

FaiR-IoT: Fairness-aware Human-in-the-Loop Reinforcement Learning for Harnessing Human Variability in Personalized IoT

Salma Elmalaki
salma.elmalaki@uci.edu
University of California, Irvine

Abstract

Thanks to the rapid growth in wearable technologies, monitoring complex human context becomes feasible, paving the way to develop human-in-the-loop IoT systems that naturally evolve to adapt to the human and environment state autonomously. Nevertheless, a central challenge in designing such personalized IoT applications arises from human variability. Such variability stems from the fact that different humans exhibit different behaviors when interacting with IoT applications (intra-human variability), the same human may change the behavior over time when interacting with the same IoT application (inter-human variability), and human behavior may be affected by the behaviors of other people in the same environment (multi-human variability). To that end, we propose **FaiR-IoT**, a general reinforcement learning-based framework for adaptive and fairness-aware human-in-the-loop IoT applications. In FaiR-IoT, three levels of reinforcement learning agents interact to continuously learn human preferences and maximize the system's performance and fairness while taking into account the intra-, inter-, and multi-human variability. We validate the proposed framework on two applications, namely (i) Human-in-the-Loop Automotive Advanced Driver Assistance Systems and (ii) Human-in-the-Loop Smart House. Results obtained on these two applications validate the generality of FaiR-IoT and its ability to provide a personalized experience while enhancing the system's performance by 40%–60% compared to non-personalized systems and enhancing the fairness of the multi-human systems by 1.5 orders of magnitude.

CCS Concepts

• **Computing methodologies** → **Multi-agent reinforcement learning**; • **Human-centered computing** → **Ubiquitous computing**; *Collaborative and social computing*.

Keywords

Fairness, Personalized IoT, Reinforcement Learning, Human Adaptation, Human-in-the-Loop

ACM Reference Format:

Salma Elmalaki. 2021. FaiR-IoT: Fairness-aware Human-in-the-Loop Reinforcement Learning for Harnessing Human Variability in Personalized IoT. In *International Conference on Internet-of-Things Design and Implementation (IoTDI '21)*, May 18–21, 2021, Charlottesville, VA, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3450268.3453525>



This work is licensed under a Creative Commons Attribution International 4.0 License. *IoTDI '21, May 18–21, 2021, Charlottesville, VA, USA*
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8354-7/21/05.
<https://doi.org/10.1145/3450268.3453525>

1 Introduction

Ubiquitous computing—that interacts and adapts to humans—is inevitable. In these pervasive systems, human reactions and behavior are observed and coupled into the loop of computation [15]. By allowing autonomy into the essence of IoT systems, these evolving systems provide services that are adaptable to the human context and intervene and take actions that are tailored to the human reaction and behavior. Despite the IoT system's ability to collect and analyze a significant amount of sensory data, traditional IoT typically depends on fixed policies and schedules to enhance user experience. However, fixed policies that do not account for variations in human mood, reactions, and expectations, fail to achieve the promised user experience. Moreover, with the continuous and inevitable interaction of the human with these systems, it becomes a pressing need to adapt to the physical environment changes and adapt to human preferences and behavior. This opens the question of how to use the monitored human state to design human-in-the-loop IoT systems that provide a personalized experience.

A personalized IoT system needs to “infer” or “learn” the human preference and continuously “adapts” and “takes actions” whether autonomously or in the form of recommendation. Hence, we need a **human-in-the-loop framework** that moves from a one-size-fits-all approach to a personalized process in which learning and adaptation agents in IoT systems are tailored towards humans' individual needs. This feedback property opens the door to design a “Reinforcement Learning (RL)” based agent. Unfortunately, applying standard RL algorithms, such as Q-learning, faces several challenges in the context of human-in-the-loop IoT applications. In particular, adapting to the human reaction and behavior poses a new set of challenges due to the **intra-human** and **inter-human** variability—for example, the same human behavior and reaction change over time. Even for a small period of time, the same human may produce different reactions based on unmodeled external effects. Similarly, different humans have different reactions under similar conditions. In addition to the variability introduced by humans, the IoT system used to infer human states suffers itself from variability (e.g., power constraints, connectivity status, classification/signal processing errors) that introduce another level of complexity. Moreover, as IoT applications are becoming more ubiquitous, multiple humans may interact within the same application space affecting each other's reaction and the way the application adapts. This **multi-human** interaction poses a new set of challenges to the IoT application relating to the “fairness” of adaptation.

In this paper, we propose a framework that can be used in conjunction with IoT applications to provide a personalized experience while addressing the aforementioned challenges. We propose **Fairness-aware Human-in-the-Loop Reinforcement Learning framework**, or **FaiR-IoT** for short, that monitors the change

in the human reaction and behavior while interacting with the IoT system and addresses the **intra-human** and **inter-human** variability, as well as, the **multi-human** interactions, to provide personalized adaptation that enhances the human's experience while ensuring fairness across all human sharing the same IoT application. The proposed framework is then used to build two human-in-the-loop IoT applications, (1) personalized advanced driver assistance system and (2) personalized smart home.

1.1 Related Work

Human-in-the-Loop IoT: Besides the fact that IoT solutions nowadays have billions of connected devices [32], humans themselves are becoming a walking sensor network equipped with wearables that are rich in sensors and network capabilities [12, 39]. Human sensing has a central role in many IoT applications. In the area of managing the energy consumption of buildings, detecting the number of occupants through human sensing has been a primary target of many energy-saving techniques, such as correlating electrical load usage with occupancy sensors [3, 43]. In automotive applications, the human driving behavior has been integrated into the loop of computation of many Advanced Driver Assistance Systems (ADAS) [11, 36], such as activate the automatic cruise control [33] or adjust the threshold of the forward collision warning [14]. In smart cities, human tracking is used for dynamic resource allocation of community services [50]. Moreover, as IoT applications are becoming more human-centric applications, efficient human modeling, human state estimation, and human adaptation are vital components for IoT applications that interact with humans [28, 53]. In this paper, we propose a general framework that can be integrated into many of these applications to provide a personalized experience while addressing the human variability that arises while interacting with these applications.

Reinforcement learning for human adaptation: This tight coupling between human behavior and computing promises a radical change in human life [41]. In the area of cognitive learning and human-in-the-loop IoT applications, reinforcement learning (RL) has proven to be adequate to monitor human intentions and responses [20, 44, 45]. Multisample RL [14] can adapt to humans and changes in their response times under various autonomous actions. Amazon has used personalized reinforcement learning to adapt to students' preferences [5]. Moreover, RL models have been used to decide residential load scheduling [55]. In this paper, we build upon work in RL literature to monitor changes in human behavior to achieve personalized IoT applications while adapting the RL models at runtime to address the different variability that arises in human-in-the-loop IoT systems.

Hierarchical reinforcement learning: Multi-layer RL has been used in the domain of parameter search [25]. In particular, RL has been used to learn policies to converge to good hyper-parameters that achieve the best performance [54]. In addition to parameter tuning, hierarchical RL was used to train multiple levels of policies or to decompose complex learning tasks into sub-goals [35]. However, these methods learn the hyper-parameters once and fit them to the model or learns the different layers of policies and fix them. In the domain of human-in-the-loop IoT systems, this is not applicable. Humans change their behavior, and there exists intra-human and inter-human variability that needs to be addressed.

Fairness in reinforcement learning: The question of fairness in RL where the agent prefers one action over another becomes more significant in multi-agent systems [22, 26]. As each agent learns its optimal policy, the notion of efficiency (system performance) and fairness may be conflicting. One approach is to train each agent independently but design the reward to be fair-efficient [24]. Another approach is the notion of envy and guilt in the study of inequity aversion to shape the fair reward function [21]. However, in the multi-human application space, a policy that is considered fair at one time may become a discriminatory policy after some time as human preferences change. In this paper, we build upon the definition of fairness proposed in literature while proposing a fairness measure that can be embedded in human adaptation models.

1.2 Paper Contribution

This paper aims to put the intrinsic variation in human behavior and reaction into the loop of computation while addressing the fairness in multi-human IoT applications. In particular, we make the following contributions:

- (1) Designing **FaiR-IoT**, an adaptation framework named "Fairness-aware Human-in-the-Loop Reinforcement Learning" that addresses the **intra-human** and **inter-human** variability in the context of IoT applications. The framework continuously monitors the human state through the IoT sensors and the changes in the environment with which the human interacts to adapt to the environment accordingly and enhance the human's experience.
- (2) Extending the framework to address the **multi-human** interaction within the same application space. The framework continuously monitors how each human behavior can affect other humans' behaviors, which changes the environment and the way it is adapting.
- (3) Proposing fairness-aware methodology for adapting to multiple humans sharing the same IoT application space.
- (4) We use the proposed framework to develop two Human-in-the-Loop IoT applications in the context of advanced driver assistance systems (ADAS) and automated smart home.

2 Fairness-aware Human-in-the-Loop Reinforcement Learning Framework

We consider one of the IoT applications in the automated smart home as a motivating example for designing the proposed framework. In particular, we focus on the application of designing smart thermostats. Current state-of-art smart thermostats can adapt the home temperature based on room occupation [3] using fixed schedules and policies [31]. In particular, preset configurations are provided by homeowners, and smart thermostat abides by these configurations. However, human needs and behavior vary across time, such as a change in the body's temperature is affected by multiple aspects, such as excitement, anxiety, physical activity, and health issues. In principle, IoT systems can play a role in detecting the human mood [1] and activity [18] to "adapt" the home temperature accordingly. This adaptation can further be automated and "tuned" by monitoring the human thermal satisfaction, which can be inferred as direct feedback from the human [42] or as an indirect measurement from another edge device, such as, black globe thermometer [6] or skin temperature monitor [48]. The end goal is to achieve a personalized experience by learning the best home

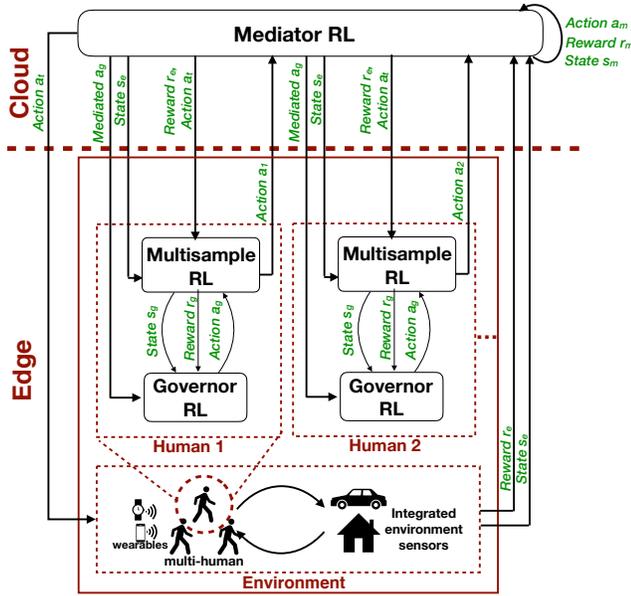


Figure 1: FaiR-IoT framework is divided into three RL agents. Multisample RL and Governor RL interact through action a_g , state s_g , and reward r_g to handle the intra-human and inter-human variability at the edge. While a third RL agent, the Mediator RL, interacts with the environment through action a_t , state s_e , and reward r_e in the cloud while providing a feedback action to both the Multisample RL and the Governor RL to handle the multi-human variability.

temperature that provides the best thermal sensation for the current human state. This “learn”, “adapt”, and “tune” feedback loop fits well with the RL paradigm. The environment in the RL setting encapsulates the interaction between the human and the IoT applications. The RL agent makes a recommendation to the human or takes action (such as changing the set-point of the thermostat) to adapt the IoT application. This action or recommendation is continuously being adapted to match the changes in human behavior (such as mood and physical activity) and reaction (such as a change in the thermal sensation). This model keeps training online or offline until it converges to the best policy, which is the best thermostat set-point for a particular human mood and physical activity. However, we argue that this setting will not achieve the overarching goal of personalized experience in IoT applications. Using the same motivating example, we list the cases in which the setting mentioned above will fall short.

- (1) **Intra-human variability:** Even under the same mood and physical activity, the same human may change the personal preference for the best room temperature. Hence, an RL model that learns a fixed policy will not be adequate.
- (2) **Inter-human variability:** Different humans may have different thermal sensations even with the same home temperature under the same mood and physical activity. Moreover, the human body response to a change in ambient temperature can differ across humans [10]. Hence, the same RL model design (same parameters and same reward function) may adapt correctly for some humans while performing poorly for others.
- (3) **Multi-human variability:** More than one human can be in the same house, each with a different mood (relaxed or stressed) and

different physical activity (sleeping or doing domestic work). One set-point that achieves the most comfortable thermal sensation for one human may not be the best one for another. Hence, an RL model that does not take the effect of the interaction between multiple humans within the same IoT application space will not satisfy the individual needs and will not account for the fairness of the adaptation policy across the humans.

Accordingly, we propose “**Human-in-the-Loop Fairness-aware Reinforcement Learning**” framework to address this variability. Our framework is divided into three components name **Multisample RL** to handle the intra-human variability, **Governor RL** to handle the inter-human variability, and **Mediator RL** to handle the multi-human variability while ensuring fairness. A conceptual figure for the proposed framework is shown in Figure 1.

2.1 Intra-Human Variability

In the domain of human-in-the-loop IoT application, the changes in the human response time and the preferences across time makes the reward for the RL agent a time-varying function. The problem of time-varying reward function was addressed by **Multisample Q-learning algorithm** [14] that divides the time horizon into three overlapping different time scales, named (1) state observing rate (T_s) where the state of the environment is observed every time new sensor data is available. This depends on the sampling rate of the sensors at the edge devices, (2) actuation rate (T_a), which is the time by which the agent should take action, and (3) evaluation rate (T_l) where the reward—corresponding to a particular taken action—should be evaluated at a relatively slow rate to take into account the delay in the environment change.

Multisample RL can handle the **intra-human variability**, where humans can change their behavior and reaction, and the continuous reward calculation tracks this change and adapts the RL action accordingly. However, it assumes that the response time of all humans is the same. By fixing the values of T_l and T_a , Multisample RL assumes that all humans have the same response time. However, T_l and T_a should be adaptive based on the human interaction with the IoT system. By adapting T_l and T_a , we can take the **inter-human variability** as well into the loop of computation. This leads to designing **Governor RL** as will be explained in the next section.

2.2 Inter-Human Variability

To account for the inter-human variability, the RL-agent needs to adapt T_l and T_a assumed by the Multisample RL algorithm to personalize the human experience. Such adaptation can be modeled as another RL-agent that observes the change in the human’s experience as the parameters T_l and T_a change. To that end, and as shown in Figure 2, we add a **Governor RL** layer that personalizes the values of T_l and T_a for Multisample RL. The Multisample RL underneath runs the Q-learning algorithm with the three overlapped time scales (T_l , T_a , and T_s) and adapts the IoT system that interacts with the human through an action or a recommendation to the human. A personalized Q-learning policy is learned based on the reward propagated from the environment. This reward is a function of the human experience and the IoT system performance. After the policy is converged (the reward value is not increasing), the Governor RL observes the cumulative reward for this particular action a_g that modifies the values of T_l and T_a and adapts them accordingly. Eventually, Governor RL will converge to the values of

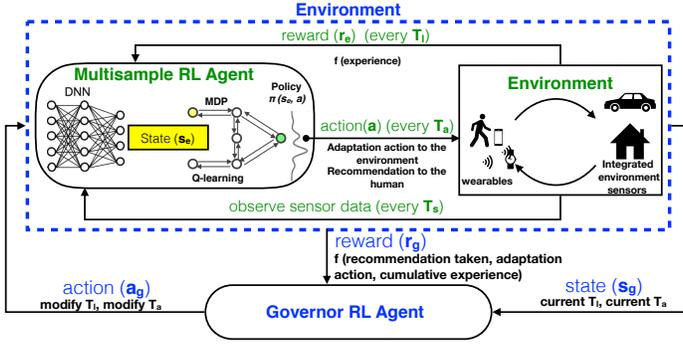


Figure 2: Governor RL interacts with Multisample RL to adapt to the inter-human variability. Governor RL adapts the values of T_l and T_a

T_l and T_a that achieve the best cumulative reward for a particular human. The algorithm and the design of the Governor RL will be discussed in Section 3.1.

2.3 Multi-Human Variability

When multiple humans interact in the same IoT application space, their preferences differ based on the number of people they interact with and the type of this interaction. Moreover, when the application adapts to one human, the environment changes accordingly to the adaptation action, and thus it has a direct effect on the other humans in the same environment. In addition to that, multiple humans may require different adaptation actions based on their different preferences. Hence, we need a mediator to intervene and manage these multiple adaptation actions from multiple humans to achieve an aggregate better state for all the humans collectively. Accordingly, we propose the third layer in our framework, the **Mediator RL**. As shown in Figure 3, the Mediator RL takes the adaptation actions (a_1, a_2, \dots, a_n) from all the Multisample RL agents.

As mentioned in Section 2.1, T_l is the evaluation rate of the RL agent, while T_a is the actuation rate that governs the time to apply the action on the environment. Since T_l captures the environment's response time, including the human response time, it should not change even while mediating multiple humans. However, the actuation rate T_a should be mediated because it controls how often we apply an action to the environment. Hence, the Mediator RL has to mediate two values, (1) a_g : which is the action taken by each Governor RL to adapt T_a , and (2) a_i : which is the action taken by each Multisample RL for each human.

Accordingly, the Mediator RL is responsible for mediating a_g by choosing the smallest value of T_a across all a_g . This will ensure the fastest actuation rate across all humans. However, by changing the actuation rate T_a , Governor RL has to be notified accordingly because now the reward that the Governor RL observes is accounted for a different value of T_a . Hence, to ensure that the reward is associated with the right action, the Mediator RL echos back the mediated a_g , which has the same T_l but a different T_a .

Next, the Mediator RL should choose the right action to apply to the environment. To that end, and not to compromise the effectiveness of the action on the human experience, we can not just choose one action from the list of actions. Hence, we use a weighted average of all the actions. In principle, the weighted average takes into account the varying degrees of importance of the

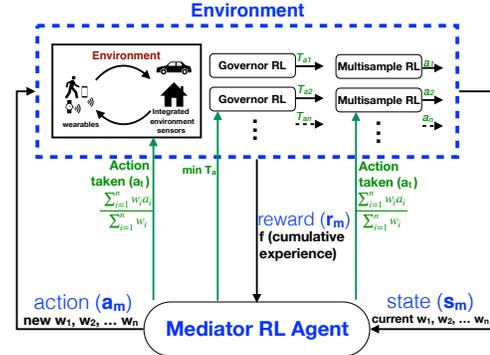


Figure 3: Mediator RL adapts to multi-human variability by adapting the weights of the adaptation actions.

numbers. However, since we do not know apriori, which action is more important than others to achieve a better cumulative experience for all humans, we can not fix these weights. Moreover, as humans' preferences, moods, and reactions are changing, one action may have more weight at a particular time while having less weight at another time. Hence, we need to learn these weights and continuously adapt them based on all humans interacting in the same environment. Accordingly, the Mediator RL agent adjusts the actions' weights and then applies the weighted average action a_t to the environment. The Mediator RL agent then observes the effect of these weights by collecting the reward r_m , a cumulative reward from all the humans' experiences in the environment. However, since the actual action taken on the environment is different from the individual actions a_i by the different Multisample RL, a_t has to be echoed back to every Multisample RL to associate the reward it is observing with the correct action. Moreover, to ensure that the Mediator RL does not keep on favoring one action over the rest. We extend the design of the Mediator RL to include a fairness measure combined with the humans' experiences to calculate the reward r_m . The design of the Mediator RL will be discussed in Section 3.

3 Algorithm Design

This section describes how the Governor RL and the Mediator RL agents are designed and how they interact with each other.

3.1 Design of the Governor RL Agent

We model the problem of adapting the values of T_l and T_a as executing an optimal policy over a Markov Decision Process (MDP) whose states are defined over T_l and T_a and whose reward function captures the IoT system performance. By executing such an optimal policy, the RL agent is guaranteed to search over the space of T_l and T_a values to maximize the system's performance. The MDP main components namely state space, action space, and reward function are designed as follows:

Governor MDP State space \mathcal{X}_G : The MDP states are the different values that T_l and T_a can take. That is, each state $s_g \in \mathcal{X}_G$ is determined by a tuple (T_l, T_a) . Since RL agents' performance depends heavily on the cardinality of the state space, we discretize the continuous values of T_l and T_a into a finite number of values where such discretization depends on the context of the application as it will be discussed in Sections 5 and 6.

Governor MDP Action space \mathcal{A}_G : The action space \mathcal{A}_G in this MDP contains all the possible combinations of changing the T_l and T_a values. Particularly, the action can be either *decrement* with some value \downarrow , *increment* with some value \uparrow , or *remain* the same value \circ . Such actions are defined for both T_l and T_a under the constraint that $T_l \geq T_a$. We designed the actions as an increment or a decrement rather than choosing a specific tuple (T_l, T_a) from all the possible combinations to make sure that there is no big sudden change in the actuator rate (T_a) which may compromise the human’s experience.

Governor MDP Reward Function R_G : Each state $s_g \in \mathcal{S}_G$ is associated with a performance p_s . Such performance p_s is a measure of the human’s experience and the performance of the IoT system, which depends on the context of the application, as it will be discussed in Section 4. The performance is calculated after running the Multisample Q -learning [14] using the values of T_l and T_a associated with state s_g . The associated performance p_s plays a role in determining the reward r_g value that accrued due to taking the action a_g at the state s_g . In particular, the reward value $r_g = R_G(s_g, a_g)$ (positively or negatively) depends on a weighted difference between the performance of the system p_s at the state s_g (before modifying the values of T_l and T_a) and the performance of the system $p_{s'}$ at the new state s'_g (after modifying the values of T_l and T_a).

In this MDP setting, the transition probabilities are known apriori since *increasing/decreasing/remain the same* T_l or T_a leads to a known state. However, the reward is unknown apriori because it depends on the human’s experience. Moreover, it can change over time. Hence, to solve the MDP when the reward values are unknown, we use Q -learning, which is summarized in Algorithm 1.

3.2 Design of the Mediator RL Agent

We model the problem of learning the best weights (w_1, w_2, \dots, w_n) for the adaptation actions collected for individual personalized performance in a multi-human setting (a_1, a_2, \dots, a_n) as executing an optimal policy over a Markov Decision Process (MDP).

Mediator MDP State space \mathcal{X}_M : The MDP states are the different values of w_i that can be summed up to 1. We discretize their continuous values into finite number of state space \mathcal{X}_M , where their values are chosen from a predefined set of values $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The state space’s size depends on the number of actions that we need to take their weighted average, reflecting the number of humans in the environment.

Mediator MDP Action space \mathcal{A}_M : The action space \mathcal{A}_M of the Mediator MDP contains all the possible jumps \nearrow to all the states with the same probability. Such a design choice allows the Mediator RL to rapidly switch between weights and converge faster to the optimal assignment of weights.

Mediator MDP Reward function R_M : Each state s_m has a performance p_s that is associated with it. The performance p is a measure of all the humans’ cumulative experience and the performance of the IoT system, which depends on the context of the application, as will be discussed in Section 6. The performance is calculated after applying the weighted average action ($a_t = \frac{\sum_{i=1}^n w_i a_i}{\sum_{i=1}^n w_i}$) at a rate of the minimum T_a across all individual humans (as chosen by their individual Governor RL). Each human will have a different

Algorithm 1 Governor RL

Hyper parameters: Learning parameters: α, γ, ϵ
Require:
 Sates $\mathcal{X}_G = \{(T_l, T_a)_1, \dots, (T_l, T_a)_x\}$
 Actions $\mathcal{A}_G = \{(\circ T_l, \circ T_a), (\circ T_l, \uparrow T_a), (\uparrow T_l, \circ T_a), (\uparrow T_l, \uparrow T_a),$
 $(\circ T_l, \downarrow T_a), (\downarrow T_l, \circ T_a), (\downarrow T_l, \downarrow T_a), (\downarrow T_l, \uparrow T_a),$
 $(\uparrow T_l, \downarrow T_a)\}$,
 Reward function $R_G : \mathcal{X}_G \times \mathcal{A}_G \rightarrow \mathbb{R}$
 Transition function $T_G : \mathcal{X}_G \times \mathcal{A}_G \rightarrow \mathcal{X}_G$
 Multisample Q -learning algorithm: $MuQL(T_l, T_a)$
 Learning rate $\alpha \in [0, 1], \alpha = 0.9$
 Discounting factor $\gamma \in [0, 1], \gamma = 0.1$
 ϵ -Greedy exploration strategy $\epsilon \in [0, 1], \epsilon = 0.2$
 Weighted Performance Difference \mathcal{W}
procedure $GovQL(\mathcal{X}_G, \mathcal{A}_G, R_G, T_G, \alpha, \gamma)$
 Initialize $Q_G : \mathcal{X}_G \times \mathcal{A}_G \rightarrow \mathbb{R}$ with 0
while true **do**
 Start in state $s_g \in \mathcal{X}_G$
 $p_s \leftarrow MuQL(s_g)$ ▷ Calculate the performance of s_g
 Calculate $\pi(s_g)$ according to Q_G and exploration strategy:
 with probability $\epsilon: \pi(s_g) \leftarrow$ choose an action a at random,
 with probability $1 - \epsilon: \pi(s_g) \leftarrow \operatorname{argmax}_a Q_G(s_g, a)$
 $a_g \leftarrow \pi(s_g)$
 $s'_g \leftarrow T_G(s_g, a_g)$ ▷ Receive the new state
 $p_{s'} \leftarrow MuQL(s'_g)$ ▷ Calculate the performance of s'_g
 $r_g \leftarrow R_G(s_g, a_g) = \mathcal{W}(p_{s'}, p_s)$ ▷ Receive the reward
 $Q_G(s'_g, a_g) \leftarrow (1 - \alpha) \cdot Q_G(s_g, a_g) + \alpha \cdot (r_g + \gamma \cdot \max_{a'} Q_G(s'_g, a'))$
 $s_g \leftarrow s'_g$
return Q_G

experience with the applied action a_t . Hence, the cumulative performance p_s associated with this state s_m will be a function of all the human experiences, which depends on the context of the application. To calculate the reward, we use the same idea of calculating the reward in Governor RL. In particular, the associated reward value r_m is computed by the reward function $R_M(s_m, a_m)$ which is a function of the relative performance between two states s_m (which corresponds to the current weights) and the next state s'_m (which corresponds to the new weights). In this MDP setting, the reward is unknown apriori as it depends on the cumulative humans’ experiences. Moreover, it can change over time. Hence, to solve the MDP when the reward values are unknown, we use the Q -learning summarized in Algorithm 2. In particular, the algorithm is divided into two procedures, $MEDQL$, which runs the Q -learning algorithm in which the reward value depends on the performance of the current state s_m and next state s'_m . The performance of each state is calculated in another procedure $CalculateStatePerformance$ that takes a state s_m as input and gets the preferred actions per human (by running $MuQL$) and the different actuation rates (by running $GovQL$) to calculate the cumulative performance p_s .

3.3 Design of Fairness-aware Mediator RL

The Mediator RL tries to maximize the total reward by choosing different weights for the different adaptation actions learned from multiple humans preferences. In one extreme case, as we will show in the results (Section 7), the Mediator RL may end up favoring one adaptation action over the others (by keeping assigning high weight to it) to increase its total reward. This means that the Mediator RL favors one human preference over the others because it

Algorithm 2 Mediator RL

Hyper parameters: Learning parameters: α, γ, ϵ

Require:

Humans $\mathcal{H}_M = \{h_1, h_2, \dots, h_n\}$

Sates $\mathcal{X}_M = \{(w_1, w_2, \dots, w_n)_1, \dots, (w_1, w_2, \dots, w_n)_x\}$

Actions $\mathcal{A}_M = \{(\nearrow 1, \nearrow 2, \dots, \nearrow x)\}$

Reward function $R_M : \mathcal{X}_M \times \mathcal{A}_M \rightarrow \mathbb{R}$

Transition function $T_M : \mathcal{X}_M \times \mathcal{A}_M \rightarrow \mathcal{X}_M$

Multisample Q -learning algorithm per human h : $MuQL_h$

Governor Q -learning algorithm per human h : $GovQL_h$

Weighted Average function: $WAVG(l, x) = \frac{\sum_{i=1}^n l(i)x(i)}{\sum_{i=1}^n x(i)}$

Learning rate $\alpha \in [0, 1], \alpha = 0.9$

Discounting factor $\gamma \in [0, 1], \gamma = 0.1$

ϵ -Greedy exploration strategy $\epsilon \in [0, 1], \epsilon = 0.2$

Weighted Performance Difference \mathcal{W}

procedure *CalculateStatePerformance*(s_m)

 Actions Array: $Act = \{\}$

 Actuation Rates Array: $ARates = \{\}$

 Human Experience Array: $HExp_{s_m} = \{\}$

$\forall h \in \mathcal{H}_M, Act.insert(a_h), s.t. a_h \leftarrow$ current action by $MuQL_h$

$\forall h \in \mathcal{H}_M, ARates.insert(T_a), s.t. T_a \leftarrow$ current T_a by $GovQL_h$

$a_{t_{s_m}} \leftarrow WAVG(Act, s_m)$

$minT_a \leftarrow \min\{ARates\}$

$\forall h \in \mathcal{H}_M, Update\ MuQL_h$ with $a_{t_{s_m}}$

$\forall h \in \mathcal{H}_M, Update\ GovQL$ with $min\ T_a$

$\forall h \in \mathcal{H}_M, HExp_{s_m}.insert(r_e),$

$s.t. r_e \leftarrow CalculatePerformance(h, a_{t_{s_m}})$

$p_s \leftarrow CalculateCumulativePerformance(HExp_{s_m})$

return p_s

procedure *MedQL*($\mathcal{X}_M, \mathcal{A}_M, R_M, T_M, \alpha, \gamma$)

 Initialize $Q_M : \mathcal{X}_M \times \mathcal{A}_M \rightarrow \mathbb{R}$ with 0

while true **do**

 Start in state $s_m \in \mathcal{X}_M$

$p_s \leftarrow CalculateStatePerformance(s_m)$

\triangleright Calculate the performance of s_m

 Calculate π according to Q_M and exploration strategy:

 with probability $\epsilon: \pi(s_m) \leftarrow$ choose an action a at random,

 with probability $1 - \epsilon: \pi(s_m) \leftarrow \arg\max_a Q_M(s_m, a)$

$a_m \leftarrow \pi(s_m)$

$s'_m \leftarrow T_M(s_m, a_m)$ \triangleright Receive the new state

$p_{s'} \leftarrow CalculateStatePerformance(s'_m)$

\triangleright Calculate the performance of s'_m

$r_m \leftarrow R_M(s_m, a_m) = \mathcal{W}(p_{s'}, p_s)$ \triangleright Receive the reward

$Q_M(s'_m, a_m) \leftarrow (1 - \alpha) \cdot Q_M(s_m, a_m)$

$+ \alpha \cdot (r_m + \gamma \cdot \max_{a'} Q_M(s'_m, a'))$

$s_m \leftarrow s'_m$

return Q_M

looks at the cumulative performance. To model this inequality, we borrow concepts from sociology which named this phenomenon the ‘‘Matthew effect’’, which is summarized as the rich get richer and the poor get poorer [7]. Accordingly, to ensure fairness across all humans, we need to avoid the ‘‘Matthew effect’’ by modifying the reward function of the Mediator RL. In particular, we model the Mediator RL as a resource distributed among multiple humans, and we want to ensure the fair-share of this resource. This leads us to use the notion of utility. We define the utility of each human adaptation h at timestep t as:

$$u_{h_t} = \frac{1}{t} \sum_{j=0}^t \frac{j}{t} w_{h_j}.$$

In particular, u_{h_t} is the average weight assigned by the Mediator RL for a particular human h over a time horizon $[0 : t]$, where the factor $\frac{j}{t}$ is used to give more value to the recent weights learnt by the Mediator RL over the ones in the past. We then measure the fairness of the Mediator RL using the coefficient of variation (cv) of the human utilities [23]:

$$cv = \sqrt{\frac{1}{n-1} \sum_{h=1}^n \frac{(u_h - \bar{u})^2}{\bar{u}^2}},$$

where \bar{u} is the average utility of all humans. The Mediator RL is said to be more fair if and only if the cv is smaller. Accordingly, we modify the reward function R_M to reflect the changes in cv . In particular, in Algorithm 2, we add a new fairness measure \mathcal{F} which denotes the difference between the cv values at s_m and s'_m . Hence, we change r_m to become:

$$r_m \leftarrow R_M(s_m, a_m) = (1 - \zeta)\mathcal{W}(p_{s'}, p_s) + \zeta\mathcal{F}(cv_{s'}, cv_s),$$

where ζ is used as a fairness scale from 0 to 1. As we will show in the results (Section 7), the objective to increase the performance measure \mathcal{W} and the objective to increase the fairness measure \mathcal{F} can be two opposing objectives.

4 Implementation and Evaluation Plan

We applied the proposed framework to two applications, the first one is in the domain of Advanced Driver Assistance System (ADAS), and in particular, we focus on the Forward Collision Warning (FCW) application. The second application is in the smart home automation system domain, and in particular, we focus on heating, ventilation, and air conditioning systems (HVAC). To ensure the safety of the real human subjects, we decided to use a simulation environment for both applications¹. In each application, we will explain the used simulator, the various human behaviors, how the performance is calculated, and how the proposed Fair-IoT framework can track the intra-human, inter-human, and multi-human variability.

5 Application 1: Human-in-the-Loop ADAS - Forward Collision Warning (FCW)

Standard FCW system measures the time-to-crash based on the distance and the relative velocity of the front object and, if the time-to-crash is below a certain threshold (signaling a possible risk of collision), it alerts the driver to apply the brakes. A human-in-the-loop FCW should take the human state and preferences while calculating this threshold. For example, if the driver is distracted, the alarm should be displayed earlier. Moreover, some drivers take more time to react to the alarm and press the brakes (response time). Hence, the alarm threshold should count for this response time as well. Personalized FCW has been addressed in the literature using offline learning of the threshold based on statistical modeling to reduce the required data [34], drivers’ expected response deceleration (ERDs) [52], parameter identification of driver behavior is proposed in [51], and adapting the time of the FCW based on the response time and the mental state of the driver [14].

¹At the time of preparing this manuscript, the COVID-19 pandemic was hindering against conducting experiments with real human subjects. Social distancing and isolation were enforced in the authors’ location.



Figure 4: Numerical results of Experiment 1. (Top) the performance of the system—obtained through brute force search—under different assignments for T_l and T_a for humans H_1 on the left, human H_2 in the middle, and human H_3 on the right. (Center) the progress of the states of the Governor RL agent across different iterations of the system showing the convergence to the states with maximum system performance. (Bottom) the performance of the system due to the adaptations of the Governor RL agent showing an increase in the performance over time.

Simulator: We used the Skoda Octavia model [49] for vehicle dynamics to simulate the physics of both the driving car and the leading car. We interfaced the vehicle model with the Matlab virtual reality toolbox (VR) to virtualize the vehicle dynamics. Each vehicle has parameters that specify its mass, the wheels’ mechanical friction on the ground, its acceleration, and braking forces.

Simulated human subjects: Following the observations in [14] to model driving behavior, we simulated three humans with three different driving behavior as follows:

- **Moderate driver behavior H_1 :** This driver has a moderate behavior with average braking intensity, acceleration intensity, average relative distance to the leading car, and average response time.
- **Aggressive driver behavior H_2 :** This driver tends to apply high intensity on both the brakes and the acceleration pedal. This is

accompanied by a small relative distance to the leading car and a short response time.

- **Slow driver behavior H_3 :** This driver tends to be more conservative. The driver tends to apply low intensity of both the brakes and the acceleration pedal. This is accompanied by a large relative distance to the leading car and a larger response time.

Adapting to the intra-human and inter-human variability using Governor RL-agent: As described in Algorithm 1, the state of MDP is a tuple of (T_l, T_a) . The response time of the driver and the behavior of the driver are unknown apriori. Different values of T_l and T_a may result in better performance and better driving experience for different humans (inter-human variability). We quantize the state space into 32 states. The sampling time T_s is fixed at 0.25 seconds while T_l and T_a are multiples of the sampling time. In particular, we choose $T_l \in [80, 90, 100, 110]$ and

$T_a \in [8, 9, 10, 11, 12, 13, 14, 15]$ multiples of T_s . A state s_g is one combination of T_l and T_a , such as (80,8). The different states capture the inter-human variability where different humans can have different response times which entail different learning rate (T_l) and different actuation rate (T_a).

Performance function: The performance is measured in terms of the following parameters:

- False Negatives (FN): The relative distance between the leading car and the driving car is below the safety distance of 7m, which is calculated based on the 2 seconds rule [37].
- False Positives (FP): The alarm with signaled unnecessarily.
- True Positives (TP): The alarm is signaled, and the driver pressed the brakes.
- True Negatives (TN): The alarm is not signaled, and the relative distance between the leading car and the driving car is above the safety distance of 7m.

Based on these parameters, the accuracy (Acc) and Matthews correlation coefficient (MCC) are then calculated for a specific state s_g . We combine both metrics (the Acc and the MCC) to measure the performance of state s_g . MCC has shown to be a good metric for binary classification evaluation [9], while accuracy gives the proportion of correct predictions. Hence, the performance p_s as indicated in Algorithm 1 is calculated as $p_s = Acc_s + MCC_s$. This value is normalized from 0 to 1.

5.1 Experiment 1: Inter-Human Variability

The performance p_s for each state s_g indicated by (T_l, T_a) for humans H_1, H_2 , and H_3 are calculated ahead of time to obtain the ground truth that allows us to validate the ability of Algorithm 1 to obtain the values of (T_l, T_a) that maximizes the performance of the system. However, when the algorithm runs, it does not know apriori the performance value of a state s_g until Multisample Q -learning runs for some time and returns the MCC and the Acc for this specific state. As shown in Figure 4 (top), the best state for H_1 is (110,12) then (80,13) with a very small difference. Similarly, the best state for H_2 is (110,13), which was the worst for H_1 , followed by (90,15). The best state for human H_3 is (80,13), then (80,8). These different states highlight the inter-human variability that needs to be accounted for to provide a personalized experience.

We run Algorithm 1 for 9000 iterations. As shown in Figure 4 (center), the algorithm converges to states (80,13), (110,13), and (80,13) for humans H_1, H_2 and H_3 , respectively. These states are the second best state for H_1 , and are the states with best performance for humans H_2 and H_3 . This is also reflected in how the performance values p_s (which is used to calculate the reward changes with the number of iterations for H_1 , as shown in Figure 4(bottom) where the performance value converges to value 0.95 which corresponds to state (80,13). As for H_2 , the performance value converges to 1, which corresponds to state (110,13), and for H_3 the performance value converges to 1 which corresponds to state (80,13), as shown in Figure 4 (bottom). These results show the proposed Governor RL's ability to track the inter-human variability and converge to the best (for H_2 and H_3) or near best (for H_1) performance for each human with different driving behavior.

5.2 Experiment 2: Intra-Human Variability

We now study the proposed Governor RL's ability to adapt to the changes in human preference across time. To that end, to show

how the algorithm can trace the change in behavior to adapt to the intra-human variability. We switch between humans on the same simulator after 10000 iterations and observe how the algorithm will adapt to the new behavior. First, we switch from H_1 to H_2 , then vice versa in Figure 5 (top) and the respective performance progress across the number of iterations are shown in Figure 5 (bottom). We repeat the same experiment but between H_1 and H_3 as seen in Figure 5 and the respective performance progress across the number of iterations in Figure 5. These figures show the ability of Governor RL to adapt to the changes in human behavior.

6 Application 2: Human-in-the-Loop Smart House - A Thermal System

Recent work in literature targets human-in-the-loop HVAC [27] while trying to assist human satisfaction and addressing fairness across all occupants [47]. Reinforcement learning has been proposed to adapt the HVAC set-point based on human activity [13]. A human-in-the-loop thermal system should take the human state and preferences into the computation loop while calculating the heater set-point. For example, the human body temperature decreases when the human goes to sleep [4], while the body temperature increases when human exercises [30] and with stress and anxiety [40]. Monitoring the human state [29], sleep cycle [38], physical activity [46] are all possible using IoT edge devices.

Simulator: We simulated a mathematical model for a thermal house. In particular, we utilize a thermodynamic model of the house that considers the geometry of the house, the number of windows, the roof pitch angle, and the type of insulation used. The house is being heated by a heater with an airflow with a temperature of $50^\circ c$. A thermostat is used to allow fluctuation of $2.5^\circ c$ above and below the desired set-point, which specifies the temperature that must be maintained indoors. The set-point is controlled by an external controller that runs the proposed Governor RL algorithm. The heat flow into the house is calculated by:

$$\frac{dQ_{heater}}{dt} = (T_{heater} - T_{room}) \cdot \dot{M} \cdot c,$$

where $\frac{dQ}{dt}$ is the heat flow from the heater into the house, c is the heat capacity of the air at constant pressure, and \dot{M} is the air mass flow rate through the heater (set at $1kg/sec$). T_{heater} and T_{room} are the temperature of hot air from the heater and the current room air temperature, respectively.

The thermal system in the house considers the heat flow from the heater and the heat loss to the environment. Heat losses to the environment depend on the thermal resistance of the house calculated by:

$$\frac{dQ_{losses}}{dt} = \frac{T_{room} - T_{out}}{R_{eq}},$$

where T_{out} is the external temperature and R_{eq} is the thermal equivalence of the house taken into consideration the thermal resistance of the wall and the windows. We assume that it is the winter season, so the T_{out} is kept at $50^\circ F$ with a random range of daily temperature variation.

We model the humans as a heat source with heat flow that depends on the average exhale breath temperature (EBT) and the respiratory minute volume (RMV) ($\frac{dQ_{human}}{dt} \propto RMV \cdot EBT$) [19].

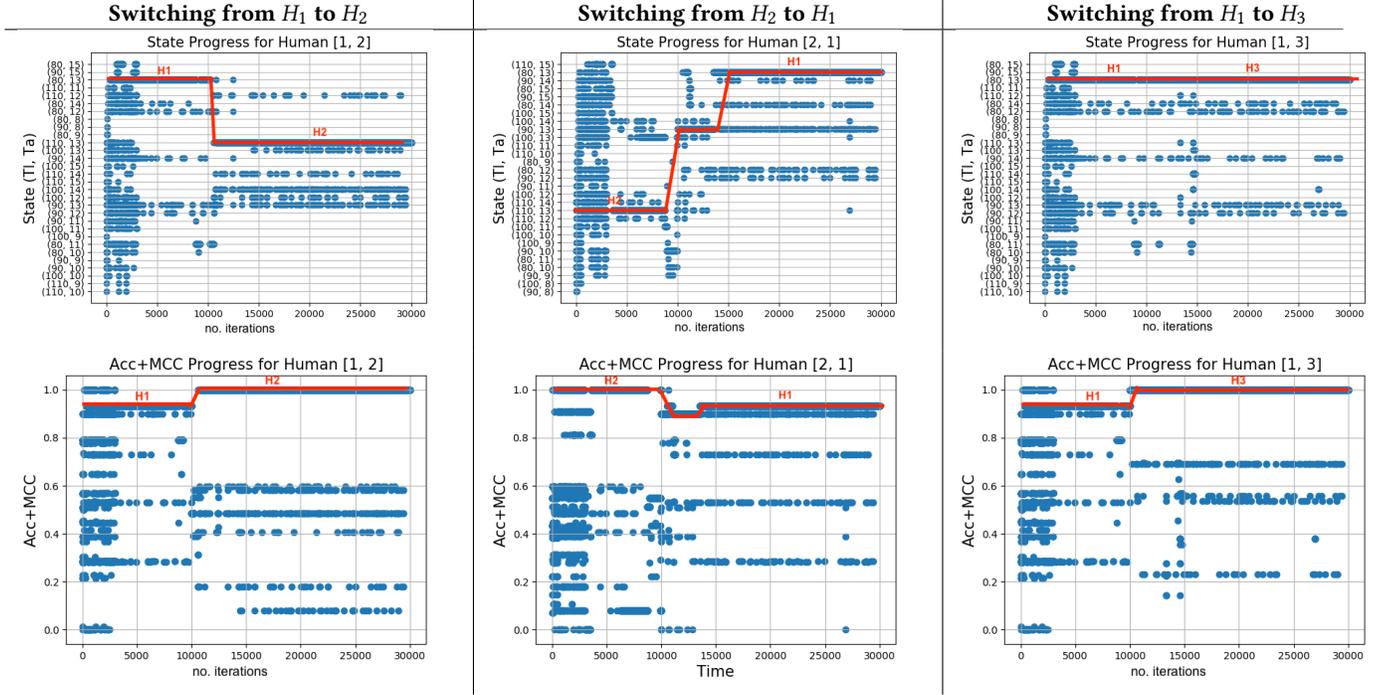


Figure 5: Numerical results of Experiment 2. (Top) the progress of the states of the Governor RL across different iterations of the system when the simulated human behavior changes across time, showing the ability of the Governor RL agent to track states with maximum system performance. (Bottom) the performance of the system when the simulated human behavior changes across time, showing the ability of the Governor RL to maximize the system performance over time.

Hr	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
H_1				*			*⊗	Δ					φ					Δ ♣	♣ ♣	♣ ♣	♣ ⊗	♣		*
H_2					*					*	⊗										Δ ♣ ♣	♣ ♣ ♣	♣ ♣ ♣	♣ ⊗
H_3						*					*	⊗		φ	φ		Δ ♣	♣ ♣ ♣	♣ ♣ ♣	♣ ♣ ♣	Δ ♣		♣ ⊗ Δ	

Table 1: Human activity across the 24 hours of the day. Sleeping:*, seated relaxed:♣, standing at rest:⊗, standing light activity:Δ, light domestic work:♣, standing medium activity:♣, washing dishes standing:♥, running:♣, and not home:φ. When multiple activities located in the same time slot, one of them is chosen randomly.

The respiratory minute volume is the product of the breathing frequency (f) and the volume of gas exchanged during the breathing cycle, which are highly dependent on human activity. For example, $RMV \approx 6 \text{ l/m}$ when the human is resting while $RMV \approx 12 \text{ l/m}$ represents a human performing moderate exercise [8]. We assume that the age/sex/time-of-day have no significance in the model.

Finally, the indoor temperature-time derivative $\frac{dT_{room}}{dt}$ is directly proportional to the difference between the heat flow rate from the heater and the heat losses to the environment while taking into consideration the heat flow from the occupants. We calculate it as follows:

$$\frac{dT_{room}}{dt} \propto \frac{dQ_{heater}}{dt} - \frac{dQ_{losses}}{dt} + \sum_{n=1}^N \frac{dQ_{human_i}}{dt},$$

where N is the number of occupants in the house. We simulated five days of a working human using this simulation environment.

Simulated human subjects: We simulated the behavior of three humans based on their activity and their stress level. The activity is simulated by the value of the RMV [8] and the metabolic rate [16]. Human stress level is simulated by an increase in the metabolic rate [40]. We simulated nine activities arranged in

the ascending order of RMV: sleeping, seated relaxed, standing at rest, standing light activity, light domestic work, standing medium activity, washing dishes standing, and running on a treadmill. We randomize the behavior by having different choices of activities in the same time slot. For example, as seen in Table 1, H_1 can be doing one of three activities (sleeping, standing at rest, running) between 6 am to 7 am. In total, we simulated five working days for each human.

- Occupant 1 (H_1) is an active human. H_1 wakes up early and exercises for one hour before leaving for work. When H_1 returns home, the activity changes according to the hour of the day.
- Occupant 2 (H_2) is a less active human. H_2 wakes up later than H_1 and does not perform any physical exercise before leaving for work.
- Occupant 3 (H_3) is a less active human with more time at home.

Adapting to the intra-human and inter-human variability using Governor RL agent: As described in Algorithm 1, the state of the MDP is a tuple of (T_i, T_a) . This state is passed to Multi-sample Q -learning as a parameter. The thermal sensation, the behavior, and the state of the human are unknown apriori to the algorithm.

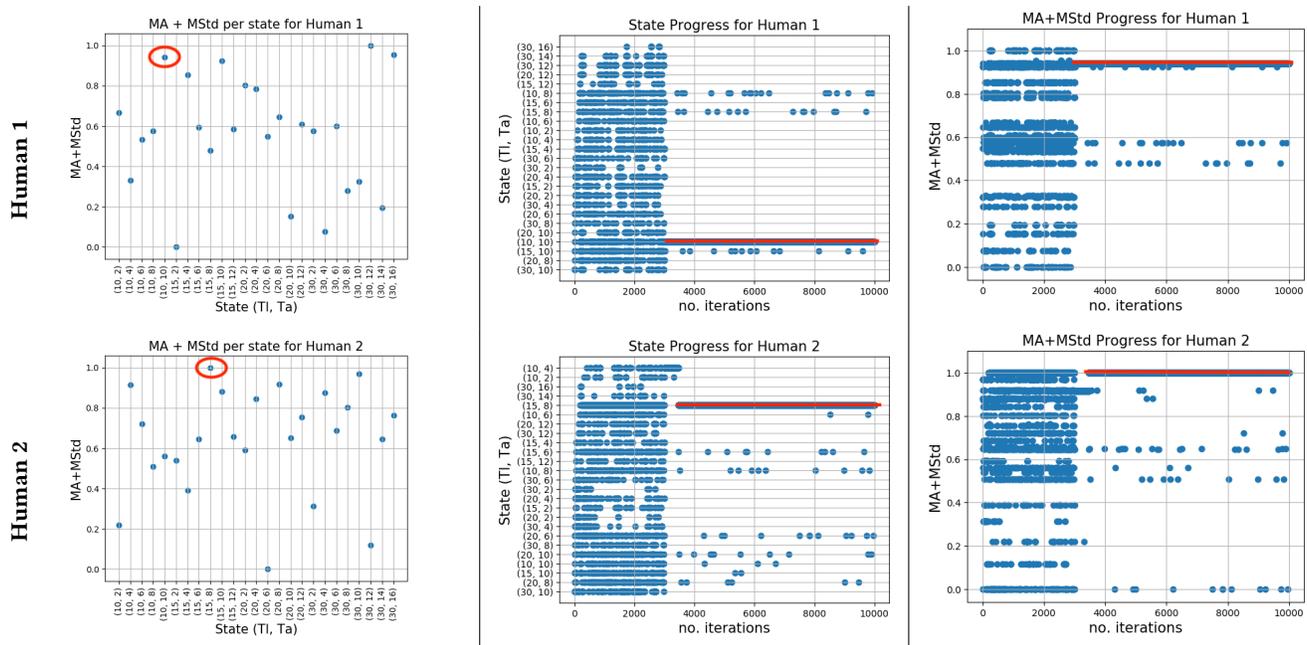


Figure 6: Numerical results of Experiment 3. (Left) The system’s performance—obtained through brute force search—under different assignments for T_l and T_a for both human H_1 on the top and human H_2 on the bottom. **(Center)** the progress of the states of the Governor RL agent across different iterations of the system showing the convergence to the states with maximum system performance. **(Right)** The system’s performance due to the adaptations of the Governor RL agent showing an increase in the performance over time.

Different values of (T_l) and (T_a) may give better thermal comfort for different humans (inter-human variability). We quantize the state space into 25 states. The sampling time is $T_s = 6$ min. T_l and T_a are multiples of the sampling time. In particular, $T_l \in [10, 15, 20, 30]$ and $T_a \in [2, 4, 6, 8, 10, 12, 14, 16]$ multiples of T_s such that, $T_l \geq T_a$. A state s is one combination of T_l and T_a , such as (15, 10). After choosing the MDP state for the Governor RL, Multisample Q -learning runs for simulated five days. We design the Multisample RL agent by modeling the human as an MDP with states corresponding to different activity (Act) and indoor temperature. A state s is defined by a tuple of T_{in} and Act . We quantize the indoor temperature (T_{in}) between 60°F to 80°F with 1°F granularity. When applying an action a on a state s , it can transition to any other state with unknown transition probability due to the change in human behavior. The action space contains the different set-points between 70°F to 80°F with 2°F granularity.

The reward function $R(s, a)$ depends on the thermal comfort of the human, which can be estimated using Prediction Mean Vote (PMV) [17]. PMV score indicates the thermal sensation of a human. It depends mainly on human activity, metabolic rate, clothing, and other environmental variables (airspeed, air temperature, mean radiant temperature, and vapor pressure). The scale of PMV ranges from -3 (very cold) to 3 (very hot). According to ISO standard ASHRAE 55 [2], a PMV in the range of -0.5 and $+0.5$ for the interior space is recommended to achieve thermal comfort. Estimation of the PMV score is calculated based on the knowledge of clothing insulation, the metabolic rate, the air vapor pressure, the air temperature, and the mean radiant temperature [17]. The human thermal sensation can change for the same state due to unmodelled external factors. Hence, the reward value r can change over

time. Moreover, each human can have a different response time (the time the human takes to feel a difference in thermal sensation). In addition to the variability introduced by the human, the response time of the thermal system (the time taken by the HVAC to actually reach the set-point) can be different and change with time due to system aging or unmodelled effects. Hence, T_l and T_a cannot be fixed and should differ from one human to another.

Adapting to the multi-human variability using Mediator RL agent: Since we modeled the human as a heat source, human activity affects the individual thermal sensation and affects indoor temperature. Moreover, when multiple humans exist in the same space and their activities are different, their thermal comfort will be at a different indoor temperature. To that end, to adapt to multiple humans in the same house, we need to measure their individual thermal comfort and then take their aggregate thermal comfort to find the HVAC set-point that can achieve a compromised thermal comfort for all the occupants. Accordingly, the HVAC set-point is controlled by running Algorithm 2 to mediate the diverse preferences of set-points from multiple humans. As mentioned in Algorithm 2, a Governor RL with Multisample RL gets the individual preferred set-point per human (a_l), as well as the individual actuation rate T_a . States of the Mediator RL models the different weights for three set-points with a total of 21 states. At each state s_m , Mediator RL takes the weighted average of the three set-points with the respective weights $a_{l_{s_m}}$ and sends it to the HVAC. For example, if the preferred set-points from the 3 humans are 72, 75, 78, respectively, and the state s_m is (0.2, 0.6, 0.4), then $a_{l_{s_m}} = 75.5$, and hence the set-point that is sent to the HVAC is 76 which is the nearest non-decimal number.

Performance function: The performance is calculated based on two parameters, (1) the moving average of the absolute value of the PMV score over five days and (2) the moving standard deviation of the PMV score. This is used to account for high fluctuation in thermal sensation. The performance value depends on a weighted sum of these parameters. In particular, the performance p_s for a particular state s is calculated as follows:

$$\overline{PMV}_{MA_s} = \frac{1}{n} \sum_{i=0}^{n-1} |pmv_{-i}|, \quad \overline{\sigma}_{MA_s} = \frac{1}{n} \sum_{i=0}^{n-1} \sigma_{-i},$$

$$p_s = \theta_1 \overline{PMV}_{MA_s} + \theta_2 \overline{\sigma}_{MA_s},$$

where $\theta_1 > \theta_2$ and n is the number of samples. Indeed, the lower the value of the performance, the worst the thermal sensation that the human experiences.

Reward functions: As described in Algorithm 1, the reward function is calculated based on the performance of the current state s_g (denoted as p_s) and the next state s'_g (denoted as $p_{s'}$) after applying an action a_g . If s'_g gives better performance than s_g , then it is a positive reward, otherwise, it is a negative reward. We use weighted performance difference between p_s and $p_{s'}$ to calculate the reward value. The reward value r_g is calculated according the following reward function $R_G(s_g, a_g)$:

$$\Delta = |p_{s'} - p_s|$$

$$\alpha = \mathcal{W}(\Delta), \text{ where } \mathcal{W} \text{ is an increasing step function.}$$

$$R_G(s_g, a_g) = \text{Sign}(p_{s'} - p_s) \times \alpha \times p_{s'}.$$

6.1 Experiment 3: Inter-Human Variability

We start by studying the proposed Governor RL agent’s ability to adapt to individual human behaviors to select the optimal (T_l, T_a) for each human. To that end, we apply the the same steps performed in application 1 in Section 5 for H_1 and H_2 . Multisample Q -learning returns \overline{PMV}_{MA_s} and $\overline{\sigma}_{MA_s}$ for a specific state s_g . By brute-forcing all combinations to discover the state with the maximum performance, we conclude that the best state for H_1 is (30,12) followed by (10,10) while the best states for H_2 are (15,8) and (30,10), as shown in Figure 6 (left). It is worth noting that the best state for H_2 results in poor performance for H_1 , highlighting again the inter-human variability that motivates the use of the proposed Governor RL. We run Algorithm 1 report the states chosen by the Governor RL across time in Figure 6 (center)². We run the algorithm for a total of 10000 iterations with adaptive ϵ that decreases when the reward value does not change for multiple consecutive iterations. The algorithm converges to the states (10,10) and (15,8) for humans H_1 and H_2 , respectively, which are the second-best and best states for the simulated humans. These results validate the designed Governor RL’s ability to generalize to multiple applications and address the inter-human variability challenge.

6.2 Experiment 4: Multi-Human Variability

Now, we study the ability of the proposed system (with both Governor RL and Mediator RL agents interacting together) to provide personalized performance when multiple occupants are present in the system. Similar to the previous experiments, we start by

²Due to space limitation, we are showing the results for H_1 and H_2 only.

brute-forcing all the states of the Governor and the Mediator RL agents to obtain the ground truth for the performance of the system under different values for T_l, T_a for each human along with the different values for the mediator weights w_1, w_2, w_3 . While the optimal values for T_l, T_a for each human (individually) was discussed in Experiment 3, we report in Figure 7a the system performance for different values of w_1, w_2, w_3 . In particular, state (0, 0, 1) shows the best performance while states (0.6, 0, 0.4) and (0.8, 0.2, 0) show the worst performances. Next, we run the whole system and compares its convergence to the state with the maximum total reward. First we run the whole system *without* fairness in Mediator RL reward by assigning ζ to value *zero* as discussed in Section 3.3.

The progress of the states is shown in Figure 7b, where the Mediator RL converges to (0, 0, 1) at iteration 3000, which is the state of maximum performance. This is also reflected in how the performance value changes in Figure 7d.

This is a counter-intuitive result since it entails that the system’s optimal performance is achieved when the HVAC set-point ultimately favors human H_3 . However, one explanation for this result is that human H_3 stays in the house for more time than both H_1 and H_2 and hence the moving average reward is maximized when human H_3 is given the highest priority. This motivates using the fairness-aware Mediator RL as explained in Section 3.3 for the next experiment.

6.3 Experiment 5: Fairness with Multi-Human Variability

We run the same experiment discussed in experiment 4 but with the fairness-aware Mediator RL by assigning ζ a value of 0.5 to balance the reward between the performance measure \mathcal{W} and the fairness measure \mathcal{F} . As seen in Figure 7f, the states that the Mediator RL chooses were a combination of different states across time. In particular, states (0, 0, 1), (0, 1, 0), (0.2, 0.6, 0.2), (0.6, 0.4, 0), (0.8, 0, 0.2), and (0.8, 0.2, 0) were selected across time. This shows that the Mediator RL changes the priority of adaptation between H_1, H_2 , and H_3 across time to ensure fairness. As a result, the performance does not converge to the maximum value state, but it changes depending on the current weights assigned by the Mediator RL. However, even though the performance does not converge to the maximum value, the fairness increases. This is reflected by the difference in the coefficient of variation (cv) values between experiment 4 (Figure 7d) and experiment 5 (Figure 7h). The lower value of cv indicates higher fairness. In particular, the cv has been improved by 1.5 orders of magnitude through integrating the fairness measure \mathcal{F} in the reward function of the Mediator RL.

6.4 Experiment 6: Personalized vs Static Smart HVAC Systems

Finally, we assess the value of personalizing the experience of the smart HVAC system compared to non-personalized HVAC that uses fixed, manually chosen set-point. Figure 8 (left) shows the PMV of all three simulated humans for fixed set-points of 70F and 76F, respectively. These figures show that the PMV of the three residents exceeds the acceptable range of comfortable thermal, which is between -1 and 1 [2]. On the other hand, and as shown in Figure 8(right), the PMV of the three residents exhibits a more favorable behavior due to the set-point adaptations provided by

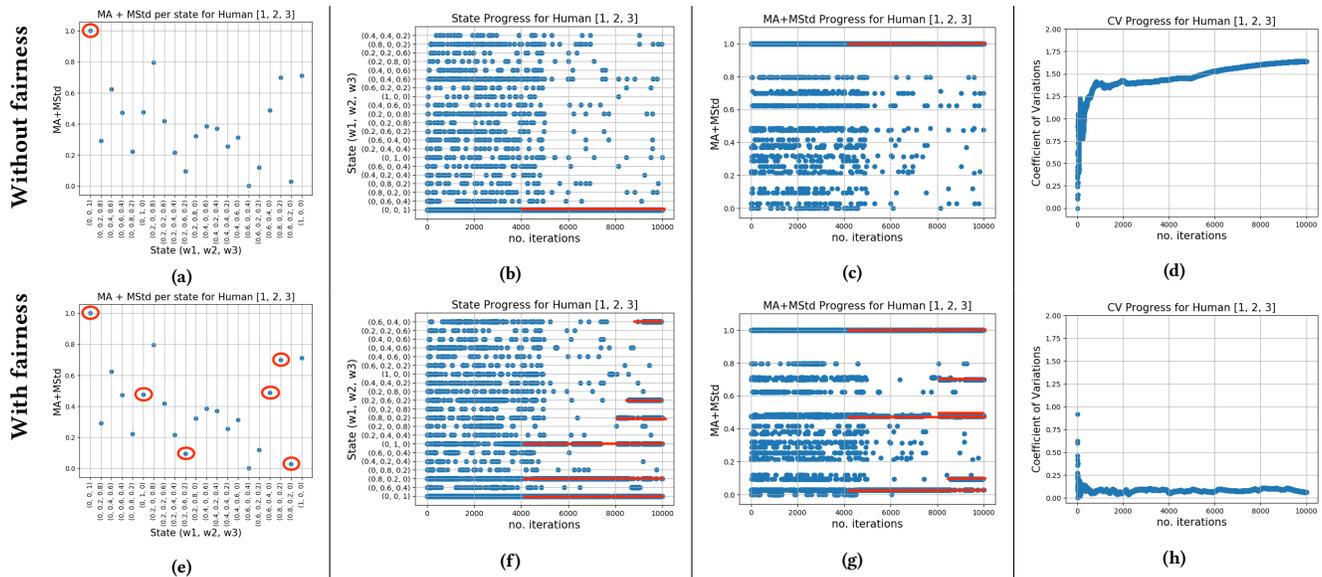


Figure 7: Numerical results of Experiment 4 and 5 by comparing the performance of the Mediator RL without/with the fairness measure. (a) and (e) show the performance of the system—obtained through brute force search—under different assignments for (w_1, w_2, w_3) . The states that the Mediator RL chooses are circled. (b) shows the progress of the states of the Mediator RL agent without the fairness measure \mathcal{F} across different iterations of the system showing the convergence to the state with maximum system performance, while (f) shows the states that Mediator RL chooses with fairness measure which are multiple states that correspond to the circled states in (e). While (c) shows the performance over time, while (g) shows that with fairness measure, the performance of the system depends on the states that the Mediator RL chooses to balance the fairness across the humans. (d) and (h) show the difference in the coefficient of variation. The low values indicate higher fairness.

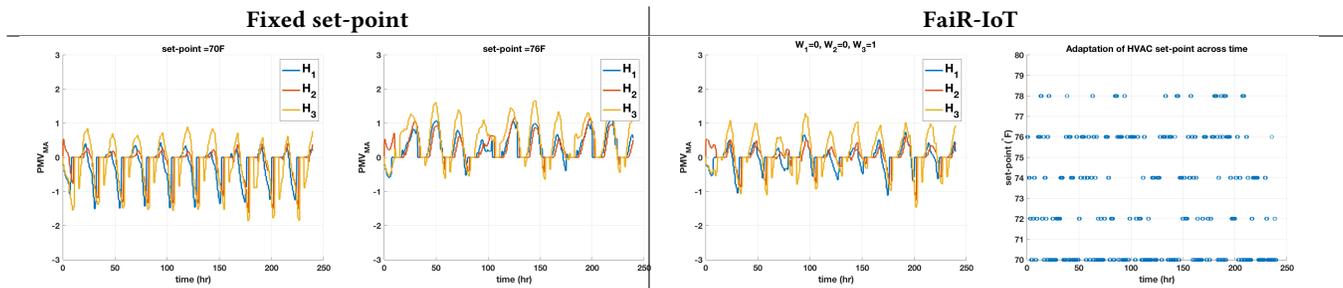


Figure 8: Numerical results of Experiment 6 comparing the PMV of the three residents when: Moving Average of PMV for the three humans at fixed HVAC set-point of $70F$ and $76F$ and Moving average of PMV for the three humans when the proposed FaiR-IoT framework is used to adapt the HVAC set-point along with the computed set points. Thanks to the proposed FaiR-IoT framework, the PMV of the three residents is maintained within the acceptable thermal comfort of $[-1, 1]$.

the proposed framework. It is worth noting that the personalized framework does utilize these two set-points $70F$ and $76F$ most of the time (as shown in Figure 8 (right)). Nevertheless, thanks to its ability to switch between these two values depending on the resident comfort, the system’s performance is improved by 41.7% and 58.96% compared to the manually fixed set-point scenarios.

7 Conclusion

Human modeling and human preference prediction hold the promise of disrupting the status-quo by designing complex IoT systems that weave advances in human sensing into the fabric of large-scale and societal IoT systems. However, with multiple humans interacting within the same IoT application space, adaptation fairness is a crucial concern. In this paper, we proposed **FaiR-IoT**, a Fairness-aware Human-in-the-Loop Reinforcement Learning framework that

addresses the challenge of human variability modeling with the fairness of adaptation. FaiR-IoT uses three hierarchical RL agents; a Governor RL with a Multisample RL to address the intra-human and inter-human variability, and a fairness-aware Mediator RL to address the multi-human variability. We showed the ability of FaiR-IoT to personalize two different IoT applications in the domain of ADAS and smart homes. By adapting to the human’s variability, FaiR-IoT was able to improve the human experience by 40% to 60% compared to the non-personalized systems and enhancing the system’s fairness by 1.5 orders of magnitude. Given that this framework focuses on establishing the intersection between human modeling and fairness in IoT applications, it opens up new research and application development directions for more fair and human-centric IoT.

References

- [1] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 837–846.
- [2] ASHRAE/ANSI Standard 55-2010 American Society of Heating, Refrigerating, and Air-Conditioning Engineers. 2010. Thermal environmental conditions for human occupancy. *Inc.Atlanta, GA, USA* (2010).
- [3] Bharathan Balaji, Jian Xu, Anthony Nwokafor, Rajesh Gupta, and Yuvraj Agarwal. 2013. Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.
- [4] Judith Barrett, Leon Lack, and Mary Morris. 1993. The sleep-evoked decrease of body temperature. *Sleep* 16, 2 (1993), 93–99.
- [5] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. 2020. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [6] Th Bedford and CG Warner. 1934. The globe thermometer in studies of heating and ventilation. *Epidemiology & Infection* 34, 4 (1934), 458–473.
- [7] Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. 2018. The Matthew effect in science funding. *Proceedings of the National Academy of Sciences* 115, 19 (2018), 4887–4890.
- [8] Robert G. Carroll. 2007. Pulmonary System. In *Elsevier's Integrated Physiology*. Elsevier, Chapter 10, 99–115.
- [9] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 1 (2020), 6.
- [10] Xiuyuan Du, Baizhan Li, Hong Liu, Dong Yang, Wei Yu, Jianke Liao, Zhichao Huang, and Kechao Xia. 2014. The response of human thermal sensation and its prediction to temperature step-change (cool-neutral-cool). *PLoS one* 9, 8 (2014), e104320.
- [11] Isha Dua, Akshay Uttama Nambi, CV Jawahar, and Venkat Padmanabhan. 2019. AutoRate: How attentive is the driver?. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.
- [12] Salma Elmalaki, Berken Utku Demirel, Mojtaba Taherisadr, Sara Stern-Nezer, Jack J. Lin, and Mohammad Abdullah Al Faruque. 2021. Towards Internet-of-Things for Wearable Neurotechnology. In *Proceedings of the The 22nd International Symposium on Quality Electronic Design (ISQED'21)*.
- [13] Salma Elmalaki, Yasser Shoukry, and Mani Srivastava. 2018. Internet of Personalized and Autonomous Things (IoPAT) Smart Homes Case Study. In *Proceedings of the 1st ACM International Workshop on Smart Cities and Fog Computing*. 35–40.
- [14] Salma Elmalaki, Huey-Ru Tsai, and Mani Srivastava. 2018. Sentio: Driver-in-the-Loop Forward Collision Warning Using Multisample Reinforcement Learning. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 28–40.
- [15] Salma Elmalaki, Lucas Wanner, and Mani Srivastava. 2015. Caredroid: Adaptation framework for android context-aware applications. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 386–399.
- [16] Engineering ToolBox. 2004. "Metabolic Rate". https://www.engineeringtoolbox.com/met-metabolic-rate-d_733.html.
- [17] Poul O Fanger. 1970. Thermal comfort. Analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering*. (1970).
- [18] Xiaonan Guo, Bo Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2017. WiFi-enabled smart human dynamics monitoring. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–13.
- [19] Jitendra K Gupta, Chao-Hsin Lin, and Qingyan Chen. 2010. Characterizing exhaled airflow from breathing and talking. *Indoor air* 20, 1 (2010), 31–39.
- [20] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*. 3909–3917.
- [21] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio Garcia Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*. 3326–3336.
- [22] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*. 1617–1626.
- [23] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. 1984. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA* (1984).
- [24] Jiechuan Jiang and Zongqing Lu. 2019. Learning fairness in multi-agent systems. In *Advances in Neural Information Processing Systems*. 13854–13865.
- [25] Hadi S Jomaa, Josif Grabocka, and Lars Schmidt-Thieme. 2019. Hyp-rl: Hyperparameter optimization by reinforcement learning. *arXiv preprint arXiv:1906.11527* (2019).
- [26] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.
- [27] Wooyoung Jung and Farrokh Jazizadeh. 2017. Towards integration of doppler radar sensors into personalized thermoregulation-based control of HVAC. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 21.
- [28] Dingwen Li, Patrick G Lyons, Chenyang Lu, and Marin Kollef. 2020. DeepAlerts: Deep Learning Based Multi-Horizon Alerts for Clinical Deterioration on Oncology Hospital Wards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 743–750.
- [29] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 389–402.
- [30] Chin Leong Lim, Chris Byrne, and Jason KW Lee. 2008. Human thermoregulation and measurement of body temperature in exercise and clinical settings. *Annals Academy of Medicine Singapore* 37, 4 (2008), 347.
- [31] Jiakang Lu, Tamim Sookoor, Vijay Srinivasan, Ge Gao, Brian Holben, John Stankovic, Eric Field, and Kamin Whitehouse. 2010. The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 211–224.
- [32] Peter Middleton, Peter Kjeldsen, and Jim Tully. 2013. Forecast: The internet of things, worldwide, 2013. *Gartner Research* (2013).
- [33] Seungwuk Moon and Kyongsu Yi. 2008. Human driving data-based design of a vehicle adaptive cruise control algorithm. *Vehicle System Dynamics* 46, 8 (2008), 661–690.
- [34] F. Muehlfeld, I. Doric, R. Ertlmeier, and T. Brandmeier. 2013. Statistical Behavior Modeling for Driver-Adaptive PreCrash Systems. *IEEE Transactions on Intelligent Transportation Systems* 14, 4 (Dec 2013), 1764–1772. <https://doi.org/10.1109/TITS.2013.2267799>
- [35] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. 2018. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*. 3303–3313.
- [36] Akshay Uttama Nambi, Shruthi Bannur, Ishit Mehta, Harshvardhan Kalra, Aditya Virmani, Venkata N Padmanabhan, Ravi Bhandari, and Bhaskaran Raman. 2018. Hams: Driver and driving monitoring using a smartphone. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 840–842.
- [37] New York State Department of Motor Vehicles. [n.d.]. Defensive Driving. <https://dmv.ny.gov/about-dmv/chapter-8-defensive-driving>.
- [38] Anh Nguyen, Raghdha Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, and Tam Vu. 2016. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 230–244.
- [39] David Sousa Nunes, Pei Zhang, and Jorge Sá Silva. 2015. A survey on human-in-the-loop applications towards an internet of all. *IEEE Communications Surveys & Tutorials* 17, 2 (2015), 944–965.
- [40] Berend Olivier, Theo Zethof, Tommy Pattij, Meg van Boogaert, Ruud van Oorschot, Christina Leahy, Ronald Oosting, Arjan Bouwknecht, Jan Veening, Jan van der Gugten, et al. 2003. Stress-induced hyperthermia and anxiety: pharmacological validation. *European journal of pharmacology* 463, 1-3 (2003), 117–132.
- [41] Rosalind W Picard. 2000. *Affective computing*. MIT press.
- [42] Marco Pritoni, Kiernan Salmon, Angela Sanguinetti, Joshua Morejohn, and Mark Modera. 2017. Occupant thermal feedback for improved efficiency in university buildings. *Energy and Buildings* 144 (2017), 241–250.
- [43] Anthony Rowe, Mario E Berges, Gaurav Bhatia, Ethan Goldman, Ragunathan Rajkumar, James H Garrett, José MF Moura, and Lucio Soibelman. 2011. Sensor Andrew: Large-scale campus-wide sensing and actuation. *IBM Journal of Research and Development* 55, 1.2 (2011), 6–1.
- [44] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems*.
- [45] Dorsa Sadigh, S Shankar Sastry, Sanjit A Seshia, and Anca Dragan. 2016. Information gathering actions over human internal state. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 66–73.
- [46] Tal Shany, Stephen J Redmond, Michael R Narayanan, and Nigel H Lovell. 2012. Sensors-based wearable systems for monitoring of human movement and falls. *IEEE Sensors Journal* 12, 3 (2012), 658–670.
- [47] Eun-Jeong Shin, Roberto Yus, Sharad Mehrotra, and Nalini Venkatasubramanian. 2017. Exploring Fairness in Participatory Thermal Comfort Control in Smart Buildings. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '17)*. ACM, 19:1–19:10.
- [48] Jai Kyoung Sim, Sunghyun Yoon, and Young-Ho Cho. 2018. Wearable sweat rate sensors for human thermal comfort monitoring. *Scientific reports* 8, 1 (2018), 1–11.

- [49] Skoda Auto. [n.d.]. Skoda Octavia model. <https://www.mathworks.com/help/sl3d/examples/vehicle-dynamics-visualization.html>.
- [50] Huey-Ru Debbie Tsai, Yasser Shoukry, Min Kyung Lee, and Vasumathi Raman. 2017. Towards a socially responsible smart city: dynamic resource allocation for smarter community service. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 13.
- [51] Jianqiang Wang, Chenfei Yu, Shengbo Eben Li, and Likun Wang. 2016. A forward collision warning algorithm with adaptation to driver behaviors. *IEEE Transactions on Intelligent Transportation Systems* 17, 4 (2016), 1157–1167.
- [52] Xuesong Wang, Ming Chen, Meixin Zhu, and Paul Tremont. 2016. Development of a Kinematic-Based Forward Collision Warning Algorithm Using an Advanced Driving Simulator. *IEEE Transactions on Intelligent Transportation Systems* 17, 9 (2016), 2583–2591.
- [53] Yi Wang, Bhaskar Krishnamachari, Qing Zhao, and Murali Annavaram. 2010. Markov-optimal sensing policy for user state estimation in mobile devices. In *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 268–278.
- [54] Chang Xu, Tao Qin, Gang Wang, and Tie-Yan Liu. 2017. Reinforcement learning for learning rate control. *arXiv preprint arXiv:1705.11159* (2017).
- [55] Chi Zhang, Sanmukh R Kuppannagari, Chuanxiu Xiong, Rajgopal Kannan, and Viktor K Prasanna. 2019. A cooperative multi-agent deep reinforcement learning framework for real-time residential load scheduling. In *Proceedings of the International Conference on Internet of Things Design and Implementation*. 59–69.