$\mathrm{F}\mathit{in}\mathrm{A}$: Fairness of Adverse Effects in Decision-Making of Human-Cyber-Physical-System

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

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

***Abstract*—Ensuring fairness in decision-making systems within Human-Cyber-Physical-Systems (HCPS) is a pressing concern, particularly when diverse individuals, each with varying behaviors and expectations, coexist within the same application space, influenced by a shared set of control actions in the system. The long-term adverse effects of these actions further pose the challenge, as historical experiences and interactions shape individual perceptions of fairness. This paper addresses the challenge of fairness from an equity perspective of adverse effects, taking into account the dynamic nature of human behavior and evolving preferences while recognizing the lasting impact of adverse effects. We formally introduce the concept of Fairness-in-Adverse-Effects (F*in*A) within the HCPS context. We put forth a comprehensive set of five formulations for F*in*A, encompassing both the instantaneous and long-term aspects of adverse effects. To empirically validate the effectiveness of our F*in*A approach, we conducted an evaluation within the domain of smart homes, a pertinent HCPS application. The outcomes of our evaluation demonstrate that the adoption of F*in*A significantly enhances the overall perception of fairness among individuals, yielding an average improvement of 66.7% when compared to the state-of-the-art method.**

***Index Terms*—human-cyber-physical-systems, fairness, decision-making, adverse-effect**

I. Introduction

The ubiquitous integration of smart technologies into our daily lives offers unprecedented opportunities but also presents a lot of challenges. A key challenge lies in understanding the interplay between humans and Cyber-Physical Systems (CPS) to shape the societal consequences of future CPS technologies. As we move towards a future defined by the coexistence of humans and smart technologies, understanding their interactions is essential, as suggested by recent studies [1], [2]. Within the realm of Human-Cyber-Physical Systems (HCPS), one central challenge emerges when CPS decisions can affect individuals with diverse preferences and perceptions within the same environment. This scenario is common in systems like smart buildings, smart cities, smart traffic management, and smart crowd control, where fairness, privacy, equity, and personalization issues intersect, sparking new societal tensions [3]–[5].

Existing research in sociology, particularly Social Exchange Theory [6], [7] and Equity Theory [8] has substantiated a direct link between "equity" and prosocial behavior. The higher an individual perceives a system as equitable, the greater the likelihood of that individual engaging in prosocial behavior [8]–[10]. This, in turn, significantly impacts the overall performance of the system. Specifically, the greater the number of individuals who engage in prosocial behavior, the higher the likelihood of compliance and acceptance of the system's decisions, ultimately leading to an enhancement in overall system performance [11]–[13]. Given that CPS decision-making often aims to optimize system performance while adhering to operational constraints, it is essential to develop formal metrics and objective functions that enhance the human perception of these decisions, ultimately fostering more prosocial behavior and improving overall system performance.

In this paper, we are interested in formalizing some of the equity objectives in decision-making HCPS. We will start by exploring a notion of fairness from the equity perspective. In particular, we will focus on what we term **Fairness-in-Adverse-Effects (F*in*A)**. Decision-making agents in HCPS employ a range of control actions. However, these control actions can have different adverse effects on a diverse population, each with its own preferences. Hence, the HCPS needs to adapt its decision-making to continuously match different populations across time and ensure that it meets the preferences of as many populations as possible. Our motivation in examining the adverse effects or the harmful impact of decision-making in HCPS stems from the psychology concept **"loss aversion**–Losses loom larger than gains" which implies that losses can be twice as powerful, psychologically, as gains [14]. This concept underscores the significance of minimizing adverse effects to promote fairness within HCPS.

II. Related Work

HCPS systems are centered on the challenge of designing adaptable, real-time decision-making processes that take into account the social context, including considerations of fairness, social welfare, ethical concerns, and societal norms [15], [16]. A substantial body of work in the field of game theory explores various facets of fairness, often framed as incentive markets among competing entities or communities striving to achieve fairness [17], [18]. In the domain of machine learning, interventions to enhance fairness have been introduced, aiming to ensure that models' decisions are devoid of discrimination [19]–[23]. In the context of decision-making systems, where agents express favoritism for one action over another, questions surrounding fairness become even more significant, especially within multi-agent systems [24]–[31]. However, imposing fairness constraints as static, one-time decisions akin to conventional supervised learning methods while neglecting

979-8-3503-6927-4/24/$31.00 ©2024 IEEE
DOI 10.1109/ICCPS61052.2024.00025

dynamic feedback and long-term consequences, particularly in sequential decision-making systems, can inadvertently lead to disparities that affect specific sub-populations [32], [33].

Recent research has also shed light on the long-term ramifications of Reinforcement Learning (RL), revealing that addressing control decisions' immediate effects in single steps does not ensure fairness in subsequent decision actions [34], [35]. Nevertheless, a significant portion of this research has predominantly focused on fairness through the lens of equality, with an emphasis on eliminating favoritism or bias within the system, and relatively less attention has been given to the concept of fairness from an equity perspective, particularly in the context of sequential decision-making [36]. Notably, ensuring fairness in sequential decision-making systems becomes increasingly complex as policies deemed fair at one point may inadvertently become discriminatory over time due to shifts in human preferences influenced by the inter- and intra-human variation [37].

The concept of "group fairness" has been introduced in the literature to tackle fairness concerns arising when the same adaptive model impacts multiple individuals. One notion is "Equalized Odds" which concentrates on achieving a level of uniform prediction accuracy across various groups, primarily within the context of binary classification tasks. The central objective is to ensure that a predictive model exhibits comparable true positive rates (sensitivity) and true negative rates (specificity) across diverse groups [38]. A second notion is "Equal opportunity,", which aims to guarantee that a predictive model affords an equal likelihood of beneficial outcomes for all groups. It places a specific requirement on the true positive rate, mandating that it should be approximately equivalent for each group [38].

Prior research in CPS has explored fairness in various ways. For instance, fairness-aware resource allocation and scheduling algorithms have been developed for CPS, addressing issues like energy consumption and real-time constraints while considering equitable distribution among system components [39]. The concept of fairness in communication protocols for CPS has been established through strategies to ensure fair access to network resources for different devices and applications [40]. Furthermore, fairness challenges in decentralized CPS environments have been examined, focusing on ensuring fair decision-making in multi-agent systems [41].

However, it's worth noting that much of the existing CPS literature concentrates on system-level performance and efficiency, often at the cost of individual-level fairness considerations. In contrast, this paper aims to delve deeper into the aspects of fairness within HCPS, addressing the interplay between human preferences, the temporal dimension of adverse effects, and perceptions of fairness. This approach allows for a more comprehensive understanding of fairness in the context of HCPS decision-making, particularly in systems where individuals' preferences and perceptions can evolve over time.

A. Paper contribution

This paper's contributions can be summarized as follows:

- **Fairness-in-Adverse-Effect (F*in*A)**: In this paper, we introduce a novel concept known as "Fairness-in-Adverse-Effect (F*in*A)." F*in*A takes a fresh perspective by considering equity in adverse effects within Human-Cyber-Physical Systems (HCPS) decision-making. We address scenarios where adaptive decisions made within HCPS can impact multiple individuals with diverse preferences. By formalizing F*in*A, we provide a means to ensure that the adverse effects of decision-making are distributed fairly among the system's users
- **Long-term effects:** Our work extends beyond the immediate effect of CPS decision-making. We delve into the temporal dimension of adverse effects, recognizing that the impact of these decisions can have lasting consequences. The interplay between human preferences, historical data, and the evolving perception of fairness adds complexity to the notion of fairness. We introduce five different approaches to formalize F*in*A within CPS decision-making to account for the relationship between the instantaneous adverse effects, long-term adverse effects, and the overall perception of fairness.
- **Generalization to different HCPS setups:** We acknowledge that the nature of adverse effects and fairness considerations can vary across different domains. Therefore, we offer a general formalization of F*in*A that can be applied to various HCPS scenarios. Additionally, we conduct thorough evaluations in the domain of smart home to illustrate the trade-offs between various interpretations and implementations of F*in*A. This demonstrates the flexibility and effectiveness of our approach across diverse HCPS settings.

The rest of the paper is organized as follows: We introduce the notion of Fairness-in-Adverse-Effects (F*in*A) within Human-Cyber-Physical Systems (HCPS) in Section III. We formally define F$in$A using five different approaches in considering the instantaneous and the long-term adverse effects while considering the human perception of fairness in Sections III-B, III-C, III-D, III-E and III-F. Afterward, we numerically evaluate these approaches using a smart home HCPS application in Section IV.

III. Approach

We consider a CPS depicted in Figure 1, which serves multiple individuals sharing the same CPS environment, each with different preferences. The control action generated by the decision-making agent in CPS can cause different adverse effects on those individuals.

To achieve Fairness-in-Adverse-Effects (F*in*A) within Human-Cyber-Physical Systems (HCPS), we propose five distinct approaches to guide CPS decision-making when selecting control actions affecting multiple humans sharing the same environment. These approaches are rooted in the recognition that individuals who exhibit pro-social attitudes might be willing to tolerate certain discrepancies between their preferences and CPS actions for a limited duration [9]. However, it is important to acknowledge that this willingness to forgive may not be indefinite, as the magnitude and persistence of discrepancies play a pivotal role in shaping individuals' perception of

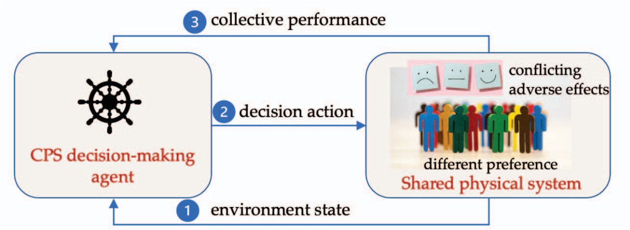


Fig. 1: Human-Cyber-Physical-System with multiple individuals sharing the same environments with different preferences. The decision-making agent control action can cause different adverse effects on those individuals.

fairness [9]. Moreover, the extent of an individual's willingness to forgive is contingent upon the CPS's responsiveness in addressing and rectifying such discrepancies [7].

Our first approach formalizes F*in*A by examining the concept of instantaneous adverse effects, focusing on the immediate impact of CPS control actions (Section III-B). In this context, we recognize that individuals may exhibit a degree of patience when faced with minor or unintentional discrepancies between their preferences and CPS actions. This approach is predicated on the assumption that in scenarios of minimal and short-term discrepancies, individuals may still perceive the CPS as acting in their best interest.

Nonetheless, the second approach acknowledges that as the severity and persistence of discrepancies increase, individuals, even those with pro-social attitudes, may become less forgiving (Section III-C). Thus, the historical aspect of adverse effects becomes a critical factor. It is during extended periods of inconsistency that individuals may experience a diminishing willingness to tolerate disparities. This approach takes into account the temporal dimension of adverse effects, recognizing that extended discrepancies may erode the perception of fairness [42].

In our third approach, we examine a balance between the instantaneous adverse effect and the historical record of adverse effects (Section III-D). By combining these two dimensions, we aim to provide a tradeoff that accounts for both short-term variations and long-term consequences. This approach is particularly valuable in situations where CPS must navigate the delicate balance between immediate and lasting impact.

Building on the well-established literature on fair resource distribution, we acknowledge that fairness is not synonymous with equal resource allocation [42]. Humans' perception of fairness is intrinsically tied to how they compare their resource distribution with that of others in the same system. In the fourth approach, we consider the collective perception of fairness among individuals who share the same CPS environment as a metric for formalizing F*in*A (Section III-E). This approach recognizes that individuals may be more forgiving of discrepancies if they perceive that others are experiencing similar variations in resource allocation.

Lastly, our fifth approach introduces a tradeoff between human perception of fairness and a budget of allowable discrepancy between individual preferences and the applied CPS control action (Section III-F). By imposing limits on the magnitude of permissible discrepancies, this approach provides a tradeoff between accommodating individual preferences and ensuring collective fairness.

A. Fairness-in-Adverse-Effect (FinA) Setup

In all of these five approaches, we consider a society S that consists of N different individuals and a CPS that serves this society by providing shared control actions a that may be tailored toward the preferences and behavior of some of those individuals. We assume that every individual $n \in N$ has a different set of g preferred actions $A_n = \{a_1^n, a_2^n, \ldots, a_g^n\}$ that can serve them better[1]. Suppose an adverse effect $v_n(a)$ on individual n occurs due to the chosen control action a. An initial mechanism to measure the **adverse effect** on each $v_n(a)$ is to assume that a preferred action $a_g^n \in A_n$ is inversely proportional to its adverse effects. That is, to measure the adverse effects of a chosen control action $v_n(a)$ on human n, we can use the distance between the set of preferred actions $\mathcal{A}_n$ and the chosen control action a, i.e.,

$$v_n(a) = \max_{a_g^n \in \mathcal{A}_n} \|a_g^n - a\|. \tag{1}$$

B. Approach I: Formalizing the definition of FinA with instantaneous effect

In our first approach, we formalize F*in*A by examining instantaneous adverse effects, delving into the immediate consequences of CPS control actions.

Hence, an initial definition of F*in*A can be:

$$\text{F}\mathit{in}\text{A} = \min_{a \in \mathcal{A}} \|\mathbf{v}(a) - \frac{1}{N}\mathbb{1}^T\mathbf{v}(a) \otimes \mathbb{1}\| + \lambda \|\frac{1}{N}\mathbb{1}^T\mathbf{v}(a)\|, \tag{2}$$

where $\mathbf{v}(a) = [v_1(a), v_2(a), \ldots, v_N(a)]^\intercal$ and $\mathcal{A} = \bigcup_{n=1}^N A_n$. In other words, Equation (2) aims to choose the control action a, out of all possible A_n, that minimizes the difference between the individual adverse effects on every individual v_n and the average of the adverse effects across all individuals $\frac{1}{N}\sum_{n=1}^N v_n(a) = \frac{1}{N}\mathbb{1}^T\mathbf{v}(a)$. Indeed, one trivial solution will be to increase the adverse effect on all individuals to achieve the same average. Therefore, the second term in Equation (2) asks that the chosen control action also aims to minimize the average adverse effect [2].

C. Approach II: Formalizing the definition of FinA using long-term temporal variations in adverse effect

In the second approach, the historical context of adverse effects plays a pivotal role. Prolonged instances of inconsistency are where individuals may exhibit a reduced willingness to endure disparities. This approach places a significant emphasis on the temporal dimension of adverse effects, acknowledging that extended discrepancies can undermine the perception of fairness [42].

We define a long-term adverse effect $\mathbf{v}_n$ by monitoring the adverse effect for every applied action a over a time horizon T on human n as follows:

[1] While in our first definition A_n we assume a discrete set of preferred actions, this can be extended to a continuous range of preferred actions.

[2] We used L^2 norm for all $\|\bullet\|$ notations.

$$\mathbf{v}_n = [v_n^0, v_n^1, ..., v_n^{T-1}]^\intercal, \tag{3}$$

where v_n^j represents the adverse effect occurred at time j for human n. However, to focus more on the recent adverse effects, we assign different weights to v_n^j and calculate an accumulated value u_n^t that represents the current history of adverse effects on human n from time $t-T$ till $t-1$. Accordingly, by assigning the weight $\frac{j}{T}$ to v_n^j, the most recent adverse effects have more contribution to the the accumulated value u_n^t.

$$u_n^t = \frac{1}{T}\sum_{j=0}^{T-1} \frac{j}{T} v_n^j, \text{ for } n = 1,2,...,N \tag{4}$$

The historical adverse effect on all N individuals can be represented by:

$$\mathbf{u} = [u_1^t, u_2^t, ..., u_N^t]^\intercal \tag{5}$$

We can then formally define F$in$A as:

$$\begin{aligned} \text{F}in\text{A} = &\min_{a\in\mathcal{A}} \mathcal{B} \\ &s.t.\ \mathbf{v}(a) < \mathcal{B} - \mathbf{u} \end{aligned} \tag{6}$$

In this equation, $\mathbf{v}(a) = [v_1(a), v_2(a), ..., v_N(a)]^\intercal$ represents the current adverse effect of action a on each individual. To elaborate, Equation (6) contains N constraints, with each constraint governing the current adverse effect $v_n(a)$ to remain below a specific budget $\mathcal{B}$. The intuition here is to use this budget $\mathcal{B}$ as the total amount of adverse effect in the past T history. However, it's essential to note that this budget $\mathcal{B}$ is gauged on the historical adverse effects of each individual, denoted as u_n^t. In this context, F$in$A tries to identify the minimum budget $\mathcal{B}$ that satisfies all N constraints. Consequently, if a human individual, say n, has a substantial historical adverse effect value, the constraint $\mathcal{B} - u_n^t$ will direct the optimization process toward finding an action a that results in a minimal $v_n(a)$. This approach is designed to steer the optimization's focus towards individuals with higher historical adverse effect values, encouraging the selection of new actions that minimize their adverse effects.

D. Approach III: Formalizing the definition of FinA as a tradeoff between instantaneous and long-term adverse effect

In our third approach, we delve into the balance between immediate adverse effects and cumulative historical adverse effects. To achieve this, we augment Equation (2) with a term that considers the historical adverse effects, denoted as u_n^t as defined in Equation (4).

The extended formulation can be expressed as:

$$\begin{aligned} \text{F}in\text{A} = \min_{a\in\mathcal{A},\mathbf{b}} \ &\alpha\Big(\|\mathbf{v}(a) - \frac{1}{G}\mathbb{1}^T\mathbf{v}(a) \otimes \mathbb{1}\| + \lambda\|\frac{1}{G}\mathbb{1}^T\mathbf{v}(a)\|\Big) \\ &+ (1-\alpha)\mathbf{u}^\intercal\mathbf{b}, \\ s.t.\ &\mathbf{v}(a) < \mathbf{b} \end{aligned} \tag{7}$$

In this formulation, $\mathbf{b} = [b_1, b_2, ... , b_N]^\intercal$ and $\mathbf{u} = [u_1^t, u_2^t, ..., u_N^t]^\intercal$. Essentially, the first term in Equation (7) is from the instantaneous adverse effect (Equation (2)), and then we introduce N new optimization variables, b_n, for each individual n, which represents the budget allocated for the adverse effect ($v_n(a)$) for each individual n. Importantly, these budgets $\mathbf{b}$ are weighted by the values of the long-term adverse effects $\mathbf{u}$.

In other words, we have N constraints for different budgets for adverse effects that are bounded for every individual ($\mathbf{v}(a) < \mathbf{b}$). Accordingly, F$in$A needs to minimize these budgets $\mathbf{b}$ to minimize the overall adverse effects. However, the upper bound for these budgets is weighted by the accumulated historical adverse effects. Hence, the term $\mathbf{u}^\intercal\mathbf{b}$ is appended to the definition of F$in$A. To modulate the trade-off between the immediate and the long-term adverse effects, we introduce parameter α where $\alpha \in [0,1]$.

E. Approach IV: Formalization the definition of FinA using human perception of fairness

Drawing upon extensive research in the realm of equitable resource distribution, we recognize that fairness does not necessarily mean equal resource allocation [42]. Instead, the perception of fairness in individuals is fundamentally linked to how they gauge their own resource distribution concerning that of their peers within the same system. In particular, fairness in this setup can be viewed generally as "variances" [42] of the "utility" shared by individuals. Hence, we exploit the notion of the coefficient of variation (CoV) of the utilities [42]. In our setup, we define this utility for human n at time t as the temporal accumulated adverse effect caused by the control actions of the CPS in the shared environment, denoted by u_n^t as expressed in Equation (4).

$$CoV_{\mathbf{u}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\frac{(u_n - \bar{\mathbf{u}})^2}{\bar{\mathbf{u}}^2}} \tag{8}$$

In Equation (8), $\bar{\mathbf{u}} = \frac{1}{N}\sum_{n=1}^{N} u_n^t$ represents the average utility (accumulated adverse effect at time t) of all N humans. The system is said to be more fair if and only if the CoV is smaller. The value of CoV can be anywhere between 0 and infinity. Hence, we use the fairness index (FI) transformation to have a value between 0 and 1 to be easily interpreted as a fairness percentage. In other words, if FI is 1, it means the system is 100% fair, otherwise, if disparity increases between individuals, this FI value will decrease [42].

$$FI_{\mathbf{u}} = \frac{1}{1 + CoV_{\mathbf{u}}^2} \tag{9}$$

Accordingly, we can define F$in$A to maximize the fairness index. The core idea is to minimize the reciprocal of this fairness index, represented as y in the optimization.

$$\begin{aligned}
&\mathrm{F}in\mathrm{A} = \min_{a \in \mathcal{A}} \; y \\
&s.t. \; y \geq 1 + \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}}\right)^2 \\
&\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \leq \epsilon, \text{ for } n = 1,2,...,N.
\end{aligned} \tag{10}$$

where:

$$u_n = u_n^t + v(a)$$

In this formulation, u_n^t is a constant value that represents the accumulated history of the adverse effect up till time $t-1$ (Equation (4)) before applying the new a that will add the new adverse effect $v(a)$ on human n. We also add the constraint $\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}} \leq \epsilon$, for $n = 1, 2, \ldots, N$ that sets a limit on how much each individual's adverse effect can deviate from the average adverse effect. The parameter ϵ defines the maximum allowable difference.

F. Approach V: Formalizing the definition of FinA using the human perception of fairness with a tradeoff for a budget for long-term adverse effect

Lastly, our fifth approach introduces a tradeoff between human perception of fairness and a budget of allowable discrepancy between individual preferences and the applied CPS control action. In particular, we combine our definition of F$in$A in Equation (6) and Equation (10) to provide a tradeoff between fairness index and adverse effect budget $\mathcal{B}$.

$$\begin{aligned}
&\mathrm{F}in\mathrm{A} = \min_{a \in \mathcal{A}} \; \alpha.y + \beta.\mathcal{B} \\
&s.t. \; y \geq 1 + \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|u_n - \bar{\mathbf{u}}|}{\bar{\mathbf{u}}}\right)^2 \\
&\mathbf{v}(a) < \mathcal{B} - \mathbf{u}
\end{aligned} \tag{11}$$

where:

$$u_n = u_n^t + v(a)$$

The α and β weights allow us to adjust the tradeoff between fairness index and adverse effect budget.

IV. Evaluation

We designed an HCPS application in the domain of smart house. Recent literature focuses on enhancing human satisfaction in smart heating, ventilation, and air conditioning (HVAC) systems by employing various techniques to adjust the set-point based on human activity and preferences [37], [43]. These HCPS systems consider the current state and individual preferences, such as body temperature changes during sleep or physical activity. To evaluate different approaches of F$in$A in this application, we consider a setup where multiple humans share a house with a single HVAC system, and their activities determine individual HVAC set-point preferences. Humans can be in the same room or separated rooms as long as they are in the same shared application space that is controlled by the same HVAC.

We exploited recent work in the literature [37] that simulated 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 running the optimization of F$in$A. A pictorial figure of the application setup is shown in Figure 5.

We implemented our proposed five approaches using CVXPY, a Python-embedded modeling language for convex optimization problems [44].

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 [45]. We simulated three humans with four activities: sleeping, relaxing, medium domestic work, and working from home. Randomness was introduced by allowing multiple activity choices during the same time slot. The different activity schedules depicted in Figures 2, 3, and 4. The humans were simulated in separate rooms as seen in Figure 5, 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, working from home, domestic activities, and relaxation. The Mathworks thermal house model was extended to include a cooling system and a human model[3].

The desired preferred action (temperature setpoint) per human $a_g^n = T_n$, for $n = 1,2$, and 3, can be obtained through fixed policy configuration. We exploit existing approaches [47] for estimating the desired HVAC setpoint based on activity and thermal comfort. 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 [48]. The shared control action space (applied temperature setpoint) is all possible temperatures ranging from the minimum preferred action to the maximum preferred action as defined in Section III-A.

A. Experiment setup

In this application, we used the difference between the desired temperature (T_d) and the applied temperature (T_a) as a measure of the adverse effect:

$$v(a) = \|T_a - T_d\|_2, \text{ where } T_a \in [60-80]°F$$

We use the most recent 100 samples for our history of adverse effects for the three humans $\mathbf{v}_1$, $\mathbf{v}_1$ and $\mathbf{v}_3$ (as explained in Equation (3)) with sampling time $t_s = 6$ min. Hence, every t_s, we compute $\mathbf{u} = [u_1, u_2, u_3]^\intercal$

[3] While more complex simulators like EnergyPlus [46] exist, considering energy consumption and electric loads, we opted for a simpler model to assess F$in$A.

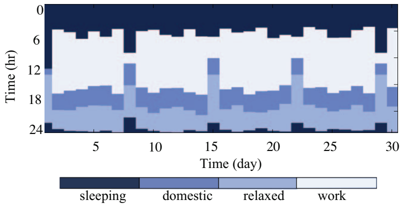


Fig. 2: Human 1 activity pattern.

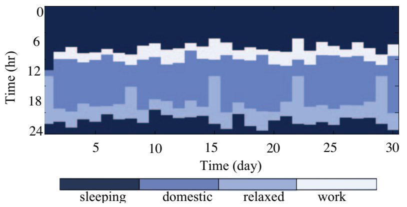


Fig. 3: Human 2 activity pattern.

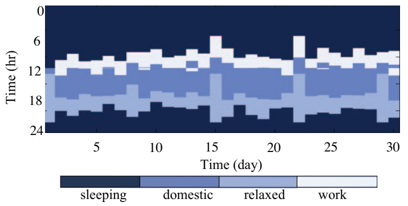


Fig. 4: Human 3 activity pattern.

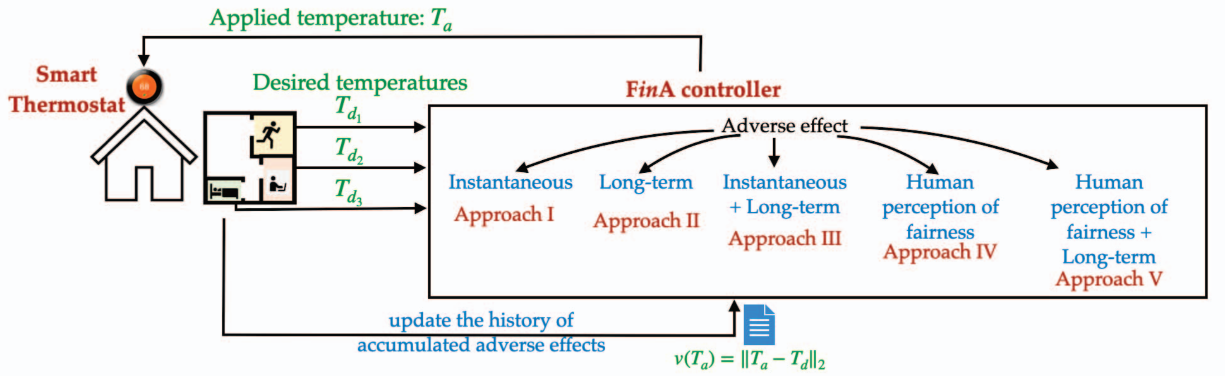

Applied temperature: T_a
Smart Thermostat
Desired temperatures
T_{d_1}
T_{d_2}
T_{d_3}
FinA controller
Adverse effect
Instantaneous
Approach I
Long-term
Approach II
Instantaneous + Long-term
Approach III
Human perception of fairness
Approach IV
Human perception of fairness + Long-term
Approach V
update the history of accumulated adverse effects
$v(T_a) = \|T_a - T_d\|_2$

Fig. 5: A smart house with three humans. Each human has a different activity, which requires different desired indoor temperature setpoints T_d. An external controller running F$in$A selects the applied setpoint T_a based on the calculations of instantaneous and long-term adverse effects.

for the three humans (as explained in Equation (5)), where $u_n = \frac{1}{100}\sum_{j=0}^{99}\frac{j}{100}v_n^j$, and $v_n^j = \|T_a^j - T_{d_1}^j\|_2$ for $n = 1,2$, and 3.

The simulation was executed using 3,000 samples, roughly equivalent to approximately 12 days. This extended duration allowed us to accurately capture changes in the behavioral patterns of the three individuals. At every time step, we check if the desired temperatures of the three humans are not identical then we run the optimization solver for F$in$A then update all the corresponding u.

We consider that the human is satisfied if the T_a is within $2.5°F$ difference from the desired temperature. We measure the satisfaction rate by considering the 100 sample window in $\mathbf{v}_n$. Hence, the satisfaction rate (SR) for human n is computed every t_s computed as:

$$SR_n = \sum_{j=0}^{99}\mathbb{1}(v(a) \leq 2.5), \text{ for all } v(a) \in \mathbf{v}_n$$

Hence, the SR can give us a measure in $\%$ since the total number of samples we consider is 100. We use this SR to compute CoV and FI as a function of the SR similar to Equations (8) and (9) respectively.

$$CoV_{\mathbf{SR}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\frac{(SR_n - \overline{\mathbf{SR}})^2}{\overline{\mathbf{SR}}^2}}, \tag{12}$$

where $N = 3$ and $\mathbf{SR} = [SR_1, SR_2, SR_3]^\intercal$ computed for the three humans every t_s. Similarly, we consider the $FI_{\mathbf{SR}}$ as follows:

$$FI_{\mathbf{SR}} = \frac{1}{1 + CoV_{\mathbf{SR}}^2} \tag{13}$$

Furthermore, we compared the five proposed approaches for F$in$A with two more approaches:

- **Mean approach:** In this case, the applied temperature T_a is the mean of the desired temperature from the three humans.
- **Round Robin**: In this case, the applied temperature T_a is selected in rotation between the desired temperature from the three humans.
- **FaiRIoT [37]**: We compare with the state-of-the-art FaiRIoT, which uses hierarchical reinforcement learning to assign weights to the desired actions to compute the applied action. Hence, $T_a = \sum_{n=1}^{3} w_n T_{d_n}$.

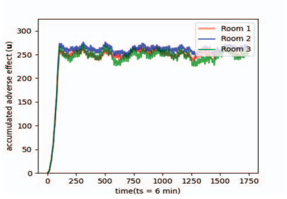
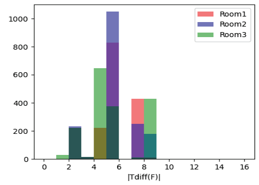
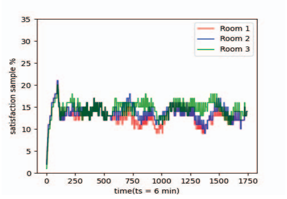
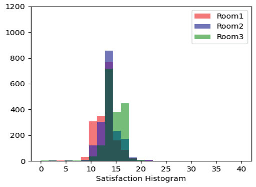
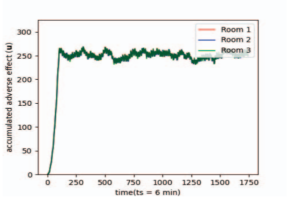
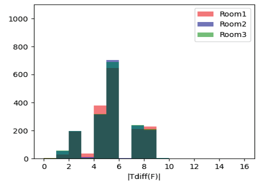
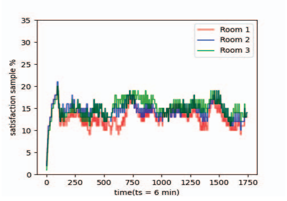
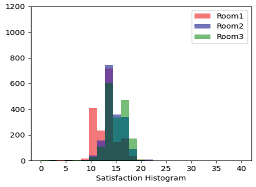
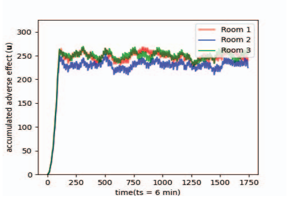
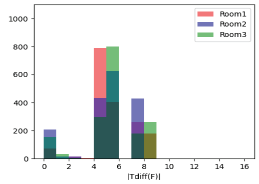
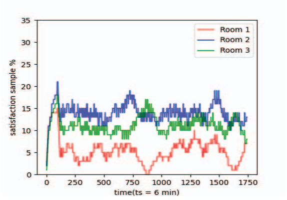
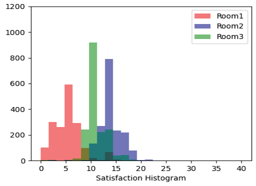
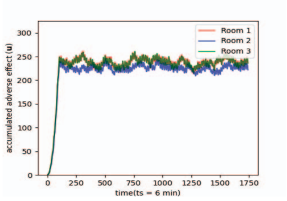
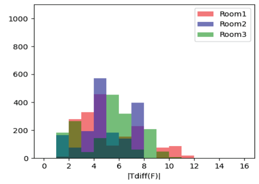
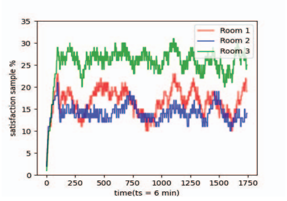
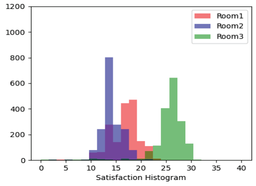
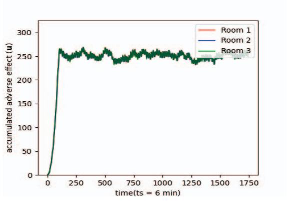
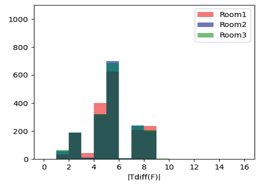
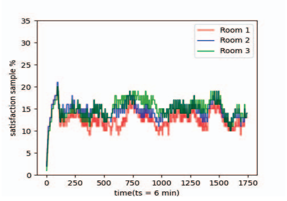
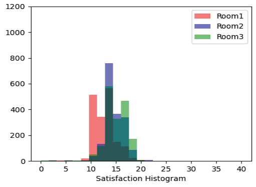
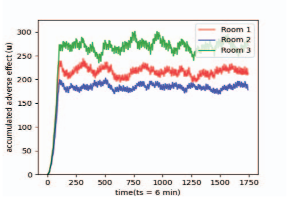
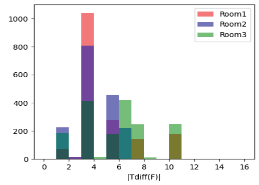
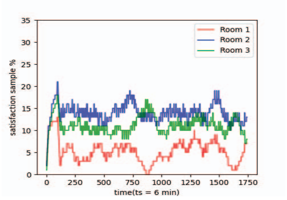
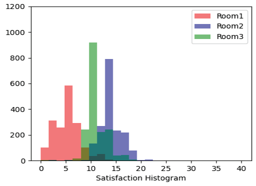
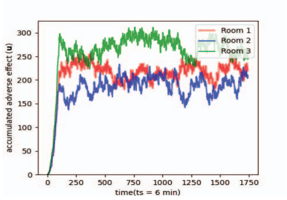
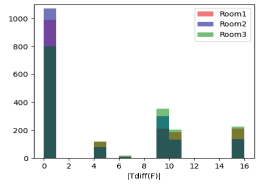
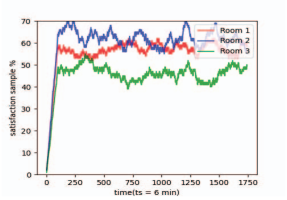
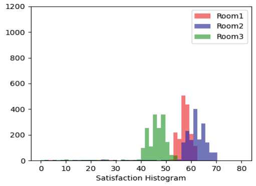
In all experiments, we set the trade-off parameter $\alpha = 0.5$, as defined in Equation 7.

B. Results

We plot in Table I, the accumulated adverse effect ($\mathbf{u} = [u_1, u_2, u_3]^\intercal$), the histogram of the absolute temperature difference between $|T_{diff}| = |T_a - T_d|$, the satisfaction rate (SR), and the histogram of the satisfaction rate (SR), across all approaches for the three individuals in three rooms. First column in Table I compares the differences in the individual's adverse effect ($\mathbf{u} = [u_1, u_2, u_3]^\intercal$) across all approaches. Approach II and V show the smallest difference which is also reflected in average $CoV_{\mathbf{u}}$ in Table II. Approach IV has a higher $CoV_{\mathbf{u}}$ (0.027) but it can bound $\mathbf{u}$ within a smaller value compared with other approaches observed in Table I.

We compare the distribution of $|T_{diff}|$ across all the approaches in Table I second column. Approach II has the highest overlap percentage 86.5% as calculated in Table II which indicates that this approach can make all 3 rooms have a more similar experience compared with other approaches.

TABLE I: Comparison between all the different five approaches of $FinA$, Mean approach, and Round Robin.

	Accumulated adverse effect($\mathbf{u}$)	Hist. temperature difference $\lvert T_{diff}\rvert$	Satisfaction rate (SR) %	Hist. satisfaction rate (SR)
Approach I				
Approach II				
Approach III				
Approach IV				
Approach V				
Mean				
Round Robin				

Round robin (RR) has a large overlap percentage due to the fact that each room can have a $|T_{diff}| = 0$ on its turn in the round. However, RR will result in significant $|T_{diff}|$, which is larger than $10°F$ in a notable number of the samples.

Table I third and fourth columns present the satisfaction rate (**SR**) across all approaches. We report the Jensen-Shannon Divergence (JSD) of the histogram for **SR** in Table II. Approach II has the lowest JSD, indicating closer SR across

TABLE II: Comparison of the overlap area percentage, Satisfaction JSD , and average Fairness Index (FI) and the average coefficient of variation (CoV) of adverse effect($\mathbf{u}$) and satisfaction($\mathbf{SR}$), respectively.

	$\|T_{diff}\|$ overlap%	SR JSD	$Avg.$ $FI_{\mathbf{u}}$	$Avg.$ $CoV_{\mathbf{u}}$	$Avg.$ $FI_{\mathbf{SR}}$	$Avg.$ $CoV_{\mathbf{SR}}$
Appr. I	22.4%	0.086	0.998	0.026	**0.994**	**0.057**
Appr. II	**86.5%**	**0.010**	**0.999**	**0.004**	**0.994**	0.066
Appr. III	37.6%	0.639	0.998	0.038	0.870	0.365
Appr. IV	19.2%	0.659	0.998	0.027	0.929	0.624
Appr. V	83.4%	0.139	**0.999**	**0.004**	0.992	0.077
Mean	24.8%	0.648	0.974	0.157	0.868	0.370
RR	68.4%	0.723	0.973	0.160	0.984	0.124

rooms. RR has the highest overall SR but it has the highest JSD indicating no fairness in the SR among 3 rooms[4].

We summarize the average values of FI and CoV for long-term adverse effect u and satisfaction rate SR in Table II with more detailed results shown in Table III. The fairness index (FI) is a metric ranging from 0 to 1, where 1 means absolute fair as explained in Section III-E. In particular, as shown in Table III, Approach I, II, III, and IV have $FI_{\mathbf{u}}$ values close to 1 and their $CoV_{\mathbf{u}}$ values are less than 0.04. On the contrary, Mean and RR have $FI_{\mathbf{u}}$ around 0.97 and a $CoV_{\mathbf{u}}$ of 0.16.

C. Comparison between these approaches

Approach II and V have the best $FI_{\mathbf{u}}$ and $CoV_{\mathbf{u}}$, which indicates better fairness in terms of long-term adverse effect u. Approach III and IV have comparable $FI_{\mathbf{u}}$ and $CoV_{\mathbf{u}}$ but their fairness in terms of satisfaction rate (**SR**), measured in metrics $FI_{\mathbf{SR}}$, and $CoV_{\mathbf{SR}}$, is not improved. Based on these analysis from Tables I, II, and III, we observe that Approach II, and Approach V provide the best results in terms of $FI_{\mathbf{u}}$, and $CoV_{\mathbf{u}}$, while Approach I provides better results in terms of $FI_{\mathbf{SR}}$, and $CoV_{\mathbf{SR}}$.

D. Implementation and Execution time

We used M1 Pro chip as the platform to run our optimization solver. We use open source solver ECOS in CVXPY [44]. Since we only call the optimization solver when the desired actions between the N humans are different, we measure the execution time only when the solver is called. Then, we averaged 1000 calls for the solver for different F*in*A approaches. The average execution time per 1000 calls for the optimization solver for Approach I, II, III, IV and V are 6.308s, 8.249s, 9.876s, 10.355s and 10.217s respectively. This indicates that around 1ms overhead on average to call the solver. We used `time()` function in Python 3 to measure the execution time.

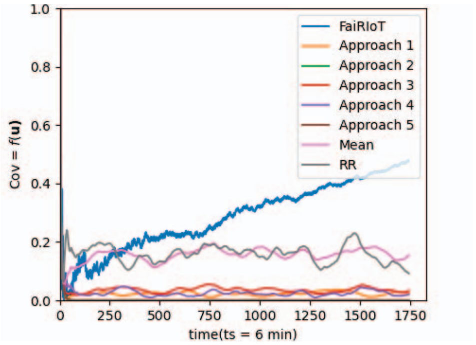


Fig. 6: Coefficient of Variation (CoV) comparison between FaiRIoT, Mean, Round Robin(RR), and all approaches of F*in*A .

E. Compare with the state of the art FaiRIoT [37]

The closest to our approach is FaiRIoT which computes the applied action through a weighted sum of all the desired actions by the N individuals $T_a = \sum_{n=1}^{N} w_n T_{d_n}$. FaiRIoT uses a notion of utility which is the average weight assigned by a layer called "Mediator RL" for a particular human h over a time horizon $[0:t]$. In particular, FaiRIoT measures the fairness of the Mediator RL using the coefficient of variation (CoV) of the human utilities. The Mediator RL is said to be more fair if and only if the CoV is smaller. Accordingly, in Figure 6, we compare the CoV in FaiRIoT with the $CoV_{\mathbf{u}}$ in all approaches in this paper. Approaches I - V achieve average CoV around 0.20, while FaiRIoT CoV is larger than 0.6. Approach II and IV has the lowest CoV at 0.04. **Hence, using F*in*A approaches improves the fairness where *CoV* is reduced by 66.7% on average.**

V. Discussion

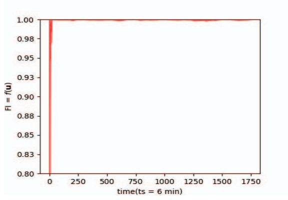
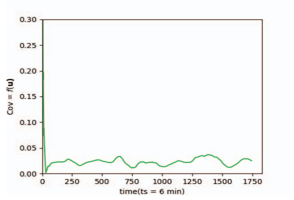
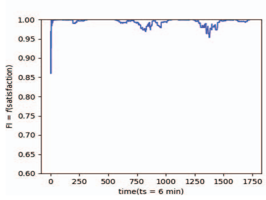
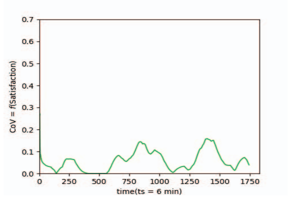
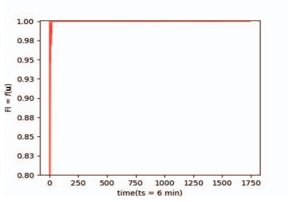
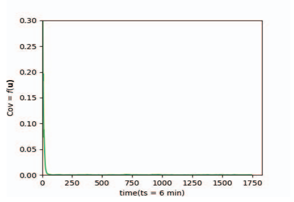
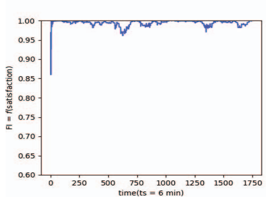
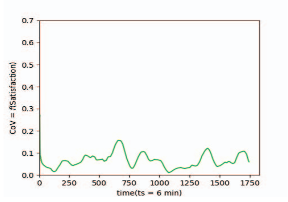
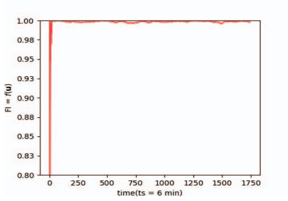
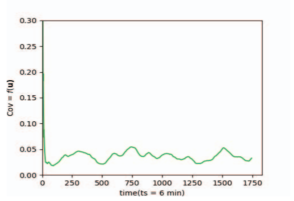
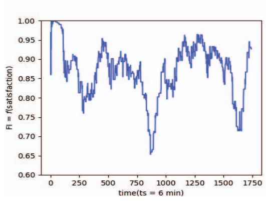
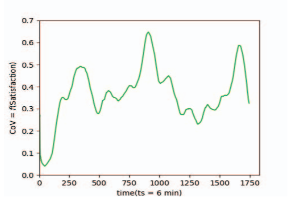
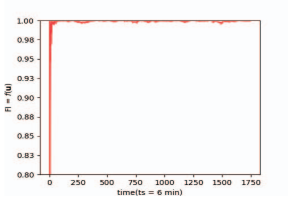
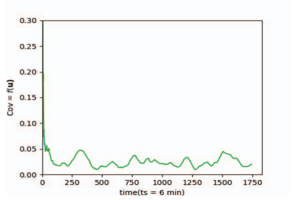
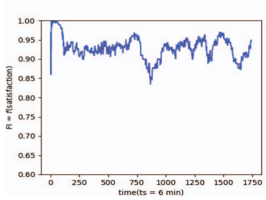
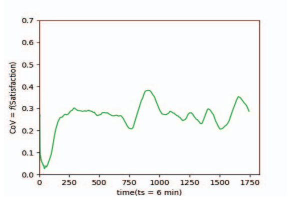
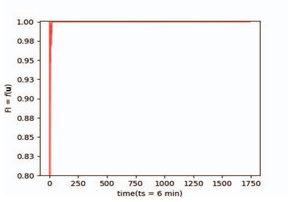
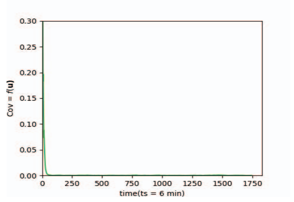
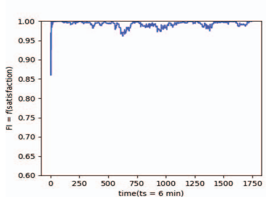
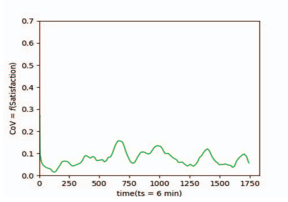
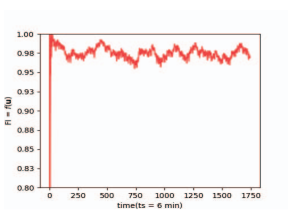
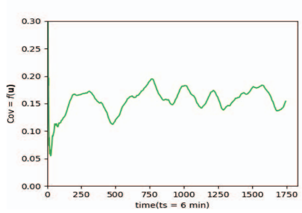
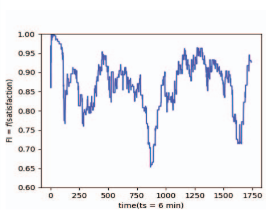
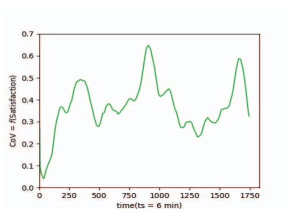
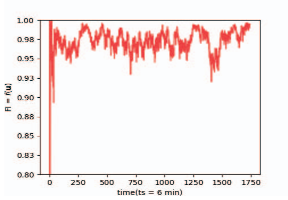
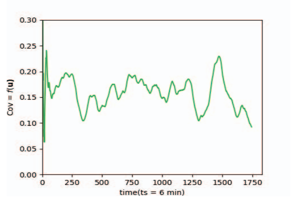
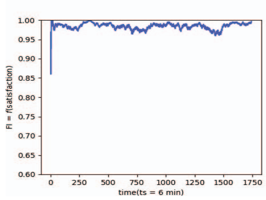
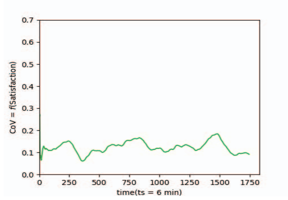
In this paper, we proposed an initial mechanism that we used to define adverse effect $v_n(a)$ of action a on human n as described in Equation 1. This assumes that the human preferred action is inversely proportional to its adverse effects. This definition can be gauged by the human perception of adverse effects. In particular, cognitive psychology offers insights on human perception. For example, Bounded Rationality Theory suggests that individuals satisfice rather than optimize decision-making [49]. Satisficing means seeking solutions that are "good enough" or satisfactory for a given situation rather than exhaustively exploring all possible options to identify the optimal choice. Hence, the adverse effect can be a time-varying function based on human perception of satisfaction.

VI. Conclusion

Addressing fairness in decision-making not only aligns with the principles of ethical AI and responsible technology, but also highlights the importance of socially-aware CPS, as individuals are more likely to cooperate with, and ultimately accept, systems that they perceive to treat them fairly. In this paper, our approaches to formalizing F*in*A within CPS decision-making capture the interplay between human preferences, the temporal dimension of adverse effects, and perceptions of fairness. Recognizing the complexities of these interactions is

[4] The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. The JSD is symmetric and always non-negative, with a value of 0 indicating that the two distributions are identical, and a value greater than 0 indicating that the two distributions are different.

TABLE III: Comparison between all the different five approaches of F$in$A, Mean approach, and Round Robin

	$FI_{\mathbf{u}}$	$CoV_{\mathbf{u}}$	$FI_{\mathbf{SR}}$	$CoV_{\mathbf{SR}}$
Approach I				
Approach II				
Approach III				
Approach IV				
Approach V				
Mean				
Round Robin				

essential for designing more equitable Human-Cyber-Physical Systems. These approaches offer a multifaceted perspective on addressing the challenges posed by the impact of CPS control actions on diverse individuals within shared environments.

ACKNOWLEDGMENT

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

References

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

[2] A. M. Annaswamy, P. P. Khargonekar, F. Lamnabhi-Lagarrigue, and S. K. Spurgeon, *Cyber-Physical-Human Systems: Fundamentals and Applications*. John Wiley Sons, Inc., 2023.

[3] M. K. Lee, J. T. Kim, and L. Lizarondo, “A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations,” in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 3365–3376.

[4] R. Wang, F. M. Harper, and H. Zhu, “Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[5] C. Stangor, “Conflict, Cooperation, Morality, and Fairness,” in *Principles of Social Psychology - 1st International Edition*. BCcampus Pressbooks, 2014.

[6] G. C. Homans, “Social behavior: Its elementary forms,” 1974.

[7] K. S. Cook and R. M. Emerson, “Social exchange theory,” 1987.

[8] J. S. Adams, “Towards an understanding of inequity.” *The journal of abnormal and social psychology*, vol. 67, no. 5, p. 422, 1963.

[9] J. W. Thibaut and H. H. Kelley, *The social psychology of groups*. Routledge, 1959.

[10] M. Redmond, “Social exchange theory,” 2015.

[11] R. Cialdini, *Influence: The Psychology of Persuasion*. Harper Collins, 2007.

[12] ——, *Presuasion: A revolutionary way to influence and persuade*. Simon and Schuster, 2016.

[13] N. M. Huijts, E. J. Molin, and L. Steg, “Psychological factors influencing sustainable energy technology acceptance: A review-based comprehensive framework,” *Renewable and sustainable energy reviews*, vol. 16, no. 1, pp. 525–531, 2012.

[14] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.

[15] J. Sztipanovits, X. Koutsoukos, G. Karsai, S. Sastry, C. Tomlin, W. Damm, M. Fränzle, J. Rieger, A. Pretschner, and F. Köster, “Science of design for societal-scale cyber-physical systems: challenges and opportunities,” *Cyber-Physical Systems*, vol. 5, no. 3, pp. 145–172, 2019.

[16] 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.

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

[18] L. J. Ratliff, R. Dong, S. Sekar, and T. Fiez, “A perspective on incentive design: Challenges and opportunities,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 1–34, 2018.

[19] 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.

[20] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, “Fairness without demographics in repeated loss minimization,” *arXiv preprint arXiv:1806.08010*, 2018.

[21] A. Chouldechova and A. Roth, “The frontiers of fairness in machine learning,” *arXiv preprint arXiv:1810.08810*, 2018.

[22] S. Kannan, J. H. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu, “A smoothed analysis of the greedy algorithm for the linear contextual bandit problem,” in *Advances in Neural Information Processing Systems*, 2018, pp. 2227–2236.

[23] N. Goel, M. Yaghini, and B. Faltings, “Non-discriminatory machine learning through convex fairness criteria,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 116–116.

[24] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, “Fairness in reinforcement learning,” in *International Conference on Machine Learning*, 2017, pp. 1617–1626.

[25] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, “Fairness in learning: Classic and contextual bandits,” in *Advances in Neural Information Processing Systems*, 2016, pp. 325–333.

[26] Y. Yu, T. Wang, and S. C. Liew, “Deep-reinforcement learning multiple access for heterogeneous wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[27] S. Gillen, C. Jung, M. Kearns, and A. Roth, “Online learning with an unknown fairness metric,” in *Advances in neural information processing systems*, 2018, pp. 2600–2609.

[28] U. Siddique, P. Weng, and M. Zimmer, “Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 8905–8915.

[29] E.-J. Shin, R. Yus, S. Mehrotra, and N. Venkatasubramanian, “Exploring fairness in participatory thermal comfort control in smart buildings,” in *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, 2017, pp. 1–10.

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

[31] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. G. Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster *et al.*, “Inequity aversion improves cooperation in intertemporal social dilemmas,” in *Advances in neural information processing systems*, 2018, pp. 3326–3336.

[32] 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.

[33] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, “Delayed impact of fair machine learning,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 3150–3158.

[34] 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.

[35] S. Milli, J. Miller, A. D. Dragan, and M. Hardt, “The social cost of strategic classification,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 230–239.

[36] 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.

[37] 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.

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

[39] A. Iosup, X. Zhu, A. Merchant, E. Kalyvianaki, M. Maggio, S. Spinner, T. Abdelzaher, O. Mengshoel, and S. Bouchenak, “Self-awareness of cloud applications,” *Self-Aware Computing Systems*, pp. 575–610, 2017.

[40] S. Huaizhou, R. V. Prasad, E. Onur, and I. Niemegeers, “Fairness in wireless networks: Issues, measures and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 5–24, 2013.

[41] F. Ho, R. Geraldes, A. Gonçalves, B. Rigault, B. Sportich, D. Kubo, M. Cavazza, and H. Prendinger, “Decentralized multi-agent path finding for uav traffic management,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 997–1008, 2020.

[42] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, “A quantitative measure of fairness and discrimination,” *ACM Transaction on Computer System*, 1984.

[43] W. Jung and F. Jazizadeh, “Towards integration of doppler radar sensors into personalized thermoregulation-based control of hvac,” in *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. ACM, 2017, p. 21.

[44] S. Diamond and S. Boyd, “CVXPY: A Python-embedded modeling language for convex optimization,” *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[45] R. G. Carroll, “Pulmonary system,” in *Elsevier’s Integrated Physiology*. Elsevier, 2007, ch. 10, pp. 99–115.

[46] M. Gerber, “energyplus energy simulation software,” 2014.

[47] 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*. ACM, 2023, pp. 262–274.

[48] P. O. Fanger, “Thermal comfort. analysis and applications in environmental engineering.” *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

[49] H. A. Simon, *Models of bounded rationality: Empirically grounded economic reason*. MIT press, 1997, vol. 3.