2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS) | 979-8-3503-6927-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCPS61052.2024.00015

FAIRO: Fairness-aware Sequential Decision Making for Human-in-the-Loop CPS

Tianyu Zhao
University of California, Irvine
tzhao15@uci.edu

Mojtaba Taherisadr
University of California, Irvine
taherisa@uci.edu

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

***Abstract*—Achieving fairness in sequential decision making systems within Human-in-the-Loop (HITL) environments is a critical concern, especially when multiple humans with different behavior and expectations are affected by the same adaptation decisions in the system. This human variability factor adds more complexity since policies deemed fair at one point in time may become discriminatory over time due to variations in human preferences resulting from inter- and intra-human variability. This paper addresses the fairness problem from an equity lens, considering human behavior variability, and the changes in human preferences over time. We propose FAIRO, a novel algorithm for fairness-aware sequential decision making in HITL adaptation, which incorporates these notions into the decision-making process. In particular, FAIRO decomposes this complex fairness task into adaptive sub-tasks based on individual human preferences through leveraging the Options reinforcement learning framework. We design FAIRO to generalize to three types of HITL application setups that have the shared adaptation decision problem.**

Furthermore, we recognize that fairness-aware policies can sometimes conflict with the application's utility. To address this challenge, we provide a fairness-utility tradeoff in FAIRO, allowing system designers to balance the objectives of fairness and utility based on specific application requirements. Extensive evaluations of FAIRO on the three HITL applications demonstrate its generalizability and effectiveness in promoting fairness while accounting for human variability. On average, FAIRO can improve fairness compared with other methods across all three applications by 35.36%.

***Index Terms*—sequential-decision making, fairness, human-in-the-loop, adaptation, equity**

I. Introduction

The emerging technologies of sensor networks and mobile computing give the promise of monitoring the humans' states and their interactions with the surroundings and have made it possible to envision the emergence of human-centered design of cyber-physical systems (CPS) applications in various domains. This tight coupling between human behavior and computing enables a radical change in human life. By continuously developing a cognition about the environment and the human state and adapting/controlling the environment accordingly, a new paradigm for CPS systems provides the user with a personalized experience, commonly named Human-in-the-Loop (HITL) systems. With the increasing number of HITL CPS applications being controlled by artificial intelligence (AI) algorithms, the algorithmic fairness of such decision-making algorithms has drawn considerable attention in the last few years [1]. Nevertheless, the unique nature of HITL CPS opens a new frontier of algorithmic fairness issues that must be carefully addressed before the wide use of such technologies. In particular, the immense challenge in designing the future HITL CPS lies in respecting human rights and values, ensuring ethics and fairness, and meeting regulatory guidelines while safeguarding our environment and natural resources [2].

The following summarizes the key distinctions between the existing literature on algorithmic fairness and the nature of Human-in-the-Loop (HITL) systems:

- **Fairness in static/singular decision-making vs fairness in dynamic/sequential decision making:** The current literature on algorithmic fairness primarily addresses the unfairness arising from biases in data and algorithms used in static systems, often employing supervised learning methods. A canonical example comes from a tool used by courts in the United States to make pretrial detention and release decisions (COMPAS) [3]. Other applications include loan applications, employment processes and markets [4]. In contrast, HITL systems are dynamic, where actions taken at one time have consequences for future states and actions. Therefore, ensuring fairness in HITL systems requires considering the impact of decisions over time, leading to a sequential decision making problem. Neglecting the dynamic feedback and long-term effects in such systems, as commonly done in static decision-making, can harm sub-populations [5].
- **Fairness in decisions (or equality) vs fairness in the impact of decisions (or equity):** Existing fairness definitions predominantly focus on equality, aiming to eliminate prejudice or favoritism based on individuals' characteristics. However, insufficient attention has been given to equity, which entails allocating resources to individuals or groups to support their success [6]. Equity becomes crucial in HITL systems. Hence, a shift from fairness defined in terms of equality to fairness based on equity is essential.

Motivated by these observations, this paper revisits fairness literature and emphasizes the importance of fairness in sequential decision making from an equity perspective for HITL systems. The main objective is to operationalize equity in the context of sequential decision making to develop improved adaptation algorithms tailored to HITL applications.

This paper introduces **FAIRO**, a novel fairness-aware adaptation framework for sequential decision making designed for HITL systems. The framework specifically tackles the issue of fairness in situations where multiple humans share the same application space and are collectively impacted by

979-8-3503-6927-4/24/$31.00 ©2024 IEEE
DOI 10.1109/ICCPS61052.2024.00015

adaptation decisions. While usually, these decisions aim to optimize overall system performance, they may inadvertently lead to undesired consequences as humans interact with the system or as the system's physical dynamics evolve.

II. Related Work

A. Fairness in decision-making systems

At the heart of HITL systems is achieving the objective of designing scalable, real-time decision-making mechanisms that are aware of the social context, such as the perceived notion of fairness, social welfare, ethics, and social norms [7]. A vast work in the game theory literature studies various notions of fairness between communities by defining incentive markets between competitors to achieve fairness [8]. Fairness-enhancing interventions have been introduced to machine learning to ensure non-discriminatory decisions by the trained models [9], [10]. In particular, the question of fairness in decision-making systems where the agent prefers one action over another [11], [12] becomes more significant in multi-agent systems [13]. However, imposing fairness constraints as a static, singular decision (as standard supervised learning methods do) while ignoring subsequent dynamic feedback or its long-term effect, especially in sequential decision making systems, can harm sub-populations [5]. Recent work investigates the long-term effects of Reinforcement Learning (RL). It shows that modeling the instantaneous effect of control decisions for single-step bias prevention does not guarantee fairness in later downstream decision actions [14]. Unfortunately, all of this work focused on fairness from the lens of equality—where the target is to ensure no favoritism or bias is present in the system—with very little work that focused on fairness from the lens of equity (mostly in singular/static decision making as opposed to sequential decision making) [6]. Indeed, achieving fairness in sequential decision making systems becomes more complex since policies deemed fair at one point may become discriminatory over time due to variations in human preferences resulting from inter- and intra-human factors [12]. This paper focuses on answering this question, especially for HITL systems.

B. Different notions of group fairness

The notion of "group fairness" is used in the literature to address the fairness problem when multiple humans are affected by the same adaptation model. While there are several definitions and approaches to defining group fairness, it's important to note that these approaches may have nuanced variations and can be interpreted differently depending on the context and specific application domain. We only summarize two widely used notions of group fairness: (1) equalized odds, which focuses on achieving similar prediction accuracy across different groups while considering binary classification tasks. It ensures that the true positive rate (sensitivity) and true negative rate (specificity) of a predictive model are comparable across different groups [15], and (2) equal opportunity: aims to ensure that the predictive model provides an equal chance of benefiting from positive outcomes for all groups. In particular, equal opportunity requires that the true positive rate for each group should be approximately equal [15]. While these two definitions primarily focus on binary classification tasks, this paper will exploit some of their ideas towards sequential decision-making and not specifically for classification tasks.

C. Multi-agent RL and hierarchical RL

Reinforcement Learning (RL) is a widely used approach for monitoring and adapting to human intentions and responses in various contexts [16]. To account for individual variability and response times, approaches like multisample RL has been proposed [17]. Hierarchical reinforcement learning (HRL) decomposes complex learning tasks into manageable components by using a hierarchical structure. The high-level policy selects optimal sub-tasks, considered high-level actions, while the lower-level policy focuses on solving these sub-tasks using reinforcement learning techniques. This decomposition strategy transforms long timescale tasks into multiple shorter timescale sub-tasks, potentially simplifying individual sub-task solving. For instance, the Option-critic Framework introduces an architecture capable of learning higher and lower-level policies without needing prior knowledge of sub-goals [18]. HRL has demonstrated superior performance in various domains, including long-horizon games, continuous control problems [19] and fairness in human-in-the-loop IoT [12]. In this paper, we will decompose the fairness problem into sub-tasks over smaller time horizons and exploit the options framework to solve these sub-tasks.

D. Paper contribution

This paper's contributions can be summarized as follows:

- **Fairness from the lens of equity:** We tackle the fairness problem in sequential decision-making systems within HITL environments by addressing the notion of equity.
- **FAIRO:** We propose FAIRO, a novel algorithm designed for fairness-aware sequential-decision making in HITL adaptation. Our approach leverages the Options RL framework to effectively incorporate fairness.
- **Generalization to different HITL application setups:** We extend FAIRO to cater to three types of HITL application setups. These setups involve multiple humans sharing the application space and being impacted by: (1) global numerical adaptation decisions, (2) shared global resources, and (3) shared global categorical adaptation decisions.
- **Evaluation on multiple HITL applications:** We conduct comprehensive evaluations of FAIRO on three different HITL applications to demonstrate its generalizability and compare with previous work in the literature.

The paper is structured as follows: Section III summarizes the Options framework, which serves as the foundation for our proposed approach. Section IV details how to incorporate fairness considerations into the decision-making process. The subsequent sections of the paper focus on evaluating our proposed approach, FAIRO, in three distinct application domains.

III. Options framework for temporal abstraction

Markov Decision Process (MDP) is widely employed for modeling sequential decision making. Various methods are utilized to solve MDPs and obtain the optimal Markov

Fig. 1: The state trajectory of an MDP with small discrete-time transitions. Options enable overlaid larger abstracted discrete events.

decision chain, including dynamic programming and reinforcement learning (RL). RL is particularly used when the transition probabilities within the MDP are unknown. Within the discrete-time finite MDP setting, the standard RL framework can be applied. In particular, an *agent* engages with an *environment* that is modeled as an MDP at discrete time steps, denoted as $t = 0, 1, 2, ...$ At each time step t, the agent observes the current state of the environment, denoted as $s_t \in \mathcal{S}$, and selects an action $a_t \in \mathcal{A}$ based on this observation. This action leads to a transition to the next state, s_{t+1}, and yields a reward value, $r_t \leftarrow \mathbb{R}$, associated with this transition. By engaging in this interaction, the agent learns a policy $\pi(s, a)$ that guides its decision-making, aiming to select the best action a for each state s to maximize the expected total reward over sequential decision actions.

The options framework was first introduced by Sutton et al. [20] to generalize primitive actions to include temporally extended courses of lower-level action. In particular, the term options represents a temporal abstraction of the lower-level actions in the MDP. A pictorial figure of options over MDP is shown in Figure 1. An MDP's state trajectory comprises small, discrete-time transitions, whereas the options enable an MDP to be abstracted and analyzed in larger temporal transitions.

Option o within the option set $\mathcal{O}$ consists of three main components: a policy $\pi(a|s, o)$ for selecting actions within option o, an initiation set $\mathcal{I} \subseteq \mathcal{S}$, a termination condition β. An option o:$(\mathcal{I},\pi,\beta)$ is available to be selected by the agent in state s_t if and only if $s_t \in \mathcal{I}$. If the option is selected, actions are selected according to the option policy π until the option terminates according to the termination condition β. When the option terminates, the agent can select another option. This definition of options makes them act as much like actions while adding the possibility that they are temporally extended[1].

In this paper, the rationale behind employing the options framework to achieve fairness in a multihuman setting stems from the inherent limitations imposed by an option's initiation set $\mathcal{I}$ and termination condition β. These constraints confine the applicability of an option's policy, π, to a subset defined by $\mathcal{I}$ rather than encompassing the entire state space $\mathcal{S}$. Consequently, options can be viewed as a means of achieving **fairness subgoals**, wherein each option's policy is adapted to enhance the attainment of its specific **subgoal**, thereby contributing to the overall fairness of the decision-making agent. The dynamic nature of the multihuman environment necessitates diverse fairness policies at different temporal instances.

[1]Options framework can be extended to include policies over options. When multiple options are available to the agent at s_t, the agent can learn which option to select using the policy over options. We consider the policy over options to be a fixed policy, and the initiation sets of all options are disjoint sets.

IV. FAIRO: Fairness using Options framework

We exploit options framework to design FAIRO to achieve fairness in sequential decision-making agents in multihuman environment [20]. As seen in Figure 2, the agent interacts in sequential discrete-time steps with an environment that has N humans $(h_1,h_2,...,h_N)$ through observing their preferences or their desired adaptation actions $(d_1,d_2,...,d_N)$ and the current fairness state of the environment s_t. Guided by the current fairness state s_t, the agent selects an appropriate option o_t from the set of N available options $\mathcal{O}$. The chosen option o_t then determines a lower-level action based on its specific option policy π_o, resulting in a global action a_{g_t} that is applied to the shared environment. This global action subsequently modifies the current fairness state, and the agent receives a reward r_{t+1}. This reward is utilized to refine the option policy. In the following subsections, we provide a detailed description of each module within the FAIRO framework.

A. Fairness state space $\mathcal{S}$

Our approach to viewing fairness from the lens of equity is by using a fairness state that encompasses the history of the positive and negative effects of the global decision action.

*1) **Satisfaction history records** $\mathbf{c}_i$:* Fairness state s_t is inferred from the history of the *satisfaction* of each human. To model the satisfaction of the human h_i, we keep a history record for each human:

$$\mathbf{c}_i = (u_i, v_i), \text{where } i \in \{1,2,...,N\}. \tag{1}$$

The value $u_i \in \mathbb{R}$ represents a record of the number of times the human h_i was unsatisfied by the applied global action a_g. In contrast, $v_i \in \mathbb{R}$ represents a record for the number of times the human h_i was satisfied by the applied global action a_g.

At time step t, every human h_i has a desired adaptation action d_{i_t}. For example, a human may prefer a particular temperature setpoint to HVAC system (Heating, ventilation, and air conditioning) in their room for thermal comfort that matches their physical activities, such as sleeping, domestic work, or sitting. Based on the difference in the values of d_{i_t} and a_{g_t}, the record $\mathbf{c}_i$ is updated to capture whether the human was satisfied or unsatisfied. For example, if this difference is within a threshold τ then we consider the human h_i is satisfied and increment v_i by a value δ.

$$\mathbf{c}_i = \begin{cases} (u_i, v_i+\delta) & \|d_{i_t} - a_{g_t}\| \leq \tau. \\ (u_i+\delta, v_i) & \|d_{i_t} - a_{g_t}\| > \tau. \end{cases} \tag{2}$$

After all the records $\mathbf{C} = (\mathbf{c}_i, i = 1,2,...,N)$ are updated, they are normalized to a unit vector. Choosing the value τ is application dependent; however, the value δ needs to be less than 1 and small enough to ensure that the unit vector direction $\mathbf{C}$ does not change drastically. Hence, we choose δ to be 0.01.

Ideally, these records $\mathbf{c}_i$ should be $(0, 1)$ indicating that the global adaptation action a_g meets the preferences of the human over time. However, as we mentioned earlier, these preferences may conflict with humans sharing the same environment. Hence, the same a_{g_t} may be perceived by one human as meeting their preference (increasing v) and by another human as not meeting theirs (increasing u).

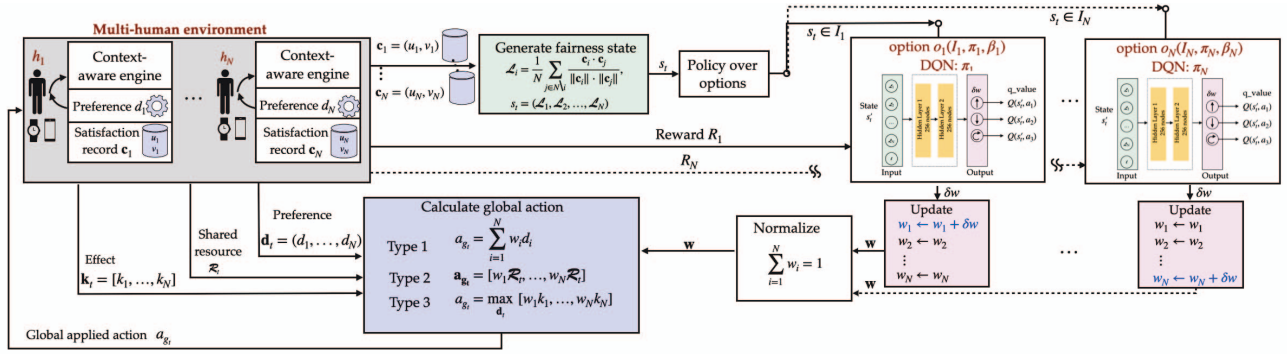


Fig. 2: FAIRO for fairness-aware framework in Human-in-the-Loop systems using options framework. FAIRO is designed for three types of applications: **Type I:** one global action based on numerical demands affecting multiple humans, **Type II:** one shared resource distributed over multiple humans, and **Type III:** one global action based on categorical preferences.

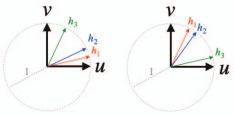


Fig. 3: Satisfaction history record $\mathbf{C}$ is represented as $2D$ unit vector, where the two components u and v represent history of "unsatisfied" and "satisfied" respectively.

A pictorial visualization of $\mathbf{C}$ is shown in Figure 3. Each $\mathbf{c}_i$ can be represented as a $2D$ vector within a unit circle. We show two examples for $\mathbf{c}_i$ for three humans where h_3 has $\mathbf{c}_3$ closer to the v axis compared to the other two humans (Figure 3-left) versus the case when h_3 has $\mathbf{c}_3$ closer to the u compared to the other two humans (Figure 3-right). Figure 3 shows an example of a relatively unfair situation, where h_3 is either treated most of the time favorably (Figure 3-left) or unfavorably (Figure 3-right).

It is worth mentioning here that $\mathbf{C}$ captures the history of the effect of the trajectory of sequential adaptation action on the shared environment. Hence, the intuition is to tune the global action in the next time step $a_{g_{t+1}}$ to either decrease the $focus$ on considering the preferences of h_3 (Figure 3-left) or vice versa (Figure 3-right). However, as mentioned in Section I, the same action affects all the humans sharing the same environment.

*2) **Fairness state** s_t:* We use the geometric intuition in Figure 3 to design our fairness state s_t to compare the directions of all N records in $\mathbf{C}$. Ideally, we would like to have all $\mathbf{c}_i$ as close as possible to each other. Hence, we define s_t to capture how close each $\mathbf{c}_i$ is to the other $N \backslash i$ records. Hence, we define (s_t) as follows:

$$\mathcal{L}_i = \frac{1}{N} \sum_{j \in N \backslash i} \frac{\mathbf{c}_i \cdot \mathbf{c}_j}{\|\mathbf{c}_i\| \cdot \|\mathbf{c}_j\|}, \qquad \mathcal{L}_i \in]0,1] \subset \mathbb{R} \tag{3}$$

$$s_t = (\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_N), \qquad s_t \in \mathcal{S} =]0,1]^N \tag{4}$$

In particular, $\mathcal{L}_i$ represents the closeness of record $\mathbf{c}_i$ to the rest of the records using the average of the cosine of the angle between pair of vectors. Hence, if the cosine value between two vectors is 1, they coincide. Since the values of $\mathbf{c}_i$ can only be positive and are normalized to a unit vector, the minimum cosine value between these vectors is 0, indicating that they are far from each other (at 90°)[2]. Equation 4 represents s_t which holds all the values of $\mathcal{L}_i$. Ideally, from our fairness point of view, the goal state s_t should be $(1,1,...1)$, which indicates that all $\mathbf{C}_h$ have the same direction, meaning that the history of the satisfaction and unsatisfaction for all the humans are close.

B. Initiation set $\mathcal{I}$ and fairness subgoals

While the ultimate goal is to learn a policy that can achieve the goal state $s_t = (1,1,...1)$, this is challenging since it is a huge state space. Accordingly, the intuition behind exploiting the options framework is to divide this goal into smaller subgoals where we learn over a subset of states or the initiation set ($\mathcal{I} \subseteq \mathcal{S}$) as explained in Section III. We divide $\mathcal{S}$ into N initiation sets $\mathcal{I}_i$ where $i \in \{1,2,...,N\}$, such that $\mathcal{I}_i$ contains all the states with $\mathcal{L}_i$ as the minimum value.

$$\mathcal{I}_i = \{s_t = \{\mathcal{L}_1, ..., \mathcal{L}_N\} \in \mathcal{S} | \mathcal{L}_i = \min(s_t)\} \tag{5}$$

Specifically, this means that each initiation set $\mathcal{I}_i$ considers only the states where h_i has received unfair adaptation either favorably or unfavorably. For example, both cases in Figure 3 are considered unfair state where $\mathcal{L}_3$ is less than $\mathcal{L}_1$ and $\mathcal{L}_2$.

C. Termination State β

Each option o_i terminates when the current state s_t reaches a terminal state for this option. Hence, in FAIRO, the set of terminal states for o_i is when $\mathcal{L}_i$ is no longer the minimum value in s_t.

$$\beta_i = \{s_t = \mathcal{L}_1, ..., \mathcal{L}_N\} \in \mathcal{S} | \mathcal{L}_i \neq \min(s_t)\} \tag{6}$$

Intuitively, this means that each option o_i will run to improve the value of $\mathcal{L}_i$ until it is no longer the minimum value which is the fairness subgoal for this option. This will trigger a new initiation set I and this option terminates and a new option starts to achieve another subgoal: improving $\mathcal{L}_i$.

D. Global action of different HITL applications

As shown in Figure 2, every human (h_i) has a desired preference (d_i). However, only one action a_g is chosen to be applied to the shared environment. In FAIRO, we identify three types of applications:

- **(Type I) Shared numerical global action:** The desired preferences d_i have numerical values and the global action a_g is a numerical value. We design each option to take

[2] $\mathbf{c}_i \in \mathbb{R}, \mathcal{L}_i$ is unlikely to reach 0 but can decrease to a very small value ϵ.

a weighted sum of these N preferences. These weights represent the contribution of each human preference d_i in the applied global action a_{g_t}. Hence, $a_{g_t} = \sum_{i=1}^{N} w_i d_i$, where $w_i \in [0,1]$. We will show an instant of this type in a simulated smart home application in Section V.

- **(Type II) Shared global resource:** The desired preferences d_i have numerical values. There is one shared time-varying resource $\mathcal{Z}$ and it has to be distributed. The global action $\mathbf{a}_{\mathbf{g}_t}$ is the weighted share of this resource $\mathcal{R}_t$ dictated by their desired preferences. Hence, $\mathbf{a}_{\mathbf{g}_t} = [w_1\mathcal{Z}_t, w_2\mathcal{Z}_t, ..., w_N\mathcal{Z}_t]$, where $w_i \in [0,1]$. We will show an instant of this type in a simulated water distribution application in Section VI.
- **(Type III) Shared categorical global action:** The desired preferences d_i have categorical values, with the global action a_g selecting one of these categorical values, which represents the maximum weighted effect of applying d_i on all other $N-i$ humans, where the effect k_i can be estimated from the context-aware engine, resulting in $\mathbf{a}_{\mathbf{g}_t} = \arg\max_{d_i}[w_1k_1, w_2k_2, ..., w_Nk_N]$, where $w_i \in [0,1]$. We will show an instant of this type in a smart education application in Section VII.

E. Learning the option policy (π_i)

Each option policy π_o learns the appropriate $\mathbf{w} = (w_1, w_2, ..., w_N)$ for the global action a_{g_t} for every state $s_t \in \mathcal{I}_i$ to reach the termination condition β_i such that $\sum_{i=1}^{N} w_i = 1$. These weights are continuous values and learning them for every $s_t \in \mathcal{I}_i$ is challenging. Hence, we opt for a simpler design of using Deep Q-Network (DQN) to reduce the search space for the appropriate weights as detailed below.

*1) **Deep Q-Network (DQN):*** DQN is a reinforcement learning algorithm that utilizes deep neural networks in combination with Q-learning within Markov Decision Processes (MDP), estimates the action-value function $Q(s,a)$, representing expected rewards for actions in states. This function is typically represented by a neural network with state inputs and estimated action values as outputs, and at each time step t, the DQN agent uses ε-greedy policy to select action, updating the Q-function based on observed states, rewards, and the next state using an update rule:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)), \quad (7)$$

where α is the learning rate, γ is the discount factor, and $\max_a Q(s_{t+1},a)$ is the estimated maximum action-value in the next state s_{t+1}. DQN aims to learn a policy that maximizes the cumulative reward over time in a given environment. It combines Q-learning with deep learning, allowing the agent to handle high-dimensional observations and non-linear function approximations. Figure 4 shows a pictorial illustration for the design of the DQN in every option o_i in FAIRO.

*2) **Input to Option DQN (s'_t):*** Every option o_i runs a DQN where the input is the s_t. However, to differentiate between the two cases shown in Figure 3 where both $\mathcal{L}_3$ is the minimum value in s_t, we append to s_t the relative location of $\mathbf{c}_3$ with respect to the $\mathbf{c}_1$ and $\mathbf{c}_2$.

$$s'_t = (s_t, l), \text{ where } s_t \in \mathcal{I}_i, \quad l = \begin{cases} 1, & v_i = \max_v(s_t). \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

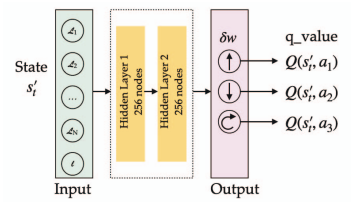


Fig. 4: DQN for option o_i. The input is the state s'_t. The output is the `q_value` for the possible three actions of weight adjustment δw.

This means that if $\mathcal{L}_i$ is the minimum in s_t, option o_i will run since $s_t \in \mathcal{I}_i$. The input to the DQN in option o_i will indicate whether $\mathcal{L}_i$ has the minimum value (unfair situation) because h_i received favorable treatment relative to all h_{Nni} (*i.e.,* $\mathbf{c}_i$ has a high v_i component) and in this case $l = 1$ in Equation 8, or $l = 0$ otherwise.

*3) **Output from Option DQN (δw_i):*** To reduce the action space of the DQN since we need to learn N weights with continuous values $[0,1]$, we designed the output from the DQN in option o_i to focus only on adjusting w_i instead of all the N weights $\mathbf{w}$. In particular, as shown in Figure 4, the output from the DQN are the Q-values ($Q(s,a)$) which decides to select between three actions to (a_1) increase the weight $w_i \uparrow$, (a_2) decrease the weight $w_i \downarrow$, or (a_3) keep the same weight $w_i \circlearrowleft$. Hence, the output from DQN (which is the selected action of the DQN as explained in Section IV-E1) is the weight adjustment for w_i by $\delta w_i = (+\delta, -\delta, 0)$ where the value δ determines how fast the DQN for option o_i changes the weight w_i.

$$w_i \leftarrow w_i + \delta w_i, \qquad \sum_{i=1}^{N} w_i = 1. \quad (9)$$

Since all the weights $\mathbf{w}$ have to be normalized to 1, all the weights will be adjusted accordingly. Using this design, we choose between 3 possible adjustments for weight w_i instead of the whole space $[0,1]$ and instead of all the N weights.

*4) **Option DQN reward ($\mathcal{R}_i$):*** Each option DQN learns the appropriate policy $\pi_i(s'_t, \delta w_i)$, which is the right weight adjustment δw_i at state s'_t. The DQN learns this policy through a notion of a feedback reward as explained in Section IV-E1. As each option o_i aims at increasing the fairness subgoal of enhancing $\mathcal{L}_i$ while improving the performance of the application, the reward function $\mathcal{R}_i$ can be expressed with two terms; a fairness term $\mathcal{F}$, and a performance term $\mathcal{P}$ using a trade-off parameter $\zeta \in [0,1]$.

$$\begin{aligned} \mathcal{R}_i &= \zeta\mathcal{F}_i + (1-\zeta)\mathcal{P}_i, \in [-1,1] \subset \mathbb{R} \\ \mathcal{F}_i &= \text{absolute fairness} + \text{option improvement} \\ &= (2 * \mathcal{L}_{i_t} - 1) + f(\mathcal{L}_{i_{t-1}}, \mathcal{L}_{i_t}), \in [-1,1] \subset \mathbb{R} \\ \mathcal{P}_i &= \text{application dependent}, \in [-1,1] \subset \mathbb{R} \end{aligned} \quad (10)$$

The fairness $\mathcal{F}_i$ considers the current value of $\mathcal{L}_i$, which we call the "absolute fairness", and the "improvement in the fairness" value of $\mathcal{L}_i$ from last time step for this particular option. It is a function (f) of the current value of $\mathcal{L}_{i_t}$ and the value from last time step $\mathcal{L}_{i_{t-1}}$ as shown in Equation 10. The overall FAIRO algorithm is listed in Algorithm 1.

V. Application Type I: Smart Home HVAC

Recent literature focuses on enhancing human satisfaction in smart heating, ventilation, and air conditioning (HVAC) systems by employing reinforcement learning (RL) techniques to adjust the set-point based on human activity and preferences [12]. These HITL systems consider the current state and individual preferences, such as body temperature changes during sleep or physical activity. To evaluate FAIRO in **Type I** applications, we consider a setup where multiple humans share a house with a single HVAC system, and their activities determine individual desired setpoint.

A. House-Human Physical Model

We used a thermodynamic model of a house incorporating the house's shape and insulation type. To regulate indoor temperature, a heater and a cooler with specific flow temperatures ($50°c$ and $10°c$) were employed. A thermostat maintained the indoor temperature within $2.5°c$ around the desired set point. An external controller controls the setpoint. The human was modeled as a heat source, with heat flow dependent on the average exhale breath temperature (EBT) and the respiratory minute volume (RMV). These parameters depend on human activity [21]. We simulated three humans with four activities: sleeping, relaxing, medium domestic work, and working from home. The different activity schedules depicted in Figure 8-Left in the Appendix. The humans were simulated in separate rooms, each exhibiting unique behavioral patterns: (1) h_1 followed an organized and repetitive weekly routine, (2) h_3 had a more random and unpredictable life pattern, and (3) h_2 displayed intermediate randomness, alternating between sleeping, being away from home, domestic activities, and relaxation. The Mathworks thermal house model was extended to include a cooling system, a human model, and an external controller running FAIRO[3].

B. Context-aware engine

A context-aware engine estimates the desired action d_i per human h_i in a smart home. The desired action d_i can be obtained through fixed policy configuration or learned policy. To focus on our main contribution and not on designing a new context-aware engine, we leverage existing RL-based approaches for estimating the desired HVAC setpoint d_i based on activity and thermal comfort [22]. The desired setpoints for the considered activities are domestic activity (72°F), relaxed activity (77°F), sleeping (62°F), and work from home (67°F). These setpoints aim to enhance thermal comfort. Thermal comfort is assessed using Prediction Mean Vote (PMV) on a scale from very cold (-3) to very hot ($+3$). Optimal indoor thermal comfort falls within the recommended range of $[-0.5,0.5]$, as per the ISO standard ASHRAE 55 [23].

C. Evaluation

We compare 5 different approaches including one of the state-of-the-art approaches FaiRIoT [12]:

[3]While more complex simulators like EnergyPlus exist, we opted for a simpler model to assess FAIRO.

Algorithm 1 FAIRO algorithm

Require:
Humans $\mathcal{H}=(h_1,...,h_N)$
Satisfaction records $\mathbf{C}=(\mathbf{c_1},...,\mathbf{c_N}),\mathbf{c}_i=(u,v)\in\mathbb{R}^2$
States $s\in\mathcal{S}=]0,1]^N\subset\mathbb{R}^N$
Initiation sets $\mathcal{I}_i\in\mathcal{I}=\{s\in\mathcal{S}\}$
Options $\mathcal{O}=(o_1,...,o_N),o_i=(\mathcal{I}_i,\pi_i,\beta_i)$
Application type $T=\{Type1,Type2,Type3\}$

procedure RUN-FAIRO
  while True **do**
    $\mathbf{d}_t \leftarrow$ context-aware-engine ($\mathcal{H}$)
    if $T==Type2$ **then**
      $\mathcal{R}_t \leftarrow$ get-current-available-resource()
    if $T==Type3$ **then**
      $\mathbf{k}_t \leftarrow$ get-current-effects($\mathbf{d}_t$)
    $s_t \leftarrow$ get-fairness-state($\mathbf{C}$)
    $s'_t \leftarrow$ append-state(s_t)
    $o_t \leftarrow$ choose-option (s_t) ▷ $o_t\in\mathcal{O}$
    $\delta w \leftarrow$ run-option(o_t) ▷ option policy $\pi_o(s_t,\delta w)$
    $\mathbf{w}_t \leftarrow$ update-normalize-weights(δw)
    $a_{g_t} \leftarrow$ calculate-global-action($\mathbf{d}_t$, $\mathbf{w}_t$,$\mathcal{R}_t$, $\mathbf{k}_t$)
      ▷ Apply global action a_{g_t} on the shared environment
    $\mathcal{R}_t \leftarrow$ receive-reward ()
    $\pi_o(s_t,\delta w) \leftarrow$ Update-option-policy (R_t)
    $\mathbf{C} \leftarrow$ Update satisfaction records ($\mathbf{d}_t$,a_{g_t})

- **FAIRO:** The global applied action is $a_{g_t}=\sum_{i=1}^{3} w_i d_i$. The reward per option i is as explained in Equation 10. We set $\zeta=0.5$. As for the performance term $\mathcal{P}_i$ in the reward, we assign a high reward when the PMV falls in the acceptable range $[-0.5,0.5]$. Further, we used the values of the satisfaction counters $\mathbf{c}_i=(u_i,v_i)$ as an indication of the performance $\mathcal{P}_i$ since their values are correlated to the desired temperature d_i which maps to the best PMV. The value $\mathcal{P}_i$ is then normalized to be $[-1,1]$:

$$f(\mathcal{L}_{i_{t-1}},\mathcal{L}_{i_t})=sign(\mathcal{L}_{i_{t-1}}-\mathcal{L}_{i_t})\times Z$$
$$Z=\begin{cases}0 & |\mathcal{L}_{i_{t-1}}-\mathcal{L}_{i_t}|\in]0,0.001]\\ 0.25 & |\mathcal{L}_{i_{t-1}}-\mathcal{L}_{i_t}|\in]0.001,0.005]\\ 0.5 & |\mathcal{L}_{i_{t-1}}-\mathcal{L}_{i_t}|\in]0.005,0.01]\\ 0.75 & |L_{i_{t-1}}-\mathcal{L}_{i_t}|\in]0.01,0.015]\\ 1 & |L_{i_{t-1}}-\mathcal{L}_{i_t}|>0.015\end{cases} \quad (11)$$
$$\mathcal{P}_i=0.2\frac{v_i}{u_i+v_i}+0.8f(\text{PMV}),\in[-1,1]\subset\mathbb{R}$$

- **Average approach:** The setpoint is the mean value of desired setpoints of all rooms. Hence, $a_{g_t}=\frac{1}{3}\sum_{i=1}^{3} d_i$.
- **Equality using round robin (RR):** The setpoint is selected from one of the desired setpoints of all rooms in a rotation. The intuition of this approach is to compare with the case where we give every room the same opportunity to use its desired setpoint across time (equality). Hence, for 3 humans, $a_{g_1}=d_1,a_{g_2}=d_2,a_{g_3}=d_3,a_{g_4}=d_1,...$, etc.
- **No subgoals using 1 DQN:** The setpoint is calculated using a single model of 1 DQN structured as Figure 4. The inputs are the three values of the fairness state ($s_t=(\mathcal{L}_1,\mathcal{L}_2,\mathcal{L}_3)$). The intuition of this approach is to evaluate whether we can achieve the overall fairness goal without the need for subgoals that FAIRO provides. The reward for the single

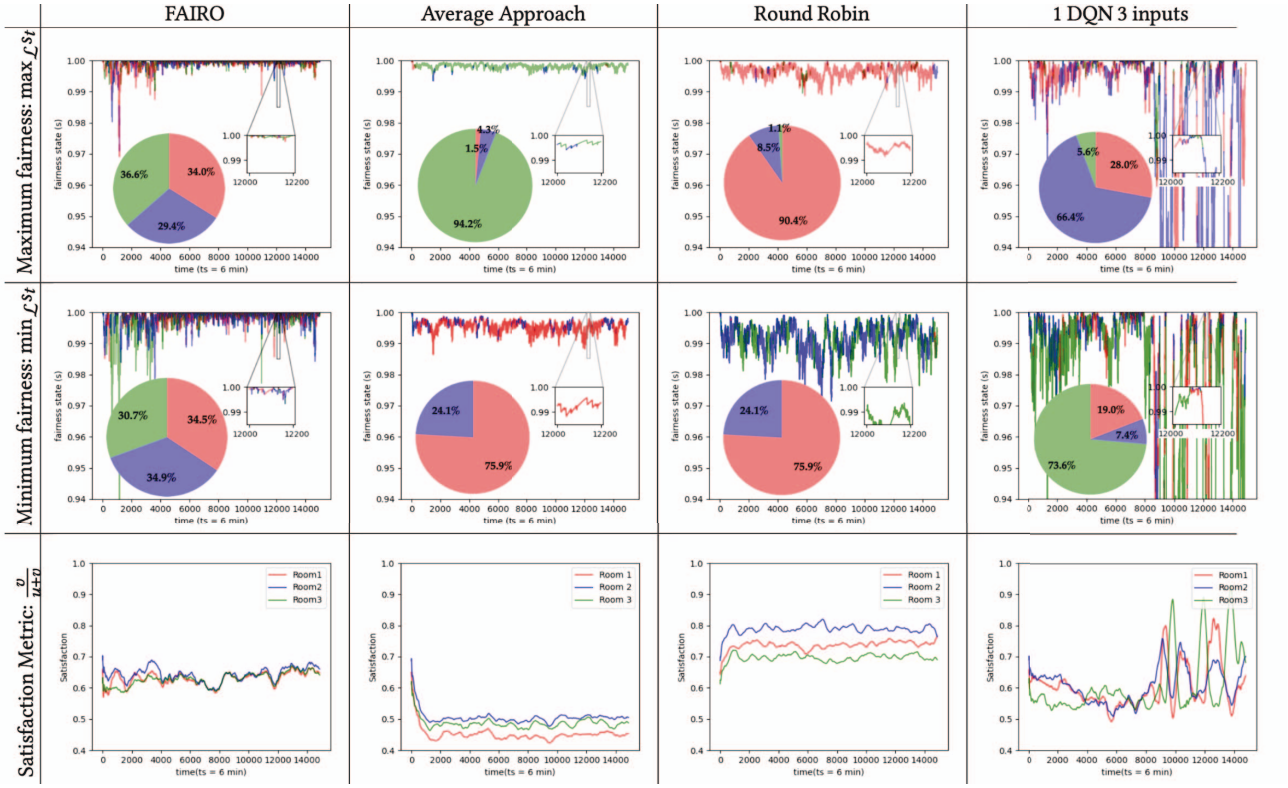


Fig. 5: Fairness state and satisfaction across all approaches in application 1 showing the probabilities of equal opportunity and equalized odds.

DQN is the average reward value across all rooms:

$$\mathcal{R}=0.5\mathcal{F}+0.5\mathcal{P}, \quad \text{where } N=3\in[-1,1]\subset\mathbb{R}$$
$$\mathcal{F}=\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i^N+f(\frac{1}{N}\sum_{i}\mathcal{L}_{i_{t-1}},\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{i_t}), \tag{12}$$
$$\mathcal{P}=0.2\frac{1}{N}\sum_{i=1}^{N}\frac{v_i}{u_i+v_i}+0.8\frac{1}{N}\sum_{i=1}^{N}f(PMV)$$

- **FaiRIoT [12]:** The closest to our approach is FaiRIoT which uses hierarchical RL.

We investigated multiple evaluation metrics to evaluate the fairness across these three rooms; 1) the values of the fairness state ($s_t=(\mathcal{L}_1,\mathcal{L}_2,\mathcal{L}_3)$, 2) a satisfaction metric, 3) PMV, and 4) the covariance cv of the learnt weights.

*1) **Fairness state and group fairness definition**:* As explained in Section IV-A2, the best fairness state should be $s_t=(\mathcal{L}_1,\mathcal{L}_2,\mathcal{L}_3)=(1,1,1)$. To measure s_t, we need to use the satisfaction history counters by updating them per to Equation 2. In this application, we set the threshold τ to be 2.5. Figure 5 shows the fairness states results with 15k samples equivalent to 62.5 simulated days.

To better get insights on how the different methods compare with respect to the values of s_t, in every time sample, we report the maximum value of $\mathcal{L}$ in s_t (Figure 5-first row) and the minimum value of $\mathcal{L}$ in s_t (Figure 5-second row). We plot the values $\mathcal{L}_1$ for Room 1(red), $\mathcal{L}_2$ for Room 2(blue), and $\mathcal{L}_3$ for Room 3(green). The results for using only 1 DQN are unstable, and have fluctuations in the fairness state values. This result is expected for 1 DQN since the fairness goal was not divided into sub-goals, unlike what the options framework provided.

The commonly used metrics for group fairness are *equal opportunity* and *equalized odds*. Although these metrics are typically applied to binary classification tasks, we adopt their original definitions to compare FAIRO with the other approaches.

- ***Equal opportunity*** aims to ensure that individuals from different groups have an equal chance or probability of experiencing positive outcomes or receiving beneficial treatment or resources. To assess the performance of FAIRO, we analyze the results presented in Figure 5 by examining the reported $\max_{\mathcal{L}} s_t$ values. Specifically, we compare the probabilities of different rooms M_i having the highest $\max_{\mathcal{L}} s_t$ values. For *equal opportunity*, these probabilities need to be close, as denoted in Equation 13.

$$p(M_i == \texttt{find}(\max_{\mathcal{L}} s_t)) \approx p(M_{N\setminus i} == \texttt{find}(\max_{\mathcal{L}} s_t)) \tag{13}$$

As observed in Figure 5-first row, these probabilities are 34%, 29.4%, and 36.6% for Room M_1, M_2, and M_3 respectively, which are closer in values compared with the other approaches. In particular, using FAIRO, the average absolute difference between the probabilities of *equal opportunity* across the 3 rooms is reduced by 57.0%, 54.8%, and 35.8% from Average Approach, RR, and 1 DQN respectively. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by 49.2% on average.**

- ***Equalized odds*** focuses on the balance between positive and negative outcomes across different groups. Hence, for the negative outcomes, we can examine the probabilities of

different rooms having the $\min_{\mathcal{L}} s_t$ values and assess whether they are approximately equal.

$$p(M_i == \texttt{find}(\min_{\mathcal{L}} s_t)) \approx p(M_{N\setminus i} == \texttt{find}(\min_{\mathcal{L}} s_t)) \quad (14)$$

As observed in Figure 5-second row, these probabilities are $34.5\%, 34.9\%$, and 30.7% for Room M_1, M_2, and M_3 respectively, which are closer in values compared with the other approaches. In particular, using FAIRO the average absolute difference between the probabilities of *equalized odds* across the 3 rooms is reduced by 47.8%, 35.8%, and 41.3% from Average Approach, RR, and 1 DQN respectively. **Across all the approaches, FAIRO improves the *equalized odds* fairness by 41.63% on average.**

*2) **Satisfaction performance**:* Figure 5-third row shows the satisfaction values ($\frac{v}{v+u}$) for all methods. FAIRO achieves close satisfaction values across the three rooms over time. Average and RR approaches have stable but different satisfaction values across rooms while 1 DQN shows more fluctuations per room. We focused on samples after FAIRO converges ($12k$ to $15k$) and examined the satisfaction value histograms. We report the Jensen-Shannon Divergence (JSD) of these histograms[4]. FAIRO has the lowest average JSD (0.013), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO reduces JSD by 92.06% on average.** More details are shown in Figure 8-right in Appendix.

To gain further insights into human satisfaction, we plot the covariance of temperature differences between the applied setpoint a_g and the desired temperature d_i for the three humans, and Figure 6 shows that FAIRO exhibits the lowest cv of 0.04, indicating better fairness.

*3) **PMV results**:* FAIRO achieves the second lowest average JSD and a comparable variance on PMV results, which indicates FAIRO can improve fairness without hurting the PMV performance. RR achieves the lowest PMV JSD average since every time step one of the rooms can get exactly its desired temperature. Hence, the 3 rooms can get almost identical PMV performance in the long term. **Across all the approaches, FAIRO's PMV JSD is reduced by 13.7% on average.** More details are shown in Figure 8-Right in the Appendix.

*4) **Comparison with the State-of-the-Art FaiRIoT [12]**:* FaiRIoT uses a notion of utility $u_{h_t} = \frac{1}{t}\sum_{j=0}^{t} \frac{j}{t} w_{h_j}$ which is the average weight assigned by a layer called "Mediator RL" for a particular human h over a time horizon $[0:t]$, where the factor $\frac{j}{t}$ is used to give more value to the recent weights learnt by the Mediator RL over the ones in the past. FaiRIoT measures the fairness of the Mediator RL using the coefficient of variation (cv) of the human utilities: $cv = \sqrt{\frac{1}{n-1}\sum_{h=1}^{n} \frac{(u_h - \bar{u})^2}{\bar{u}^2}}$, where $\bar{u}$ is the average utility of all humans. The Mediator RL is said to be more fair if and only if the cv is smaller. Accordingly, we compare the cv in FaiRIoT and FAIRO in Figure 7. FAIRO achieves cv around 0.15, while FaiRIoT cv is larger than 0.6. **Hence, FAIRO improves the fairness where *cv* is reduced by 75%.**

[4]The Jensen–Shannon divergence (JSD) is a symmetric measure of similarity between two probability distributions, always non-negative, with 0 denoting identical distributions and any value above 0 indicating differences.

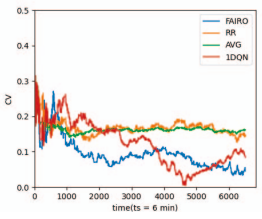


Fig. 6: Coeff. of variation (cv) of the temperature differences.

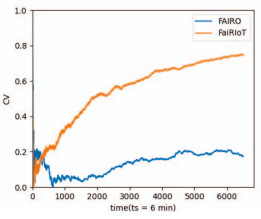


Fig. 7: Coeff. of variation (cv) of the utility u_t over $\mathbf{w}$.

VI. Application Type II: Water Supply Application

In the context of global climate change and rapid population growth, Water Demand Models (WDMs) are crucial for gaining insights into water consumption behavior and forecasting demand, aiding decision-making in water distribution system (WDS) operation policies and infrastructure planning. However, managing limited water resources while ensuring equitable distribution among households remains a challenge, with various pricing and non-pricing policies explored in the literature [24]. For evaluating FAIRO in application Type II, it is assumed that WDM estimates desired water demand for three households from a time-varying, insufficient shared water resource.

A. Household-Water Demand Physical Model

We utilized a WDM to model water demand based on residents' activity patterns, matching the activity patterns presented in Section V-A to water demand behavior [25]. The water demand patterns for three households are illustrated in Figure 11-Left in the Appendix. Each household has a water reservoir $\mathcal{T}$ and aims to meet its specific water demand d_i, with the shared water resource ($\mathcal{Z}_t$) being variable and insufficient to satisfy all demands. Assuming the availability of the WDM, we set $\mathcal{Z}_t$ to 1.5 times the maximum demand among the three profiles, with the water demand profiles being independent due to distinct human activities. We expanded the Mathworks water supply physical model to accommodate these household water demand profiles and multiple houses [26].

B. Context-aware engine

Our main contribution is not designing a new water consumption behavior or forecasting demand models, hence, the context-aware engine provides the desired action for every household h_i, which is the current water demand d_i based on the water demand pattern. This demand is supposed to be satisfied via the water supply s_i from the shared limited resource $\mathcal{Z}_t$ and the current reserved water in the household water reservoir $\mathcal{T}_i$. Hence, we define the application performance as the percentage of balancing the demand and the supply.

$$\text{Balance Rate (BR)} = \frac{\text{supply } s_i + \text{reserve } \mathcal{T}_i}{\text{demand } d_i} \quad (15)$$

C. Evaluation

We compare the same methods as in the first application.

- **FAIRO:** The supply s_i for each household h_i is calculated by FAIRO. In particular, s_i receives $w_i \mathcal{Z}_t$ as explained in Section IV-D where the global action a_{g_t} is the weighted distribution of this resource $\mathcal{Z}_t$. The reward is the same as explained in the first application (Equation 11), where performance $\mathcal{P}_i$ is based on the BR ($\mathcal{P}_i = 0.2\frac{v_i}{u_i+v_i} + 0.8 f(\text{BR})$).

- **Average approach:** Each household h_i has the same amount of supply. Hence, $s_i = \frac{1}{3}\ \mathcal{Z}_t$. We further augment the average approach in this application by using a **Weighted Average approach**. In particular, each household supply is proportional to its demand. Hence, $s_i = \frac{d_i}{\sum_{i=1}^{3} d_i}\mathcal{Z}_t$.
- **Round Robin (RR):** Each time step, one of the households in a rotation will be guaranteed sufficient supply to cover its demand. Leftovers from the resource will be shared equally with all households. For example, at time $t = 1$, $s_1 = d_1$ while $s_2 = s_3 = \frac{\mathcal{Z}_t - d_1}{2}$. We further augmented the RR to consider a **Weighted RR**. In this case, supplies are calculated similarly to RR, but the leftover water will be distributed proportionally to their demands as in the weighted average approach.
- **No subgoals using 1 DQN :** Supply for each household is calculated by 1 DQN structure as explained in Section V-C.

We investigated multiple evaluation metrics as follows:

*1) **Fairness state and group fairness definition**:* We simulated 62.5 simulated days equivalent to $15k$ samples. To measure s_t, we need to use the satisfaction history counters by updating them per Equation 2. In this application, we set the $\|d_i - (s_i + t_i)\| \leq \tau$ with τ equals 20% of the demand d_i which means BR is 80%. Similar to the analysis we did in application 1, FAIRO achieves *equal opportunity* probabilities 31.9%, 35.4%, and 32.7% for household $\#1$, $\#2$, and $\#3$ respectively, which are closer in values than other methods. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by $\mathbf{26.38}\%$ on average.** As for *equalized odds* probabilities, FAIRO reports 34.4%, 30.1%, and 35.5% for households $\#1, \#2$, and $\#3$ respectively, which are also closer in values compared to the other approaches. **Across all the approaches, FAIRO improves the *equalized odds* fairness by $\mathbf{32.1}\%$ on average.** More numerical details are shown in Figure 9 in the Appendix.

*2) **Satisfaction performance**:* We use $3k$ samples after FAIRO converge ($12k$ to $15k$) to examine the satisfaction value histograms. We report the Jensen-Shannon Divergence (JSD) of these histograms. FAIRO has the lowest average JSD (0.017), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO JSD is reduced by $\mathbf{91.06}\%$ on average.** More numerical details are shown in Figure 11-Right in the Appendix.

*3) **Balance rate (BR) results**:* We report the average number of samples that have $>80\%$ BR across all households and average it over $3k$ samples. Samples have $> 80\%$ BR correspond to satisfactory samples. **Across all the approaches, FAIRO's average $>\mathbf{80}\%$ BR is improved by $\mathbf{46.66}\%$ on average.** More numerical details are shown in Figure 11-Right in the Appendix.

VII. Application Type III: Smart learning system

Monitoring human learning state and performance is crucial for assessing progress and personalizing instruction [27]. Immersive technologies like virtual reality (VR) in education and workforce training show promise for improving learning experiences. However, over extended online or VR education periods, human performance can decline due to distractions, drowsiness, and fatigue [27]. This application examines multiple users sharing the same educational environment, with a HITL learning system that adapts this environment by enabling an immersive VR experience to improve the learning experience.

A. Human-Learning VR Model

We used a real human dataset from recent literature studying VR's impact on learning [22]. The dataset describes human learning experience through three features: alertness, fatigue, and vertigo, resulting in 8 states with binary values (e.g., Alert: 1, Not Alert: 0). A HITL learning environment adapts based on these states using actions like (1) a_1 giving a break, (2) a_2 enabling VR, or (3) a_3 disabling VR.

We analyzed this dataset to categorize 15 participants into three groups based on VR tolerance: the most tolerant, those with some cybersickness, and those with the least VR tolerance. We model these groups' behavior using MDP models (see Figure 10 in Appendix), with state 8 representing the best human state and state 1 the worst. Three humans, one from each profile, were instantiated for a daily class schedule, with their states initialized to state 3 or state 1 randomly at the beginning of each day.

B. Context-aware engine

Unlike the previous two applications, this application has a non-numerical action space, as mentioned in Section VII-A. Hence, to use FAIRO we need to quantify this categorical action in terms of its effect as explained in Section IV-D. We exploit the MDP model in Figure 10 in the Appendix to quantify these actions. In particular, as S_8 and S_0 encode the best and the worst state, respectively, we assign a value to each state linearly with S_8 as the highest value. Hence, the desired action d_i that can make a transition to a better state in the MDP has a high numerical value. However, every human may be in a different state. Hence, for every desired action d_i, we calculate the *effect* denoted as k_i of applying d_i on the *other* humans. Using the MDP models, if the desired action d_i from human h_i were to be applied in the shared environment, we check the state transition it will cause on all the other humans $h_{N\setminus i}$. The difference in the values of the two states in the transition is used to measure its *effect*. After applying an adaptation action, the human learning experience can be measured as the improvement in the human state values and the current human state value.

$$\text{Learn Exp. (LE)} = \text{state improvement} + \text{state value} \in [-1,1] \quad (16)$$

C. Evaluation

- **FAIRO:** The global action $a_{g_t} = d_i$ is the desired action of the human$_i$ with the maximum weighted effect as explained in Section IV-D. The reward is the same as explained in the first application (Equation 11), where performance $\mathcal{P}_i$ is based on the learning experience (LE) ($\mathcal{P}_i = 0.2\frac{v_i}{u_i + v_i} + 0.8f(\text{LE})$).
- **Average approach:** The global action $a_{g_t} = d_i$ is the desired action of the human$_i$ with the median weighted effect.

- **Round Robin (RR):** The global action a_{g_t} is selected from one of the desired actions of all humans in a rotation.
- **No subgoals using 1 DQN:** The weighted effects are determined by using 1 DQN structure (Section V-C).

We investigated multiple evaluation metrics as follows:

*1) **Fairness state results and group fairness definition**:* To measure s_t, we need to use the satisfaction history counters by updating them per Equation 2. If the learning experience (LE) of the human is >0 as measured in Equation 16, it will be considered satisfied. **Across all the approaches, FAIRO improves the *equal opportunity* fairness by 36.35% and *equalized odds* fairness by 30.55% on average.** More details are shown in Figure 12 in the Appendix.

*2) **Satisfaction performance**:* We use $3k$ samples after FAIRO convergence($12k$ to $15k$) to examine the satisfaction value ($\frac{v}{v+u}$) histograms. FAIRO has the lowest average JSD (0.021), indicating closer satisfaction values across rooms. **Across all the approaches, FAIRO satisfaction JSD is reduced by 83.53% on average.** More numerical details are shown in Table I in the Appendix.

*3) **Learning experience (LE) results**:* We count the number of samples with positive LE and average over $3k$ samples. FAIRO achieves comparable LE results. FAIRO LE is improved by 11.4% from Average Approach, reduced by 3.4% from RR, and reduced by 3.2% from 1 DQN 3 inputs.

VIII. Discussion, Limitation, and Conclusion

This paper addresses the challenge of ensuring fairness in Human-in-the-Loop (HITL) systems by breaking it down into manageable subgoals that span a timeline for fairness-aware sequential decision-making. We introduce the FAIRO framework, which is tailored to account for the dynamic nature of human variability and preferences as they evolve over time. Our approach centers on a novel fairness state dedicated to achieving satisfaction equity, grounded in meeting human preferences. This concept of fairness is designed to enhance CPS that involve human interactions by ensuring equitable satisfaction levels. Nevertheless, our current model of fairness presents a limitation as it does not fully consider how interpersonal interactions may influence individuals' perceptions of satisfaction, indicating an area for future enhancement.

Acknowledgment

This research was partially supported by the National Science Foundation (NSF) awards 2105084 and 2339266.

References

[1] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in *Aea papers and proceedings*, vol. 108, 2018, pp. 22–27.

[2] A. Annaswamy, K. Johansson, and G. Pappas, "Control for societal-scale challenges roadmap 2030," 2023.

[3] Northpointe, "Practitioner's guide to COMPAS core," https://www.equivant.com/practitioners-guide-to-compas-core/, 2015.

[4] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, "Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management," *International Transactions in operational research*, vol. 9, no. 5, pp. 583–597, 2002.

[5] E. Creager, D. Madras, T. Pitassi, and R. Zemel, "Causal modeling for fairness in dynamical systems," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2185–2195.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[7] P. P. Khargonekar and M. Sampath, "A framework for ethics in cyber-physical-human systems," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 17 008–17 015, 2020.

[8] L. J. Ratliff and T. Fiez, "Adaptive incentive design," *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3871–3878, 2020.

[9] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 329–338.

[10] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1929–1938.

[11] Y. Zhao, Z. An, X. Gao, A. Mukhopadhyay, and M. Ma, "Fairguard: Harness logic-based fairness rules in smart cities," in *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 2023, pp. 105–116.

[12] S. Elmalaki, "Fair-iot: Fairness-aware human-in-the-loop reinforcement learning for harnessing human variability in personalized iot," in *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, 2021, pp. 119–132.

[13] J. Jiang and Z. Lu, "Learning fairness in multi-agent systems," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 854–13 865.

[14] S. Kannan, A. Roth, and J. Ziani, "Downstream effects of affirmative action," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 240–248.

[15] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[16] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions." in *Robotics: Science and Systems*, 2017.

[17] S. Elmalaki, H.-R. Tsai, and M. Srivastava, "Sentio: Driver-in-the-loop forward collision warning using multisample reinforcement learning," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018, pp. 28–40.

[18] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[19] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 299–308.

[20] R. S. Sutton, D. Precup, and S. Singh, "Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning," *Artificial intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.

[21] R. G. Carroll, "Pulmonary system," in *Elsevier's Integrated Physiology*. Elsevier, 2007, ch. 10, pp. 99–115.

[22] M. Taherisadr, S. A. Stavroulakis, and S. Elmalaki, "adaparl: Adaptive privacy-aware reinforcement learning for sequential decision making human-in-the-loop systems," in *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 2023, pp. 262–274.

[23] ASHRAE/ANSI Standard 55-2010 American Society of Heating, Refrigerating, and Air-Conditioning Engineers, "Thermal environmental conditions for human occupancy," *Inc.Atlanta, GA, USA*, 2010.

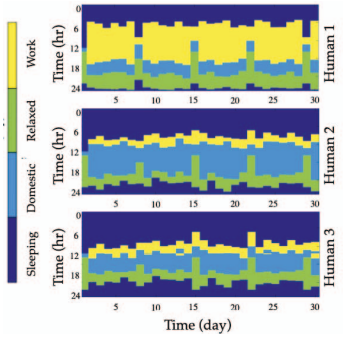
[24] G. M. Sechi, R. Zucca, and P. Zuddas, "Water costs allocation in complex systems using a cooperative game theory approach," *Water Resources Management*, vol. 27, pp. 1781–1796, 2013.

[25] Philadelphia Government, "Gallons used per person per day," https://water.phila.gov/pool/files/home-water-use-ig5.pdf, 2023.

[26] MATLAB, "Water distribution system scheduling using reinforcement learning," https://www.mathworks.com/help/reinforcement-learning/ug/water-distribution-scheduling-system.html, Mar. 2023.

[27] S. Terai, S. Shirai, M. Alizadeh, R. Kawamura, N. Takemura, Y. Uranishi, H. Takemura, and H. Nagahara, "Detecting learner drowsiness based on facial expressions and head movements in online courses," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 2020, pp. 124–125.

APPENDIX



Histogram	Rooms	FAIRO	Average	RR	1 DQN
Satisfaction Gaussian fitting μ,σ^2	Room 1	0.642, 0.047	0.452, 0.030	0.748, 0.021	0.657, 0.142
	Room 2	0.649, 0.050	0.506, 0.027	0.785, 0.023	0.641, 0.084
	Room 3	0.641, 0.047	0.487, 0.030	0.698, 0.024	0.680, 0.152
Satisfaction JSD	Room 1&2	0.022 bits	0.399 bits	0.269 bits	0.115 bits
	Room 1&3	**0.002** bits	0.175 bits	0.443 bits	0.020 bits
	Room 2&3	**0.014** bits	0.080 bits	0.835 bits	0.121 bits
	avg. JSD	**0.013** bits	0.218 bits	0.516 bits	0.085 bits
PMV Gaussian fitting μ,σ^2	Room 1	0.346, 0.490	0.369, 0.430	0.367, 0.590	0.336, 0.502
	Room 2	-0.014, 0.612	0.021, 0.498	0.021, 0.701	-0.030, 0.642
	Room 3	-0.099, 0.673	-0.065, 0.564	-0.052, 0.732	-0.120, 0.723
PMV JSD	Room 1&2	0.088 bits	0.112 bits	**0.071** bits	0.089 bits
	Room 1&3	0.119 bits	0.152 bits	**0.099** bits	0.119 bits
	Room 2&3	0.012 bits	0.023 bits	**0.014** bits	0.014 bits
	avg. JSD	0.073 bits	0.096 bits	**0.063** bits	0.074 bits

Fig. 8: Left: Human activities, Right: Comparison of the satisfaction values $\frac{v}{u+v}$ and the PMV across all the approaches in application 1. Compared with all methods, the JSD of FAIRO satisfaction is reduced by 94.0% from Average Approach, reduced by 97.5% from RR, and reduced by 84.7% from 1 DQN. Similarly, FAIRO's PMV JSD is reduced by 24.0% from Average Approach, increased by 15.9% from RR, and reduced by 1.4% from 1 DQN.

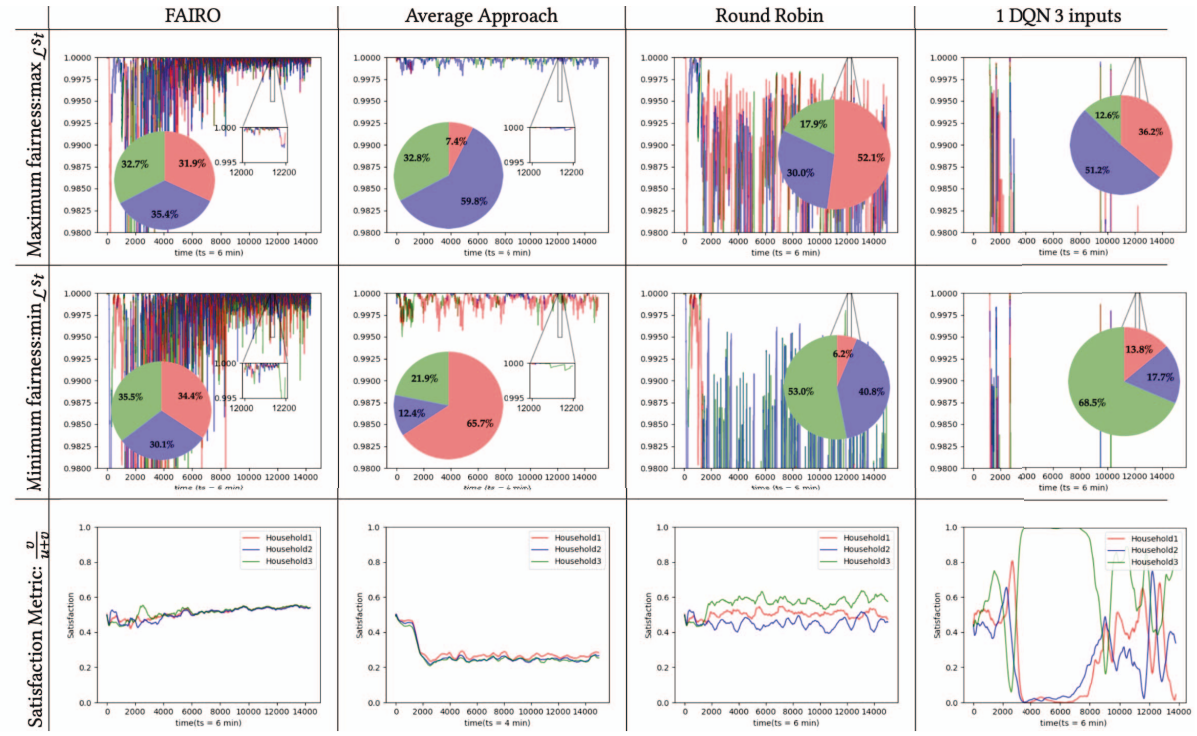


Fig. 9: Fairness state and satisfaction across all approaches for application 2. First and second rows show the maximum and minimum value of s_t. The fairness states results with $15k$ samples equivalent to 62.5 simulated days. Using FAIRO, the average absolute difference between the probabilities of *equalized odds* across the 3 households is reduced by 31.9%, 31.0%, 37.1%, 27.6%, and 32.9% from Weighted Average, Weighted RR, Average, RR,and 1 DQN 3 inputs respectively with an average of 32.1. Third row reports the satisfaction values across the three households. Satisfaction values in FAIRO are closer compared with other approaches with satisfaction values around 0.55. RR satisfaction values are from 0.6 to 0.4 with different between households. The Weighted Average Approach satisfaction values are close, but values stay below 0.3. 1 DQN-3 inputs has unstable fluctuations across 3 households.

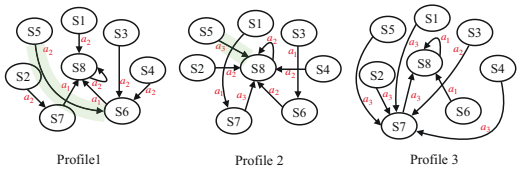
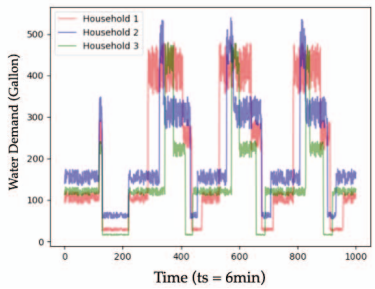


Fig. 10: MDP for three human profiles in VR learning environment.



Histogram	Rooms	FAIRO	Weighted Average	Weighted RR	Average	RR	1 DQN
Satisfaction Gaussian fitting μ,σ^2	Room 1	**0.540**, 0.037	0.267, 0.021	0.459, 0.014	0.214, 0.100	0.496, 0.061	0.342, 0.280
	Room 2	**0.540**, 0.029	0.249, 0.022	0.445, 0.014	0.205, 0.085	0.462, 0.050	0.315, 0.257
	Room 3	0.535, 0.048	0.244, 0.020	0.429, 0.013	0.495, 0.105	0.596, 0.052	**0.636**, 0.241
Satisfaction JSD	Room 1&2	**0.016** bits	0.091 bits	0.079 bits	0.088 bits	0.094 bits	0.048 bits
	Room 1&3	**0.011** bits	0.150 bits	0.427 bits	0.654 bits	0.390 bits	0.237 bits
	Room 2&3	0.024 bits	**0.010** bits	0.165 bits	0.756 bits	0.644 bits	0.288 bits
	JS Average	**0.017** bits	0.084 bits	0.224 bits	0.499 bits	0.376 bits	0.191 bits
Balance Rate (BR)	avg.>80%BR (# samples > 80%BR)	**0.535(1606)**	0.254(763)	0.446(1336)	0.305(915)	0.517(1552)	0.432(1296)

Fig. 11: Left: Water demand profile for three households during three days, Right: Comparison of the satisfaction values $\frac{v}{u+v}$ and the balance rate (BR) across all the approaches in application 1. In particular, for satisfaction, compared with other methods, FAIRO JSD is reduced by 79.8%, 92.4%, 96.5%, 95.5%, and 91.1% from Weighted Average Approach, Weighted RR, Average Approach, RR, and 1 DQN 3 inputs methods respectively with an average of 91.06%. In terms of BR, FAIRO has the highest average $>80\%$ BR value (0.535). Compared with other methods, FAIRO's average $>80\%$ BR is improved by 110.6%, 20.0%, 75.4%, 3.5%, and 23.8% from Weighted Average Approach, Weighted RR, Average Approach, RR, and 1 DQN 3 inputs methods respectively with an average of 46.66%.

Fig. 12: Fairness state and satisfaction metric across all different approaches for application 3. First and second rows show the maximum and minimum value of s_t. FAIRO achieves *equal opportunity* probabilities 42.1%, 24.0%, and 33.9% for human #1, #2, and #3 respectively. As for *equalized odds* probabilities, FAIRO reports 24.3%, 49.4%, and 26.3% for humans #1, #2, and #3 respectively. While the difference in these values across the three humans is not as small as we had in the other two applications, we observe that this difference in FAIRO is better than the other approaches.

Histogram	Rooms	FAIRO	Average	RR	1 DQN
Satisfaction Gaussian fitting μ,σ^2	Room 1	0.639, 0.017	0.761, 0.033	0.594, 0.015	0.692, 0.062
	Room 2	0.647, 0.020	0.821, 0.020	0.671, 0.018	0.787, 0.050
	Room 3	0.638, 0.018	0.554, 0.020	0.583, 0.012	0.604, 0.094
Satisfaction JSD	Room 1&2	**0.022** bits	0.528 bits	0.801 bits	0.396 bits
	Room 1&3	**0.000** bits	1 bits	0.060 bits	0.229 bits
	Room 2&3	**0.021** bits	1 bits	0.884 bits	0.657 bits
	JS Average	**0.094** bits	0.843 bits	0.582 bits	0.427 bits
Learning Experience (LE)	Overlap Samples %	0.488	0.438	0.507	**0.497**

TABLE I: Comparison of the satisfaction values $\frac{v}{u+v}$ and the learning experience (LE) in application 3. FAIRO satisfaction JSD is reduced by 88.8%, 83.8%, and 78.0% from Average Approach, RR, and 1 DQN respectively.