POSSIBLE IMPACTS OF SMART HOMES ON HUMAN BEHAVIOR: A REINFORCEMENT LEARNING APPROACH

Shashi Suman, Ali Etemad, Francois Rivest
1 Dept. ECE and Ingenuity Labs Research Institute, Queen’s University
2 Dept. of Mathematics and Computer Science, Royal Military College of Canada
{shashi.suman, ali.etemad}@queensu.ca, francois.rivest@mail.mcgill.ca, rmc.ca

ABSTRACT

Smart homes are becoming increasingly popular as a result of advances in machine learning and cloud computing. Devices such as smart thermostats and speakers are now capable of learning from user feedback and adaptively adjust their settings to human preferences. Nonetheless, these devices might in turn impact human behaviour. To investigate the potential impacts of smart homes on human behaviour we simulate a series of Hierarchical-Reinforcement Learning-based human models capable of performing various activities namely setting temperature and humidity for thermal comfort inside a Q-Learning-based smart home model. We then investigate the possibility of the human models’ behaviors being altered as a result of the smart home and the human model adapting to one another. For our human model, the activities are based on Hierarchical-Reinforcement Learning. This allows the human to learn how long it must continue a given activity and decide when to leave to pursue a different one. We then integrate our human model in the environment along with the smart home model and perform rigorous experiments considering various scenarios involving a single human model and two human models with the smart home. Our experiments show that with the smart home, the human model can exhibit unexpected behaviors like frequent changing of activities and an increase in the time required to modify the thermal preferences. With two human models, we interestingly observe that certain combinations of models result in normal behaviours, while other combinations exhibit the same unexpected behaviours as those observed from the single human experiment.

1 Introduction

Smart Home Systems (SHS) have drastically evolved in recent years due to advances in AI, enhanced connectivity, affordability, and the Internet of Things (IoT). For instance, smart thermostats [3] and smart device schedulers [25] are increasing in popularity everyday. These devices and the ecosystems which they create, aim to enhance human quality of life by saving time and costs, and by increasing comfort. On the other hand, the increased use of AI and automation in our daily lives has begun to impact human performance [23]. For instance, it was shown in [31] that with increased efficiency in human performance, which can be the result of automation and the use of AI, happiness also increases.

Reinforcement learning (RL) has been widely used to enable agents to learn from feedback provided either by other agents or the environment [39]. As a result, it has been extensively used in intelligent/automated agents [22].

While RL-based agents generally learn to optimize their performance for maximizing received rewards, the maximized received reward may only be optimal from the perspective of the agent without achieving the goal of the environment. For example, it was shown in [10] that a smart agent can exploit the environment to gain significant rewards by taking advantage of an edge-case in the rules of the environment without completing the game.

In recent years, SHS have also begun to utilize RL for learning from human interactions and feedback in order to provide customized and personalized user experience [34] [3]. While intelligent agents such as SHS generally provide value and increase quality of life, we question whether it is possible for an SHS powered by RL to learn to exploit the intricacies of human behaviours in order to maximize its own reward regardless of its implications on the humans.
In this paper, we tackle this problem by investigating if it is possible for an adaptive RL-based smart home to exploit assumptions in the environment and consequently change the behaviors of the human agents while trying to maximize its received reward by controlling thermal comfort parameters. To do so, we run a set of experiments on an RL-based simulated smart home and human agents in three steps. In the first step, we train a Hierarchical Reinforcement Learning based human model to pursue a set of activities while also setting temperature and humidity (TH) parameters to gain maximum activity-based rewards and thermal comfort rewards. In the second step, we train the smart home model with our trained human model. Here, the SHS learns to anticipate the thermal preference of the human model for each activity from the received feedback. In the final step, we integrate the trained human model into the smart home, and run/evaluate them together. Here, the SHS receives the temperature, humidity, and activity as observations, and sets the temperature and humidity in order to provide maximum comfort to the human model for that activity, based on its learned policy. For the smart home to learn to adapt to human preferences based on activities, we use Q-Learning [49].

For the human model, we utilize Hierarchical Reinforcement Learning, which is capable of efficiently modeling the ability of the human agent to switch between activities [7]. Figure 1 shows an overview of our study. Our simulations show that indeed unintended consequences can be caused in certain scenarios, as we observe behavioral changes exhibited by the human agent such as frequently switching between the activities along with increased time-steps to set the TH settings. We also perform multi-human experiments where two human agents are simultaneously integrated into a single home. Interestingly, we observe that in this scenario, human models are able to change their task order to minimize the differences between their TH preferences as to reduce the time spent in changing TH and maximize their comfort.

In summary, our contributions in this paper are as follows. (1) We model a smart home with RL to learn to optimize ambient parameters, namely temperature and humidity for maximizing human agents’ comfort. We then model a simulated human agent capable of pursuing and switching between various activities. The agent can also control the temperature and humidity levels in order to manage its own comfort. (2) Through our simulations, we find that in an environment that follows well-established physics laws of heating and humidity, the co-existence of the RL-based SHS and the human model can lead to unexpected changes in the optimal policy that would otherwise not be present without the SHS, resulting in a change in the behaviour of the human model. (3) We then perform additional experiments to consider a scenario where two human models simultaneously pursue their activities and try to optimize their comfort in parallel in a home environment. We discover that unintended behaviors are present with models that have less overlapping comfort ranges and different reward structures.

2 Related Work

In this section, we review some of the key papers in three main areas related to this paper. First, we review papers that use AI in smart environments. Next, we explore prior works in the area of human-RL interaction. These two surveys are necessary as our study integrates AI in the form of RL in a simulated smart home environment and aims to analyse the interaction between RL-based SHS and simulated human agents. Lastly, we review past works that have used RL to model human behaviour as we aim to simulate humans in the smart home using RL. It is therefore critical to base our human models on accepted approaches that have used RL for modeling human agents.
AI in Smart Environments. Recent advances in machine learning have enabled smart devices to learn user behaviors and preferences for enhanced user experience and comfort. For instance, in [3] smart thermostats were capable of learning user thermal preferences in real-time using Bayesian networks. In [29], neural networks were used to predict user behaviors in smart homes to adjust the ambient thermal conditions.

RL algorithms have also begun to make their way into SHS. In [34], value iteration, a reward-based algorithm, was used to learn user preferences about lighting, conditional on the number of occupants, time of day, and surrounding illumination. Multi-agent deep Q-Networks have been used in [35] to learn the energy cost of smart home appliances. Deep RL was employed in [16] for energy minimization along with a neural network that predicted user thermal comfort. This was further extended in [42] where air quality for a particular temperature was controlled using the Double Deep Q-Network (Double DQN) algorithm while also minimizing the energy consumption. Similarly in [3] Q-Learning was used to control the indoor ventilation to maintain air quality by learning the rate of CO2 generation from individuals. An autonomous ventilation system was designed in [19] that learned to maintain the indoor air quality using DQN in a simulated environment, managing to reduce the energy consumption as well. Similarly in [21], an air quality sensing system was developed where robots sensed the air quality via sensors and optimized their navigation path using Partial Observable Markov Decision Process (POMDP) and Double DQN.

Human-RL Interaction. A number of studies have explored the interaction between humans and RL systems. For instance, [35] discusses the notion of placing humans in the loop to provide feedback while an agent plays Atari games. With this approach, the learning stability of the agent increased as the number of iterations required for convergence was reduced. Similarly in [28], bots learned to converse through feedback from humans using the Reinforce algorithm, resulting in an improved learning curve.

Based on studies such as the above, it can be observed that including humans in RL loops generally results in better AI learning. Nonetheless, some studies have interestingly discovered potential negative or unintended consequences of human-RL interactions. For instance, it was shown in [35] that in complex environments where the human is unsure about the nature of feedback it has to provide to the agents, AI learning frequently fails. Moreover, in some other studies, it has been shown that RL agents can exploit certain rules in the environment to maximize their reward regardless of its implications on the overall system. For example in [10] the agent exploited the environment to obtain indefinite rewards, while in [5], the agent exploited the physics engine of the environment to win the game.

RL for Modeling Humans. Modeling human behavior as an artificial agent is a challenging task due to the existence of a large number of factors that govern our behaviors. As a result, its underlying learning mechanisms and associated reward systems are still open research problems. However, since the discovery of links between dopamine and the reward prediction error signal in RL [37], RL models have become key in studying multiple facets of human behaviours including learning, motivation, addiction, decision making, and numerous diseases [30]. RL models of learning, behaviours, and dopamine are now well established and used in a large number of publications [24].

Authors in [15] showed hierarchical RL to be more accurate in contextualizing and generalizing tasks, and measuring variations in performance between the users when compared with flat RL. In [35], hierarchical RL showed high accuracy in sentiment classification, noise reduction, and prioritizing certain sentiments in a given text document. Authors in [14] [8] [36] showed that humans in general divide a given task into sub-tasks and solve them in a hierarchical order that would return the highest reward at the end. Similarly, authors in [52] talked about the presence of hierarchical structures in the brain while making decisions, suggesting that we possess a cognitive control of sub-tasks based on their positions in the hierarchy order. These studies, among others, back the plausibility of using HRL-based models for simulating human decision making and behaviour in interaction with smart agents or environments.

3 Methods

Problem Setup. We aim to investigate whether for human agent \( H \) capable of interleaving a set of activities through policy \( \pi_H \), there exists an RL-based SHS, \( M \), that can change the human behavior in an attempt to maximize comfort. Specifically, for a policy \( \pi_H \) which \( H \) learns in the absence of \( M \), we aim to obtain \( \pi_{H'} \neq \pi_H \) when \( M \) is integrated into the environment. To do so, we assume \( M \) controls TH combined with an existing HRL model for \( H \). We design \( H \) capable of carrying out the activities rest, leisure, and physical workout, which are three common categories of activities in normal human life, through policy \( \pi_H \). \( M \) will learn to anticipate the TH preferences of \( H \), such that \( H \) would feel more comfortable while completing the activities through a different policy \( \pi_{H'} \). In the next two sections, we describe \( M \) and four variations of \( H \) to evaluate whether our hypothesis on the possibility of obtaining \( \pi_{H'} \neq \pi_H \) is correct.

Smart Home System. We model the SHS using a Markov Decision Process (MDP) defined by the tuple \( < S, A, P, R, \gamma > \), where \( S \) is the set of discretized states, \( A \) is the set of actions, \( P(s'|s,a) \) is the probability to go to state \( s' \) when taking action \( a \) in state \( s \), \( R(s,a) \) is the expected reward given when taking action \( a \) in state \( s \),
and $\gamma$ is the discounted factor devaluing rewards received later in time. The objective is to learn a policy $\pi(s)$ which returns the action to take in state $s$ that maximises the expected sum of future rewards given as $\sum_{t=1}^{\infty} \gamma^t r_t$, where $r_t$ is the reward received after taking the $t^{th}$ action. The expected sum of future rewards received for taking action $a$ in state $s$ by following policy $\pi$ is given by the Bellman equation:

$$Q^\pi(s,a) = \sum_{s'\in S} P(s'|s,a)[R(s,a) + \gamma Q^\pi(s',\pi(s'))].$$

(1)

The optimal policy $\pi^*(s) = \arg \max_{a \in A} \{Q(s,a)\}$ is obtained using Q-Learning. Given a transition from state $s$ to state $s'$ using action $a$ and reward $r$, the update rule is given by:

$$Q(s,a) \leftarrow (1 - \alpha)Q(s,a) + \alpha \left[ r + \gamma \max_{a' \in A}\{Q(s',a')\} \right],$$

(2)

where $\alpha$ is the learning rate. While learning, actions are selected using a decaying $\epsilon$-Greedy policy to keep a balance between exploration and exploitation. The $\epsilon$-Greedy policy is given by

$$\pi(s) = \begin{cases} \arg \max_{a \in A} \{Q(s,a)\} & \text{with probability } 1 - \epsilon \\ a \in A \text{ with probability } \epsilon/|A| \end{cases}.$$  

(3)

Based on the $\epsilon$ value, the agent decides between a random versus a greedy action, where the value of $\epsilon$ decays with each iteration.

The smart home receives state $s$ that consists of temperature, humidity, and the current activity that our human model is pursuing as a discrete state. Given that many recent wearable devices that track human activity can now be integrated into smart-home applications, we include the human activity in the observation that the smart home receives as a state to ensure the smart home can incorporate the activity when setting the temperature thus learning preference of the model for a specific activity. Based on this observation, the smart home may take an action to increase, maintain, or decrease the temperature or humidity respectively. The smart home receives a negative reward (-1) every time the human changes for a specific activity. Based on this observation, the smart home may take an action to increase, maintain, or decrease the temperature or humidity respectively. The smart home receives a negative reward (-1) every time the human changes the TH, thus, encouraging the smart home to set the TH to the human preference.

To simulate realistic changes in temperature and humidity of the environment, we use thermal calculations based on Newton’s law of cooling assuming the dew-point to be constant because of low deviation in temperatures to keep Newton’s law from breaking. If we consider our Human temperature preference to be $T_h$ and the surrounding temperature at time $t$ as $T_t$, then the temperature at the next time-step can be expressed as:

$$T_{t+1} = T_h + (T_t - T_h)e^{-k},$$

(4)

where $t$ is the number of time-steps and $k = 0.8$ is the decay constant that can be used to decide how fast or slow the desired temperature is achieved. For simplicity, we ignore the heat generated by the human body’s metabolism which varies due to sweat evaporation, radiation, convection, and conduction. Hence, we keep the mean radiant temperature at 22 degrees to reduce the complexity.

To simulate the changes in humidity within the environment when the user makes changes in the temperature, we use the August-Roche-Magnus approximation. As per definition, relative humidity in terms of dry bulb temperature and dew-point temperature can be approximately expressed as:

$$R_h = 100 \times \frac{17.625 \times T_d}{17.625 + T_d},$$

(5)

where $T$ is the room temperature (Eq. (4)) and $T_d$ is the dew-point temperature. Similarly, when the human agent or the smart home sets the humidity, changes in temperature can be defined approximately as:

$$T = 243.04 \times \frac{17.625 \times T_h}{243.04 + T_h} - \ln \frac{R_h}{100} - \frac{17.625 \times T_d}{243.04 + T_d},$$

(6)

where $T$ is the approximate temperature in °C given current humidity $R_h$ and dew-point temperature $T_d$. Since most homes are thermally insulated, very little variation is observed in temperature and humidity, making the dew-point fairly stable. In this paper, the simulated temperature varies between 15°C and 30°C. The dew-point does not vary much within this range, thus we consider a fixed dew-point of $T_d = 4°C$ for our simulations.

**Human Thermal Comfort.** To represent human thermal comfort, we use the Fanger model that provides an empirical solution. One of the important parameters described by the model is called the Predicted Mean Vote (PMV)
We divide the hierarchy for the human model representation for our simulated human. Rectangles are composite SMDP subroutines and ovals are primitive actions. Here T stands for Temperature and H stands for Humidity.

giving an understanding of the human sensation towards thermal comfort. This is the mean response for a large group of people for various thermal conditions. The PMV index spans from -3 to 3 with 3 being the hottest feeling, -3 being the coldest feeling, and 0 being the most comfortable feeling. PMV depends on six factors, which are the dry-bulb temperature, relative humidity, clothing index, wind speed, and metabolism index (specific to each activity). Based on [2], the clothing index for rest, leisure, and physical workout is 0.5, 0.67, and 0.36 respectively. Similarly, we have set wind speed in the surroundings to be 0. The optimal PMV range for the best comfort lies between -0.5 to 0.5. For our human model, we have used the PMV equations which can be found in [2] to express the intrinsic reward to the human agent.

**Human Basic Model.** We need a human model \( \mathcal{H} \) that can: (i) represent approximately how humans adapt based on reward; (ii) carry out or switch between tasks; and (iii) set TH to gain comfort and reward. Accordingly, we use reinforcement learning for the human model as such approaches have frequently been used for modeling human behaviors and adaptation based on rewards [18] [7] [17] [6] [9]. In particular, an HRL model capable of capturing human behaviors in selecting, completing, and switching among different tasks has been proposed in [17] [7] [6].

We design \( \mathcal{H} \) such that within a limited time window \( N \), it learns to select or leave the available or ongoing activities to maximize reward. We use a discounted Hierarchically-MAXQ (HMAXQ) algorithm from [1], which was shown to simulate well-known human behavior of task-interleaving [17]. In this formulation, the goal can be broken down into a hierarchy of sub-problems where each sub-problem is a separate Semi-Markov Decision Process (SMDP) that the human model has to learn. This allows us to combine primitive actions to create more composite ones. We select three possible types of activity classes based on activity intensity for the human agent: ‘rest’ for low intensity, ‘leisure’ for medium intensity, and ‘physical workout’ for high intensity. Each activity class has a different cost function, state-space, and PMV preferences. The entire hierarchy has four composite actions and six primitive (non-SMDP) actions as shown in Figure 2. An SMDP is similar to an MDP, with the exception that some actions can take multiple time-steps to be completed, and hence the agent receives delayed rewards. This is due to the fact that some actions are represented as macros consisting of sequences of actions.

We divide the hierarchy for \( \mathcal{H} \) into three distinct levels. The first level is the root node SMDP. The root subroutine chooses an activity that returns a maximum cumulative reward. The second level consists of tasks that the human agent must perform. Each task is an SMDP subroutine that models the progress within an activity based on the time spent in it so far. This helps us to generalize the interleaving ability of humans within activities. In each activity, the human can either continue or leave the activity, or decide to change either the temperature or humidity as the third level in the hierarchy.

To model our HMAXQ-based human agent, we use a three-part value function decomposition as described in [1][17] and given by:

\[
Q_r(a, s) = \begin{cases} 
\sum_{s'} P(s'|s,a)R(a,s') & \text{if } a \text{ is primitive} \\
Q_r(\text{child}(i), s) + Q_c(\text{child}(i), s, \pi_{\text{child}(i)}(s)) & \text{if } a = i \text{ is composite}
\end{cases}
\]

\[
Q_c(i, s, a) = \sum_{ss'} P'(s'|s,t,a)\gamma^t |Q_c(\pi_i(s'), s') + Q_c(i, s', \pi_i(s'))|
\]
Hierarchical Reinforcement learning is categorized into two notions for an optimal solution. i) Recursive Optimal and ii) Hierarchical Optimal. In Recursive Optimal or MAXQ [13], the expected return for performing a subtask $a$ is $Q^*(i, s, a) = Q_e(a, s) + Q_c(i, s, a)$. The expected external reward $Q_e(i, s, a)$ after the completion of subtask $i \neq 0$ is not a part of the value function which makes MAXQ a Recursive Optimal solution. In contrast, in Hierarchical Optimal [1], when solving the entire hierarchy, the policy is hierarchically optimal given the rewards from the external subroutines (Eq. 9). Each task may or may not be locally optimal. The Bellman equation for Hierarchical Optimal is given in Eq. [10].

In this paper, we have taken $H$ as a Hierarchical Optimal agent to imitate the switching ability of humans where they have context about the external rewards from other available tasks. The learning rules can be found in [11].

**Reward Structure.** We design the reward structure for all three activity types based on the generalized assumption a given task can be divided into sub-tasks, which in turn can help humans switch between them based on given set of priorities [14, 8]. In [17] this is modeled by rewarding the time-steps required to complete milestone within an activity, and by penalizing leaving of the activity prior to achieving the milestone. The reward represents the motivation that the human agent senses for a given activity in a given state [47]. Similarly, the penalty models the time required for humans to recall the forgotten information after an interruption [45].

For our simulations, we have three activities, namely, 'rest', 'leisure', and 'physical workout' as shown in Figure 4. For the 'rest' activity (Activity 0), as supported by [11], we model the reward $r(s)$ such that each period of rest increases the reward up to a maximum point, while providing smaller rewards for the subsequent periods (Figure 4(a)). For the 'leisure' activity (Activity 1), similar to [38, 56], we assume that the reward decreases over time. In the 'physical

![Diagram of Hierarchical Representation](image-url)
workout’ activity (Activity 2), spending time on workout increases the performance up to an optimal point, after which it starts decreasing as modeled by [41]. Occasionally, individuals can extend their workout and go beyond the fatigue threshold to achieve a considerable milestone and obtain a high reward as shown in [20, 12] and modeled in our reward function (Figure 4(c)).

While pursuing any of the activities, it is important for the human model to show reluctance in switching between tasks and make an effort to continue the current activity. To elaborate, in [46], it was demonstrated that the task-switching ability of humans is based on cognitive factors like task attractiveness, priority, engagement, interest, and others, which are difficult to model since each person might express these feelings differently. We model this by penalizing the ‘leave’ action (Figure 4), as well as the root to enforce reluctance in frequent task-switching. The root is penalized with a reward which is given as:

$$ R(0, s, a) = -c(task(s)) $$

where $task(s)$ returns the state of the child SMDP corresponding to root state $s$, and $c(s)$ is the penalty for leaving the task. Figure 4 shows the penalty (in red) for leaving the task or activity.

Finally, we integrate a penalty to represent discomfort when the TH is outside the comfort zone, encouraging the model to make adjustments when needed. Our final reward structure for the human model $H$ is given by:

$$ R_1(s, a) = \begin{cases} 
  r(s) - d & \text{if } a \text{ is continue} \\
  -c(s) & \text{if } a \text{ is leave}
\end{cases} $$

Since we want to show that differences between the expected human model used when training the SHS and the actual human model could lead to issues, we define a second reward function for $H$ as:

$$ R_2(s, a) = \begin{cases} 
  r(s) - d & \text{if } a \text{ is continue} \\
  -c(s) + d & \text{if } a \text{ is leave}
\end{cases} $$

where $d$ is defined as:

$$ d_{0.5} = \begin{cases} 
  |pmv(s)| & \text{if } |pmv(s)| > 0.5 \\
  0 & \text{if } |pmv(s)| < 0.5
\end{cases} $$

Here $pmv(s)$ is the intrinsic comfort reward given for the current state $s$, while $r(s)$ and $c(s)$ are the reward and penalty functions for pursuing and leaving the activity respectively as shown in Figure 4. In both equations $pmv(s)$ returns the thermal reward of the state as given in Eq. 14. In Eq. 13, it represents the sense of relief when humans leave an uncomfortable environment for a more comfortable one. In the next section, we will use Eq. 12 and Eq. 13 to represent the reward function of various human agents.

4 Experiments and Results

In this section, we describe our experiments and present the results. First, we utilize only one human agent in a smart home, while in the second sub-section, we integrate two human agents simultaneously into a single smart home to perform the studies for a multi-human setting.
Figure 5: Sample plots of activities over time for each model ($\mathcal{H}_A$, $\mathcal{H}_B$, $\mathcal{H}_C$, $\mathcal{H}_D$) with and without the SHS. In (a), (c), (e), each model learns to complete the tasks without interruption. In (b) and (f), the SHS anticipates human preferences, speeding up the time for human models $\mathcal{H}_A$, $\mathcal{H}_B$, and $\mathcal{H}_D$ respectively to complete the activities. In (d), Model $\mathcal{H}_C$ (different internal reward structure) behaves erratically in the presence of SHS. ‘Set $i$’ denotes the action of setting TH for the corresponding activity $i$.

4.1 SHS with a single Human model in the environment

We perform the following experiments to evaluate different assumptions regarding the human model $\mathcal{H}$ while interacting with the SHS. We train each human model for 350 episodes, which are empirically enough for convergence. Next, we train the human models with the SHS for an additional 150 episodes. We repeat our experiments 50 times to evaluate how the SHS performs with different combinations of TH settings when initializing the environment. Following we describe the different human models and experiments along with the obtained results.

First, we explore whether the SHS can learn the preferences of a given general human model. Second, we test whether such SHS could adapt to a human agent with a slightly different set of TH preferences. Finally, we test what would happen if the SHS interacts with a human agent that has a different reward structure than the general human model.

4.1.1 Experiment 1: Baseline

In this experiment, we explore a baseline scenario where an SHS is trained for an average human model using the existing thermal comfort sensational model [2]. Here, we aim to show that (1) our human model can indeed learn to complete each available activity, (2) the SHS can learn to anticipate the human TH preferences, and (3) the human performance does not exhibit any unexpected behaviors when integrated into the environment with the SHS.

Let Model $\mathcal{H}_A$ be a generic human model used to train the SHS with the optimal comfort PMV indices (i.e. the SHS initial settings) between −0.5 and 0.5 as per [2]. The metabolic rates used by the PMV function for the 3 activities are [1.0, 1.3, 1.8] respectively as defined in [2]. The reward function for $\mathcal{H}_A$ is given by Eq. [12].

As seen in Figures [5](a) and (b), this experiment shows that Model $\mathcal{H}_A$ is capable of completing each activity along with setting the optimal TH settings. As expected, with the SHS, the number of time-steps required for $\mathcal{H}_A$ to correct the TH is reduced as shown in Figure [6](a), which indicates that the SHS learns to quickly adjust the TH to the human preference for the current task, and the human to expect the SHS in doing so.
With Experiments 1 and 2, we can conclude that the human models in the RL-based smart home can successfully
interact with the SHS, the human model takes fewer time-steps to correct the TH. Figure 6(a) and (b) show that for the human
model to adjust the TH is reduced, without negatively affecting the completion of activities by the human agent. As expected,
with the SHS, the human model takes fewer time-steps to correct the TH. Figure 6(a) and (b) show that for the human
agent with the SHS (in green), the time-steps are reduced as compared to the human model without the SHS (in red)
for both Models $H_A$ and $H_B$. This concludes that the SHS can learn to adapt to Model $H_B$ with a different thermal
preference than that of $H_A$. This works well for the TH preferences of an average model $H_A$ as well as for $H_B$ with
different preferences, showing that the SHS model can adapt to both generic and specific human model preferences.
With Experiments 1 and 2, we can conclude that the human models in the RL-based smart home can successfully
complete the activities without abruptly leaving any of them while requiring fewer time-steps to set TH as shown in
Figure 6.

4.1.2 Experiment 2: Control

In this experiment, our goal is to validate that the SHS can learn to adapt to a human model with a thermal preference
different to that of our average human model (Model $H_A$). Here, we define Model $H_B$ as an agent with different
activity-specific TH preferences compared to Model $H_A$, but with the same reward functions defined by Eq. 12. The
metabolic values within the PMV for the three activities of Model $H_B$ are set to [1.10, 1.35, 1.75] based on [2],
which also changes the TH preference as the activity-specific metabolism rates are associated with different TH values
for optimal comfort.

In Figure 5(a), we can see that both human models $H_A$ and $H_B$, once trained, converge to the expected behaviors by
completing all 3 activities, and spending a few time-steps tuning the TH at the beginning of each activity. Then, Figure
5(b) shows that once the SHS is integrated into the environment, the amount of time spent by the human model in
adjusting the TH is reduced, without negatively affecting the completion of activities by the human agent. As expected,
with the SHS, the human model takes fewer time-steps to correct the TH. Figure 5(a) and (b) show that for the human
agent with the SHS (in green), the time-steps are reduced as compared to the human model without the SHS (in red)
for both Models $H_A$ and $H_B$. This concludes that the SHS can learn to adapt to Model $H_B$ with a different thermal
preference than that of $H_A$. This works well for the TH preferences of an average model $H_A$ as well as for $H_B$ with
different preferences, showing that the SHS model can adapt to both generic and specific human model preferences.
With Experiments 1 and 2, we can conclude that the human models in the RL-based smart home can successfully
complete the activities without abruptly leaving any of them while requiring fewer time-steps to set TH as shown in
Figure 6.

4.1.3 Experiment 3: Demonstration

In this experiment, we look into the prospect of the SHS learning the preference of a human that has a different intrinsic
thermal reward structure than the ones in Models $H_A$ and $H_B$. Here we introduce two human models namely Model
$H_C$ and $H_D$, where both models differ from $H_A$ and $H_B$ by having an extra reward term for leaving uncomfortable
thermal conditions [7]. We define the reward structures of Models $H_C$ and $H_D$ by Eq. 13.

The metabolism indices of Model $H_C$ are set to [1.80, 1.35, 1.15] for activity 0, activity 1, and activity 2 respectively
which is moderately different from the baseline (Model $H_A$). Similarly, for Model $H_D$, we set it to [1.15, 1.25, 1.85]
for each activity respectively which is similar to the baseline model. This changes the TH preference for Model $H_D$
during different activities. We are here considering the case where human models with different reward function might
have similar or different thermal preferences than the baseline model. This helps to learn about the factors that may
possibly impact human model’s behavior in the presence of SHS. The comfortable PMV range for both models is
between -0.25 to 0.25 ($d_{0.25}$, Eq. 14) to simulate a more constrictive range of PMV for our human agents.

Figure 5(c) (in blue) shows the experiment results where Model $H_C$ converges to similar behaviors as Models $H_A$
and $H_B$ without the smart home. However, with the smart home, Model $H_C$ deviates from the expected behavior
(which is completing the activities without abruptly leaving any of them, as well as setting TH with reduced time-steps).
Figure 5(c) (in green) shows that with the SHS, Model $H_C$ takes more time to set the TH than without the SHS (in red).
While trying to complete the tasks, we also observe that $H_C$ occasionally leaves the activities before completing them
(Figure 5(d)).

To study this deviation in behavior, we look into the $Q$ values of $H_C$ as shown in Figure 8 and Table I. We observe
that the overall reward available in external subroutines ($Q_e$) increases, making $H_C$ seek rewards from other activities
instead of continuing the current activity. Parallel to this, the expected reward to continue within the same subroutine
($Q_c$) decreases in the presence of SHS indicating less enthusiasm to continue and more eagerness to exit to gain rewards.
from other activities. From Table 1, we can observe a decrease in the mean $Q_c$ and an increase in the mean $Q_e$ for each activity with the SHS when compared to without the SHS, indicating the availability of higher rewards from external activities (subroutines). For any state, if the $Q_e$ value is higher than the $Q_c$ value, Model $H_C$ decides to leave the current activity to maximize the rewards by pursuing external activities. Model $H_C$ switches to other activities between subtasks since it involves no penalty. Subtasks are small achievements within an overall activity as shown in Figure 4. The time-step where our penalty is zero marks the completion of the subtask within an activity. Moreover, activity switching, loss in activity performance, and loss in effort (similar to $Q_c$) to continue with incorrect temperature and humidity were also observed in [26].

With the occasional switching between activities, the SHS fails to learn the TH preference of Model $H_C$ during a particular activity which inherently increases the overall time-steps needed to learn the TH preference of the human model. After evaluating Model $H_C$ (see Table 3), we can see that for 50 trials, the mean reward for Model $H_C$ has
Table 1: Mean $Q_c$ and $Q_e$ values of Model $H_C$ for each activity with and without the SHS.

| Activity | Without SHS | With SHS |
|----------|-------------|----------|
|          | $Q_c$ | $Q_e$ | $Q_c$ | $Q_e$ |
| Activity 0 | 0.119 | 0.0003 | 0.034 | 0.018 |
| Activity 1 | 0.063 | 0.0002 | 0.028 | 0.012 |
| Activity 2 | 0.117 | 0.0005 | 0.045 | 0.014 |

Table 2: Mean $Q_c$ and $Q_e$ values of Model $H_D$ for each activity with and without the SHS.

| Activity | Without SHS | With SHS |
|----------|-------------|----------|
|          | $Q_c$ | $Q_e$ | $Q_c$ | $Q_e$ |
| Activity 0 | 0.1298 | 6.34e-05 | 0.0304 | 0.0184 |
| Activity 1 | 0.0467 | 7.57e-05 | 0.0234 | 0.0161 |
| Activity 2 | 0.1642 | 5.50e-05 | 0.0186 | 0.0186 |

decreased from 264 to 197 showing a decrease in the performance of SHS to learn the preference of Model $H_C$ with different intrinsic rewards and metabolism indices. Meanwhile, the standard deviation increased from 3.3 to 18.2, similar to our observation in Figure 5(c), which may suggest additional instability in the overall system.

Next, we compare Model $H_C$ with Model $H_D$ that has similar metabolism indices for the activities as the baseline model $H_A$. Figure 5(e) shows that without the SHS, Model $H_D$ is able to complete each task and set optimal TH settings. Figure 5(f) shows the progress of Model $H_D$ successfully completing each activity in the presence of SHS with occasional switching. Figure 6(d) shows the change in the PMV while completing all activities versus the required time-steps. As can be seen, with the SHS (green plot), the number of time-steps required to set the TH is reduced only by a small margin for Model $H_D$ indicating that a different reward function impacts the performance of the human model as well as the SHS. With the SHS, for $H_D$, we observe a decrease in the expected $Q$-value to continue the activity ($Q_c$) and an increase in the expected $Q$-value available from external activities ($Q_e$) (see Table 2). Evaluations of Model $H_D$ (see Table 3) shows that with the SHS, the mean reward increase while the time-steps decrease.

With the above experiments, we arrive at the following conclusions.

- Our smart home model can successfully learn to anticipate thermal preference of a given model (i.e., $H_A$), which does not show an anomaly in the presence of smart home. When integrated with the SHS, $H_A$ spends fewer time-steps to change the TH without any unintended behaviour like excessive switching between activities.
- When a human model with a slightly different thermal preference ($H_B$) is introduced in the environment, the SHS can learn and adapt according to its preference, thus also managing to reduce the number of time-steps needed to changed the TH.
- When a model with a different reward function and different TH preference ($H_C$) is integrated into the environment, it exhibits an unexpected behavior by frequently switching between activities, resulting in an increase in time-steps.
- When a model with a different reward function but similar TH preference ($H_D$) compared to the baseline model is integrated into the SHS, frequent switching between activities is also observed, and the time-steps to set the TH are relatively high.

4.2 SHS with Multiple Human models in the Environment

In the previous experiments, we explored the impacts of SHS on human performance in an environment with a single human agent. To further extend the above experiments and explore the behavioral performance of multi-human models with and RL-based SHS, we also consider a case of two human models pursuing their activities in the same environment. Our aim with these experiments is to explore whether the SHS can learn the preferences of (a) two human models with similar thermal preferences; (b) two human models with different intrinsic reward and thermal preferences; (c) two human models with same reward function but different from (a); and (d) two human models with different intrinsic rewards but similar thermal preferences.

We first train the the two HRL-based human models such that each can pursue its activities and set TH to obtain optimal comfort within the same environment. Then we integrate the SHS with the trained human models and train it until convergence. It is important to note that the order of activity execution is not fixed for these simulations, which means that each model is free to start with an activity that has a maximum $Q$ value.
Figure 9: Experiment 4 Top Row: Model $\mathcal{H}_A$. Bottom Row: Model $\mathcal{H}_B$. Activity execution and its corresponding reward (in green). Activity# is the activity that the human agent is performing (in blue). Set#: Changing TH for activity # (in blue).

Figure 10: Experiment 5 Top Row: Model $\mathcal{H}_A$. Bottom Row: Model $\mathcal{H}_C$. Activity execution and its corresponding reward (in green). Activity# is the activity that the human agent is performing (in blue). Set#: Changing TH for activity # (in blue).
Table 3: Mean Time-Steps (MTS) required to change TH, Mean Reward (MR) ± standard deviation for each model with and without SHS.

| Model   | MTS  | MR     | With SHS          | MTS  | MR    |
|---------|------|--------|-------------------|------|-------|
| Model $H_A$ | 21   | 253±3.2| 11  | 260±5.6 |
| Model $H_B$ | 13   | 255±4.3| 08  | 256±9.7 |
| Model $H_C$ | 26   | 264±3.3| 36  | 197±18.2|
| Model $H_D$ | 22   | 255±5.6| 26  | 201±12.3|

Table 4: Example of Possible Temperature and Humidity combination in PMV range [-0.25, 0.25] for Model $H_C$ with metabolism rate of 1.0.

| Temperature °C | Humidity % |
|---------------|------------|
| 25            | 65         |
| 25            | 70         |
| 26            | 30         |
| 26            | 35         |
| 26            | 40         |
| 26            | 45         |
| 26            | 50         |
| 26            | 55         |
| 26            | 60         |
| 26            | 65         |
| 26            | 70         |
| 27            | 30         |

4.2.1 Experiment 4: $H_A$ & $H_B$

In this experiment, we consider a scenario of having two human models with similar thermal preferences pursuing their activities at the same time and an SHS trying to learn their preferences. Here we integrate Models $H_A$ and $H_B$ simultaneously into the SHS. Both models have a similar thermal preference for the same activities along with the same intrinsic reward function. Similar to Experiment 1, we keep the metabolic indices for Model $H_A$ as [1.0, 1.3, 1.8] for activity 0, activity 1, and activity 2, respectively, and [1.10, 1.35, 1.75] for Model $H_B$ for the same sequence of activities. The reward function for both models can be define by Eq. 12.

In Figures 9(a) and (c), after convergence, both $H_A$ and $H_B$ are able to complete each activity without frequent switching (in blue). Both agents set the TH before starting their activities (Set# in Figure 9) similar to what we might expect a human would do in a real-life scenario. The activity-reward plot in Figures 9(b) and (d) show that in the presence of the SHS, both models can still manage to pursue activities without any change in the activity sequence and with reduced time-steps required to change TH preference as shown in Figure 16. The models pursue the activity in which the metabolism index difference among the two is at a minimum, so that the SHS does not have to learn to anticipate their different TH preference. Without the SHS, Model $H_A$ took a mean of 6 time-steps to set the TH and a mean of 3 time-steps with the SHS. Similarly, without SHS, Model $H_B$ took a mean of 11 time-steps, while this was reduced to 5 time-steps with the SHS. Here, we can observe that Model $H_B$ gives more feedback (11 times) while Model $H_A$ gives less feedback (6 times). The SHS also receives the activity that the model is pursuing while changing the TH thus it becomes more biased towards Model $H_B$ than Model $H_A$ as it receives more feedback from $H_B$ which results in a large difference between the time-steps with and without the SHS for $H_B$ when compared with $H_A$. With this experiment, we observe that the SHS can become biased towards the human model which provides more feedback than the other. In the end, we can conclude this experiment as follows: (1) Both Models $H_A$ and $H_B$ can pursue their activities together, (2) the SHS can learn each model’s TH preference showing bias towards the model that gives more feedback, and (3) the performance of the two human models is generally not impacted when integrated with the SHS.

4.2.2 Experiment 5: $H_A$ & $H_C$

In this experiment, we look into the scenario of where a baseline model and model with different intrinsic reward function pursue activities simultaneously in the same environment. Here our aim is to learn whether the SHS can learn the comfort preference for each model without changing their behavior. To do so, we utilize Model $H_C$ which was introduced in Experiment 3, along with Model $H_A$. The reward functions for Model $H_A$ and $H_C$ is given by Eq. 12 and Eq. 13 respectively.
Activity 0 with and without the SHS. When comparing the metabolism rate of their activities at any time-step, we find that well with as the SHS failing to learn its comfort policy. Interestingly, here we observe that Model H well spends a bit more time setting the TH at the beginning of the activity. With the SHS, both models converge without any anomalies. A possible explanation is that in this experiment there is a second human model (H), working well with the SHS, that may result in a more optimal behaviour. A deeper analysis reveals that having the human models share the same environment is what accounts for the largest difference between the two models need to change the TH settings as when either of the models change the TH to their preference. However, together (experiment 5), they only need 16 time-steps (7 and 9 respectively). When adding the SHS, the number of time-steps required to adjust the TH is down to 12 (6 for each model). Therefore, it seems that the absence of anomalies in behaviour of Model H showed behavior anomalies due to the different intrinsic reward function with respect to the baseline model, as well as the SHS failing to learn its comfort policy. Interestingly, here we observe that Model H pursues activities in parallel to Model H without any anomalies. A possible explanation is that in this experiment there is a second human model (H), working well with the SHS, that may result in a more optimal behaviour. A deeper analysis reveals that having the human models share the same environment is what accounts for the largest difference between the two experiments (3 and 5). In the single human experiment, Models H, H, and H need 21 and 26 time-steps respectively (on average) to adjust the TH to their preference. However, together (experiment 5), they only need 16 time-steps (7 and 9 respectively). When adding the SHS, the number of time-steps required to adjust the TH is down to 12 (6 for each model). Therefore, it seems that the absence of anomalies in behaviour of H is mostly due to the fact that it interacts well with H, despite their different preferences.

By looking at Figures 10(a) and (c), we can observe that the human models use a different activity order from each other which allows them to synchronize their activities with matching TH preferences. This can be seen in both cases, with and without the SHS. When comparing the metabolism rate of their activities at any time-step, we find that the difference between their indices is minimum. Accordingly, when H pursues Activity 0, H pursues Activity 2. Similarly when H pursues Activity 1, H pursues Activity 1. And finally when H pursues Activity 2, H pursues Activity 0. Interestingly, the same behavior can be observed when both models are integrated with the SHS. Having minimal difference in the metabolism rate results in similar TH preferences for their activities. This in turn reduces the required time-steps by the models need to change the TH settings as when either of the models change the TH settings towards their preference, it also advances the other agent towards the state with optimal comfort.

To further assess that it is the positive interaction between the two human models that eliminates the behavioural anomalies of Model H, visible in Experiment 3, we rerun Experiment 5 with a more restrictive preference range (d0.25
Figure 12: Experiment 5 Model $\mathcal{H}_C$. Expected Q value to complete (in blue) or exit (in orange) the task. In comparison to Experiment 3, with the SHS, the expected Q value to complete the subroutine is higher for $\mathcal{H}_C$ when engaged with $\mathcal{H}_A$.

Figure 13: Experiment 5: Similar PMV trajectory of Model $\mathcal{H}_A$ and Model $\mathcal{H}_C$ with the SHS performing activities in parallel and switching between activities with similar thermal comfort to gain maximum rewards. Because of switching between activities having similar thermal comfort, changes in TH made by $\mathcal{H}_A$ benefits $\mathcal{H}_C$ as well.
Figure 14: Experiment 6 PMV Range [-0.25, 0.25] Top Row: Model $\mathcal{H}_C$, Bottom Row: Model $\mathcal{H}_C'$. Activity execution and its corresponding reward (in green). Activity# is the activity that the human agent is performing (in blue). Set#: Changing TH for activity # (in blue).

Figure 15: Experiment 7 Top Row: Model $\mathcal{H}_A$, Bottom Row: Model $\mathcal{H}_D$. Activity execution and its corresponding reward (in green). Activity# is the activity that the human agent is performing (in blue). Set#: Changing TH for activity # (in blue).
Figure 16: Experiment 4: Model $H_A$ and Model $H_B$ PMV trajectory while completing all activities with and without the SHS.

Figure 17: Experiment 5: Model $H_A$ and Model $H_C$ PMV trajectory while completing all activities with and without the SHS.

instead of $d_{0.50}$) for both models. The result is shown in Figure 11. In this case, we can see that without the SHS, both models show more frequent task switching than in Figure 10 with Model $H_C$ being more affected. When the SHS is added, model $H_C$ switches tasks even more frequently, as found in Experiment 3.

4.2.3 Experiment 6: $H_C$ & $H_C'$

In this experiment, we hypothesize a scenario where two simulated thermally sensitive models pursue activities within the same environment with the SHS at the same time. The intrinsic reward function for both models can be defined as Eq. 13 which is different from the reward function of baseline models. Here we include Model $H_C$ (from experiment 3) and Model $H_C'$. Model $H_C'$ is similar to $H_C$, except that it has a slightly different thermal preference for each activity. For Model $H_C$ the metabolism indices in $[1.80, 1.35, 1.15]$, while for Model Model $H_C'$, it is $[1.75, 1.30, 1.15]$. We represent the thermal sensitivity by setting the PMV range between $-0.25$ and $0.25$ which is rewarded by $d_{0.25}$. In this range, the number of TH combinations for optimal comfort is 12, thus providing more options (as shown in Table 4) where both human agents can feel comfortable. To obtain these combinations, we used the thermal comfort tool [40] to get the TH combinations for the above mentioned PMV range, empirically the thermal step is set to 1 degree, and the humidity step is set to 5%. Narrowing down the PMV range would also reduce the combinations of comfortable TH making it more difficult for the agents to have a common preference among the TH combinations. With this experiment, we aim to learn (1) if both thermally sensitive models can learn to pursue their activities and also set TH, and (2) if the SHS can learn the comfort policy for both models without any anomaly for the human agents.

In Figures 14(a) and (b), we observe that Model $H_C$ converges to an optimal policy to complete the activities without frequent switching. Then in Figures 14(c) and (d), we observe that Model $H_C'$ also converges to an optimal policy to complete the activities without frequent switching. The stability of both sensitive models is due to the high number of comfortable TH combinations and similar intrinsic reward function (Eq. 13). From Figure 18, we observe that in the PMV trajectory for $H_C$ and $H_C'$ the number of time-steps required to set the TH is reduced. Table 7 presents the mean time-steps (MTS), mean reward (MR), and its standard deviation, with and without the SHS. From Figure 14(b) and (d) we observe that two that two thermally sensitive human models $H_C$ and $H_C'$ can remain stable with the SHS without showing any behavior anomaly like frequent switching between activities and increase in time-steps to set TH.
4.2.4 Experiment 7: $\mathcal{H}_A$ & $\mathcal{H}_D$

In this experiment, we simulate a scenario where models have different intrinsic reward functions but similar thermal preferences. We use Model $\mathcal{H}_D$ from Experiment 3 whose intrinsic reward function can be expressed by Eq. [13]. In Experiment 3, we observed behavior anomalies with Model $\mathcal{H}_D$ when integrated with the SHS as shown in Figure 6 and Table 3. Thus, with this simulated experiment, we aim to learn whether two human models can display behavioral anomalies where the models have different reward functions but similar thermal preference. To keep similar thermal preferences, we kept the metabolism indices for Model $\mathcal{H}_A$ and Model $\mathcal{H}_D$ as $[1.0, 1.3, 1.8]$ and $[1.15, 1.25, 1.85]$ respectively.

Figures 15(a) and (c) show the activity plots for $\mathcal{H}_A$ and $\mathcal{H}_D$ without the SHS (in blue). No occasional switching was observed because the comfortable TH options for $\mathcal{H}_D$ is a subset of the TH options for $\mathcal{H}_A$ as the thermal preference is similar resulting in pursuing same activities. From Table 8, we observe that $\mathcal{H}_D$ gives more feedback (more time-steps) to the SHS, making the SHS biased towards the model.

The activity performance for the two models is presented in Figures 15(b) and (d) (in blue), where we observe that neither model encounters frequent switching, with or without the SHS. Even though the intrinsic reward function of the two models is different, the comfortable PMV range for Model $\mathcal{H}_D$ ($-0.25$ to $0.25$) lies within the PMV range of Model $\mathcal{H}_A$ ($-0.5$ to $0.5$), as well as having a similar thermal preference. Moreover, having a similar metabolism rate for the same activities results in similar thermal preference. Thus, attaining the comfortable TH state for Model $\mathcal{H}_D$ (the more sensitive model because of the constrictive PMV range) would also achieve a comfortable state for Model $\mathcal{H}_A$. With two human agents in the same environment, the number of time-steps for each model is reduced. This can be compared from Table 5 and Table 8 where for Model $\mathcal{H}_D$ as a single agent takes 26 time-steps with the SHS to set TH along with occasional switching, whereas with two-human agent, Model $\mathcal{H}_D$ only takes 5 time-steps. Therefore when human models with similar thermal preference are integrated with the SHS, the SHS doesn’t have to learn a different comfort policy for each model, thus the SHS’s learned policy that provides comfort is similar for both human models. Although, bias towards the model that gives more feedback can also observed. Figure 19 illustrates the improvements in $\mathcal{H}_A$ and $\mathcal{H}_D$ with the SHS, where the trajectory for $\mathcal{H}_D$ seems to have better improvement as compared to $\mathcal{H}_A$ due to $\mathcal{H}_D$ providing more feedback. The slope of the PMV curve improves with the SHS versus without it, achieving the
comfortable TH within fewer time-steps. We can also observe from Figure [15] (in blue) that with the SHS, both models mostly switch to the same activities, thus having similar metabolisms most of the time to reduce the thermal difference and avoid time spent to change the TH.

5 Conclusion

We investigate the interaction between simulated human models and smart homes in the context of setting thermal preferences. To do so, we simulate human models based on hierarchical reinforcement learning capable of continuing and leaving a set of predefined activities and also setting thermal preferences. Our smart home is based on Q-learning and learns the preferences of the human model for each activity. While the SHS could reduce the time spent by the human model toward setting the thermal preferences, we observe that human models with reward functions different than the one our smart home is trained with, show behavior anomalies such as frequent switching and an increase in the time required to set the thermal preferences. Similar results were found when two human models were sharing the same environment. Depending on the two selected human models, these models could see improved performance together and with the SHS. But again, a model different than the one for which the SHS was designed for could lead to frequent activity switching and frequent change to the temperature and humidity. Our study highlights the importance of validating and testing smart-home systems against a wide variety of end-users to ensure that the pre-trained smart home systems and products perform as desired without negative impacts on the users.

References

[1] D. Andre and S. J. Russell. State abstraction for programmable reinforcement learning agents. In AAAI/IAAI, pages 119–125, 2002.
[2] S. ASHRAE et al. Standard 55-2010, thermal environmental conditions for human occupancy. American Society of Heating, Refrigerating and Air Conditioning Engineers, 2010.
[3] F. Auffenberg, S. Stein, and A. Rogers. A personalised thermal comfort model using a bayesian network. International Joint Conference on Artificial Intelligence, 2015.
[4] S. Baghaee and I. Ulusoy. User comfort and energy efficiency in hvac systems by q-learning. In 2018 26th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2018.
[5] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528, 2019.

[6] M. Botvinick and A. Weinstein. Model-based hierarchical reinforcement learning and human action control. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1655):20130480, 2014.

[7] M. M. Botvinick. Hierarchical reinforcement learning and decision making. Current Opinion in Neurobiology, 22(6):956–962, 2012.

[8] M. M. Botvinick, Y. Niv, and A. G. Barto. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition, 113(3):262–280, 2009.

[9] E. S. Bromberg-Martin, M. Matsumoto, and O. Hikosaka. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron, 68(5):815–834, 2010.

[10] P. Chrabaszcz, I. Loshchilov, and F. Hutter. Back to basics: Benchmarking canonical evolution strategies for playing atari. arXiv preprint arXiv:1802.08842, 2018.

[11] A. Dahlgren, G. Kecklund, and T. Åkerstedt. Different levels of work-related stress and the effects on sleep, fatigue and cortisol. Scandinavian journal of work, environment & health, pages 277–285, 2005.

[12] S. DePasque and E. Tricomi. Effects of intrinsic motivation on feedback processing during learning. NeuroImage, 119:175–186, 2015.

[13] T. G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. Journal of artificial intelligence research, 13:227–303, 2000.

[14] C. Diuk, A. Schapiro, N. Córdova, J. Ribas-Fernandes, Y. Niv, and M. Botvinick. Divide and conquer: hierarchical reinforcement learning and task decomposition in humans. In Computational and robotic models of the hierarchical organization of behavior, pages 271–291. Springer, 2013.

[15] M. K. Eckstein and A. G. Collins. Computational evidence for hierarchically structured reinforcement learning in humans. Proceedings of the National Academy of Sciences, 117(47):29381–29389, 2020.

[16] G. Gao, J. Li, and Y. Wen. Energy-efficient thermal comfort control in smart buildings via deep reinforcement learning. arXiv preprint arXiv:1901.04693, 2019.

[17] C. Gebhardt, A. Oulasvirta, and O. Hilliges. Hierarchical reinforcement learning as a model of human task interleaving. arXiv preprint arXiv:2001.02122, 2020.

[18] P. W. Glimcher. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences, 108(Supplement 3):15647–15654, 2011.

[19] S. Heo, K. Nam, J. Loy-Benitez, Q. Li, S. Lee, and C. Yoo. A deep reinforcement learning-based autonomous ventilation control system for smart indoor air quality management in a subway station. Energy and Buildings, 202:109440, 2019.

[20] M. B. Herlambang, F. Cnossen, and N. A. Taatgen. The effects of intrinsic motivation on mental fatigue. Plos one, 16(1):e0243754, 2021.

[21] Z. Hu, S. Cong, T. Song, K. Bian, and L. Song. Airscope: Mobile robots-assisted cooperative indoor air quality sensing by distributed deep reinforcement learning. IEEE Internet of Things Journal, 7(9):9189–9200, 2020.

[22] A. Juliani, V.-P. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange. Unity: A general platform for intelligent agents. arXiv preprint arXiv:1809.02627, 2018.

[23] D. B. Kaber and M. R. Endsley. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. Theoretical Issues in Ergonomics Science, 5(2):113–153, 2004.

[24] R. Keiflin and P. Janak. Dopamine prediction errors in reward learning and addiction: From theory to neural circuitry. Neuron, 88(2):247–263, 2015.

[25] A. Khalili, C. Wu, and H. Aghajan. Autonomous learning of user’s preference of music and light services in smart home applications. In Behavior Monitoring and Interpretation Workshop at German AI Conf, page 12, 2009.

[26] L. Lan, P. Wargocki, D. P. Wyon, and Z. Lian. Effects of thermal discomfort in an office on perceived air quality, sbs symptoms, physiological responses, and human performance. Indoor air, 21(5):376–390, 2011.

[27] M. G. Lawrence. The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. Bulletin of the American Meteorological Society, 86(2):225–234, 2005.

[28] J. Li, A. H. Miller, S. Chopra, M. Ranzato, and J. Weston. Dialogue learning with human-in-the-loop. arXiv preprint arXiv:1611.09823, 2016.

[29] T. Liang, B. Zeng, J. Liu, L. Ye, and C. Zou. An unsupervised user behavior prediction algorithm based on machine learning and neural network for smart home. IEEE Access, 6:49237–49247, 2018.

[30] P. R. Montague, S. E. Hyman, and J. D. Cohen. Computational roles for dopamine in behavioural control. Nature, 431(7010):760–767, 2004.

[31] A. J. Oswald, E. Proto, and D. Sgroi. Happiness and productivity. Journal of Labor Economics, 33(4):789–822, 2015.
[32] J. P. O’Doherty, S. W. Lee, and D. McNamee. The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1:94–100, 2015.

[33] C. T. O’Sullivan. Newton’s law of cooling—a critical assessment. *American Journal of Physics*, 58(10):956–960, 1990.

[34] J. Y. Park, T. Dougherty, H. Fritz, and Z. Nagy. Lightlearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. *Building and Environment*, 147:397–414, 2019.

[35] W. Saunders, G. Sastry, A. Stuhlmueller, and O. Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

[36] D. W. Schneider and G. D. Logan. Hierarchical control of cognitive processes: switching tasks in sequences. *Journal of Experimental Psychology: General*, 135(4):623, 2006.

[37] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[38] J. Shin and A. M. Grant. Bored by interest: How intrinsic motivation in one task can reduce performance on other tasks. *Academy of Management Journal*, 62(2):415–436, 2019.

[39] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[40] F. Tartarini, S. Schiavon, T. Cheung, and T. Hoyt. Cbe thermal comfort tool: Online tool for thermal comfort calculations and visualizations. *SoftwareX*, 12:100563, 2020.

[41] K. H. Teigen. Yerkes-dodson: A law for all seasons. *Theory & Psychology*, 4(4):525–547, 1994.

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

[43] J. Wang, C. Sun, S. Li, J. Wang, L. Si, M. Zhang, X. Liu, and G. Zhou. Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning. *arXiv preprint arXiv:1910.09260*, 2019.

[44] C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[45] C. Wickens and J. McCarley. Executive control: Attention switching, interruptions, and task management. *Applied attention theory*, pages 145–160, 2008.

[46] C. D. Wickens, R. S. Gutzwiller, and A. Santamaria. Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human-Computer Studies*, 79:79–84, 2015.

[47] C. D. Wickens and J. S. McCarley. *Applied attention theory*. CRC press, 2019.

[48] Y. Yang, J. Hao, Y. Zheng, and C. Yu. Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework. In *IJCAI*, pages 630–636, 2019.

[49] R. Zhang, F. Torabi, L. Guan, D. H. Ballard, and P. Stone. Leveraging human guidance for deep reinforcement learning tasks. *International Joint Conference on Artificial Intelligence*, 2019.