Potential Impacts of Smart Homes with Reinforcement Learning on Human Behavior

Shashi Suman[1], **Ali Etemad**[1] and **Francois Rivest**[2]

[1]Department of Electrical and Computer Engineering, Queen's University
[2]Department of Mathematics and Computer Science, Royal Military College

Introduction

Smart Home Systems (SHS) have drastically evolved in recent years due to advances in AI, enhanced connectivity, and affordability. On the other hand, the increased use of AI and automation in our daily lives has begun to impact human performance.
Reinforcement learning (RL) has been widely used to enable agents to learn from feedback. While RL-based agents generally learn to optimize their performance for maximizing received rewards, such reward maximization may only be optimal from the perspective of the agent without solving the environment. a smart agent can exploit the environment to gain significant rewards by exploiting an edge-case in the rules of the environment.

Goal and Objectives

Goal: To demonstrate that a discrepancy in the intrinsic reward of the human model used to devise a Reinforcement Learning-based smart home system could lead to undesirable human behavior.

Objective: Simulate a Smart Home model capable to learn human thermal comfort preference. Simulate a Human model capable to pursue or leave activity based on internal and external rewards while also setting thermal parameters to attain comfort.

Human Thermal Model

- Thermal comfort is defined by human satisfaction for a particular temperature, humidity of its respective environment.
- Subjective evaluation of thermal comfort can be calculated to assign a numerical value as defined by American Society of Heating, Refrigerating and Air-Conditioning Engineers 2017
- Thermal comfort score defined as *Predicted Mean Vote* (PMV)
- Ranges from -3 to 3. -3 being too cold, 3 being too hot, 0 being optimal comfort as in Table 1.
- Factors assumed for this paper is temperature and humidity as these parameters are majorly responsible for human comfort.

Scale	Condition
+3	Too Hot
+2	Warm
+1	Slightly Warm
0	Neutral
-1	Slightly Cool
-2	Cool
-3	Too Cold

Table. 1: Scale of PMV values with comfort feeling.

Environment Models

Smart Home System
The SHS assumes a Markov Decision Process (MDP). The optimal policy for the SHS is found using Q-Learning.

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

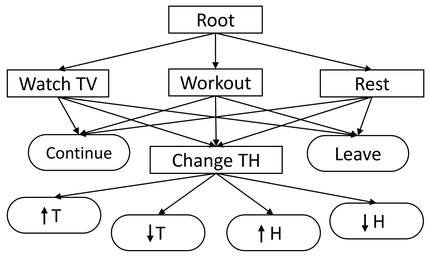

Fig. 1: Hierarchical model representation of Human model. Rectangles are SMDP subroutines and ovals are primitive actions.

Human Model
We model the human agent using MAXQ HRL [2] [3] for activities. In this formulation, the MDP is broken down into a hierarchy of Semi-Markov Decision Processes (SMDP) as shown in Figure 1. We use a three-part value function de-composition similar to [1] as given by:

$$Q(i,s,a) = Q_r(i,s,a) + Q_c(i,s,a) + Q_e(i,s,a), \tag{2}$$

Here $Q_r(i,s,a)$ is the expected discounted reward for taking action a in state s of task i, $Q_c(i,s,a)$ is the expected discounted reward for completing the rest of the subroutine i. $Q_e(i,s,a)$ refers to all the rewards external to the current subroutine of the agent i.

- Model A: An average optimal human model used by the manufacturers to train the SHS with the optimal comfort PMV indices.
- Model B: Optimal human model with a slightly different thermal preference than Model A for the activities, but with the same comfort reward functions.
- Model C: Optimal human model with different thermal preference than Model A but with different intrinsic reward function.

Result

- Both human models A and B, once trained, converge to the expected behaviors by completing all 3 activities as shown in figure 2,
- Models A and B spend a little amount of time tuning the TH at the beginning of each activity.
- With integration of SHS, the amount of time spent by the human model in adjusting the TH is reduced
- The SHS successfully adapts according to Model B with slight different thermal preference without changing the expected behavior of human model.
- Model C converges to the exact same behaviors as ModelA and B without the smart home.
- Model C's behavior becomes erratic when put in the presence of the SHS.
- More steps is required to complete the tasks, and C also switches between tasks before completing them as shown in Table 2

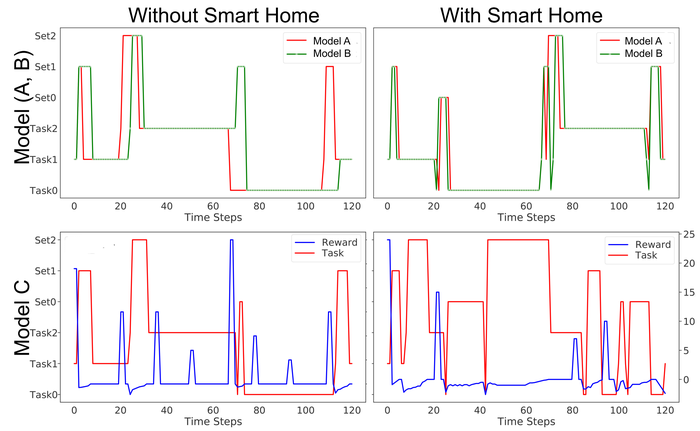

Fig. 2: Plot of activities through time for each model (A, B, C) with and without the SHS. Left column: Each model can learn to complete the tasks without interruption. Top right graph: The SHS anticipates human preferences speeding up the time for human models A and B to complete the activities. Bottom right graph: Model C (different internal reward structure) behaves erratically when integrating the SHS. *Set#* denotes the action of setting TH for *Task#*.

	Without SHS		With SHS	
	Time Steps	Reward	Time Steps	Reward
Model A	14	231.65	10	244.50
Model B	16	228.76	12	236.01
Model C	20	234.20	100	97.67

Table. 2: Number of time steps spent changing TH and average episodic reward.

References

[1] David Andre and Stuart J Russell. "State abstraction for programmable reinforcement learning agents". In: *AAAI/IAAI*. 2002, pp. 119–125.

[2] Thomas G Dietterich. "Hierarchical reinforcement learning with the MAXQ value function decomposition". In: *Journal of artificial intelligence research* 13 (2000), pp. 227–303.

[3] Christoph Gebhardt, Antti Oulasvirta, and Otmar Hilliges. "Hierarchical Reinforcement Learning as a Model of Human Task Interleaving". In: *arXiv preprint arXiv:2001.02122* (2020).