# adaPARL: Adaptive Privacy-Aware Reinforcement Learning for Sequential Decision Making Human-in-the-Loop Systems

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

Stelios Andrew Stavroulakis
University of California Irvine
Irvine, CA, USA
sstavrou@uci.edu

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

## ABSTRACT

Reinforcement learning (RL) presents numerous benefits compared to rule-based approaches in various applications. Privacy concerns have grown with the widespread use of RL trained with privacy-sensitive data in IoT devices, especially for human-in-the-loop systems. On the one hand, RL methods enhance the user experience by trying to adapt to the highly dynamic nature of humans. On the other hand, trained policies can leak the user's private information. Recent attention has been drawn to designing privacy-aware RL algorithms while maintaining an acceptable system utility. A central challenge in designing privacy-aware RL, especially for human-in-the-loop systems, is that humans have intrinsic variability, and their preferences and behavior evolve. The effect of one privacy leak mitigation can differ for the same human or across different humans over time. Hence, we can not design one fixed model for privacy-aware RL that fits all. To that end, we propose **adaPARL**, an adaptive approach for privacy-aware RL, especially for human-in-the-loop IoT systems. **adaPARL** provides a personalized privacy-utility trade-off depending on human behavior and preference. We validate the proposed **adaPARL** on two IoT applications, namely (i) Human-in-the-Loop Smart Home and (ii) Human-in-the-Loop Virtual Reality (VR) Smart Classroom. Results obtained on these two applications validate the generality of **adaPARL** and its ability to provide a personalized privacy-utility trade-off. On average, **adaPARL** improves the utility by 57% while reducing the privacy leak by 23% on average.

## CCS CONCEPTS

• **Computing methodologies** → **Reinforcement learning**; *Virtual reality*; • **Security and privacy**; • **Human-centered computing**;

## KEYWORDS

Privacy, Reinforcement Learning, Human-in-the-Loop, IoT

**ACM Reference Format:**
Mojtaba Taherisadr, Stelios Andrew Stavroulakis, and Salma Elmalaki. 2023. **adaPARL**: Adaptive Privacy-Aware Reinforcement Learning for Sequential Decision Making Human-in-the-Loop Systems. In *International Conference on Internet-of-Things Design and Implementation (IoTDI '23), May 09–12, 2023, San Antonio, TX, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3576842.3582325



*IoTDI '23, May 09–12, 2023, San Antonio, TX, USA*

ACM ISBN 979-8-4007-0037-8/23/05.
https://doi.org/10.1145/3576842.3582325

## 1 INTRODUCTION

The emerging technologies of sensor networks and mobile computing give the promise of monitoring the humans' states and their interactions with the surroundings [40] and have made it possible to envision the emergence of human-centered design of Internet-of-Things (IoT) 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 the environment accordingly, a new paradigm for IoT systems provides the user with a personalized experience, commonly named Human-in-the-Loop (HITL) systems.

The fundamental essence of designing HITL applications is learning the best adaptation to the environment, which is subjective to the human interaction and response to this adaptation which vary from one human to another [12]. Reinforcement Learning (RL) has proven to be adequate for monitoring human intentions and responses to provide such personalized adaptations [13, 26, 47]. Multisample RL and adaptive scaling RL (ADAS-RL) can adapt to inter-and intra-human variability among humans and the changes in their response times under different autonomous actions [1, 15]. Amazon has used personalized RL to adapt to students' preferences for adaptive class schedules [4]. Advances in deep learning with RL have been used to decide which content to present to students at any given time based on their cognitive memory models [45].

This increasing adoption of RL-based models in various HITL applications has paved the way to reformulate the trained policies with constraints to address fairness [12], risk-sensitivity [21], safety under exploration [27], and human variability [15]. Adapting to the human often leads to systems where increased sophistication comes at the expense of more privacy weaknesses. In particular, RL has the added benefit of adapting to human variations to provide a personalized experience. However, privacy concerns are raised since the optimal trained policy holds a tight correlation between the human private state and the adaptation actions provided by the RL-based HITL system. For example, a smart NEST thermostat can automatically turn on and off the HVAC equipment based on users' presence or domestic activity [3]. Such coupling between human behaviors and decisions taken by the HITL system can open a side channel, leaking sensitive information about users' daily behavioral patterns. In particular, a malicious eavesdropper can infer a user's private information only by monitoring time-series data of the adaptation actions [14, 16].

One of the critical challenges in designing HITL systems stems from the fact that the system's utility might be at odds with human expectations and privacy-preserving needs. While previous work in the literature addressing the problem of privacy leaks in learning-based adaptation engines, especially within RL-based models, through studying the fundamental privacy-utility trade-off [55],

HITL systems hold a different challenge. In particular, we argue that privacy leak mitigation techniques should not be oblivious to the fact that there are intrinsic human variations. One privacy-aware RL model may cause severe degradation in the system utility from one human's perspective, while the same model can have acceptable utility for another human. This stems from the fact that every human's response, behavior, and interaction with the HITL application is different (**inter-human variation**). Moreover, human behavior may change over time, so even if we design an RL-based model that preserves the privacy of a particular human, it may not be adequate after some time as the human changes (**intra-human variations**). Hence, the concept of one privacy-aware model that fits all is inadequate for HITL systems.

In this paper, we propose an **adaptive privacy-aware algorithm for RL-based models to provide personalized privacy-preserving human-in-the-loop IoT systems**. We borrow from the established theoretical underpinning of RL and information theory to formulate the problem as a sequential decision-making problem that maximizes the system utility with a personalized tunable regularizer that limits private information leakage due to human adaptation. We evaluate our proposed algorithm on two HITL IoT applications; the first one in the domain of smart homes and the second one in the domain of smart classrooms using Virtual Reality (VR).

## 2 RELATED WORK AND CONTRIBUTION

Privacy has been a matter of concern for decades [44]. Indeed, a plethora of work in the literature addresses privacy leaks and mitigation using a multitude of approaches. Game-theoretical approaches have been used to formulate an objective function that maximizes the utility and minimizes the privacy leaks [29]. Data encryption has been proposed to mitigate side-channel attacks on the communication links between the edge and cloud services [38]. **Our motivation behind the focus on RL methods stems from the following two fundamental properties:**

- **Computational complexity and scalability:** RL enjoy favorable computational scalability compared to other techniques especially game theoretic approaches. In particular, several game-theoretic approaches for sequential decision-making are known to be intractable [20].
- **Generalizability:** RL enjoys a unique ability to directly model the impact of taken decisions, leverage temporal feedback in learning, and improve the decision-making policy performance for a wide set of systems which is particularly important for HITL systems.

### 2.1 Privacy-preserving RL

Various aspects of privacy-preserving RL problems have been considered and tackled, such as online learning with bandit feedback [36], linear contextual bandits [22],and deep RL (DRL) [42]. Garcelon et al. [23] formulated an algorithm that guarantees regret and privacy for the tabular setting. In the continuous state context, Wang et al. [54] developed a variant of Q-learning that can find a policy where the reward function satisfies the differential privacy constraints.

The work by Erdemir et al. [16] studied the privacy-utility trade-off (PUT) in time-series data sharing. Existing approaches to PUT mainly focus on a single data point; however, temporal correlations in time-series data introduce new challenges. Methods that preserve privacy for the current time may leak a significant amount of information at the trace level as the adversary can exploit temporal correlations in a trace. They considered sharing the distorted version of a user's true data sequence with an untrusted third party.

Liu et al. [35] worked on the privacy of the reward function in RL systems. They tried to make it difficult for an observer to determine the reward function used. They presented two models for privacy-preserving rewards. These models are based on dissimulation – a deception that 'hides the truth.' They evaluated their models both computationally and via human behavioral experiments. The assumption in this study and other RL-based studies considering HITL systems is that the human state and available actions are finite and limited; otherwise, the implementation would be infeasible.

In this paper, we build upon work in the literature that exploited the correlation in the time-series data between the human state and adaptation action in RL with the assumption that HITL has finite and limited human states and adaptation actions. However, we differ from the work in the literature by arguing that the RL preserving algorithms should be adaptive to human variability in HITL IoT systems.

The closest to our approach is the work proposed by Cundy et al. [11]. They proposed a regularizer based on the mutual information (MI) between the sensitive state and the actions at a given time step for sequential decision-making. They use an upper bound as an estimation of MI to guarantee that the policy derived from an RL algorithm satisfies the privacy constraint. They mathematically prove the correctness of the algorithm and then implement it on publicly available real-world data sets. The main difference between our proposed work (**adaPARL**) and their approach is the adaptability of the privacy policy to inter-human and intra-human variations. In our experiments, we illustrate that using a constant upper bound privacy constraint, without considering the human variations, cannot efficiently satisfy a personalized privacy-utility trade-off in the HITL IoT systems.

### 2.2 Paper Contributions

In this paper, we focus on sequential decision-making human-in-the-loop IoT systems with finite state/action pairs. In particular, we aim to address the potential privacy concerns that arise from sequential decision-making systems that interact with humans whose behavior and preference vary across time. Our contributions can be summarized as follows:

- Designing **adaPARL** - a privacy-aware RL-based algorithm for sequential decision-making HITL IoT systems that mitigates the privacy leaks adaptively based on human variability.
- Providing general design parameters in our proposed **adaPARL** algorithm that can be tuned based on the application domain.
- Implementing the proposed **adaPARL** algorithm on two different HITL IoT systems in the domains of a smart home (simulation) and smart classroom with Virtual Reality (VR) (real-world).
- Personalizing the trade-off between privacy mitigation and the application's utility.

## 3 HUMAN MODELING IN REINFORCEMENT LEARNING

In the standard RL framework, a learning agent continuously interacts with an environment. The agent selects an action based on the current environment state, and the environment responds to this

action by presenting a new state to the agent. This response is in the form of a feedback reward presented to the agent. The agent seeks to maximize the reward over time through its sequential decisions of actions [50]. More formally, an RL agent interacts with an environment modeled as the Markov Decision Process (MDP) over a series of time steps $t \in \{0,1,2,...\}$. At each time step, the RL agent takes action $a_t \in \mathcal{A}$ based on the current environment state $s_t \in \mathcal{S}$ and receives a reward $r_t : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ in the same time step[1]. The dynamics underlying the environment can be described as an MDP with state-to-state transition probabilities, $p(s'|s,a) \doteq Pr\{\mathcal{S}_{t+1} = s' | \mathcal{S}_t = s, \mathcal{A}_t = a\}$ and expected rewards for state-action pairs as: $r(s,a) \doteq \mathbb{E}\{\mathcal{R}_t | \mathcal{S}_t = s, \mathcal{A}_t = a\}$. Through repeated interaction with the environment, the agent tries to learn a state-action policy, $\pi(s,a) \doteq Pr\{\mathcal{A}_t = a | \mathcal{S}_t = s\}$ that maximizes the estimated reward over time. In the special case of deterministic policy, $\pi(s) \doteq a$ with probability= 1 for $\mathcal{S}_t = s$.

### 3.1 Human as a Markov Decision Process

Unique to the HITL systems is the integration of humans with the environment. Modeling the human in a way that captures the change in behavior and preference is an open, challenging research question. Borrowing up from the psychology literature, the behavior of the changes in the human decision historically was modeled through the expected utility theorem (EUT) [39], which is based on an axiomatic framework defined as *completeness*, *transitivity*, *independence*, and *continuity*. Human is said to be rational if these four axioms hold. However, the EUT-based models have shown that these axioms are *unrealistic* and that human decisions tend to deviate from the axioms of the EUT [53]. Another approach to modeling the human is using the Partially Observed Markov Decision Process (POMDP) based on the fact that even with advanced sensing technology, the actual human state can not be measured [46]. Although POMDP aims to capture the human state's uncertainty, POMDP-based RL algorithms are computationally intractable, hindering their practical use [41]. Hence, in this work, we model the change in the human state as a Markov Decision Process (MDP) with unknown transition probabilities $p(s'|s,a)$. This uncertainty in the transition from one state to another can model the uncertainty and variability in the human state, which is essential in HITL systems.

### 3.2 Q-learning Reinforcement Learning

Learning the optimal policy $\pi(s,a)$ —action per state that maximizes the total reward—when the transition probabilities of the MDP model are unknown can be solved using RL. By applying an action in a particular state and observing the next state, the RL converges to the optimal policy that maximizes the reward function. This type of RL technique is called the Q-learning algorithm. The Q-learning algorithm assigns a value for every state-action pair. For each state $s$, the Q-learning algorithm chooses an action $a$ (among the set of allowable actions) according to a particular policy. After an action $a$ is chosen and applied to the environment, the Q-learning algorithm observes the next state $s'$ of the environment and updates the `q-value` of the pair $(s,a)$ based on the observed reward $r(s,a)$ as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r(s,a) + \gamma \max_a Q(s',a) - Q(s,a)] \quad (1)$$

[1]Some RL convention expresses the reward for action $a_t$ at time step $t$ in the next time step $r_{t+1}$.

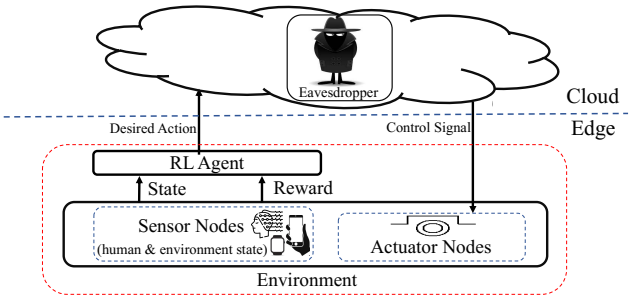


**Figure 1: Threat Model for RL-based HITL IoT application.**

The hyperparameters $\gamma$ and $\alpha$ are known as the discount factor and the learning step size, respectively. To choose an action, $a$ at each state $s$, an $\epsilon$-greedy policy can be adopted. In the $\epsilon$-greedy policy, the RL agent chooses the action that it believes has the best long-term effect with probability $1-\epsilon$, and it picks an action uniformly at random; otherwise. In other words, at each time step, the RL agent flips a biased coin and chooses the action with the maximum `q-value` with probability $1-\epsilon$ or a random action with probability $\epsilon$. This hyperparameter $\epsilon$ (also known as the exploration vs. exploitation parameter) controls how much the RL agent is willing to explore new actions that were not taken before versus relying on the best action that has been learned so far. By updating $Q(s,a)$, it is guaranteed that the optimal policy $\pi$ will converge to a deterministic action $a$ per state $s$ that provides the maximum reward $r(s,a)$ in a finite time steps $T$ [50].

## 4 THREAT MODEL IN RL-BASED HITL IOT

Recent advances in edge devices' power and memory capabilities paved the way to accomplish relatively intensive computing on the edge. Accordingly, in **adaPARL**, we can assume that the edge layer can handle relatively intensive data processing, including raw data processing and running an RL algorithm. Furthermore, synchronizing many IoT applications, especially in ubiquitous environments, requires a central decision-making server at a cloud-based level. Hence, the edge does not send the control actions directly to the environment and has to share it first with the cloud for other constraints that the cloud may need to enforce, such as synchronizing multiple IoT applications. This IoT edge-cloud computing model is a typical architecture for many pervasive and ubiquitous IoT applications [5, 43, 48]. A pictorial figure for our proposed threat model is shown in Figure 1, which can be summarized as follows:

- An HITL IoT application collects information from the environment and the human interacting with it through multiple sensors on edge devices.
- An RL agent — as explained in Section 3 — runs on the edge to infer the human and environment state and recommends the desired adaptation action based on human preference and behavior.
- Only the desired action recommended by the RL agent is propagated to a cloud-based server. In particular, the inferred environment or human states are not shared with the cloud. The edge is a trusted entity.
- The cloud elects the appropriate control signals that can be based on other enforced constraints and sends them to the actuator nodes (edge devices) to adapt to the environment.

Hence, based on this threat model, our attack vector is as follows:

- An eavesdropper is mounted in the cloud and has no access to the edge devices (sensors or actuators) or the communication channel between the edge and the cloud.
- The eavesdropper has access to the time-series data of the desired actions in the cloud.
- The eavesdropper can run any machine learning model exploiting this time-series data.
- The eavesdropper has prior knowledge of the application domain.

In this attack vector, even if the communication channel between the edge and the cloud may be secured (e.g., by encryption), the cloud still needs to decrypt the transferred packet — which contains the desired action — to elect the control signals that are sent back to the actuators at the edge for ubiquitous IoT applications. Hence, the eavesdropper can observe and record the time-series data of the desired actions of this IoT application.

Our goal in **adaPARL** is to provide a guarantee that limits the ability of an eavesdropper to infer the private state, even with unlimited computational power and complete knowledge of the application domain.

# 5 ADAPARL: ADAPTIVE PRIVACY-AWARE RL

As explained in Section 3, the state in MDP holds information about the environment, including sensitive information about the human interacting with it, such as human mental state, location, gender, or behavioral activity. The fact that the optimal policy $\pi(s,a)$ is a function of the state-action pairs can open a side channel that leaks the private human state $s$. We formulate the problem of privacy-aware RL as a Markov decision process (MDP) that ensures privacy constraints. In particular, we aim to learn a policy $\pi(s,a)$ at the edge, which maximizes the cumulative reward while constraining the privacy leak of the sensitive private state to an eavesdropper mounted in the cloud.

## 5.1 State-Action Dependence

We draw on the information theory literature and leverage *mutual information* (MI) to quantify the amount of correlation (or dependence) between two random variables. In our problem, we use the MI between *state* and *action* to measure how certain an eavesdropper can infer a *state* from observed *actions*. MI provides a theoretical bound on the inference capability of *any* learning algorithm. Generally speaking, the lower the MI between state and action is, the lower the accuracy of *any* inference algorithm. Push into one extreme; if the MI is zero, then *no* algorithm can infer the state from monitored actions. Hence, we consider the MI measure $I_\pi(a_t;s_t)$ for a particular policy $\pi(s,a)$ as a quantifiable bound on the ability to obtain the highest amount of information on $s_t$ by observing $a_t$. In particular, the amount of information leaked to the eavesdropper is bounded by $I_\pi(a_t;s_t)$[2].

Hence, we formulate the privacy-aware policy RL problem as a regularized optimization problem that maximizes the cumulative reward while maintaining a bound on them $I_\pi(a_t;s_t)$. In particular, we add an adaptive regularizer that penalizes the reward through a constraint on the value of $I_\pi(a_t;s_t)$. Hence, the reward value $r(s,a)$ to calculate the `q-value` as explained in Equation 1 can be formulated as:

$$r(s,a)=\begin{cases}\mathbb{E}\{\mathcal{R}_t|\mathcal{S}_t=s,\mathcal{A}_t=a\}, & \text{if } I_\pi(a_t;s_t)<\lambda_t\\ (1-\zeta_t)\mathbb{E}\{\mathcal{R}_t|\mathcal{S}_t=s,\mathcal{A}_t=a\}-\zeta_t I_\pi(a_t;s_t), & \text{otherwise}\end{cases} \tag{2}$$

Where $\zeta_t \in [0,1]$ is a design parameter regulating the trade-off between the privacy leak mitigation and the application utility, commonly known as the privacy-utility tradeoff (PUT). The constraint $I_\pi(a_t;s_t)<\lambda_t$ is used to set a boundary on the rise of MI by penalizing the reward function. Unique to **adaPARL** is that the boundary $\lambda_t$ is adaptive and is human and application dependent, as we will explain in Section 6 and 7.

## 5.2 Adaptive Privacy-Aware Constraint

It is worth mentioning that penalizing the reward with $I_\pi$ should be adaptive based on human variability. In particular, we make the following observations:

- **High intra-human variability:** If the human behavior and preferences frequently change with no particular pattern, the reward received by the Q-learning agent will be a dynamic and time-varying reward. This means the optimal policy $\pi(s,a)$ will take more time to converge. Consequently, the value of $I_\pi$ will rise very slowly as the agent learns the personalized policy $\pi(s,a)$.
- **Low intra-human variability:** In contrast, the reward will be less dynamic if the human has repeated patterns with the same expected behavior and preference. Hence, the Q-learning agent will learn the personalized policy $\pi(s,a)$ in less time. Consequently, the value of $I_\pi$ will rise quickly.
- The maximum value of $I_\pi$ for a particular policy $\pi(s,a)$ depends on the time-series of the state $s_t$ and the action $a_t$ which are application dependent.

Based on these observations, the upper bound $\lambda_t$ should be adaptive based on how the MI increases, which is correlated to human variability. In particular, in **adaPARL**, we propose the following adaption strategy to find the appropriate $\lambda_t$.

As the agent learns the optimal policy $\pi(s,a)$, it keeps track of a time series of $I_\pi(a_t;s_t)$. This time series is used to fit a higher-order polynomial function of degree 2. Based on this fitted function, we can approximate the growth rate and the maximum value that the $I_\pi(a_t;s_t)$ can reach. The upper bound $\lambda_t$ is then set at a particular percentage of this maximum value. This percentage – which we call $\lambda_{percent}$ – is a design parameter, as we will show in our evaluation. Hence, the constraint on $I_\pi(a_t;s_t)$ is not based on an absolute value but rather on how the agent learns, which is human-dependent (inter-human variability). Intuitively, this means that we allow the agent to learn to provide acceptable utility before it is penalized through the privacy constraints $I_\pi(a_t;s_t)<\lambda_t$. Indeed, the fitted curve is corrected over time through more interaction with the environment and to track any changes in human behavior. Hence, the value of $\lambda_t$ is also time-varying depending on the changes in human behavior (intra-human variability)

## 5.3 adaPARL Algorithm

Algorithm 1 summarizes the general **adaPARL** algorithm. Indeed the reward function $\mathcal{R}(s_t,a_t)$ in **adaPARL** algorithm is application dependent. Moreover, it is worth mentioning that the MI value ($I_t$) in

[2]Since we focus on HITL IoT applications, the number of states and actions are finite and limited as mentioned in Section 2. Hence, estimating the MI in this setup is tractable.

the algorithm is calculated over a time series of states ($\mathcal{H}_S$) and actions ($\mathcal{H}_A$) which are based on the sequential decision-making of the RL-agent at every time step $t$. Hence, by using the reward shaping approach in the **adaPARL** algorithm with the adaptive MI regularizer, the **adaPARL** agent learns to consider the consequences of choosing an action $a_t$ at time step $t$ on the distribution of states $s$ at the future time steps, which is human dependent. Eventually, **adaPARL** agent will choose future action $a_t$ that decreases the dependency on $s_t$ for a particular human. To evaluate the proposed **adaPARL** algorithm, we design HITL RL-based IoT applications and show that the adaptation to human preferences is achieved. Afterward, we discuss the privacy leaks that may occur due to the RL-based adaptation, which is application dependent. Ultimately, we show how **adaPARL** can mitigate privacy leaks and provide the privacy-utility trade-off (PUT) that varies across different humans. Accordingly, we discuss two IoT applications. The first one is in the domain of smart house (Section 6), a simulation-based application to evaluate the different design parameters in a controlled simulated environment. The second one is in the domain of smart classrooms (Section 7) using Virtual Reality (VR), which is a real-world experiment. Through providing these two different application scenarios, we aim to evaluate the applicability and adaptability of the **adaPARL** in different situations and on different people (inter- and intra-human variability) and to emphasize the generalizability of **adaPARL**.

*5.3.1 Generalizability of **adaPARL**.* The generalizability of **adaPARL** comes from the two designed parameters, $\zeta_t$ and $\lambda_t$. As explained in Section 5.2, the $\lambda_t$ is set at a particular percentage of the maximum value of $I_\pi(a_t;s_t)$. This percentage – which we call $\lambda_{percent}$ – is a design parameter correlated to human behavioral variations and provides the notion of personalized adaptation of privacy leak mitigation. Moreover, the privacy-utility trade-off can be tuned through the parameter $\zeta_t$, as explained in Section 5.1. Accordingly, the **adaPARL** algorithm can be used and implemented in various applications that need tuning of the privacy-utility trade-off.

**Algorithm 1 adaPARL** algorithm

**Q-Learning hyperparameters:** $\alpha, \gamma, \epsilon$
**adaPARL design parameters:** $\zeta, \lambda_{percent}$
**Require:**
States $\mathcal{S} = \{1, \dots, S_n\}$, Actions $\mathcal{A} = \{1, \dots, a_n\}$
Reward function $\mathcal{R} = \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
Transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$
Privacy-Utility trade-off $\zeta \in [0,1]$
Privacy mitigation upper bound $\lambda_{percent} \in [0,1]$
Mutual Information Queue $\mathcal{MI} = []$
States History Queue $\mathcal{H}_S = []$, Actions History Queue $\mathcal{H}_A = []$
Learning rate $\alpha \in [0,1]$, $\alpha = 0.01$
Discounting factor $\gamma \in [0,1]$, $\gamma = 0.001$
$\epsilon$-Greedy exploration strategy with decay $\epsilon \in [0,1]$,
max $\epsilon = 0.9$, min $\epsilon = 0.1$, decay$= 0.01$
**procedure** ADAPARL($\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma, \alpha, \epsilon, \zeta, \lambda_{percent}$)
Initialize $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with 0
time sample $t = 0$
Observe initial state $s_t \in \mathcal{S}$
**while** true **do**
Apply $\pi(s)$ according to the exploration strategy:
with probability $\epsilon$: $\pi(s) \leftarrow$ choose $a \in \mathcal{A}$ at random,
with probability $1-\epsilon$: $\pi(s) \leftarrow \text{argmax}_a Q(s,a)$)
$a_t \leftarrow \pi(s_t)$ ▷ Choose a desired action
push $a_t$ to $\mathcal{H}_A$ and push $s_t$ to $\mathcal{H}_S$
calculate $I_t(\mathcal{H}_S;\mathcal{H}_A)$ ▷ Calculate Mutual Information
push $I_t$ to $\mathcal{MI}$
$f(I_t)$ = polynomial function of degree 2 fitted to $\mathcal{MI}$
$\lambda_t = f_{max}(I_t) \times \lambda_{percent}$
**if** $I_t < \lambda_t$ **then**
$r(s_t,a_t) = \mathcal{R}(s_t,a_t)$ ▷ Receive the performance reward
**else** ▷ Penalize based on the privacy leak
$r(s_t,a_t) = (1-\zeta)\mathcal{R}(s_t,a_t) - \zeta I_t(a_t;s_t)$
**end if**
$s'_t \leftarrow \mathcal{T}(s_t,a_t)$ ▷ Observe the next state
$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma \max_a Q(s'_t, a) - Q(s_t,a_t)]$
$s_t \leftarrow s'_t$
**end while**
**end procedure**

## 6 APPLICATION 1: HUMAN-IN-THE-LOOP SMART HOME- A THERMAL SYSTEM

Recent work in the literature targets human-in-the-loop smart heating, ventilation, and air conditioning system (HVAC) while trying to assist human satisfaction [30]. RL has been proposed to adapt the HVAC set-point based on human activity [12]. A human-in-the-loop HVAC system should take the human state and preferences into the computation loop while calculating the HVAC set-point. For example, the human body temperature decreases when the human goes to sleep, while the body temperature increases when the human exercises and with stress and anxiety. Monitoring the human state, sleep cycle, and physical activity are all possible with IoT edge devices [34]. While the main purpose of this section is to evaluate the privacy-utility trade-off provided by **adaPARL**, we first describe the environment design and the RL agent used for this application.

### 6.1 System Design & Implementation

***Environment Design:*** We simulated thermodynamic model of a house that considers the house's geometry, the number of windows, the roof pitch angle, and the type of insulation used. The house is heated by a heater with an airflow of temperature $50°c$ and cooled by a cooler with an airflow of temperature $10°$. A thermostat allows fluctuation of $2.5°c$ above and below the desired set-point, specifying the temperature that must be maintained indoors [37]. The desired set-point is controlled by an external controller that runs the proposed **adaPARL** algorithm.

***Simulated Human Model:*** We model the humans as a heat source with heat flow that depends on the average exhale breath temperature ($EBT$) and the respiratory minute volume ($RMV$) [12]. The $RMV$ is the product of the breathing frequency ($f$) and the volume of gas exchanged during the breathing cycle, which is highly dependent on human activity. For example, $RMV \approx 6\ l/m$ when the human is resting while $RMV \approx 12\ l/m$ represents a human performing moderate exercise [8]. We simulated the behavior of three humans based on their activity. The human activity is simulated by different values of the $RMV$ [9] and the metabolic rate. We simulated four activity classes, including three in-home activities and a "not at home" state. Since some activities have close $RMV$ and their differences do not

affect the primary goal of this study, we categorized the normal human activities inside the house into three groups. The three in-home activity categories arranged in ascending order of *RMV* are sleeping, relaxed (sitting, standing, reading, and watching TV), and medium domestic work (washing dishes, cooking, and cleaning). We randomize the behavior by having different choices of activities in the same time slot to design different human daily behavior. We assume that the age/sex/time of day has no significance in the model. We extended the thermal house model by Mathworks [37] to include a cooling system and a human model[3].

## 6.2 RL Design

To adapt the HVAC set-point based on the human activity and thermal comfort level, we designed an RL as described below.

*6.2.1* ***State and Action Space.*** State Space $\mathcal{S}$: All combinations of four distinct human activities, $\mathcal{S} = \{(act) : act \in [1,4]\}$, where $act$ is the current human activity as mentioned in Section 6.1. Action Space $\mathcal{A}$: A discrete value for the set-point within the temperature range [60,80] with a heater option or a cooling option: $\mathcal{A} = \{a : a \in [60,80]\}$.

*6.2.2* ***Designing the reward function.*** Reward $\mathcal{R}$: We use the Prediction Mean Vote (PMV) as an estimation for the human thermal comfort [18]. The scale of PMV ranges from −3 (very cold) to 3 (very hot). According to ISO standard ASHRAE 55 [2], a PMV in the range of [−0.5,0.5] for the interior space is recommended to achieve thermal comfort. Estimating the PMV score is calculated based on the knowledge of clothing insulation, the metabolic rate, the air vapor pressure, the air temperature, and the mean radiant temperature [18]. We use a simple reward value based on the comfort value of the human dictated by the PMV. In particular, the comfortable thermal sensation $PMV = [-0.5, 0.5]$ receives positive and higher rewards, and the discomfort levels of PMV receive negative rewards. In practice, the PMV value can be estimated using edge devices, such as black globe thermometers [7].

*6.2.3* ***Hyperparameters selection.*** We briefly list some of the hyperparameters in the design of the RL agent.

- ***Discount factor $\gamma$:*** In Equation 1, $\gamma$ determines how much the RL agent cares about rewards it receives in the distant future relative to the immediate reward. In our design, the `q-value` updates only when the indoor temperature reaches the selected set-point (the selected action $a$), independently of how long it takes for the indoor temperature to reach this set-point which is dictated by the thermal dynamics of the house. Hence, a low discount factor $\gamma = 0.001$ is selected.
- ***Exploration vs. Exploitation $\epsilon$:*** Exploration is critical due to the inter- and intra-human variability. Once the agent has the appropriate information through interaction with the HITL environment, it is better to lower the exploration rate. Hence, every time `q_value` is updated, as explained in Equation 1, we gradually lower $\epsilon$ following an exponential decay of 0.01.

## 6.3 Human-in-the-Loop RL Adaptation

We simulated three different humans. Each human has been designed to have different life patterns. $H_1$'s life pattern follows an organized pattern and is repeated weekly with limited randomness. $H_3$ has a more random life pattern meaning that activities do not follow a specific daily or weekly routine, and it contains numerous unexpected changes. $H_2$ contains medium randomness, which in terms of randomness stands in between $H_1$ and $H_3$. For example, a pictorial image for the behavioral pattern of $H_2$ is shown in Figure 2, where the human behavior alternates between 4 main activities (sleeping, not at home, domestic activity, and relaxed). We run the simulator for 8000 time step ($T_s = 6min$ simulation time). To elaborate on how the designed RL-agent learned the best set-point per human activity, we split the RL-agent actions into three different plots (for clarity) based on the human activity at a particular time step, as shown in Figure 4 for $H_2$. In particular, as Figure 4 presents, approximately after 350 hours (15 simulated days), the random selection of the set-point decreased, indicating that the RL-agent started to learn the appropriate set-point for this activity.

[3]While there are more sophisticated simulators for smart houses and smart buildings that consider the energy consumption and the electric loads, such as EnergyPlus [24], we opt for a simpler model of the thermal house to evaluate **adaPARL**.

## 6.4 Information Leak

We evaluate the threat of the private information leak in this application by assuming an eavesdropper who can monitor the time series of actions decided by the RL agent. This is possible assuming a smart thermostat system that uses a cloud-based service, such as NEST [52] with a mounted spyware eavesdropper in the cloud.

State in the RL model (human activity) can be determined from sensor nodes inside the house or wearable devices. RL model runs at the edge (such as a mobile phone). The RL model sends the desired set-point (action) to the cloud engine. In this model, the attacker is located in the cloud engine, where the control signal is generated and sent to the HVAC thermostat to control the environment (house).

Since the eavesdropper has no prior knowledge of human behavior inside the house, unsupervised learning techniques can be used to infer the hidden patterns. For example, if the eavesdropper uses a clustering algorithm, such as $K$-means, we can show that sensitive information, such as occupancy and sleeping time, can be leaked. Since the eavesdropper has no prior knowledge of the human activity, the number of clusters is unknown. Hence, a common technique an eavesdropper can do is to use the elbow point to determine the best number of clusters. Figure 3 shows the elbow point result. Four clusters are the most dominant result for the clustering numbers, which equals the actual number of human activities in the simulation model.

Accordingly, we compare the ground truth (the actual behavior of $H_2$ human in the simulation model) and the clustering results by an eavesdropper (4 clusters) for each day respectively in Figure 2 top and bottom, respectively. As Figure 2 (bottom) illustrates, the eavesdropper can cluster the actions (set-point) meaningfully, which is correlated with the pattern of human activity. In this case, the eavesdropper can achieve a clustering accuracy of 86%. This clustering results by the eavesdropper can show that approximately after 15 days (as illustrated in Figure 4), there are some information leaks on the behavioral pattern of the human. Indeed the exact human activity is not inferred. However, we show here that the eavesdropper can infer some behavioral pattern of the human with the knowledge of the application domain, such as when the human most likely goes to sleep or when the human leaves the house or any changes in the normal human daily behavior.

## 6.5 Privacy Leak Mitigation with adaPARL

We evaluated **adaPARL** in adaptively mitigating the privacy leak based on human behavior while providing a privacy-utility trade-off.

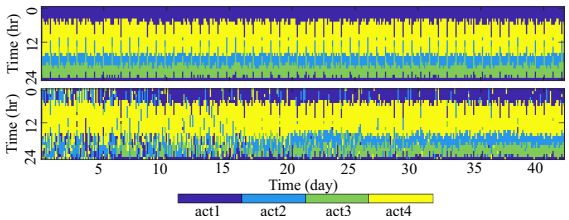


**Figure 2: Top: A human daily activity ($H_2$) inside a house across the 24 hours of the day for 40 days. act1, act2, act3, and act4 are sleeping, not at home, domestic activity, and relaxed, respectively. Bottom: The clustering of the RL actions (set-points) as seen by the eavesdropper with no access to the actual human behavior ground truth.**

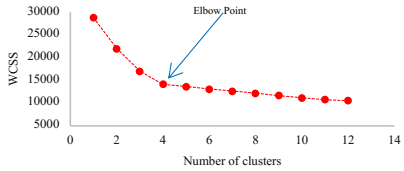


**Figure 3: WCSS (Within-Cluster Sum of Square) vs. the number of clusters. The elbow point presents the most efficient number of clusters.**

In particular, we compare **adaPARL** with two approaches. The first is a naive approach of randomizing some actions propagated to the cloud by adding noise (sending random action instead of the RL decided action). The second approach is what we call a "fixed privacy" mitigation approach [11] which will be the baseline to which we compare **adaPARL**. In all three approaches; randomization, fixed privacy mitigation, and **adaPARL**; the main objective is to reduce the MI to limit the eavesdropper's ability to infer the correlation between the state and the action pairs independently of the machine learning algorithm used by the eavesdropper. Figure 8 shows the summary for this comparison across different human behavior (as explained in Section 6.3) which we explain in detail in this section.

*6.5.1* ***Mitigation 1: Randomization.*** A biased coin with a probability $p$ is used to decide whether to randomize the currently selected action $a_t$. In particular, if $p=0.5$, then 50% of the selected actions are masked through randomization before being sent to the cloud. In Figure 5 (Top), we show the clustering results as seen by an eavesdropper in the cloud with $p=0.5$. The accuracy of clustering dropped to 65% compared to 86% before randomization. Figure 6 (left) presents the effect of adding randomization on the human comfort level $PMV$. After randomization, the $PMV$ histogram expands toward values outside the acceptable range of human thermal comfort. As expected, adding randomization decreases the eavesdropper's ability to predict the human's daily behavior. However, it comes at the cost of reducing the application utility measured by the $PMV$, where human experiences more uncomfortable thermal comfort moments.

We evaluated different values of $p$ and their effect on the MI on different human behavior. As seen in Figure 8 (row 2), the MI before and after adding randomization (with $p=0.5$) to the RL actions across 8000 simulation time steps, where each time step is equivalent to a simulated 6-minute in the system model we described in Section 6.1. As expected, adding randomization limits the MI. In particular, the MI reaches 1.45 $bits$ on average for the simulated 3 humans without mitigation. By adding randomization, the MI is decreased to a value less than 1 $bits$ on average[4].

To evaluate the privacy-utility trade-off using this approach, we use the standard deviation (STD) of the $PMV$ as a measurement of utility. In particular, as the value of the STD of the $PMV$ increases, it indicates a low utility (more spread of $PMV$ value). We plot the clustering accuracy as an indication of privacy leak vs. the STD of the $PMV$ for different randomization values $p$ as shown in Figure 8 (row 3). Increasing the randomization leads to better mitigation of the privacy leak. However, this privacy leak mitigation costs higher STD for $PMV$, meaning that human experiences discomfort, indicating low application utility. While the clustering accuracy drops to approximately 65% at $p=0.5$, the STD of the $PMV$ exceeds 1, which means $PMV$ values are more than 1 or less than −1, indicating high discomfort levels of the $PMV$. We chose not to increase $p$ more than 0.5 due to the increase in the STD of $PMV$ beyond 1.

*6.5.2* ***Mitigation 2: adaPARL.*** While the first approach of adding randomization can achieve acceptable privacy leak mitigation, we argue that this mitigation has to be adaptive and not oblivious to human behavior. We compare the MI using **adaPARL** with the randomization approach in the same Figure 8 (row 2) using $\lambda_{percent}=80\%$. The MI is reduced and limited to $\sim 1$ $bits$ on average for the three humans. As expected, the value of $\lambda$ differs for the three humans. Across the three humans, the MI before mitigation is different due to the different behavioral patterns (**inter-human variability**). Behavior patterns represent the complexity and non-uniformity of human activity in terms of repetition and order during the day. A more complex life pattern results in MI that grow slower ($H_3$), and the RL agent requires a longer time to learn the policy $\pi(s,a)$ in contrast with a more repetitive behavior ($H_1$). Hence, with $\lambda_{percent}=80\%$, the value of $\lambda$ differs depending on human behavior to provide adequate and personalized regularization to the RL reward value. Below we evaluate the design parameters $\lambda_{percent}$ and $\zeta$ as explained in **adaPARL**.

- **Tuning $\lambda_{percent}$:** Using the same metrics we explained in Section 6.5.1 to evaluate privacy-utility trade-off, we evaluated the effect of different values of the design parameter $\lambda_{percent}$ for the three humans as shown in Figure 8 (row 1) with $\zeta$ at 0.6. We observe that $\lambda_{percent}=80\%$ shows a good compromise between privacy mitigation and utility. Regarding privacy mitigation, Figure 5 (Bottom) presents the effect of **adaPARL** with $\lambda_{percent}=80\%$ on clustering results compared with the randomization approach. As Figure 5 (Bottom) demonstrates, the eavesdropper is less likely to predict human activity after mitigation. The clustering accuracy drops to 50% for the human subject with regular activity ($H_2$) with $\zeta=0.6$. As for the application utility, Figure 6 (right) shows the histogram of the total $PMV$ before mitigation and after using **adaPARL** for the human with the regular activity schedule ($H_2$). After using **adaPARL** with $\lambda_{percent}=80\%$, the $PMV$ histogram around zero (best thermal comfort) does not change dramatically, and also the histogram values outside of the comfort zone decrease (in comparison with randomization), which indicates that while the

[4]The unit of MI value depends on the base of the logarithm. If base 2 is used, MI is measured in bits.

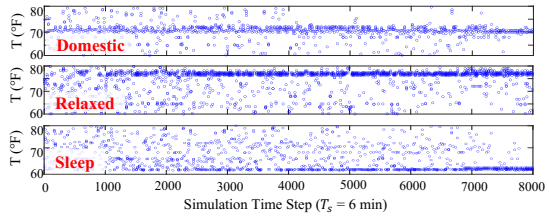


**Figure 4: RL Actions (set-point) divided into three plots. Set-point converges to $\approx 60^{\circ}F$ for sleeping, $\approx 77^{\circ}F$ for relaxed, and $\approx 71^{\circ}F$ for domestic activity.**

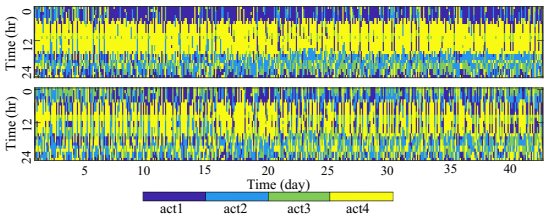


**Figure 5: Clustering before and after using mitigations. Top: clustering results with randomization mitigation with $p = 0.5$. Bottom: clustering result of adaPARL with $\lambda_{percent} = 80\%$ as seen by the eavesdropper with no access to the actual human behavior.**

privacy of the human subject is preserved, the thermal comfort of the subject is affected.

- **Tuning $\zeta$:** Similarly, as Figure 8 (row 4) illustrates, by increasing $\zeta$, the utility (the STD of $PMV$) increases while clustering accuracy drops on average across the three humans to 50% before the STD of the $PMV$ exceeds 1. Hence, $\zeta = 0.6$ provides a good compromise for the privacy-utility trade-off.
- **Tracking intra-human variability:** As human behavior may change over time, the value of $\lambda$ has to adapt and can not be fixed even for a single human. Figure 7 shows the MI for a human with a routine behavioral life pattern with a growing MI at the beginning. Human behavioral pattern changes to follow another growing MI pattern. In particular, to simulate this change in behavior, we switched between the behavior of $H_2$ and $H_1$ at runtime. As seen in the figure, $\lambda$ follows a newly fitted MI curve to follow these changes in the human behavioral pattern.

*6.5.3* ***Mitigation 3: Comparison with the baseline.*** We compared **adaPARL** with a baseline method proposed in [11] when the reward value is regularized (penalized) by the MI from the beginning, regardless of the shape of the MI (human behavior). In our setup, this means $\lambda_{percent} = 0$. We call this approach a fixed privacy approach. Figure 8 (row 5) provides this comparison across different values of $\zeta$. This early penalization hinders the RL agent from learning the human comfort zone partially, and the STD of the $PMV$ grows faster than **adaPARL**. In **adaPARL**, by tuning the parameter $\lambda$, RL has time to learn the human comfort zone, and the human experiences less uncomfortable duration (with slower $PMV$ STD growth). This pattern is followed for all three humans with different behavioral patterns, which means that **adaPARL** can adapt to human variability with less loss in the system utility compared to the baseline method.

## 6.6 Observations

In this section, we summarize the observations from application 1. Using **adaPARL** with personalized $\lambda$, we could achieve privacy mitigation close to the baseline method with a smaller loss on the application utility per human. In particular, **adaPARL** is able to enhance the utility by 46% (STD of PMV) compared to randomization and by 57% compared to the baseline method on average across three different human behavior. In terms of privacy leak mitigation, **adaPARL** reduces the privacy leak by 16% (the clustering accuracy) compared to the baseline method on average. Furthermore, compared with the randomization, the privacy leak in **adaPARL** is decreased by 38%. Moreover, by using the parameters $\lambda$ and $\zeta$, **adaPARL** can adapt to intra- and inter-human variability and regulate the privacy utility trade-off. Results showed that when privacy leak is highly mitigated (high values of $\zeta$, with $\zeta = 0.8$), the utility is in the acceptable range ($-0.5 \leq PMV \leq 0.5$). Hence, **adaPARL** was able to achieve an acceptable performance even with high privacy requirements.

# 7 APPLICATION 2: HUMAN-IN-THE-LOOP VIRTUAL REALITY SMART CLASSROOM

The first experiment provided insights into the effect of different design parameters in **adaPARL** in a controlled simulated environment. Next, we design a real-world VR application to evaluate **adaPARL**. Inspired by the recent paradigm shift in the education system post-COVID-19 era and the need for personalized and remote education setup, we selected a smart classroom IoT application using remote instruction with VR. During elongated training/education periods, especially in an online or remote environment, human performance is prone to significantly decline [51] due to distractions, drowsiness, and fatigue. In this experiment, an RL agent monitors these changes in the human state and provides personalized feedback to improve human learning performance. The eavesdropper is located in the cloud and has access to the actions taken by the RL model. As motivated in Experiment 1 (Section 6), these adaptation actions are correlated to the human private state (learning performance and mental state).

We first design the application and then show the correlation between the RL agent taken actions and the human mental state. Lastly, we apply mitigation techniques, including **adaPARL**, to mitigate the private data leak.

## 7.1 System Design and Implementation with VR

We incorporated 2 presentation modes (2D and 3D) to present the lecture contents to the participants. We used the Virtual Reality (VR) technologies for the 3D presentation mode because recent studies showed that these new technologies would have a significant impact on the learning [28], and workforce training sectors. Using a VR device (Oculus device), we provided the 3D presentation, and with the regular laptop screen, we provided the 2D visualization mode. We chose the lecture contents from Khan Academy along with their quizzes that cover topics on biology [31], chemistry [33], and physics [32]. We asked 15 participants, all within the age range of 20−30, to watch these lectures. Each lecture is stand-alone and does not require any prior knowledge from the participants to be understood. The participants' main task was to watch the lecture and pay attention to answer the questions regarding the content at the end of the lecture. We provided the 3D version of these lectures by converting them from 2D to 3D for VR presentation mode. Each

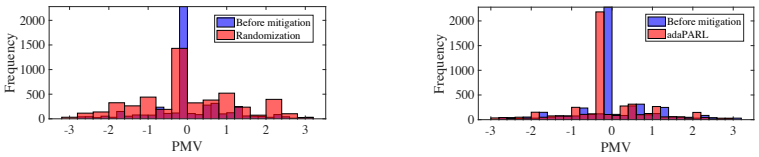


**Figure 6: *PMV* histogram before and after mitigation. Left: randomization mitigation with $p=0.5$. Right: adaPARL mitigation with $\lambda_{percent}=80\%$.**

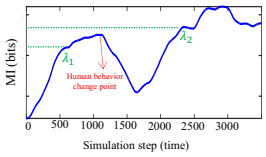


**Figure 7: Human changes the behavior which causes the change in MI curves.**

**Table 1: Human state is one of 8 states depending on the alertness level (AL), fatigue level (FL), and vertigo level (VL).**

| **AL** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|
| **FL** | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| **VL** | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| **State** | **S8** | **S7** | **S6** | **S5** | **S4** | **S3** | **S2** | **S1** |

lecture is ∼55 minutes, in narrative style, and does not include any quizzes or other interruptions in the middle of them.

We used EEG wearable devices to monitor the human EEG signals to infer the mental state related to learning. We used an EMOTIV EPOC_X 14 channels portable EEG device. Before the experiment, we presented a 10 minutes 2D video presentation on a laptop screen to measure their baseline mental state signals. We use every 10 minute duration of the EEG raw data to infer the human state regarding their alertness and readiness to learn. While several wearable devices measure various physiological signals that can be used to infer the human alert level or drowsiness level, we choose the EEG signal due to the recent studies that showed the frontal lobe activation of the brain could be used to infer the human ability to learn and cognitive performance [17]. We divided each lecture into 10 minutes videos that we call stages, with approximately 5 stages each lecture[5]. Figure 9 shows the setup of the application.

## 7.2 RL Design

*7.2.1* ***Human State Space.*** Albeit the beneficial aspects of VR technologies in education and workplace training, humans react differently to the VR environment. Some humans report vertigo and cybersickness symptoms during exposure to VR [6], which can affect the learning experience. We designed the human state as a combination of three features, including alertness level (AL), fatigue level (FL), and vertigo level (VL). Accordingly, we classify these three features into binary classes to reduce our state space. In particular, $AL$ is classified as "Alert (1)" versus "Not Alert (0)", $FL$ is classified as "Fatigue (1)" versus "Vigor (0)", and $VL$ is classified as "Not Vertigo (1)" versus "Hypocalcemia (0)". $\mathcal{S}$ refers to a tuple of 3 features: $\mathcal{S} = \{(AL,FL,VL) : AL \in \{0,1\}, FL \in \{0,1\}, VL \in \{0,1\}\}$. Table 1 illustrates the state space. Each feature is classified based on a corresponding threshold $\delta_{AL}$, $\delta_{FL}$,*and* $\delta_{VL}$. If the feature's measured value passes the threshold, the class for the given feature is 1; otherwise, it is 0. Accordingly, the best human state for learning is $S_8$, where the human is alert and is not experiencing cybersickness (no fatigue and vertigo). In contrast, the worst state is $S_1$, where the human is not alert and experiencing cybersickness (fatigue and vertigo).

Indeed, humans can transition between any of these states. It is worth mentioning that these thresholds can be tuned based on the application and participant. Below we describe how we infer the human state and calculate these thresholds.

[5]The average attention span of the human is 10 to 15 minutes.

- **Alertness level (brain engagement) ($AL$):** To measure human alertness and engagement during the learning process, spectro-temporal EEG signal analysis can be used. One method recently implemented to analyze the EEG signal during the learning process is fractal dimension [19]. Various methods have been developed to calculate the fractal dimension, mainly based on the entropy concept. In this experiment, we used the box-counting method to calculate the fractal dimension [19] on the recorded EEG time series. Since the frontal lobe of the brain is responsible for cognitive functions such as memory and problem solving [17], we used the $F3$ and $F4$ channels of the EEG device (located in the frontal lobe). We averaged the outcomes of these channels for our analysis. After preprocessing (filtering and noise removal) and fractal dimensions analysis, we observed that the fractal dimension of the EEG signal is higher in the 3D compared to the 2D videos. Since the fractal dimension reflects the complexity of the signal, this result indicates that the EEG signal is more complex in response to 3D visual stimuli than 2D visual stimuli. In other words, the human brain becomes more engaged with a stimulus when presented in the 3D compared to the 2D. We used this measurement as the $AL$ state.
- **Fatigue Level ($FL$):** We define a measure for $FL$ that exploits the EEG signals. In particular, we use Wavelet Packet Decomposition (WPD) to decompose the EEG signal into its spectral sub-bands with 1 $Hz$ resolution. Recent work in the literature showed that fatigue and drowsiness correlate with the range of $8-14\,Hz$, which is mostly the $\alpha$ band of the EEG signal [10]. Accordingly, we define $FL$ as the power spectrum of the $\alpha$ band. Similar to $AL$, we collected the data from $F3$ & $F4$ channels and averaged the resulting $FL$ from both channels.
- **Vertigo Level ($VL$):** The most commonly reported measure of vertigo and cybersickness symptoms is the Simulator Sickness Questionnaire (SSQ). The SSQ was derived directly from the Pensacola Motion Sickness Questionnaire (MSQ) [25]. The MSQ consists of a list of 25 to 30 symptoms, such as spinning, vertigo, and may vomit. Symptoms severity are rated on four levels, "none" (0), "slight" (1), "moderate" (2), and "severe" (3). A total score was computed by summing item scores. The highest score was determined to specify emesis as the worst case of sickness. In particular, the SSQ is a selection of 16-items from the MSQ with a different scoring scheme. Based on three main subfactors of cybersickness, including Nausea ($N$), Oculomotor ($O$), and Disorientation ($D$), a Total Score (TS) is computed, representing the overall severity of cybersickness experienced by the subject. In particular, $TS$ can range from 0 to 235.62 [49]. Participants were asked to fill out the SSQ questionnaire, and we used it to calculate the $TS$. The threshold for $TS$ to consider it vertigo depends on the application [49]).

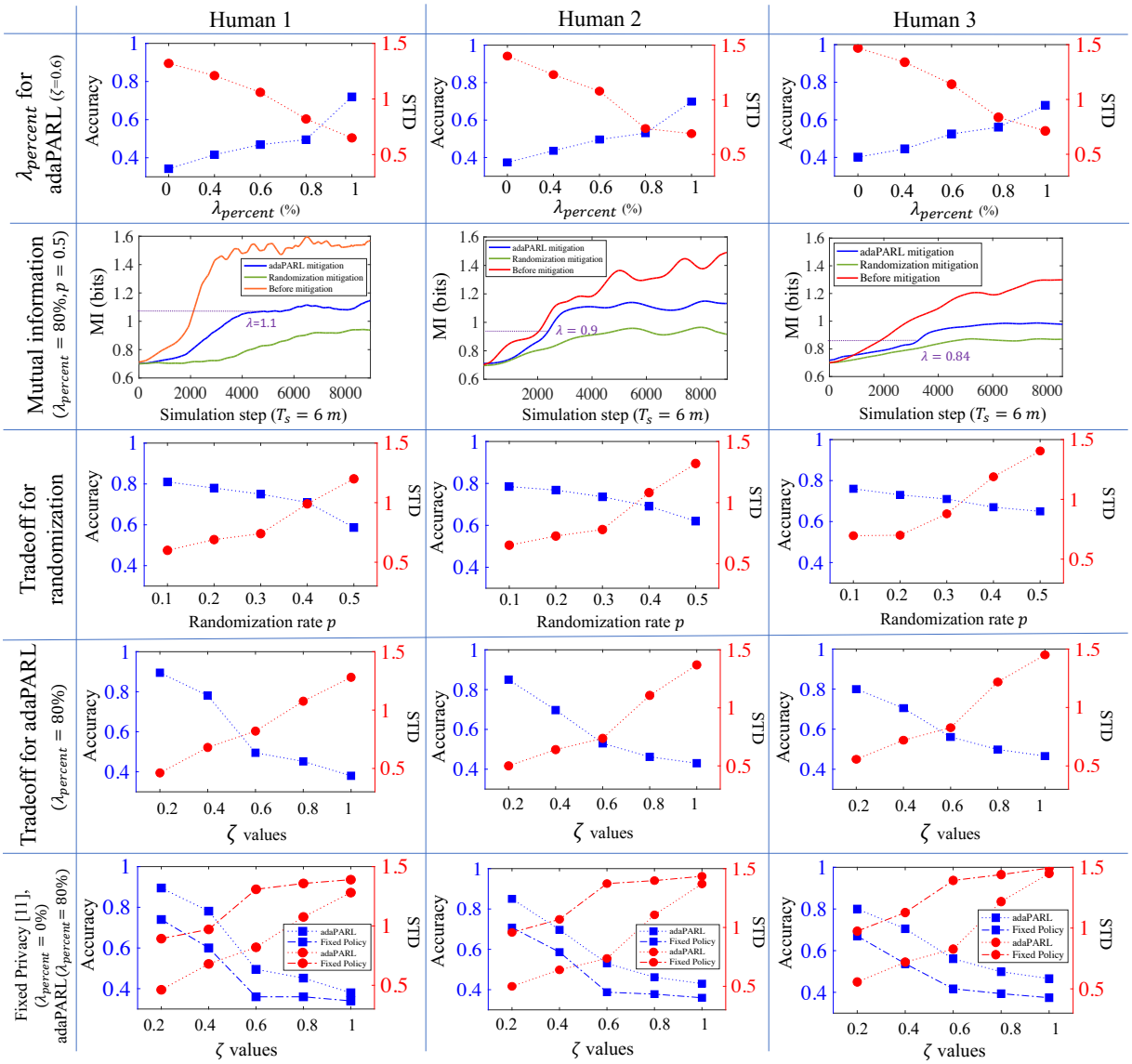


**Figure 8: Proposed adaPARL algorithm parameters analysis and privacy vs. utility trade-off. The first row provides privacy vs. utility trade-off by considering different values of $\lambda_{percent}$ for standard deviation (STD) of $PMV$ on the right vertical axis and clustering accuracy on the left vertical axis. The second row depicts the MI growth rate for different mitigation algorithms. The third row presents the privacy vs. utility trade-off for the randomization mitigation approach. The fourth row illustrates the privacy vs. utility trade-off for the adaPARL algorithm. On the right vertical axis, it presents the STD of the $PMV$ and clustering accuracy on the left vertical axis across different values of design parameter $\zeta$ on the horizontal axis. The last row shows the privacy vs. utility trade-off comparison between adaPARL and baseline algorithm proposed in [11].**

- **Choosing the thresholds ($\delta_{AL}, \delta_{FL}, \delta_{VL}$):** $\delta_{AL}$ was defined based on the calculation of the $AL$. $AL$ correlates with the fractal dimension values calculated for the EEG signal. We compared each calculated fractal dimension value with the baseline (measured before the experiment). If the $AL$ exceeds the baseline ($\delta_{AL}$), the human is classified as "Alert" (1), else the human is classified as "Not Alert"(0).
  We used the spectral power of the $\alpha$ band to define the threshold for $FL$. $\delta_{FL}$ is the midpoint of the $t$ measure, calculated based on the $p-Values$ of the comparison of the $\alpha$ band of the baseline EEG and the current stage EEG signals. If the $t$ measure exceeds $\delta_{FL}$, then the human is classified as "Fatigue" (0), or else the human is classified as "Not Fatigue"(1).
  We defined $\delta_{VL}$ to be $\frac{TS_{max}}{4}$ where $TS_{max}$ = 235.62. If the measured $TS$ is bigger than the $\delta_{VL}$, the human is classified as "Not Vertigo" (0), or else the human is classified as "Vertigo"(1). At the end of each stage of the experiment, participants are asked to fill out the SSQ, and using the questionnaire, we calculate $TS$.

*7.2.2* ***Action Space.*** The action space $\mathcal{A}$ in this VR application includes the following actions: (1): Give a break to the human, (2): Enable VR mode by switching from 2D to 3D, (3): Disable VR mode


**Figure 9: Left: EMOTIV Epoc_x and Oculus device worn by a participant. The laptop on the left casts the participant's view on the Oculus device, and the laptop on the right presents the live EEG signal collected from the EMOTIV device. Right: Screenshot of the view of the participant while watching biology content on an Oculus device with an office background.**

by switching from 3D to 2D, (4): Changing the content of the presentation, and (5): No change to the learning environment.

In particular, enabling the VR mode increases brain engagement and enhances learning performance. However, some humans may experience cybersickness with exposure to VR; hence the RL agent may need to switch back to regular 2D to reduce cybersickness symptoms. Moreover, a break during a learning session may also be needed to reduce drowsiness, cybersickness symptoms, or cognitive load. Hence, $\mathcal{A}$ is bounded and discrete and can be within the range $[1,5]$: $\mathcal{A} = \{a : a \in [1,5], a \in \mathbb{N}\}$.

*7.2.3* ***Reward.*** We used the same definition of the reward function explained in Section 5. The human performance in a quiz dictates the reward value after every learning module, where the score in this quiz is measured as a percentage. Quiz quantification applies to 10 multiple-choice questions uniformly as the 10/10 (100%) and 0/10 (0%) scores receive rewards 100 and 0, respectively.

## 7.3 Private Information Leak

As described in our threat model in Section 4, eavesdroppers in the cloud can access the RL desired actions and run machine learning algorithms to infer the participant's private information. The human state is calculated at the edge. The human mental state is private information, and any gain by the attacker to this information is considered a privacy intrusion. Similar to the analysis we did in the first application in Section 6, we monitor the MI between the RL action and the participants' state. Figure 10 illustrates the MI between taken actions and states of a participant (solid red line referring to "before mitigation"). The MI reaches approximately $\approx 1.5$ $bits$. A similar approach for an eavesdropper we designed in the first application in Section 6.4 can be deployed in this application. An eavesdropper can use unsupervised learning techniques, such as clustering. Hence, we will evaluate **adaPARL**'s ability to mitigate the private leak.

## 7.4 Privacy Leak Mitigation with adaPARL

As mentioned in the previous experiment, we aim to mitigate this information leak using two approaches, action randomization, and **adaPARL**.

*7.4.1* ***Mitigation 1: Randomization.*** As a representative result, by choosing random actions ($p = 0.5$), Figure 10 depicts the randomization effect on MI between RL actions and one of the participant's states, which shows the decrease in MI from $\approx 1.5$ $bits$ to less than $\approx 1$ $bits$. We evaluate the privacy-utility trade-off using action randomization. In particular, we use the drop in the performance in the quizzes as a utility metric. We use the human state prediction accuracy using clustering (similar to application 1) as the privacy leak measure. We emphasize here that even though the eavesdropper may not know the actual human state, the change in the human state (learning pattern) through observed actions can be inferred, which can leak private information, such as the human attention span.

To study the effect of randomization $p$ without asking the participant to repeat the experiment 5 times, which may bias the results, we used the data we collected from the online experiment with the 15 participants at $p = 0.5$. In particular, from the online experiment, we knew the quiz performance of a participant given a state-action pair. Hence, we change the value of $p$ offline to generate different actions per state and record the expected quiz performance. Figure 11 shows the effect of parameter $p$ on utility-privacy trade-off for the randomization algorithm for the participants (dotted lines) and the average (solid line), which illustrates a reduction in the state prediction accuracy on average to 60% with a 60% drop in utility (performance). Adding more random actions leads to lower performance in the quiz as the chosen presentation mode is less likely to be optimal for the participant's current state. This trade-off shows how privacy protection increases according to the randomization of the actions. It also demonstrates that performance drops quickly after increasing randomization.

*7.4.2* ***Mitigation 2: adaPARL.*** We tuned the parameter $\lambda_{percent}$ on 5 participants offline using a similar approach to application 1. We used the same category of lecture content (i.e., biology) but not the same content as the one we used in the online experiment to prevent any content bias on the participants. The offline test resulted in $\lambda_{percent} = 85\%$. Hence, we set $\lambda_{percent} = 85\%$ across the 15 participants during the online test. In Figure 10 (solid blue line), we show the MI between the actions and one of the participant's states using **adaPARL**. As Figure 10 illustrates, the MI increases as the agent learns to take corrective actions and then using $\lambda = 1.06$ ($\lambda_{percent} = 85\%$) **adaPARL** regularized the growth of the MI as explained in Section 5.2.

Similar to the approach we used to study the effect of $p$ in the randomization, we study the effect of $\zeta$. Figure 12 presents the percentage of performance drop and participants' state prediction accuracy by changing the values of $\zeta$ for the participants (dotted lines) and its average (solid line). Increasing the parameter $\zeta$ decreases the eavesdropper's ability to predict the participants' state. The prediction ability of the eavesdropper for $\zeta = 0.6$ decreases by $\sim 50\%$. At the same time, the participant's performance is higher than 75% (performance drop is less than 25%), showing improvement in the trade-off compared to the randomized actions. Table 2 provides the numeric details of the trade-off parameter ($\zeta$) and $\lambda$ ($\lambda_{percent} = 85\%$) values specific to each participant. **adaPARL** chooses different values of $\lambda$ for each participant (inter-human variability).

## 7.5 Observations

Results from this application show that using the parameter $\zeta$ **adaPARL** can mitigate the eavesdropper's ability to predict the participants' state. Our results show that the prediction ability of the eavesdropper for $\zeta = 0.6$ decreases by $\sim 50\%$. On the other hand, the utility (participant's performance in a quiz) is higher than 75%, meaning that **adaPARL** can improve the trade-off compared to using the randomization approach. Furthermore, **adaPARL** uses different numeric values of $\lambda$ for different participants, which shows that **adaPARL** can provide adaption to inter-human variability. On average, the accuracy of the private state detection decreases to 44%

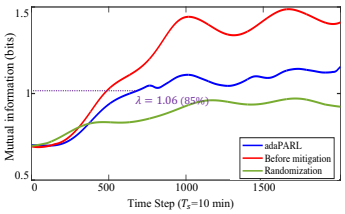


**Figure 10: MI between actions and participant's states before and after adding mitigation using randomization and adaPARL at $\lambda = 1.06$ (85%).**

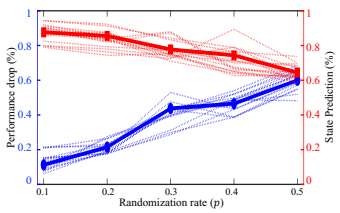


**Figure 11: State prediction accuracy (%) and performance drop on the quiz (%) trade-off in randomization. The dotted line presents one participant, and the solid lines present the average.**

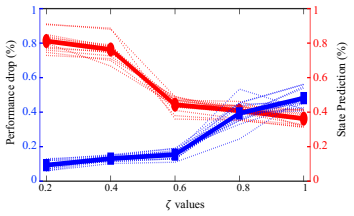


**Figure 12: Trade-off between participants' state prediction accuracy (%) and performance drop on the quiz (%) using adaPARL ($\lambda_{percent} = 85\%$).**

**Table 2: Privacy (State Prediction) vs. Utility (Performance Drop) trade-off and $\lambda$ parameter results of adaPARL for all the 15 participants.**

| Parameter | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 | P13 | P14 | P15 | Ave |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda$ **(with** $\lambda_{percent} = 85\%$**)** | 1.06 | 1.11 | 0.98 | 0.97 | 1.03 | 0.93 | 1.02 | 1.01 | 1.04 | 0.95 | 0.91 | 1.02 | 1.16 | 0.89 | 0.96 | **1.01** |
| **State Prediction (%)** | 0.41 | 0.38 | 0.46 | 0.47 | 0.43 | 0.48 | 0.43 | 0.44 | 0.43 | 0.47 | 0.49 | 0.44 | 0.36 | 0.49 | 0.47 | **0.44** |
| **Performance Drop (%)** | 0.15 | 0.13 | 0.18 | 0.17 | 0.15 | 0.16 | 0.14 | 0.16 | 0.15 | 0.17 | 0.19 | 0.13 | 0.11 | 0.19 | 0.13 | **0.15** |

before the drop in participants' performance (utility) passes 15%. In contrast, in the randomization approach, the 20% drop in the utility leads to a high privacy cost of 80% as indicated in the prediction accuracy. In randomization, the privacy leak, which is the prediction accuracy, never exceeds 65%, even with the degradation in the utility. For instance, when the participants' performance drops by over 60%, the privacy leak is still 65%.

## 8 CONCLUSION

In this paper, we proposed **adaPARL**, an adaptive human-in-the-loop privacy-aware RL algorithm that addresses the privacy challenges associated with human variability in RL-based systems trained with privacy-sensitive data in IoT applications. We adopted a typical edge-cloud threat model architecture where all the sensitive human state inference is calculated on a trusted edge that hosts the RL agent, and the cloud has only access to the desired control actions where an eavesdropper is mounted. **adaPARL** provided an adaptive and personalized threshold ($\lambda$) to regularize the reward function of the RL agent, which changes its value at runtime based on the changes in human behavior to mitigate the privacy leak. We validated **adaPARL** on two Human-in-the-Loop IoT applications in simulated (smart house) and real-world (VR smart classroom) environments. We showed that **adaPARL** could achieve a personalized privacy-utility trade-off through two tunable design parameters, $\zeta$, which provides the privacy-utility trade-off, and $\lambda$, which provides the adaptation to inter-human variability. In the first application, on average, **adaPARL** improved the application utility over the randomization approach by 43% and over the baseline approach by 57%. Furthermore, **adaPARL** reduced the privacy leak on average by 23%. We implemented **adaPARL** in a real-world application and demonstrated how **adaPARL** is capable of adapting to inter-human variability. Thanks to the flexibility of its parameter ($\lambda$), **adaPARL** was able to adapt to 15 different human participants.

## ACKNOWLEDGMENTS

This research was partially supported by NSF award # CNS-2105084.

## REFERENCES

[1] Armand Ahadi-Sarkani and Salma Elmalaki. 2021. ADAS-RL: Adaptive vector scaling reinforcement learning for human-in-the-loop lane departure warning. In *Proceedings of the First International Workshop on Cyber-Physical-Human System Design and Implementation*. 13–18.

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

[3] Bharathan Balaji, Jian Xu, Anthony Nwokafor, Rajesh Gupta, and Yuvraj Agarwal. 2013. Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[4] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. 2020. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[5] Phuthipong Bovornkeeratiroj, Srinivasan Iyengar, Stephen Lee, David Irwin, and Prashant Shenoy. 2020. Repel: A utility-preserving privacy system for iot-based energy meters. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 79–91.

[6] Kjell Brunnström, Elijs Dima, Tahir Qureshi, Mathias Johanson, Mattias Andersson, and Mårten Sjöström. 2020. Latency impact on Quality of Experience in a virtual reality simulator for remote control of machines. *Signal Processing: Image Communication* 89 (2020), 116005.

[7] campbellsci. 2022. black globe thermometer. Retrieved June 10, 2022 from https://www.campbellsci.com/blackglobe

[8] Robert G Carroll. 2006. *Elsevier's Integrated Physiology E-Book*. Elsevier Health Sciences.

[9] Robert G. Carroll. 2007. Pulmonary System. In *Elsevier's Integrated Physiology*. Elsevier, Chapter 10, 99–115.

[10] Agustina Garcés Correa, Lorena Orosco, and Eric Laciar. 2014. Automatic detection of drowsiness in EEG records based on multimodal analysis. *Medical engineering & physics* 36, 2 (2014), 244–249.

[11] Chris Cundy and Stefano Ermon. 2020. Privacy-constrained policies via mutual information regularized policy gradients. *arXiv preprint arXiv:2012.15019* (2020).

[12] Salma Elmalaki. 2021. 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*. 119–132.

[13] Salma Elmalaki. 2022. MAConAuto: Framework for Mobile-Assisted Human-in-the-Loop Automotive System. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 740–749.

[14] Salma Elmalaki, Bo-Jhang Ho, Moustafa Alzantot, Yasser Shoukry, and Mani Srivastava. 2022. VindiCo: Privacy Safeguard Against Adaptation Based Spyware in Human-in-the-Loop IoT. *arXiv preprint arXiv:2202.01348* (2022).

[15] Salma Elmalaki, Huey-Ru Tsai, and Mani Srivastava. 2018. Sentio: Driver-in-the-Loop Forward Collision Warning Using Multisample Reinforcement Learning. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 28–40.

[16] Ecenaz Erdemir, Pier Luigi Dragotti, and Deniz Gündüz. 2020. Privacy-Aware Time-Series Data Sharing With Deep Reinforcement Learning. *IEEE Transactions on Information Forensics and Security* 16 (2020), 389–401.

[17] Paul J Eslinger and Antonio R Damasio. 1985. Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. *Neurology* 35, 12 (1985), 1731–1731.

[18] Poul O Fanger. 1970. Thermal comfort. Analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.* (1970).

[19] Kayhan Foroutan-pour, Pierre Dutilleul, and Donald L Smith. 1999. Advances in the implementation of the box-counting method of fractal dimension estimation. *Applied mathematics and computation* 105, 2-3 (1999), 195–210.

[20] Aviezri S Fraenkel. 2004. Complexity, appeal and challenges of combinatorial games. *Theoretical Computer Science* 313, 3 (2004), 393–415.

[21] Christopher Gagne and Peter Dayan. 2021. Two steps to risk sensitivity. *Advances in Neural Information Processing Systems* 34 (2021), 22209–22220.

[22] Evrard Garcelon, Kamalika Chaudhuri, Vianney Perchet, and Matteo Pirotta. 2022. Privacy amplification via shuffling for linear contextual bandits. In *International Conference on Algorithmic Learning Theory*. PMLR, 381–407.

[23] Evrard Garcelon, Vianney Perchet, Ciara Pike-Burke, and Matteo Pirotta. 2021. Local differential privacy for regret minimization in reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 10561–10573.

[24] Michael Gerber. 2014. energyplus energy Simulation Software. (2014).

[25] John F Golding. 1998. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain research bulletin* 47, 5 (1998), 507–516.

[26] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*. 3909–3917.

[27] Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft. 2008. Safe exploration for reinforcement learning.. In *ESANN*. Citeseer, 143–148.

[28] María Blanca Ibáñez, Ángela Di Serio, Diego Villarán, and Carlos Delgado Kloos. 2014. Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education* 71 (2014), 1–13.

[29] Richeng Jin, Xiaofan He, and Huaiyu Dai. 2017. On the tradeoff between privacy and utility in collaborative intrusion detection systems-a game theoretical approach. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*. 45–51.

[30] Wooyoung Jung and Farrokh Jazizadeh. 2017. 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, 21.

[31] KhanAcademy. 2021. AP College Biology. https://www.khanacademy.org/science/ap-biology. Accessed: 2021-10-07.

[32] KhanAcademy. 2021. AP College Physics. https://www.khanacademy.org/science/ap-physics-2. Accessed: 2021-10-07.

[33] KhanAcademy. 2021. Organic Chemistry. https://www.khanacademy.org/science/organic-chemistry. Accessed: 2021-10-07.

[34] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 389–402.

[35] Zhengshang Liu, Yue Yang, Tim Miller, and Peta Masters. 2021. Deceptive Reinforcement Learning for Privacy-Preserving Planning. *arXiv preprint arXiv:2102.03022* (2021).

[36] Mohammad Malekzadeh, Dimitrios Athanasakis, Hamed Haddadi, and Ben Livshits. 2020. Privacy-preserving bandits. *Proceedings of Machine Learning and Systems* 2 (2020), 350–362.

[37] MATLAB. 2022. Thermal Model of a House. Retrieved June 10, 2022 from https://www.mathworks.com/help/simulink/slref/thermal-model-of-a-house.html

[38] Shilpi Mishra, Divyapratap Singh, Divyansh Pant, and Akash Rawat. 2022. Secure Data Communication Using Information Hiding and Encryption Algorithms. In *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 1448–1452.

[39] Oskar Morgenstern and John Von Neumann. 1953. *Theory of games and economic behavior*. Princeton university press.

[40] Subhas Chandra Mukhopadhyay. 2015. Wearable sensors for human activity monitoring: A review. *IEEE sensors journal* 15, 3 (2015), 1321–1330.

[41] Kevin P Murphy. 2000. A survey of POMDP solution techniques. *environment* 2 (2000), X3.

[42] Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, and Dawn Song. 2019. How You Act Tells a Lot: Privacy-Leaking Attack on Deep Reinforcement Learning. *AAMAS* (2019).

[43] Franz Papst, Naomi Stricker, Rahim Entezari, and Olga Saukh. 2022. To Share or Not to Share: On Location Privacy in IoT Sensor Data. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 128–140.

[44] Maria Petrescu and Anjala S Krishen. 2018. Analyzing the analytics: data privacy concerns. , 41–43 pages.

[45] Siddharth Reddy, Sergey Levine, and Anca Dragan. 2017. Accelerating human learning with deep reinforcement learning. In *NIPS workshop: teaching machines, robots, and humans*.

[46] Stephanie Rosenthal and Manuela Veloso. 2011. Modeling humans as observation providers using pomdps. In *RO-MAN, 2011 IEEE*. IEEE, 53–58.

[47] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. 2017. Active Preference-Based Learning of Reward Functions.. In *Robotics: Science and Systems*.

[48] Sopicha Stirapongsasuti, Wataru Sasaki, and Keiichi Yasumoto. 2019. Decision making support for privacy data upload in smart home. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 214–217.

[49] William Bruce Stone III. 2017. *Psychometric evaluation of the Simulator Sickness Questionnaire as a measure of cybersickness*. Ph. D. Dissertation. Iowa State University.

[50] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[51] Shogo Terai, Shizuka Shirai, Mehrasa Alizadeh, Ryosuke Kawamura, Noriko Takemura, Yuki Uranishi, Haruo Takemura, and Hajime Nagahara. 2020. 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*. 124–125.

[52] The New York Times. 2022. The Best Smart Thermostat. https://www.nytimes.com/wirecutter/reviews/the-best-thermostat/. Accessed: 2022-06-10.

[53] Amos Tversky and Daniel Kahneman. 1979. Prospect Theory - Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–292.

[54] Baoxiang Wang and Nidhi Hegde. 2019. Privacy-preserving q-learning with functional noise in continuous spaces. *Advances in Neural Information Processing Systems* 32 (2019).

[55] Wenjing Zhang, Bo Jiang, Ming Li, and Xiaodong Lin. 2022. Privacy-Preserving Aggregate Mobility Data Release: An Information-Theoretic Deep Reinforcement Learning Approach. *IEEE Transactions on Information Forensics and Security* 17 (2022), 849–864.