Towards Adversarially Robust Human-in-the-Loop Learning for HVAC Systems

Aditya Kumar[1] *, Hanqing Yang[2] *, Wai Shun Lee, Jet Wei Goh[1], Carlee Joe-Wong[2], Marie Siew[1]

[1]Singapore University of Technology and Design
[2]Electrical and Computer Engineering, Carnegie Mellon University.
aditya_kumar@mymail.sutd.edu.sg, hanqing3@andrew.cmu.edu, waishun_lee@alumni.sutd.edu.sg,
goh_jetwei@mymail.sutd.edu.sg, cjoewong@andrew.cmu.edu, marie_siew@sutd.edu.sg
*indicates equal contribution *

Abstract

Reducing global energy consumption is an urgent step towards slowing the pace of global warming and climate change. Many opportunities to do so lie in the building sector, which is a major energy occupant, particularly Heating, Ventilation and Air-Conditioning (HVAC) systems. Centralized air conditioning in large communal spaces (e.g. malls, offices, cinemas, libraries) often aims to cool spaces to very low temperatures, leading to significant energy consumption that may not even be necessary for users' comfort. To address this, human-in-the-loop learning (HIL), a network-enabled crowdsourced reinforcement learning (RL) framework has been proposed. This framework leverages direct thermal comfort feedback from occupants to optimize energy efficiency and thermal comfort in HVAC systems in public buildings. Nevertheless, in HIL, control systems may receive unreliable feedback from adversarial or irrational users. Therefore, in this work we work towards increasing the safety and robustness of HIL frameworks in HVAC systems. We propose **RARL_HIL**, in which a primary agent is jointly trained with an adversarial agent which aims to destabilize the system via generating 'false' feedback. The primary agent learns to operate effectively in challenging and destabilizing environments. Simulation results shows that our algorithm outperforms a traditional human in the loop RL algorithm, in unseen test environments involving adversarial or irrational user feedback.

Introduction

Human-in-the-loop learning (HIL) with crowdsourced user feedback is a network-enabled AI framework that uses human input to adjust settings in control systems. Human-in-the-loop techniques have the potential to enable control systems to make more user-friendly, equitable and explainable decisions (Mosqueira-Rey et al. 2022). At the same time, human-in-the-loop techniques can also have an edge over manual modeling methods in terms of capturing nuanced human complexities. However, HIL learning also raises significant challenges in handling possibly adversarial, irrational, or missing feedback from unreliable human users.

One application of HIL learning is in Heating, Ventilation and Air-Conditioning (HVAC) systems. HVAC systems are the largest contributor (Katili, Boukhanouf, and Wilson 2015) of energy consumption in the building sector, accounting for 36% of total global energy consumption (Santamouris and Vasilakopoulou 2021). In public buildings, it is estimated that cooling accounts for more than 50% of energy usage (Katili, Boukhanouf, and Wilson 2015). Therefore, optimizing temperature settings in HVAC systems is important for better energy efficiency. Nevertheless, in tropical regions near the equator, effective cooling systems should also ensure the thermal comfort of building occupants. Thus, it is crucial to balance energy efficiency with the thermal comfort of users (García, Prett, and Morari 1989; Kwadzogah, Zhou, and Li 2013).

Traditional techniques for controlling temperatures in HVAC systems (Wang and Ma 2008) include model predictive control (Kwadzogah, Zhou, and Li 2013) and optimization techniques (Huang et al. 2016; Mossolly, Ghali, and Ghaddar 2008). There has been a shift towards learning-based methods, in which explicit environment models are not pre-defined. Instead, the learning agents learn optimal temperature settings through interaction with the environment. Reinforcement learning (RL) is often proposed as a solution, where the agent learns a policy which determines the optimal actions given different environmental settings (Mozer 1998; Biemann et al. 2021; Gao, Li, and Wen 2019).

However, RL for HVAC comes with challenges such as *slow convergence, sample inefficiency*, and difficulties in *generalizing to unseen environments* with data distributions that differ from those encountered during training. RL-based meta-learning and transfer learning (Lissa, Schukat, and Barrett 2020) have been proposed to deal with the challenges. Human-in-the-loop learning (HIL) for HVAC temperature optimization has also been proposed (Chen, Meng, and Zhang 2023), for incorporating real-time user feedback into the RL process. This is especially useful for *public spaces* with centralized air-conditioning systems, such as malls, cinemas, libraries, or communal office spaces. Unlike individual households or offices with thermostats and temperature switches, public spaces often do not allow users to control the temperature, partly due to the presence of many users with potentially conflicting preferences. In HIL for HVAC, wireless networks play a pivotal role in enabling

*This work is supported by the SUTD Undergraduate Research Opportunities Program (UROP), and the SUTD-MOE Early Career Award under the Singapore MOE START Scheme.

communication between users, sensors, HVAC devices, and central controllers. A human-in-the-loop RL algorithm in this context aggregates **real-time crowd-sourced** occupant feedback, e.g. via apps on wireless mobile devices, and learns how to adjust the temperature in a way that balances optimizing energy savings with user comfort.

Nevertheless, for crowd-sourced human-in-the-loop algorithms to be successfully implemented in real-world HVAC systems or other cyber-physical systems in smart cities, it is crucial to work towards improving their **robustness and safety**, because humans can be unreliable or malicious. There may be adversarial users who provide random feedback to destabilize the system, or intentionally give misleading feedback not corresponding to their true preferences, to guide the system towards specific outcomes (Harris 2023). Additionally, humans are not perfectly rational and may non-maliciously and unintentionally provide feedback that does not align with their true internal preference states. Integrating such feedback introduces unreliable or adversarial data that could lead RL agents to make suboptimal decisions. At the same time, it may be difficult to detect and deal with such feedback, given that no objective ground truth exists in human-in-the-loop systems.

This paper seeks to address a critical gap by introducing **RARL_HIL**, a framework designed to enhance the robustness of human-in-the-loop systems under realistic and adversarial conditions. Our work is the first to explicitly study the implications of irrational or adversarial human feedback in HVAC control settings. **RARL_HIL** integrates robust adversarial reinforcement learning (Pinto et al. 2017) with human-in-the-loop dynamics, introducing an adversarial agent that actively generates 'false' feedback to simulate potential real-world adversarial scenarios. The primary agent learns to operate effectively in this challenging environment, developing strategies to counteract the destabilizing effects of adversarial feedback. While our focus is on HVAC optimization, the modular design of **RARL_HIL** makes it broadly applicable to other crowd-sourced human-in-the-loop systems, establishing a foundation for robust operation in adversarial environments.

Our contributions are as follows:

1. We present a novel system model for human-in-the-loop reinforcement learning-based temperature control in HVAC systems, in which we model adversarial and irrational human input.
2. To be robust towards adversarial and irrational human input, we present **RARL_HIL**, an adversarial reinforcement learning algorithm. In our algorithm, the primary and adversarial agents are jointly trained in a zero-sum framework, to improve the robustness of the primary agent in difficult environments.
3. We present experimental results which analyze the impact of adversarial and irrational user input, on a vanilla deep Q-learning network (DQN) human-in-the-loop algorithm. Finally, we show that in unseen online scenarios, the converged policy of our algorithm **RARL_HIL** is more robust than the converged policy of the vanilla DQN human-in-the-loop algorithm.

Human-in-the-loop Framework

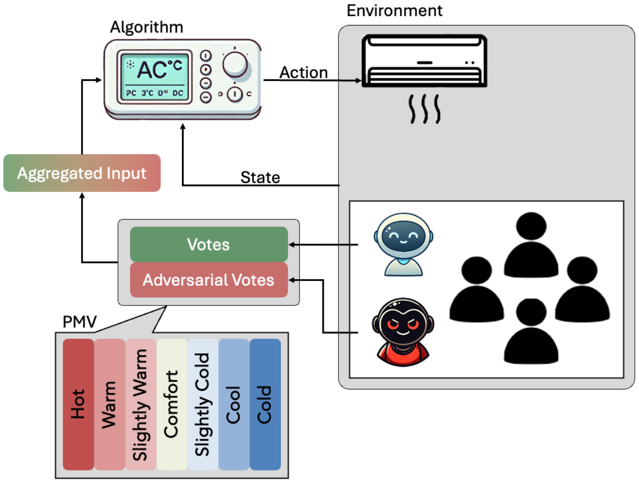


Figure 1: Human-in-the-Loop (HIL) RL for HVAC Systems. Users are allowed to provide timely feedback on thermal comfort via voting. The RARL_HIL robustly makes real-time decisions based on the environment state and human input, regardless of the input being adversarial, balancing human comfort and energy usage.

We propose a human-in-the-loop RL framework that incorporates occupant thermal comfort feedback to balance the dual objectives of user comfort and energy efficiency, as shown in Fig. 1. In this framework, building occupants provide scheduled *thermal perception* feedback (e.g. via an app on mobile devices), on the current indoor temperature of the system, which is then aggregated. The RL agent uses this aggregated feedback, combined with information about the building's thermal state, to determine the optimal temperature for the next interval. To illustrate this, we modeled a public space where occupants with diverse thermal profiles enter and exit the building at different times, and where indoor temperatures vary hourly based on the time of day.

Occupant Simulation

The simulation of occupancy level O at each time slot t is modeled based on time-of-day occupancy probabilities. The simulation dynamically generates an occupancy level O based on the time of day, with different occupancy patterns observed during specific periods. The occupancy level O at each time slot t is modeled using the following function:

$$O(t) = \begin{cases} \text{Uniform}(0.5, 0.8) & \text{for } 9 \le h < 12 \text{ or} \\ & 13 \le h < 17 \\ \text{Uniform}(0.8, 1.0) & \text{for } 12 \le h < 13 \\ \text{Uniform}(0.2, 0.5) & \text{for } 8 \le h < 9 \text{ or} \\ & 17 \le h < 18 \\ \text{Uniform}(0.0, 0.1) & \text{for other hours} \end{cases} \quad (1)$$

During normal work hours (9:00 AM to 12:00 PM and 1:00 PM to 5:00 PM), occupancy level O is randomly set between 50% and 80%. During the lunch hour (12:00 PM to 1:00 PM), occupancy level O reaches its peak, with levels ranging from 80% to 100%. In the transition hours (8:00 AM to 9:00 AM and 5:00 PM to 6:00 PM), occupancy level O is lower, ranging from 20% to 50%. During off hours, outside of typical working hours, occupancy level O are very low, ranging from 0% to 10%. All cases of the occupancy level O function follows a random uniform distribution.

Predicted Mean Vote of Occupants

The Predicted Mean Vote (PMV) is an index that predicts the average thermal sensation of individuals (ISO 2005; ASHRAE 2017). Its calculation is influenced by six factors: metabolic rate (met), clothing insulation (clo), dry bulb air temperature, mean radiant temperature, relative humidity, and relative air velocity (Dyvia and Arif 2021). Among these six features, metabolic rate (met), clothing insulation (clo) capture *inter-human* differences. In our simulations, each occupant's thermal profile is defined by the features *metabolic rate* (range: 1–1.5) and *clothing insulation index* (range: 0–1), both drawn from a random uniform distribution. Based on their thermal profile and the external environmental conditions, occupants provide feedback which is aggregated and fed into the RL algorithm, acting as a communal thermostat to guide the temperature setting process.

Towards Realistic Modeling of Human Input

Human input is often unreliable, particularly in crowdsourced applications. Unreliable or even adversarial feedback has the potential to skew the RL agent's decision-making, leading to suboptimal outcomes. In this framework, we model two distinct scenarios in which the aggregated PMV collected from human inputs may deviate from the true PMV values. In the *Experiment and Results* section, we will evaluate how our framework performs under the scenarios where a portion of users are adversarial or irrational.

Scenario 1: Adversarial Input

In this scenario, occupants intentionally provide an erroneous PMV score

$$PMV_{bias} = \begin{cases} -1.0 & \text{if } PMV_{True} > 0 \\ 1.0 & \text{otherwise} \end{cases}$$

$$PMV_{Observed} = PMV_{True} + PMV_{bias} \cdot f(t, o, h) \quad (2)$$

where $f(t, o, h)$ represents adjustments based on time of day, occupancy, and historical PMV trends. Occupants give adversarial PMV values inconsistently, with feedback becoming more extreme during peak hours or when the system attempts corrections. This behavior can skew the agent's policy, as it relies on aggregated PMV feedback to estimate the thermal profile of the population.

Scenario 2: Irrational Human Input

Building on and modifying the irrational preference models in (Ruiz–Medina and Miranda 2021), we consider a scenario where occupants choose a PMV value with a diminishing probability as it deviates from their true PMV. The probability that $PMV_{Observed} = v_i$, given PMV_{true} is given by:

$$P(v_i|PMV_{True}) = \frac{e^{-B|PMV_{True} - v_i|}}{\sum_{j=-3}^{3} e^{-B|PMV_{True} - v_j|}} \quad (3)$$

where B is an irrationality constant. A value of $B = 0$ implies fully random PMV choices, irrespective of the true PMV, while $B \to \infty$ implies occupants are almost certain to report their true PMV.

Markov Decision Process

We formulate the HVAC temperature optimization environment as a Markov Decision Process (MDP). An MDP is defined by the tuple $< \mathbf{S}, \mathbf{A}, p, r >$ where:

1. The state space $\mathbf{S}$ describes the HVAC environment (consisting of the building and its occupants). It is a three-dimensional array:
   (a) Outdoor Temperature, $T_{outdoor}$: Continuous variable representing the outdoor temperature in $°C$.
   (b) Occupancy Coefficient, $c_{occupancy}$: An index ranging from 0 to 1, representing the proportion of building occupancy at any given time.
   (c) Aggregated Predicted Mean Vote, PMV_{agg}: A continuous variable ranging from -3 to +3, capturing the average thermal comfort level of occupants, aggregated from individual PMV values.
2. The action space $\mathbf{A}$ is the set of discrete temperatures (T_{indoor}) that the agent can choose, from within the range of 22-30$°C$.
3. The transition probability function $p : S \times A \to R$, where $p(s'|s, a) = p(s_{t+1} = s'|s_t = s, a_t = a)$ represents the probability of transitioning to state $s' \in S$ given the action $a \in A$ in the state $s \in S$. The transition probability accounts for the dynamics brought about by the arrivals and departures of user with heterogeneous thermal profiles, which impact PMV_{agg} in a way which cannot be tractably modelled.
4. The reward function $r : S \times A \to R$, where $r(s, a)$ indicates the cost/reward associated with the selected temperature. The reward function is a weighted sum of the **user comfort** and **energy savings**, described below.

Our objective is to maximize the reward function:

$$\begin{aligned} \textit{Reward}(t) = w_c \textit{UserComfort}(user_1, ...user_i, ...user_N, \\ temp_t) - w_e \cdot \textit{energy}(temp_t), \end{aligned} \quad (4)$$

where w_c and w_e are weights representing the relative importance of **user comfort** and **energy savings**, respectively. Adjusting these weights allows the algorithm to shift its focus between prioritizing occupant comfort and energy efficiency. $N = |O(t)|$ is the number of users currently in the system.

UserComfort: This function is modeled after the Predicted Mean Vote (PMV) index, which rates comfort levels on a scale from -3 (very cold) to +3 (very hot). $\textit{UserComfort}(user_1, ...user_i, ...user_N, temp_t)$ is given by the average PMV, PMV_{agg} of all building occupants.

EnergyUsage: The Energy Usage function is derived from the heat transfer equation (Holman 1986):

$$\textit{energy} = \frac{mc_a \Delta T}{\textit{EER}}, \quad (5)$$

where m is the mass of the room, c_a is the specific heat capacity of dry air, ΔT is the temperature difference between the outdoor (ambient) temperature and the indoor (air-conditioned) temperature, and *EER* (Energy Efficiency Ratio) is the ratio of cooling capacity to power input.

Deep Q-learning

We input this state space into a deep Q-learning algorithm (Mnih et al. 2015). The neural network's output is the **action** taken: the air-con temperature. In Q-learning, the Q-function $Q(s, a)$ represents the value (sum of expected discounted reward) of taking action a at state s, and the algorithm's aim is to iterate until convergence at the true Q-value.

$$Q(s, a) \leftarrow (1-\alpha)Q(s, a)+\alpha(Reward(t)+\gamma \max_{a} Q(s', a)), \tag{6}$$

where α is the learning rate, and γ is the discount factor. In deep Q-learning, the neural network parameterizes the Q-function. The loss function which the neural network optimizes is

$$L = [Reward + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta)]^2. \tag{7}$$

Robust Human-in-the-Loop Learning Algorithm

In this section we present our algorithm, **RARL_HIL**, which is inspired by (Pinto et al. 2017). Fig 2 shows a visualisation of its architecture. In this reinforcement learning algorithm, we jointly train 2 agents: the main learning agent (θ_t^l), and the adversarial agent (θ_t^a). The adversarial agent's goal is to perturb the system, in the hope of minimizing the learning agent's rewards. These two agents get rewards which are of the opposite sign of each other, i.e. $r_i^{a,t} = -r_i^{a,t}$, making it a zero sum game. For the main agent, its action is the indoor temperatures T_{indoor}, as mentioned in the earlier section. The adversarial agent's actions is the extent of deviation from the true aggregated PMV_{agg}, hence disturbing the system via these 'false' aggregated PMVs. This serves as a proxy for adversarial or irrational users and helps train the main agent to be robust to real world conditions.

Our algorithm is presented in Algorithm 1. The flow of our algorithm is as follows: for T_l steps we train the learning agent θ_t^l and optimize its parameters, while keeping the parameters of the adversarial agent fixed. During this process, the environment runs, and both agents take actions according to their current policies derived from their respective Q-networks. Next, for T_a steps, we train the adversarial agent θ_t^a and optimize its parameters, while keeping the parameters of the learning agent fixed. This alternating process repeats. As our learning agent learns in an environment where our adversary distrubs the system via the aggregated PMV values, the learning agent may give less emphasis to the energy consumption. Therefore, we add a regularization term in the reward, to give additional emphasis to the energy consumption if it falls above a threshold E_{th}.

Simulation to reality: Training is performed offline. The converged algorithm will be tested in real-life environments with distrubances, in this case, adversarial or irrational users, who do not report their true PMVs.

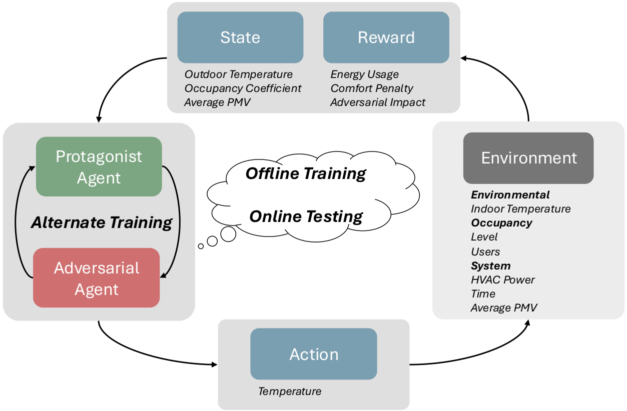
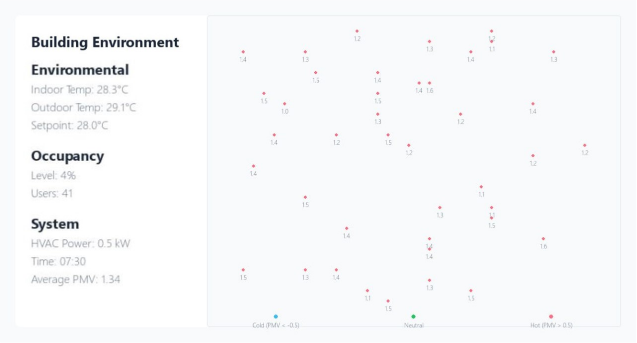


Figure 2: Architecture of RARL_HIL. The training of the *Protagonist Agent* and the *Adversarial Agent* alternates every T_l/T_a episodes. The *Protagonist Agent* aims to learn a robust policy that can withstand adversarial input, while the *Adversarial Agent* aims to disrupt the learning process of the *Protagonist Agent* during its training.

Experiment and Results

Simulation Setup An open space is modeled to mimic a single-level office building with an area of 4800 m^2 and a maximum capacity of 1000 consumers as shown in Fig 3. The outdoor environment is simulated to match the summer conditions of a tropical climate, via using a cosine function, with base temperature 28°C.



Building Environment

Environmental

Indoor Temp: 28.3°C

Outdoor Temp: 29.1°C

Setpoint: 28.0°C

Occupancy

Level: 4%

Users: 41

System

HVAC Power: 0.5 kW

Time: 07:30

Average PMV: 1.34

Figure 3: A visualisation of the simulation of a single-level office building where we will be running our experiments. Each red dot represents a person in the building at that point in time, with a label of their individual PMV.

Comparison of vanilla DQN HIL with Set-point Control For the baseline policy, we use the *Set-point Control* (fixed policy) technique of VRF systems (Kim et al. 2020), as VRF systems are often used in large buildings. Just like (Kim et al. 2020), we use a constant temperature of 26°C, which is within ASHRAE 55-2017's (ASHRAE 2017) recommended summertime thermal comfort range.

Fig. 4 shows how the learned reward value changes during the training process of a vanilla DQN HIL algorithm.

Algorithm 1: Robust Human-in-the-Loop Learning Algorithm (RARL-HIL)

```
Input: Energy threshold $E_{th}$, penalty factor $\lambda$
Initialization: Neural network parameters for the learning agent $\theta_0^l$ and the adversarial agent $\theta_0^a$
for $t = 1, 2, \ldots, T$ do
    $\theta_t^l \leftarrow \theta_{t-1}^l$
    for $i = 1, 2, \ldots, T_l$ do
        $\{(s_i^t, a_i^{l,t}, a_i^{a,t}, r_i^{l,t}, r_i^{a,t})\} \leftarrow Env_{\text{HVAC}}(\pi_{\theta_t^l}, \pi_{\theta_t^a})$
        if $energy_i > E_{th}$ then
            $r_i^{l,t} \leftarrow r_i^{l,t} - \lambda(energy_i - E_{th})/1000$
        end if
        $\theta_t^l \leftarrow OptimizePolicy(\{s_i^t, a_i^{l,t}, r_i^{l,t}\}, \theta_t^l)$
    end for
    $\theta_t^a \leftarrow \theta_{t-1}^a$
    for $i = 1, 2, \ldots, T_a$ do
        $\{(s_i^t, a_i^{l,t}, a_i^{a,t}, r_i^{l,t}, r_i^{a,t})\} \leftarrow Env_{\text{HVAC}}(\pi_{\theta_t^l}, \pi_{\theta_t^a})$
        if $energy_i > E_{th}$ then
            $r_i^{a,t} \leftarrow r_i^{a,t} - \lambda(energy_i - E_{th})/1000$
        end if
        $\theta_t^a \leftarrow OptimizePolicy(\{s_i^t, a_i^{a,t}, r_i^{a,t}\}, \theta_t^a)$
    end for
end for
```

Each episode represents a day-long simulation with rewards calculated at every 30 minutes interval. Here, there are no irrational or adversarial users. As seen in Fig. 4, vanilla DQN HIL outperforms the set point policy (fixed temperature of $26°C$) and converges after around 250 episodes.

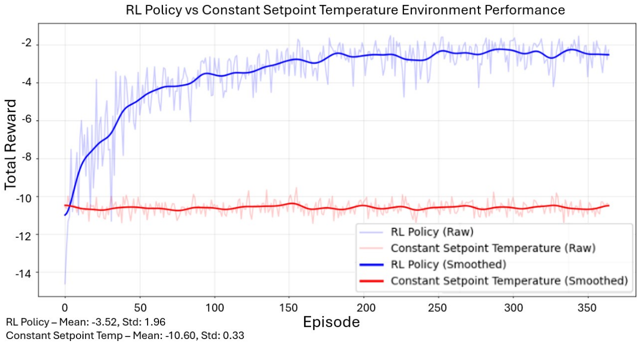


Figure 4: Cumulative reward curves for vanilla DQN HIL agent vs set-point temperature, without adversarial or irrational users.

Effect of Unreliable Human Input on a Vanilla DQN HIL algorithm.

We study the impact of adversarial and irrational input on a vanilla DQN human-in-the-loop algorithm, where the state and action space follows that mentioned in this paper.

Effect of Adversarial Input We evaluate the impact of adversarial occupant inputs on the vanilla DQN HIL model's ability to identify a policy that optimizes rewards. To this end, we train the model in the same environment as above, while varying the proportion of adversarial inputs. The comfort rewards component is computed using user-reported PMV regardless of whether they are adversarial or not, this is because the system has no knowledge of who is adversarial. We conduct this experiment for adversarial fractions of $[0.1, 0.3, 0.5]$, representing the proportion of adversarial inputs among all occupants. Each configuration is repeated 5 times, and the mean and variance of the cumulative reward are depicted in Fig. 5. As expected, an increase in the adversarial fraction leads to a corresponding decline in rewards. This decline is attributable to the distortion of aggregated PMV signals caused by a higher proportion of adversarial inputs, impairing the model's ability to make optimal temperature decisions.

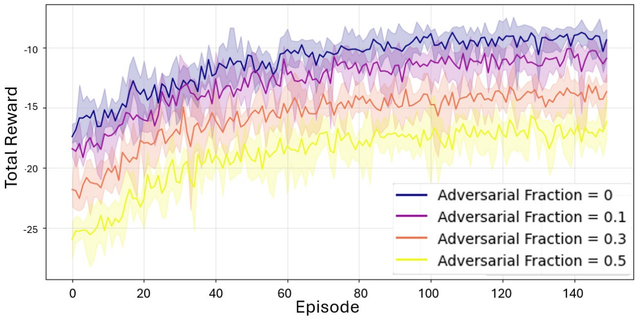


Figure 5: Cumulative reward curves for the vanilla DQN HIL agent under varying adversarial fractions.

Effect of Irrational Input Likewise, we evaluate the impact of irrational inputs on the vanilla DQN HIL human-in-the-loop algorithm. As mentioned in our system model section, a value of $B = 0$ implies fully random PMV choices, while $B \to \infty$ implies occupants are almost certain to report their true PMV. The comfort rewards component is calculated using the rewards prior to irrationality being applied so that we optimize for user's true comfort level. Fig. 6 shows a positive correlation between the irrationality constant, B, and cumulative rewards, indicating that increasing irrationality (lower B) of occupants lead to worsening performance of the vanilla DQN HIL agent in choosing the optimal temperature.

Performance of RARL_HIL vs Vanilla DQN HIL with Unreliable Human Input

As seen above, a vanilla DQN human-in-the-loop algorithm may not be able to deal with unreliable human input, especially as the human input increases in unreliability. In this section, we apply the *converged policies* of our algorithm RARL_HIL and a vanilla DQN HIL algorithm, in online scenarios where a fraction of users are adversarial or irrational.

First, we test both policies on an environment in which the the converged policy of the adversarial agent $\pi_{\theta_t^a}$ is used to make disturbances to the system. The results are presented in Fig. 7, averaged over 5 runs of experiments. As seen, our

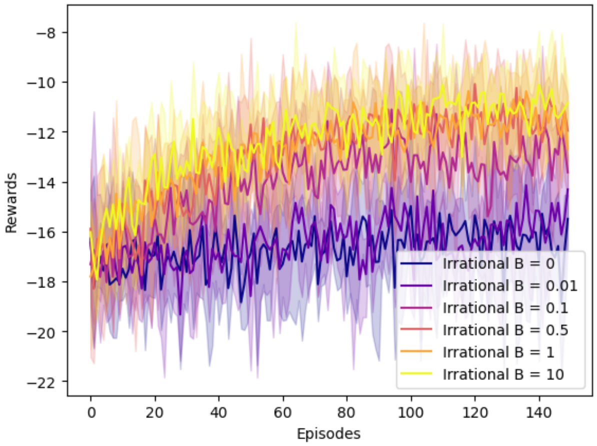


Figure 6: Cumulative reward curves for vanilla DQN HIL agent with varying irrational coefficient B.

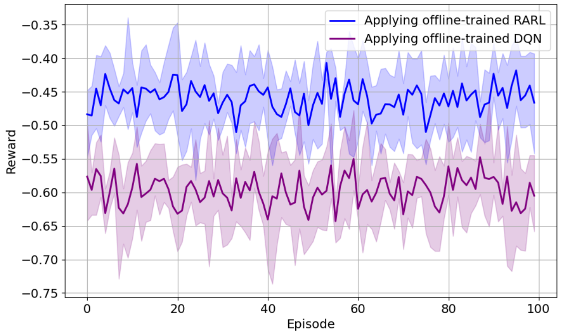


Figure 7: Comparison of rewards, when the converged policies of RARL_HIL and vanilla DQN HIL are applied, in an environment with the adversarial agent.

algorithm RARL_HIL outperforms vanilla DQN human-in-the-loop, in this adversarial environment with 'false' PMVs. Next, we will test the two converged policies in unseen online environments.

Testing in Unseen Environments: Adversarial users: We evaluate and compare the performance of the converged policies of a trained RARL_HIL and a trained vanilla DQN HIL in environments with simulated adversarial occupant input. To simulate adversarial users, we implemented a dynamic bias calculation system that adapts to both system state and user behavior. The bias combines multiple factors: a base directional bias opposite to the current PMV, a temporal trend factor based on PMV history, occupancy-based scaling, and time-of-day amplification during business hours (9:00-17:00). The system amplifies the bias by $1.5\times$ when detecting potential control countermeasures (pmv$_{direction}$ $*$ pmv$_{current}$ < 0). We systematically evaluate both models across multiple adversarial ratios (0, 0.5, 0.75 and 1.0), running three independent tests per ratio. Results presented in Fig. 8 and Table 1 demonstrate that as the proportion of adversarial users increases, our algorithm's performance advantage over the vanilla DQN baseline becomes more pronounced, achieving up to 14.8% improvement with a fully adversarial population.

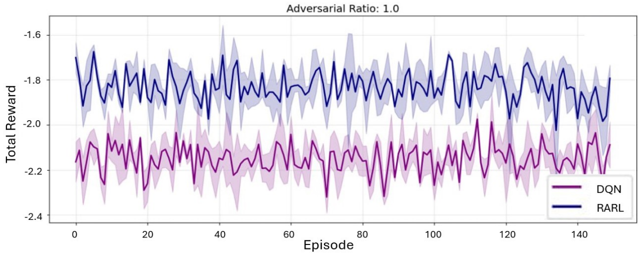


Figure 8: Comparison of rewards when the converged models of RARL_HIL and a vanilla DQN HIL agent are applied in a previously-unseen online setting with simulated adversarial user input.

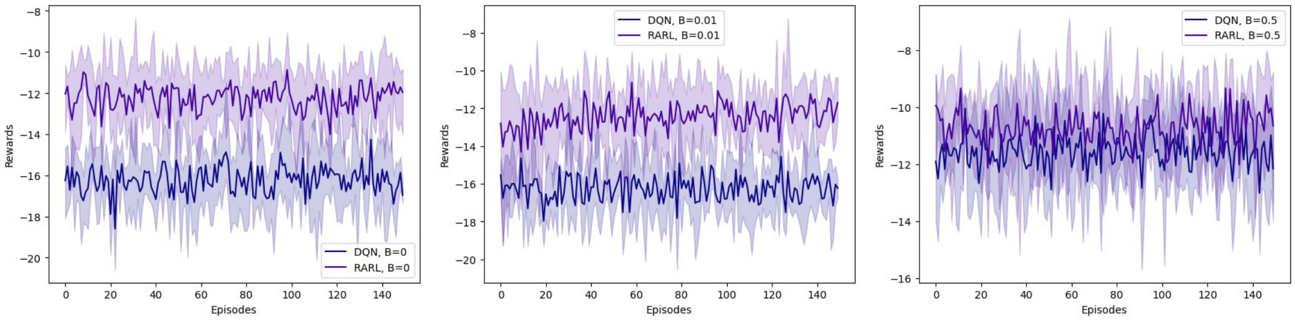
	Ratio: 0.75	Ratio: 1.0
Vanilla DQN HIL	-1.6509 ± 0.0055	-2.1502 ± 0.0079
RARL_HIL	-1.5686 ± 0.0092	-1.8323 ± 0.0066
% improvement	5.0%	14.8%

Table 1: Tabular comparison of average reward statistics, when the converged models of RARL_HIL and a vanilla DQN HIL agent are applied in a previously-unseen online setting, with varying ratio of adversarial user input.

Testing in Unseen Environments: Irrational Users. Likewise, we evaluate and compare the performance of a trained RARL_HIL and a trained vanilla DQN HIL with simulated irrational occupant input. This simulated irrational occupant input follows Eq. (3). The irrationality constant B indicates the likelihood of users reporting their true PMV. As $B \rightarrow \infty$, they are almost certain to report their true PMV; as $B \rightarrow 0$, users tend towards more and more random PMV choices. We set $B = 0.5, 0.01, 0$. Results are presented in Fig. 9. It can be seen that as the irrationality of users increases (i.e. B decreases), our algorithm outperforms the vanilla DQN baseline by a greater extent.

Conclusion

Human-in-the-loop (HIL) crowdsourced feedback is a network-enabled AI application across mobile devices, HVAC devices, and central controllers. In HVAC systems, this approach functions analogously as a thermostat for temperature control, in large public spaces with multiple users, helping to jointly optimize energy efficiency with user comfort. User feedback is fed into the reinforcement learning algorithm, which adjusts the temperature settings. For HIL



(a) B=0.

(b) B=0.01.

(c) B=0.5.

Figure 9: Comparison of rewards, when the converged models of RARL_HIL and a vanilla DQN HIL are applied in an online setting with irrational user input. The irrationality constant B is varied.

algorithms to be implemented in practice, it is crucial to improve their safety and robustness towards malicious or unreliable user input. In this work, we propose a framework, **RARL_HIL**, which enhances the robustness of HIL systems under realistic and adversarial conditions. Our framework involves jointly training a reinforcement learning agent along with an adversarial agent that seeks to disturb the system, in a zero-sum setting. Our method outperforms vanilla DQN in unseen online environments with adversarial and irrational human input. We will release this building simulator, with RARL_HIL algorithm integrated in, for open use.

References

2005. Ergonomics of the thermal environment — Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria.

ASHRAE, A. 2017. Standard 55-2017. *Thermal environmental conditions for human occupancy*.

Biemann, M.; Scheller, F.; Liu, X.; and Huang, L. 2021. Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control.

Chen, L.; Meng, F.; and Zhang, Y. 2023. Fast Human-in-the-Loop Control for HVAC Systems via Meta-Learning and Model-Based Offline Reinforcement Learning.

Dyvia, H.; and Arif, C. 2021. Analysis of thermal comfort with predicted mean vote (PMV) index using artificial neural network. In *IOP Conference Series: Earth and Environmental Science*, volume 622, 012019. IOP Publishing.

Gao, G.; Li, J.; and Wen, Y. 2019. Energy-Efficient Thermal Comfort Control in Smart Buildings via Deep Reinforcement Learning.

García, C. E.; Prett, D. M.; and Morari, M. 1989. Model predictive control: Theory and practice—A survey.

Harris, C. G. 2023. Evaluating Mitigation Approaches for Adversarial Attacks in Crowdwork. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 113–119.

Holman, J. P. 1986. *Heat transfer*. McGraw Hill.

Huang, Y.; Khajepour, A.; Ding, H.; Bagheri, F.; and Bahrami, M. 2016. An energy-saving set-point optimizer with a sliding mode controller for automotive air-conditioning/refrigeration systems. *Applied Energy*.

Katili, A. R.; Boukhanouf, R.; and Wilson, R. 2015. Space cooling in buildings in hot and humid climates—a review of the effect of humidity on the applicability of existing cooling techniques. In *14th International Conference on Sustainable Energy Technologies â€“SET*.

Kim, J.; Song, D.; Kim, S.; Park, S.; Choi, Y.; and Lim, H. 2020. Energy-saving potential of extending temperature set-points in a VRF air-conditioned building. *Energies*, 13(9): 2160.

Kwadzogah, R.; Zhou, M.; and Li, S. 2013. Model predictive control for HVAC systems — A review.

Lissa, P.; Schukat, M.; and Barrett, E. 2020. Transfer Learning Applied to Reinforcement Learning-Based HVAC Control.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533.

Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; and Fernández-Leal, 2022. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4): 3005–3054.

Mossolly, M.; Ghali, K.; and Ghaddar, N. 2008. Optimal control strategy for a multi-zone air conditioning system using a genetic algorithm. *Energy*.

Mozer, M. C. 1998. The Neural Network House: An Environment that Adapts to its Inhabitants.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *International conference on machine learning*, 2817–2826. PMLR.

Ruiz–Medina, M.; and Miranda, D. 2021. Bayesian surface regression versus spatial spectral nonparametric curve regression.

Santamouris, M.; and Vasilakopoulou, K. 2021. Present and future energy consumption of buildings: Challenges and opportunities towards decarbonisation. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 1: 100002.

Wang, S.; and Ma, Z. 2008. Supervisory and optimal control of building HVAC systems: A review. *Hvac&R Research*, 14(1): 3–32.