 2019. URL: https://cran.r-project.org/web/packages/pmml/pmml.pdf (cit. on p. 83).

[WMC14] L.T. Wong, K.W. Mui and C.T. Cheung. 'Bayesian thermal comfort model'. In: *Building and Environment* 82 (2014), pp. 171–179. DOI: http://dx.doi.org/10.1016/j.buildenv.2014.08.018. URL: http://www.sciencedirect.com/science/article/pii/S0360132314002704 (cit. on p. 46).

[YB14] Zheng Yang and Burcin Becerik-Gerber. 'Modeling personalized occupancy profiles for representing long term patterns by using ambient context'. In: *Building and Environment* 78 (2014), pp. 23–35 (cit. on p. 40).

[YLL09] Runming Yao, Baizan Li and Jing Liu. 'A theoretical adaptive model of thermal comfort- Adaptive Predictive Mean Vote (aPMV)'. In: *Building and Environment* 44 (2009), pp. 2089–2096 (cit. on pp. 37, 39, 47).

[You08] Derek S Young. 'An overview of mixture models'. In: *arXiv preprint arXiv:0808.0383* (2008) (cit. on p. 71).

[ZRM10] Sofia Zaidenberg, Patrick Reignier and Nadine Mandran. 'Learning User Preferences in Ubiquitous Systems: A User Study and a Reinforcement Learning Approach'. In: *Artificial Intelligence Applications and Innovations* 339 (2010), pp. 336–343 (cit. on p. 21).

[Zha+05] H Zhang, C Huizenga, Edward A Arens and T Yu. 'Modeling thermal comfort in stratified environments'. In: *Indoor Air, Beijing* (2005) (cit. on pp. 36, 59, 60).

[Zha+10a] Hui Zhang, Edward Arens, Charlie Huizenga and Taeyoung Han. 'Thermal sensation and comfort models for non-uniform and transient environments, Part III: Whole-body sensation and comfort'. In: *Building and Environment* 45.2 (2010), pp. 399–410 (cit. on pp. 37, 53, 55, 60).

[Zha+10b] Hui Zhang, Edward Arens, Charlie Huizenga and Taeyoung Han. 'Thermal sensation and comfort models for non-uniform and transient environments: Part II: Local comfort of individual body parts'. In: *Building and Environment* 45 (2010), pp. 389–398. DOI: 10.1016/j.buildenv.2009.06.015 (cit. on p. 37).

[Zha+14a] Qianchuan Zhao, Yin Zhao, Fulin Wang, Yi Jiang and Fan Zhang. 'Preliminary study of learning individual thermal complaint behavior using one-class classifier for indoor environment control'. In: *Building and Environment* 72.2014 (2014), pp. 201–211. DOI: 10.1016/j.buildenv.2013.11.009. URL: http://dx.doi.org/10.1016/j.buildenv.2013.11.009 (cit. on pp. 54, 191).

[Zha+14b] Yin Zhao, Hui Zhang, Edward A Arens and Qianchuan Zhao. 'Thermal sensation and comfort models for non-uniform and transient environ-

ments, Part IV: Adaptive neutral setpoints and smoothed whole-body sensation model'. In: *Building and Environment* 72 (2014), pp. 300–308 (cit. on pp. 39, 45).

[Zie+08] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell and Anind K Dey. 'Maximum Entropy Inverse Reinforcement Learning.' In: *AAAI*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438 (cit. on p. 31).

A

IMPLEMENTATIONS OF DEMONSTRATION

Argall et al. [Arg+09] state that when learning from demonstrations, there are two fundamental phases: example gathering and policy derivation. The first phase relies on human-robot interaction and does not necessarily require expert knowledge. The second phase represents a process that relies on strictly expert understanding. Subsequently, RL algorithms are applied in the second phase and rely on internal feedback. This feedback takes the form of numerical reward that shows how desirable a particular state is to visit. The subject (not necessarily an expert) is involved in the act of demonstration helping the agent focus on a subset of the state space. The reward comes as a penalty for states that have not been visited during demonstration and prevent the agent from preferring a set of states and actions that trigger a high reward.

Due to demonstrations being noisy and sub-optimal, exploration is encouraged to supplement for states that have not been previously encountered. The user can define the reward function manually, leading to sparse rewards (there are a few states in which the rewards are different from zero). In this case, demonstration proves to be an advantage as it signals the areas of interest in the state-space, preventing the agent from conducting extensive exploration and facilitating the discovery of rewards. On the other hand, the engineered reward can trigger penalties for actions executed by the agent that are not found in demonstration as the teacher cannot provide actions for all possible states. The agent can choose the state-action pairs that achieve a local maximum and not explore alternative pairs.

Leon et al. [Leo+11; LTM13] conducted a series of tests using a Kinect sensor in order to get the motion traces from a user that changes the place of an object. The agent, in this case the robot, needs to reproduce the demonstrated actions. In the event that the robot succeeds in reproducing the action, the position and orientation of the object are occasionally different. This is because the starting position of the robot was different to that of the demonstrator.

Leon et al. [Leo+11; LTM13] also identified key disadvantages such as the demonstration can be noisy and sub-optimal, the user can be inexperienced, the necessary hardware too complex and the experimental conditions need to be strictly controlled. Berlin et al. [Ber+06] developed several experiments in action reproduction, having the user provide demonstrations that are incomplete in order to see if the agent was capable of following the human demonstration or executing the given task. The agent was trained to adopt the belief system of the human diverged from the set goal and reproduced incomplete demonstrations.

To overcome the sparsity of the reward, exploration can be encouraged for the Reinforcement Learning (RL) algorithm. The existing data can be used for the states that have not been covered by demonstration. Torrey et al. [Tor+10] used demonstrations of previous robotic actions and tactics translated into a programming language in order to construct the knowledge base with which the agent is trained, generalising the state and action pairs. Nicolescu et al. [NM03] improved task learning by using the refinement of existing behaviours. The network of abstract behaviour was based on the information collected by the robot's sensors during demonstration. The assumption is that whilst the teachers are not experts in knowing how the robot learns to execute the task, they know about the position of the sensors as well as the skills that the robot possesses. Using generalisation in this case helps identify the steps that are useful for execution (most frequently executed steps). On the other hand, omissions can be generated due to sensor limitations or the learner can be biased due to irrelevant steps in task demonstration.

The second option is to use inverse reinforcement in order to obtain a learned reward. Abbeel et al. [AN] aimed to find a policy that performs similarly to the expectations of an expert that triggers the reward. The reward function is derived

from the observed behaviour of the expert and is correctly learned with the proper estimation of the count of features. The only downside represents that the function is assumed to be linear with respect to the known features.

Another alternative represents having rewards aimed at how similar to the demonstration is the executed behaviour. This can either trigger adaptation to a different task by assigning high values when the agent is close to a goal or by a model of Bayesian learning that uses feature vectors. In the case of Ollis et al. [OHH07] the feature vectors were obtained during joystick training in order to transverse different types of terrain. Additional feedback to the initial demonstrations increases the learning performance. This can be done by updating the policy in real time using rewards [Leo+11; LTM13] during the actual execution of the task, or feedback is provided as direct corrections or advice [NM03] in the case of continuous state spaces.

B

IMPLEMENTATIONS OF ADVICE

Taylor et al. [Tay+14] used multiple methods such as: early advising, alternating advice, importance advising, mistake correcting and predictive advising on two Reinforcement Learning (RL) agents in a teacher-student framework, one providing the advice and the other learning from it. Compared to a standard RL algorithm with no advice, all the proposed methods had a higher performance. Among these, mistake correcting had a larger impact on the performance of the learning agent. This is because advice was given only in significant states and in small amounts. Advice is a valuable means to prioritise actions by giving hints towards the preferred or expected behaviour.

Taylor et al. [Tay+14] stated that advice has different effects on the agent depending on when it is provided in the learning process. The user cannot provide continuous information while the learning takes place (the advice is limited) and the agent cannot execute all the suggested actions immediately. When teachers notice the particular mistakes that the students make, the advice can be provided in an effective manner regardless of the algorithms used for learning.

The introduction of advice to an RL agent can take the form either of a series of action commands ([Lin92], [UC91]) directly affecting the policy or critique [Whi91] of the agent's actions affecting the reward function.

In the work of Maclin et al. [Mac+05], the advice directly affects the policy (Figure 2.6). For a specific environmental state the users give advice about which action is preferred. The advice is represented by if-then rules encoded by the advisor— an experienced programmer. Function approximation is done using a

type of knowledge-based kernel regression that instead of focusing on the Q values, indicates action preference for the condition part of the rule (Preference Knowledge-Based Kernel Regression (Pref-KBKR)).

Previously Maclin et al. [MS96] made use of a simple programming language to allow the user to give advice in order to maximize the agent's reward. The advice is directly inserted into the utility function of the agent using techniques based on artificial neural networks. This allows the system to accept certain advice at any time during the learning process. The agent avoided the cases in which it was deliberately given erroneous advice by making associations. When further advice was provided, the learner was able explore and refine the actions.

Torrey et al. [Tor+10] used existing knowledge bases from robot soccer matches as advice. The authors used inductive logic programming to transfer skills learnt by the agent, further allowing the user to give advice in order to maximise the agent's reward. The advice was directly inserted into the utility function of the agent using the Knowledge-Based Kernel Regression (KBKR) algorithm. The role of the human was to map between the tasks used as examples and the current task, further identifying the differences between them. It is not necessary to provide constant advice as long as it is provided for certain states that are deemed by the user to be significant. What is more when knowing the root of the mistakes performed by the agent, the user can improve the quality of the advice in order to address the respective problem. Daswani et al. [DSH14] adapted an algorithm called RLAdvice which uses information about the expected return of state-action pairs in the form of advice from an oracle.

The value of the advice is used by the RL agent to learn a well-performing policy for the respective environment (Arcade Learning). RLAdvice is an adaptation of the Dataset Aggregation algorithm (DAgger) [RGB11], characteristic of a supervised learning technique called imitation learning [RB10] with the difference that for each action, a set of weights is generated. These weights contribute to the function approximation value that serves as guidance for the agent. In this case, the policy is directly affected by the advice provided by the oracle.

Leon et al. [LTM13] explored the possibility of additional feedback via brief commands (sentences that can produce a change in action) and critiques (approval or disapproval of a specific action). The verbal cues (fixed words and sentences) have a corresponding associated reward.

C

ALTERNATIVE ADAPTIVE BEHAVIOURS AND FACTORS

C.1 CLOTHING CHANGES

Clothing insulation is one of the personal parameters that can be included in thermal comfort models [Dam+16; Ruz11; Sta94; Wal+06]. It varies between 0.35-0.5 Clo for a sedentary person in an enclosed environment [Sta94; Ruz11]. The clothing insulation is assumed to be constant through the trial duration and cover uniformly the entire body [MH09]. Conversely, the amount of clothing an occupant wears depends on the external environment [Ruz11]. Additionally, the activity level that a person achieves before entering the car can impact the level of discomfort. This personal parameter is included in models such as Predicted Mean Vote (PMV).

In the building environment, an example of adaptive behaviour is removing or adding clothes when the occupants feel uncomfortable [Par02; LWG15; LWG16]. For vehicles, little research has been conducted on clothing changes [Hod13]. The effect of clothing on occupant thermal comfort is determined by using manikins [NH03; Nil07] wearing a fixed amount of clothing and without exhibiting any simulated actions.

There are two aspects that can influence clothing changes: i) either the occupant feels uncomfortable and decides to remove or add more layers of clothing, or ii) the clothes impede driving performance. Chang [Cha+11] developed clothes that have incorporated an air conditioning system for car and motorcycle occupants.

C.2 USE OF WINDOWS AND ADDITIONAL HEATED SURFACES

Similar to clothing, windows have a high rate of protection against UV light. What is more, tinted windows improve comfort and prevent the negative effect of glazing on driving performance [AB14].

Current literature identifies two different motivations for the use of windows: i) to combat sleepiness, ii) smoking. The first motivation comes when drivers are feeling sleepy after driving for a long period of time. Opening the window is reported to be the first counter measure in 20-35% of the cases as cold air is directly blown towards the face and subsequently helps the driver to be alert [Asa+12; HR99; NS07; Pyl+15]. The second motivation is related to smoking in the car cabin. Occupants use the windows to prevent exposure to second hand smoke despite it being an ineffective measure [Hit+12; Jon+09]. There is little to no literature linking window opening in the car cabin with thermal comfort as a primary motivation, hence this action will not be considered in this thesis.

Contact with heated or cool surfaces produces local and overall discomfort [Ruz11]. Extensive research has been conducted in the design of seats (material, size, position) in order to improve the comfort of the occupants [BP99; CB07; KC98; KST04; Sal01]. As a further improvement, heated seats are reported to increase both thermal comfort and sensation (especially at the foot region) [Oi+12]. Heated or cooled seats are controlled either by activating them on and off through the Heating, Ventilation and Air Conditioning (HVAC) control panel [Oi+12], or a separate control placed at the back of the seat and cushion [BP99]. Despite this fact, there is no record on preferences for the heating or cooling or how and when people choose to activate them.

Due to the fact that drivers perceive discomfort at the hand and arm region [Che+15; Ruz11], several patents propose heated, and cooled steering wheels [Gar08; Mye04; NWS08; NKM87; PLD02]. As of yet, there is not much investigation of people's comfort evaluations and preferences when using these types of steering wheels.

C.3 ADDITIONAL FACTORS IMPACTING THERMAL BEHAVIOUR

While humidity is included in the PMV model, it is not for other comfort models (e.g Nilsson, Zhang). Nonetheless, Zhao et al. [Zha+14a] state that high levels of humidity impact occupants' perceptions of hot and cold environments, decreasing sweat evaporation.

One of the most important factors that influence passenger behaviour is skin sensitivity. Whilst the connection between thermal sensation and mean skin temperature has been explored through the rules, according to Liu et al. [Liu+14], skin is more sensitive to cold stimuli, the changes being noted for subjects entering from hot to neutral environments. Liu et al. [Liu+14] noted that the most sensitive human body parts are the head, chest, back and calf. Therefore body part equivalent temperature can be considered.

Sensitivity in the building environment is linked to gender, with women being more sensitive to temperature changes than males. Karjalainen [KK07] outlined that women exhibit a preference for higher room temperatures, feel more often uncomfortable and are less satisfied with the temperatures of their surroundings. According to Kim et al. [Kim+13] women are more sensitive to the conditions provided by the HVAC system and are prone to complain in relation to their thermal comfort. A total of 30.5% of women are dissatisfied with room temperatures compared to 21.1% of the males. Despite this fact, Karjalainen [KK07] stated that the use of thermostats in housing can be linked to males. Nonetheless, according to Frontczak et al. [FW11] gender, job satisfaction, and interpersonal relationships can influence comfort but it is not conclusive as to how.

What is more, a topic that is well-known in the vehicle industry to cause discomfort and further influence occupant's actions is HVAC noise (Jaeger et al. [Jag+08], Eilemann [Eil99] , and Hohls et al. [Hoh+14]). Hohls et al. found that roughness, sharpness and the articulation index have an impact on how occupants perceive blower noise, with loudness specifically affecting their use of the system. Noise discomfort is associated with the blowers as its source [Eil99], hence passengers reduce the levels of the blower speed to minimise noise. Jaeger et al. [Jag+08]

found that not only blower noise but air ducts and vents can also produce noise discomfort, which can cause the occupants not to use the respective functions.

One aspect that cannot be ignored when highlighting motivation behind interactions with an HVAC system is the individuality of the occupant. Ruzic [Ruz11] stated that comfort is impacted by variations in clothing, gender, age, mood and other individual differences, whereas Karjalainen et al. [Kar07] maintained the idea that HVAC systems and culture impact personal control. Karjalainen [KK07] further brings into context the possibility that not all individuals can be satisfied with their thermal environment. Rammsayer et al. [RBN95] maintained that this can become an advantage as the differences between individuals can impact small samples of data bringing new discoveries to light.

D

ALTERNATIVE MODEL FOR RULE 2

R2 can be converted from a subset of binary classification problems to a multi-class problem. The recorded outputs for each selection can be combined into a single unique class (table D.1). The same set of multi-class models (section 4.2.3) were used for training.

The model with the highest accuracy and kappa is the neural network model displaying a good estimation of the combination of settings compared to the other models (table D.2).

An alternative method for testing the performance of the models is to reduce them to binary classes. Among the available strategies of binary reduction, One-vs-Rest and One-vs-One [Gal+11]. The One-vs-Rest technique is based on assigning a classifier for each class (e.g. a classifier for temperature selection, blower selection, and vent selection). When the samples of the class are positive (e.g. temperature has been selected), the rest of the samples are considered negative (including the other settings that are available). Therefore each class is compared against the others. The

Class	Temperature	Blower	Vent
Class1	0	0	1
Class2	0	1	0
Class3	0	1	1
Class4	1	0	0
Class5	1	0	1
Class6	1	1	0
Class7	1	1	1

Table D.1: Class labels determined by the combination of selected settings.

Model	Accuracy	Kappa
nnet	0.7	0.51
PART	0.55	0.27
gbm	0.55	0.29
rf	0.5	0.18
svmRadial	0.5	0.19
rpart	0.45	0
svmLinear2	0.3	-0.02
svmLinear3	0.45	0.14
ctree	0.45	0
knn	0.4	0.06

Table D.2: The highest performing model in terms of accuracy and Cohen's kappa, for the classification of setting selections is the neural network model.

problem with this classification technique is that there can be ambiguous regions (different scales for the confidence scores factoring in the decision), with class imbalance emerging in the training set as the negative samples have higher rates than the positive ones.

The One-vs-One technique can be suitable for this rule as it involves training a set of binary classifiers for each case of setting selection (in this case 7 sets in which at least one setting is selected), the classifier estimating the probability that setting combination is selected or not. After training each classifier, at the prediction stage all the classifiers are given a testing sample. The classifier with the highest amount of positive predictions is chosen as output. Using this procedure the same neural network model has the highest mean area under the curve of 0.68 compared to the alternative models.

By combining the three selection combinations into a single output the risk of estimating no selection for all the settings (when R1 is activated) is eliminated. Moreover the complexity of the hybrid model (figure D.1) is reduced, enabling direct estimation for the value selection when all settings are selected.

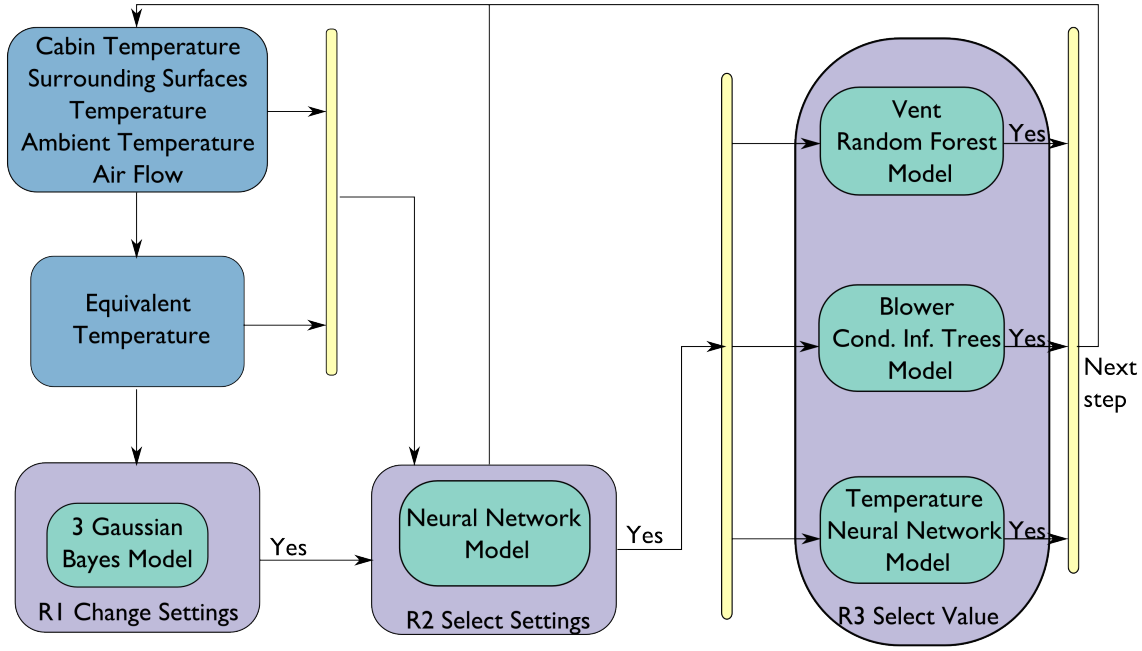

Figure D.1: The architecture of the hybrid model using a single classifier for estimating setting selections (R2).

E

MODEL CODE LISTINGS

The User-Based Module (UBM) code (figure E.1) was developed in Java using the jpmml library [Ruu18] for activating the six Predictive Model Markup Language (PMML) classifiers and the hand-coded Bayesian classifier. The classifiers have been trained, validated and tested in R-Studio using the caret package [KJ13] and the Gaussian distributions for the Bayesian model were fitted using the mixtools package [Ben+09]. The code has approximately 700 lines, including the calls for the classifiers and testing the model, for further information on obtaining the model please contact the author.

The code for the UBM is integrated into the cabin environment file titled "SimpleCabinEnvironment" which is connected to the simulation through simulation platform "SarsaSimulation" (figure E.2). The state of the environment is passed to the Reinforcement Learning (RL) agent that is trained with the various State-Action-Reward-State-Action (SARSA) algorithms (Expected, Double, and Double Expected). The agent selects the actions of the Heating, Ventilation and Air Conditioning (HVAC) controller based on the fitness function (reward function) and the algorithm that is running. The actions are passed back to the environment as the HVAC outputs of temperature, air velocity and recirculation (figure E.3).

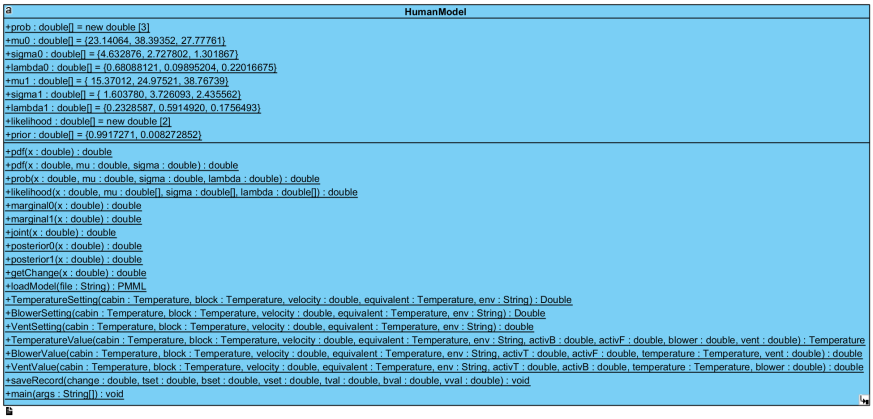

Figure E.1: UBM diagram overview of the main functions and parameters.

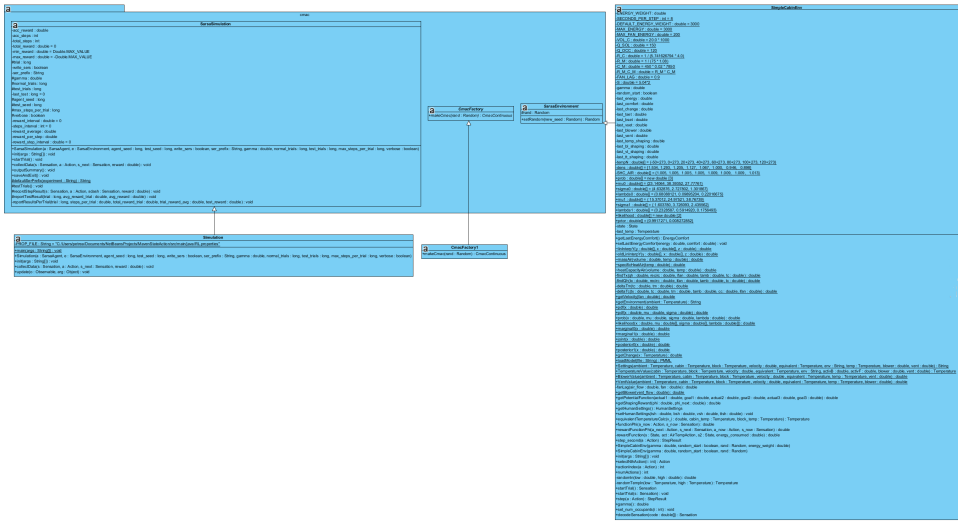

Figure E.2: Overview diagram of the architecture of the car cabin environment that includes the UBM as a simulated occupant, and the lumped capacitance model, and is connected to the additional simulation files.

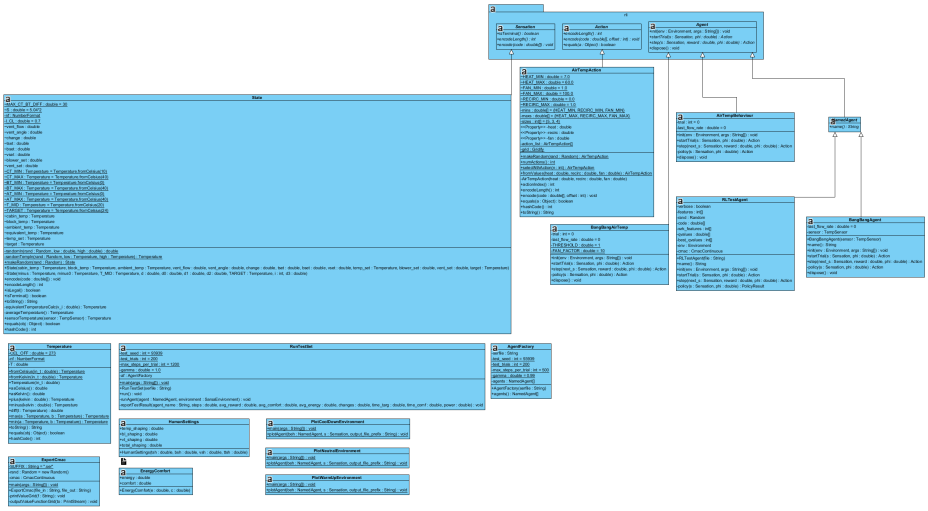

Figure E.3: Overview diagram of the RL agent and its connections to the state of the environment and the actions of the HVAC controller.