A Framework for Effective AI Recommendations in Cyber-Physical-Human Systems

Aditya Dave, *Member, IEEE*, Heeseung Bang, *Student Member, IEEE*, and Andreas A. Malikopoulos, *Senior Member, IEEE*

***Abstract*—Many cyber-physical-human systems (CPHSs) involve a human decision-maker who acts using recommendations from an artificial intelligence (AI) platform. In such CPHS applications, the human decision-maker may depart from an optimal recommended decision and instead implement a different one for various reasons, resulting in a loss in performance. In this letter, we develop a rigorous framework to overcome this challenge. In our framework, humans may deviate from AI recommendations as they interpret the system's state differently to the AI platform. We establish the structural properties of optimal recommendation strategies and develop an approximate human model (AHM) used by the AI. We provide theoretical bounds on the optimality gap that arises from an AHM and illustrate the efficacy of our results in a numerical example.**

***Index Terms*—Cyber-physical human systems, human-AI interaction, human model, recommender systems.**

I. Introduction

IN SEVERAL cyber-physical-human systems (CPHSs), e.g., aircraft co-pilot [1], autonomous driving [2], and social media [3], a human decision-maker may receive recommendations from an artificial intelligence (AI) platform while holding the ultimate responsibility of making decisions. For example, consider a traffic environment [4] where a human driver receives a route recommendation from an AI platform that runs a central traffic management system. In such applications, the human's actual decision may depart from a recommended decision for various reasons [5], including: (1) different interpretations of the system's state to the AI platform; (2) different objectives than those designated for the AI; or (3) high confidence in their inherent decision-making ability. This human influence during decision-making [6] renders CPHSs challenging to control [7].

To better understand this phenomenon, there has been recent interest in learning [8] and empirically developing models for human behavior [9] during collaborations with AI platforms. It has been established that humans are likely to adhere to recommendations that are easy to interpret and reaffirm their preconceived opinions [10]. Alternatively, humans may disregard recommendations that cause discomfort [11] or misinterpret recommendations [12], worsening the overall performance [13]. In response to these findings, many research efforts have focused on increasing human trust in AI [14] and adoption of recommendations [15]. However, there remains a need to design AI recommendations that account for human behavior.

The adherence-aware Markov decision process is one approach to formalize human-AI interactions by limiting human behavior to two choices: they may either accept or reject AI suggestions, as dictated by their adherence probability [16]. In this context, optimal recommendations can be derived for humans using Q-learning [17]. Furthermore, this framework has motivated reinforcement learning approaches that consider whether an AI platform should abstain from recommending decisions [18]. While promising, these results rely upon the simple model of human behavior. Human-AI collaboration has also been analyzed under other specific models for human behavior, including opinion aggregation [12] and trust evolution [19]. However, the resulting recommendations typically require knowledge of an "internal state space" of the human, and require observations of the internal state to learn a human model. Practically, we may not have access to human internal states nor to a reliable model for their evolution in most CPHS applications. Thus, we need a more general approach to AI recommendations with relaxed assumptions.

In this letter, we present such a general framework for effective AI recommendations to humans in CPHSs. We impose minimal assumptions on human behavior and develop our theory to support both empirical modeling and learning from interactions. Our contributions in this letter are (1) a framework for learning-based recommendations in CPHSs that generalizes many state-of-the-art models for human-behavior [12], [17], [19] through a human-AI POMDP (Lemma 1) and the structure of optimal recommendation strategies (Theorem 1); and (2) the introduction of an "approximate human model" (Definition 1) that yields approximate recommendation strategies with performance guarantees (Theorem 2). We illustrate the efficacy of our framework in a numerical example.

The remainder of this letter proceeds as follows. In Section II, we present our formulation. In Section III, we

Manuscript received 8 March 2024; revised 11 May 2024; accepted 29 May 2024. Date of publication 5 June 2024; date of current version 26 June 2024. This work was supported by the NSF under Grant CNS-2149520 and Grant CMMI-2219761. Recommended by Senior Editor K. Savla. *(Corresponding author: Aditya Dave.)*

The authors are with the School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: a.dave@cornell.edu; hb489@cornell.edu; amaliko@cornell.edu).

Digital Object Identifier 10.1109/LCSYS.2024.3410145

analyze the structure of optimal recommendations, propose an approximate human model, and derive approximation bounds. In Section IV, we present a numerical example, and in Section V, we draw concluding remarks.

A. Notation

We denote random variables by upper case letters and their realizations by the corresponding lower case letters. The random variable X is said to take values in a set $\mathcal{X}$ if its realizations are restricted to $\mathcal{X}$. For integers $a < b$, $X_{a:b}$ is shorthand for $(X_a, X_{a+1}, \dots, X_b)$. We denote the probability of a random variable X taking a realization $x \in \mathcal{X}$ given Y takes realization $y \in \mathcal{Y}$ concisely as $\mathsf{P}(X = x \mid Y = y) = \mathsf{P}(X = x \mid y) = \mathsf{P}(x|y)$. Similarly, $\mathsf{P}(X \mid y)$ represents the complete distribution on X given realization y. Finally, $\mathsf{P}(X|Y)$ is itself a random variable taking realizations in the space $\Delta(\mathcal{X})$ of distributions on X. The indicator function $\mathsf{I}[\cdot]$ returns 0 if the condition within $[\cdot]$ is true and 0 otherwise.

II. Problem Formulation

We consider an AI platform that recommends decisions to a human in a CPHS. The human is responsible for implementing actions that influence the system's evolution. In this context, the human implements a decision by incorporating the platform's recommendations with an instinctive understanding of the situation, as illustrated in Fig. 1. Thus, the AI platform must account for the possibility that a human may re-interpret or disregard the recommended actions. The CPHS has a finite state space $\mathcal{X}$, and the human selects actions from a finite feasible set $\mathcal{U}$. The system evolves over discrete time steps until a finite horizon $T \in \mathbb{N}$. At each time $t \in \mathcal{T} = \{0, 1, \dots, T\}$, the state of the system is denoted by the random variable $X_t \in \mathcal{X}$ and the action implemented by the human is denoted by the random variable $U^{\mathrm{h}}_t \in \mathcal{U}$. Starting at the initial state $X_0 \in \mathcal{X}$, the system evolves for all t as $X_{t+1} = f(X_t, U^{\mathrm{h}}_t, W_t)$, where $W_t \in \mathcal{W}$ is a random variable that corresponds to an uncontrollable disturbance. The disturbances form a sequence of independent random variables $\{W_t : t \in \mathcal{T}\}$ that are also independent of the initial state X_0.

The system state X_t is observed by both the human and the AI platform at each $t \in \mathcal{T}$. The platform generates a recommendation to guide the human's eventual action. This recommendation is denoted by a random variable U^{ai}_t that takes values in the human's space of feasible actions $\mathcal{U}$. At each $t \in \mathcal{T}$, the platform provides U^{ai}_t based on the history $H_t = (X_{0:t}, U^{\mathrm{h}}_{0:t-1}, U^{\mathrm{ai}}_{0:t-1})$ taking values in the set $\mathcal{H}_t = \mathcal{X}^{(t+1)} \times \mathcal{U}^{2t}$, where $H_0 = X_0$. The recommendation strategy is $\boldsymbol{g}^{\mathrm{ai}} = (g^{\mathrm{ai}}_0, \dots, g^{\mathrm{ai}}_T)$, with recommendation laws $g^{\mathrm{ai}}_t : \mathcal{H}_t \to \mathcal{U}$ leading to $U^{\mathrm{ai}}_t = g^{\mathrm{ai}}_t(H_t)$ for all $t \in \mathcal{T}$.

At each $t \in \mathcal{T}$, the human receives the recommendation before deciding which action to implement. This decision is also affected by their own internal state, denoted by the random variable S_t taking values in a finite space $\mathcal{S}$. An internal state represents a combination of the human's interpretation of the system state, amenability towards AI suggestions, self-confidence, or a variety of other factors affecting the human's choices. Starting at $S_0 = f^{\mathrm{h}}_0(X_0, N_0) \in \mathcal{S}$, the internal state evolves for all $t \in \mathcal{T}$ as $S_{t+1} = f^{\mathrm{h}}(S_t, U^{\mathrm{ai}}_t, X_{t+1}, N_{t+1})$, where

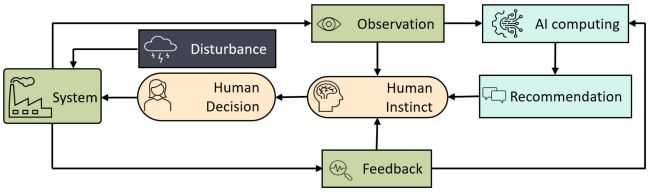

Fig. 1. Control loop of the recommendation problem.

N_t is an uncontrolled disturbance that takes values in a finite set $\mathcal{N}$ for all $t \in \mathcal{T}$. This disturbance represents stochastic uncertainties in the evolution of the human's internal state. The sequence of uncertainties $\{N_t : t \in \mathcal{T}\}$ are independent of each other, of X_0, and of the disturbances $\{W_t : t \in \mathcal{T}\}$. Then, the human uses a control law $g^{\mathrm{h}} : \mathcal{S} \times \mathcal{U} \to \mathcal{U}$ to implement the action $U^{\mathrm{h}}_t = g^{\mathrm{h}}(S_t, U^{\mathrm{ai}}_t)$ at each $t \in \mathcal{T}$. Note that we consider the influence of the history H_t in the human's action implicitly through the internal state S_t, and we consider the influence of the internal state S_t on the system's next state X_{t+1} indirectly through the human's action U^{h}_t. Subsequently, both the human and the AI platform receive shared feedback from the system, generated using the reward function $r : \mathcal{X} \times \mathcal{U} \to [r^{\min}, r^{\max}]$, where $r^{\min}, r^{\max} \in \mathbb{R}$. We denote this feedback by the random variable $R_t = r(X_t, U^{\mathrm{h}}_t) = r(X_t, g^{\mathrm{h}}(S_t, U^{\mathrm{ai}}_t))$. The AI platform seeks to maximize the expected total reward:

$$J\left(\boldsymbol{g}^{\mathrm{ai}}\right) = \mathsf{E}^{\boldsymbol{g}^{\mathrm{ai}}}\left[\sum_{t=0}^{T} \gamma^t \cdot r\left(X_t, g^{\mathrm{h}}\left(S_t, U^{\mathrm{ai}}_t\right)\right)\right], \tag{1}$$

where $\mathsf{E}^{\boldsymbol{g}^{\mathrm{ai}}}[\cdot]$ is the expectation with respect to the joint distribution imposed by strategy $\boldsymbol{g}^{\mathrm{ai}}$, when human actions use the control law g^{h}, and $\gamma \in (0, 1)$ is a discount factor. This yields the following optimization problem for the platform.

Problem 1: The AI platform seeks an optimal recommendation strategy $\boldsymbol{g}^{*\mathrm{ai}}$, such that $J(\boldsymbol{g}^{*\mathrm{ai}}) \geq J(\boldsymbol{g}^{\mathrm{ai}})$, given the sets $\{\mathcal{X}, \mathcal{W}, \mathcal{U}\}$ and function f.

An optimal strategy $\boldsymbol{g}^{*\mathrm{ai}}$ exists because all variables are finite valued, but it may not be computable without knowledge of $\mathcal{S}$, g^{h}, and f^{h}. We impose the following assumptions.

Assumption 1: The action of the human U^{h}_t and the reward R_t are perfectly observed by the AI platform at each $t \in \mathcal{T}$.

Assumption 1 implies that the human and the AI platform receive consistent rewards and allows the platform to anticipate the human. Additional analysis is needed when humans interpret rewards differently to a platform, e.g., economic systems [9]. Similarly, the AI platform's observation of human actions facilitates our learning approach and in its absence, the platform may be unable to identify the human's influence [18].

III. Recommendation Framework

In this section, we develop our theoretical framework to compute optimal recommendations. In Section III-A, we analyze an AI platform with access to the true model for a human's behavior. This analysis yields a structural form for optimal AI recommendations. Building upon this structure and taking inspiration from recent work in partially observed reinforcement learning [20], [21], we define the notion of an approximate human model (AHM) in Section III-B. We

show that an AI platform can use an AHM to compute recommendations with performance guarantees. Finally, we propose an approach to construct an AHM in Section III-C.

A. Optimal Recommendation Strategies

We start our exposition by considering that the AI platform knows a priori an exact human model consisting of the set of internal states $\mathcal{S}$, an initial distribution on S_0, the function $f^{\mathrm{h}}(\cdot)$, and the human's control law $g^{\mathrm{h}}(\cdot)$. However, the platform does not observe S_t at any $t \in \mathcal{T}$. Next, we prove that such a system constitutes a partially observable Markov decision process (named the human-AI POMDP) for the platform.

Lemma 1: Given a human model, Problem 1 is equivalent to computing the optimal strategy in a POMDP with state $(X_t, S_t) \in \mathcal{X} \times \mathcal{S}$, input $U_t^{ai} \in \mathcal{U}$, observation $(X_t, U_{t-1}^{h}) \in \mathcal{X} \times \mathcal{U}$, and reward $R_t \in [r^{\min}, r^{\max}]$ for all $t \in \mathcal{T}$.

Proof: We establish that $\mathcal{X} \times \mathcal{S}$ is the state space for the POMDP by showing that it predicts (i) the reward and (ii) the joint distribution on the next state and observation. For (i), recall from Section II that $R_t = r(X_t, g^{\mathrm{h}}(S_t, U_t^{\mathrm{ai}}))$ at each $t \in \mathcal{T}$. For (ii), for all $t \in \mathcal{T}$, consider any jointly feasible realization $m_t = (x_{0:t}, s_{0:t}, u_{0:t-1}^{\mathrm{h}}, u_{0:t}^{\mathrm{ai}})$ of the associated random variable $M_t = (X_{0:t}, S_{0:t}, U_{0:t-1}^{\mathrm{h}}, U_{0:t}^{\mathrm{ai}})$. Using the Markov property, we state that $\mathsf{P}(x_{t+1}, s_{t+1}, u_t^{\mathrm{h}} \mid m_t) = \mathsf{I}(u_t^{\mathrm{h}} = g^{\mathrm{h}}(s_t, u_t^{\mathrm{ai}})) \cdot \mathsf{P}^{f^{\mathrm{h}}}(s_{t+1} \mid s_t, x_{t+1}, u_t^{\mathrm{ai}}) \cdot \mathsf{P}^{g^{\mathrm{h}}}(x_{t+1} \mid x_t, g^{\mathrm{h}}(s_t, u_t^{\mathrm{ai}})) = \mathsf{P}(x_{t+1}, s_{t+1}, u_t^{\mathrm{h}} \mid x_t, s_t, u_t^{\mathrm{ai}})$, where $\mathsf{I}(\cdot)$ is the indicator function. Thus, $\mathcal{X} \times \mathcal{S}$ satisfies the Markov property for the state and next observation, rendering it a valid state space for the POMDP. Finally, the expected total discounted reward under any strategy $\boldsymbol{g}^{\mathrm{ai}}$ in this POMDP is the same as (1), implying that this POMDP yields the solution to Problem 1. ∎

We can construct a dynamic programming (DP) decomposition for the human-AI POMDP in Lemma 1 using the history H_t at each $t \in \mathcal{T}$. To this end, for each realization $h_t \in \mathcal{H}_t$ and $u_t^{\mathrm{ai}} \in \mathcal{U}$, for all $t \in \mathcal{T}$, we define the value functions

$$Q_t\left(h_t, u_t^{\mathrm{ai}}\right) := \mathsf{E}\left[r\left(X_t, U_t^{\mathrm{h}}\right) + \gamma \cdot V_{t+1}(H_{t+1}) \mid h_t, u_t^{\mathrm{ai}}\right], \quad (2)$$

$$V_t(h_t) := \max_{u_t^{\mathrm{ai}} \in \mathcal{U}} Q_t\left(h_t, u_t^{\mathrm{ai}}\right), \quad (3)$$

where, $V_{T+1}(h_{T+1}) := 0$ identically, $U_t^{\mathrm{h}} = g^{\mathrm{h}}(S_t, U_t^{\mathrm{ai}})$, and $H_{t+1} = (H_t, X_{t+1}, U_t^{\mathrm{h}}, U_t^{\mathrm{ai}})$ for all t. The recommendation law computed by this DP at each $t \in \mathcal{T}$ is $g_t^{*\mathrm{ai}}(h_t) := \arg\max_{u_t^{\mathrm{ai}}} Q_t(h_t, u_t^{\mathrm{ai}})$. Standard arguments for POMDPs can be used to prove that the resulting recommendation strategy $\boldsymbol{g}^{*\mathrm{ai}} := g_{0:T}^{*\mathrm{ai}}$ is an optimal solution to the POMDP and consequently, to Problem 1 [22]. However, this DP decomposition suffers from an increase in computational complexity as the history grows in size with time t. Furthermore, it does not provide insights into the underlying structure of optimal recommendation strategies. Typically, these challenges are overcome in POMDPs using an information state that compresses the history into a sufficient statistic [23]. Thus, we construct an information state for the human-AI POMDP. To begin, we define a sufficient statistic for all $t \in \mathcal{T}$ as the AI's belief on the internal state $B_t := \mathsf{P}(S_t \mid H_t) \in \Delta(\mathcal{S})$. Note that the belief is itself a random variable taking values in the space of distributions. We denote its realization by the lowercase $b_t \in \Delta(\mathcal{S})$. Next, we prove two important properties of the sufficient statistic.

Lemma 2: For any given realization $h_t = (\tilde{x}_{0:t}, \tilde{u}_{0:t-1}^{\mathrm{ai}}, \tilde{u}_{0:t-1}^{\mathrm{h}}) \in \mathcal{H}_t$ of the history at time $t \in \mathcal{T}$, the internal state and system state are conditionally independent, i.e., for any $s_t \in \mathcal{S}$ and $x_t \in \mathcal{X}$:

$$\mathsf{P}(S_t = s_t, X_t = x_t \mid h_t) = b_t(s_t) \cdot \mathsf{I}\left(x_t = \tilde{x}_t\right). \quad (4)$$

Proof: Let $h_t = (\tilde{x}_{0:t}, \tilde{u}_{0:t-1}^{\mathrm{ai}}, \tilde{u}_{0:t-1}^{\mathrm{h}}) \in \mathcal{H}_t$, $s_t \in \mathcal{S}$ and $x_t \in \mathcal{X}$ denote the realizations of the associated random variables for all $t \in \mathcal{T}$. The result holds directly from the fact that the realization of X_t is included in the history h_t and using the definition of the realized belief $b_t(s_t) = \mathsf{P}(S_t = s_t \mid h_t)$. ∎

Lemma 3: We can construct a function $\psi : \Delta(\mathcal{S}) \times \mathcal{U} \times \mathcal{X} \to \Delta(\mathcal{S})$ independent of the choice of $\boldsymbol{g}^{ai}$ (given a control action U_t^{ai}), such that

$$B_{t+1} = \psi\left(B_t, U_t^{ai}, X_{t+1}\right), \quad \forall t \in \mathcal{T}. \quad (5)$$

Proof: For all $t \in \mathcal{T}$ and any realizations $s_{t+1} \in \mathcal{S}_t$ and $h_{t+1} = (h_t, x_{t+1}, u_t^{\mathrm{h}}, u_t^{\mathrm{ai}}) \in \mathcal{H}_{t+1}$, using the law of total probability we obtain $b_{t+1}(s_{t+1}) = \mathsf{P}(s_{t+1} \mid h_t, x_{t+1}, u_t^{\mathrm{h}}, u_t^{\mathrm{ai}}) = \sum_{\tilde{s}_t} \mathsf{P}(s_{t+1} \mid \tilde{s}_t, u_t^{\mathrm{ai}}, x_{t+1}) \cdot \mathsf{P}(\tilde{s}_t \mid h_t) =: \psi(b_t, u_t^{\mathrm{ai}}, x_{t+1})(s_{t+1})$. Thus, we can construct ψ that satisfies (5) independent of the choice of $\boldsymbol{g}^{\mathrm{ai}}$. ∎

Next, we propose an information state as $\Pi_t := (B_t, X_t)$ for all $t \in \mathcal{T}$, and establish that Π_t can evaluate the expected cost and next observations in the human AI POMDP.

Lemma 4: For all $t \in \mathcal{T}$, given realizations $h_t = (x_{0:t}, u_{0:t-1}^{\mathrm{ai}}, u_{0:t-1}^{\mathrm{h}}) \in \mathcal{H}_t$, $u_t^{ai} \in \mathcal{U}$, and $\pi_t = (b_t, x_t)$, we have that $\mathsf{E}[r(X_t, U_t^{\mathrm{h}}) \mid h_t, u_t^{ai}] = \mathsf{E}[r(X_t, U_t^{\mathrm{h}}) \mid \pi_t, u_t^{ai}]$.

Proof: At any $t \in \mathcal{T}$, we can write that $\mathsf{E}[r(X_t, U_t^{\mathrm{h}}) \mid h_t, u_t^{\mathrm{ai}}] = \sum_{s_t} r(x_t, g^{\mathrm{h}}(s_t, u_t^{\mathrm{ai}})) \cdot \mathsf{P}(s_t \mid h_t, u_t^{\mathrm{ai}}) = \sum_{s_t} r(x_t, g^{\mathrm{h}}(s_t, u_t^{\mathrm{ai}})) \cdot b_t(s_t) = \mathsf{E}[r(X_t, U_t^{\mathrm{h}}) \mid \pi_t, u_t^{\mathrm{ai}}]$ where, we use Lemma 2 in the second equality. ∎

Lemma 5: For all $t \in \mathcal{T}$, for any realizations $h_t = (x_{0:t}, u_{0:t-1}^{\mathrm{ai}}, u_{0:t-1}^{\mathrm{h}}) \in \mathcal{H}_t$ and $u_t^{ai} \in \mathcal{U}$, the corresponding realization π_t of Π_t satisfies for all $x_{t+1} \in \mathcal{X}$ and $u_t^{\mathrm{h}} \in \mathcal{U}$:

$$\mathsf{P}\left(X_{t+1} = x_{t+1}, U_t^{h} = u_t^{\mathrm{h}} \mid h_t, u_t^{ai}\right) = \mathsf{P}\left(X_{t+1} = x_{t+1}, U_t^{h} = u_t^{\mathrm{h}} \mid \pi_t, u_t^{ai}\right). \quad (6)$$

Proof: To prove the result, consider the realizations $x_{t+1} \in \mathcal{X}$ and $u_t^{\mathrm{h}} \in \mathcal{U}$ for any $t \in \mathcal{T}$. Using the law of total probability and Markov property, we can expand the probability in (6) as $\mathsf{P}(x_{t+1}, u_t^{\mathrm{h}} \mid h_t, u_t^{\mathrm{ai}}) = \mathsf{P}(x_{t+1} \mid x_t, u_t^{\mathrm{h}}) \cdot \sum_{\tilde{s}_t} \mathsf{I}[u_t^{\mathrm{h}} = g_t^{\mathrm{h}}(\tilde{s}_t, u_t^{\mathrm{ai}})] \cdot b_t(\tilde{s}_t) = \mathsf{P}(x_{t+1}, u_t^{\mathrm{h}} \mid \pi_t, u_t^{\mathrm{ai}})$, where, we use Lemma 2 in the first equality. ∎

Using the preceding results, we establish that Π_t is an information state that it yields an optimal DP decomposition.

Theorem 1: For all $t \in \mathcal{T}$, the random variable $\Pi_t = (B_t, X_t)$ is an information state of the human-AI POMDP. Furthermore, for all realizations $\pi_t \in \Delta(\mathcal{S}) \times \mathcal{X}$ and $u_t^{ai} \in \mathcal{U}$, let $\bar{Q}_t(\pi_t, u_t^{ai}) := \mathsf{E}[r(X_t, U_t^{h}) + \gamma \cdot \bar{V}_{t+1}(\Pi_{t+1}) \mid \pi_t, u_t^{ai}]$ and $\bar{V}_t(\pi_t) := \max_{u_t^{ai} \in \mathcal{U}} \bar{Q}_t(\pi_t, u_t^{ai})$, where $\bar{V}_{T+1}(\pi_{T+1}) := 0$. Then, an optimal recommendation law in Problem 1 is $\bar{g}_t^{*ai}(\pi_t) := \arg\max_{u_t^{ai}} \bar{Q}_t(\pi_t, u_t^{ai})$ for all t.

Proof: Lemmas 3 - 5 establish that Π_t is sufficient to evaluate the expected cost, evolves in a state-like manner, and

is sufficient to predict future observations for all $t \in \mathcal{T}$, hence it satisfies the standard conditions reported in [20, Definition 3] of an information state. As a direct consequence of the properties [20, Th. 5] and Lemma 1, the recommendation strategy $\bar{\boldsymbol{g}}^{*\mathrm{ai}} = \bar{g}^{*\mathrm{ai}}_{0:T}$ is an optimal solution to Problem 1. ■

Theorem 1 establishes that there is no loss of optimality when the AI platform utilizes the belief B_t to compute optimal recommendations at each $t \in \mathcal{T}$. However, in most applications, the platform will not have access to an exact model for human behavior to compute B_t. Thus, in the next subsection, we define the notion of an AHM that can either be designed heuristically or learned from data to compute approximate recommendations.

Remark 1: In Problem 1, consider instead that the system's state X_t is partially observed by both the human and the AI platform, i.e., both the AI and human receive an observation $Y_t = h(X_t, Z_t) \in \mathcal{Y}$ with noise $Z_t \in \mathcal{Z}$. Then, the internal state's evolution can be expressed as $S_{t+1} = f^{\mathrm{h}}(S_t, U^{\mathrm{ai}}_t, Y_{t+1}, N_t)$ for $t \in \mathcal{T}$ and $S_0 = f^{\mathrm{h}}_0(Y_0)$. Similarly, the history at any $t \in \mathcal{T}$ is $H_t = (Y_{0:t}, U^{\mathrm{ai}}_{0:t-1}, U^{\mathrm{h}}_{0:t-1})$. Here, we can use a similar sequence of arguments as in Theorem 1 to prove that (B_t, Q_t) is an information state for Problem 1, where $Q_t = \mathsf{P}(X_t \mid H_t) \in \Delta(\mathcal{X})$ is a belief on the state X_t at time t held independent of the belief B_t on the internal state S_t. This result is proved in our online preprint [24].

B. Approximate Human Model

In this subsection, we define the notion of an AHM that can be used by an AI platform instead of an exact human model.

Definition 1: An *approximate human model* consists of a Borel space $\hat{\mathcal{S}}$, an evolution equation $\hat{\sigma}_t : \mathcal{H}_t \to \hat{\mathcal{S}}$, and a probability mass function $\hat{\mu} : \hat{\mathcal{S}} \times \mathcal{U} \to \Delta(\mathcal{U})$, such that the approximate internal state $\hat{S}_t := \hat{\sigma}_t(H_t)$ satisfies for all $t \in \mathcal{T}$:

1) Evolution in a belief-like manner: There exists a function $\hat{\psi} : \hat{\mathcal{S}} \times \mathcal{U} \times \mathcal{X} \to \hat{\mathcal{S}}$ independent of the choice of recommendation strategy $\boldsymbol{g}^{\mathrm{ai}}$, such that

$$\hat{S}_{t+1} = \hat{\psi}\left(\hat{S}_t, U^{\mathrm{ai}}_t, X_{t+1}\right). \tag{7}$$

2) Approximate prediction of human actions: For any realization $h_t \in \mathcal{H}_t$ and $u^{\mathrm{ai}}_t \in \mathcal{U}$, the probability distribution induced by $\hat{\mu}$ is such that for some $\varepsilon > 0$:

$$\delta^{\mathrm{TV}}\left(\mathsf{P}^{g^{\mathrm{h}}}\left(U^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t\right), \hat{\mu}\left(U^{\mathrm{h}}_t \mid \hat{\sigma}_t(h_t), u^{\mathrm{ai}}_t\right)\right) \le \varepsilon, \tag{8}$$

where $\delta^{\mathrm{TV}}(\cdot,\cdot)$ is the total variation distance and $\mathsf{P}^{g^{\mathrm{h}}}(\cdot)$ is the conditional probability distribution induced on U^{h}_t by the human's choice of control law g^{h}.

Remark 2: The total variation distance between any two probability mass functions P and Q on a finite set $\mathcal{A}$ is defined as $\delta^{\mathrm{TV}}(\mathsf{P},\mathsf{Q}) := \frac{1}{2}\sum_{a\in\mathcal{A}} |\mathsf{P}(a) - \mathsf{Q}(a)|$.

Remark 3: The AHM is directly inspired by the properties of the belief B_t in Section III-A. The first property imposes the structure in Lemma 3 and the second property is essential to approximate the results of Lemmas 4 - 5 later in Lemma 6.

Remark 4: From Definition 1, any empirically designed or learned model qualifies as an AHM if it satisfies the conditions (7) and (8). Note that (7) can be ensured for an AHM by construction, and (8) can be verified using an empirical distribution from sampled observations of U^{h}_t without knowledge of the true distribution $\mathsf{P}^{g^{\mathrm{h}}}(U^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t)$.

Given an AHM, we define the random variable $\hat{\Pi}_t := (\hat{S}_t, X_t)$ for all $t \in \mathcal{T}$. Next, we prove that $\hat{\Pi}_t$ approximates the information state of the human-AI POMDP at each t, and it yields an approximate recommendation strategy using the following DP decomposition. For all $t \in \mathcal{T}$, for all realizations $\hat{\pi}_t \in \hat{\mathcal{S}} \times \mathcal{X}$ and $u^{\mathrm{ai}}_t \in \mathcal{U}$, we recursively define

$$\hat{Q}_t\left(\hat{\pi}_t, u^{\mathrm{ai}}_t\right) := \mathsf{E}\left[r\left(X_t, U^{\mathrm{h}}_t\right) + \gamma \hat{V}_{t+1}\left(\hat{\Pi}_{t+1}\right) \mid \hat{\pi}_t, u^{\mathrm{ai}}_t\right], \tag{9}$$

$$\hat{V}_t(\hat{\pi}_t) := \max_{u^{\mathrm{ai}}_t \in \mathcal{U}} \hat{Q}_t\left(\hat{\pi}_t, u^{\mathrm{ai}}_t\right), \tag{10}$$

where $\hat{V}_{T+1}(\hat{\pi}_{T+1}) := 0$ identically. Then, the corresponding recommendation law is $\hat{g}^{*\mathrm{ai}}_t(\hat{\pi}_t) := \arg\max_{u^{\mathrm{ai}}_t} \hat{Q}_t(\hat{\pi}_t, u^{\mathrm{ai}}_t)$ for all $t \in \mathcal{T}$. In (9), the distribution on U^{h}_t is $\hat{\mu}_t$ from the AHM. Note that the DP in (9)–(10) takes the same structural form as (2)–(3) while utilizing $\hat{\pi}_t$ instead of h_t at each t. Next, we prove an essential property.

Lemma 6: At any $t \in \mathcal{T}$, for any realizations $h_t \in \mathcal{H}_t$ and $u^{\mathrm{ai}}_t \in \mathcal{U}$, the corresponding realization $\hat{\pi}_t \in \hat{\mathcal{S}} \times \mathcal{X}$ satisfies:

$$a) \quad \left|\mathsf{E}^{g^h}\left[r\left(X_t, U^h_t\right) \mid h_t, u^{ai}_t\right] - \mathsf{E}^{\hat{\mu}}\left[r\left(X_t, U^h_t\right) \mid \hat{\pi}_t, u^{ai}_t\right]\right| \le 2r^{\max} \cdot \varepsilon, \tag{11}$$

$$b) \quad \delta^{TV}\left(\mathsf{P}^{g^h}\left(X_{t+1}, U^h_t | h_t, u^{ai}_t\right), \mathsf{P}^{\hat{\mu}}\left(X_{t+1}, U^h_t | \hat{\pi}_t, u^{ai}_t\right)\right) \le \varepsilon. \tag{12}$$

Proof: At any $t \in \mathcal{T}$, for a given realization $h_t = (x_{0:t}, u^{\mathrm{ai}}_{0:t-1}, u^{\mathrm{h}}_{0:t-1}) \in \mathcal{H}_t$ of the history, $\hat{\pi}_t = (\hat{\sigma}_t(h_t), x_t)$.

a) To prove (11), we expand the expected rewards under the distributions generated by $\mathsf{P}^{g^{\mathrm{h}}}$ and $\hat{\mu}$, i.e., $|\mathsf{E}^{g^{\mathrm{h}}}[r(X_t, U^{\mathrm{h}}_t) \mid h_t, u^{\mathrm{ai}}_t] - \mathsf{E}^{\hat{\mu}}[r(X_t, U^{\mathrm{h}}_t) \mid \hat{\sigma}_t(h_t), x_t, u^{\mathrm{ai}}_t]| = |\sum_{\tilde{u}^{\mathrm{h}}_t} r(x_t, \tilde{u}^{\mathrm{h}}_t) \cdot \mathsf{P}^{g^{\mathrm{h}}}(u^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t) - \sum_{\tilde{u}^{\mathrm{h}}_t} r(x_t, \tilde{u}^{\mathrm{h}}_t) \cdot \hat{\mu}(u^{\mathrm{h}}_t \mid \hat{\sigma}_t(h_t), u^{\mathrm{ai}}_t)| \le 2r^{\max} \cdot \varepsilon$, where, in the inequality, we use the definition of total variation distance in Remark 2, and the fact that $r^{\max}$ is an upper bound on the reward.

b) To prove (12), we first use the definition of the total variation distance and Bayes' law to write that $\delta^{\mathrm{TV}}(\mathsf{P}^{g^{\mathrm{h}}}(X_{t+1}, U^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t), \mathsf{P}^{\hat{\mu}}(X_{t+1}, U^{\mathrm{h}}_t \mid \hat{\pi}_t, u^{\mathrm{ai}}_t)) = \sum_{\tilde{x}_{t+1}, \tilde{u}^{\mathrm{h}}_t} \frac{1}{2}|\mathsf{P}^{g^{\mathrm{h}}}(\tilde{x}_{t+1}, \tilde{u}^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t) - \mathsf{P}^{\hat{\mu}}(\tilde{x}_{t+1}, \tilde{u}^{\mathrm{h}}_t \mid \hat{\pi}_t, u^{\mathrm{ai}}_t)| = \sum_{\tilde{x}_{t+1}, \tilde{u}^{\mathrm{h}}_t} \frac{1}{2}|\mathsf{P}^{g^{\mathrm{h}}}(\tilde{x}_{t+1} \mid h_t, \tilde{u}^{\mathrm{h}}_t) \cdot \mathsf{P}^{g^{\mathrm{h}}}(\tilde{u}^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t) - \mathsf{P}^{\hat{\mu}}(\tilde{x}_{t+1} \mid h_t, \tilde{u}^{\mathrm{h}}_t) \cdot \hat{\mu}(\tilde{u}^{\mathrm{h}}_t \mid \hat{\sigma}_t(h_t), u^{\mathrm{ai}}_t)|$. Here, using the Markov property, $\mathsf{P}^{g^{\mathrm{h}}}(\tilde{x}_{t+1} \mid h_t, \tilde{u}^{\mathrm{h}}_t) = \mathsf{P}(\tilde{x}_{t+1}|x_t, u^{\mathrm{h}}_t) = \mathsf{P}(\tilde{x}_{t+1}|\hat{\pi}_t, u^{\mathrm{h}}_t) = \mathsf{P}^{\hat{\mu}}(\tilde{x}_{t+1}|\hat{\pi}_t, u^{\mathrm{h}}_t)$, where, in the second equality, we use the fact that $\hat{\pi}_t$ contains x_t as a component; and in the fourth equality, we note that the transition probability is independent of $\hat{\mu}$. Substituting this result, we have that $\delta^{\mathrm{TV}}(\mathsf{P}^{g^{\mathrm{h}}}(X_{t+1}, U^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t), \mathsf{P}^{\hat{\mu}}(X_{t+1}, U^{\mathrm{h}}_t \mid \hat{\pi}_t, u^{\mathrm{ai}}_t)) \le \frac{1}{2}\sum_{\tilde{x}_{t+1}, \tilde{u}^{\mathrm{h}}_t} \mathsf{P}(\tilde{x}_{t+1}|\hat{\pi}_t, \tilde{u}^{\mathrm{h}}_t) \cdot |\mathsf{P}^{g^{\mathrm{h}}}(\tilde{u}^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t) - \hat{\mu}(\tilde{u}^{\mathrm{h}}_t \mid \hat{\sigma}_t(h_t), u^{\mathrm{ai}}_t)| \le \delta^{\mathrm{TV}}(\mathsf{P}^{g^{\mathrm{h}}}(U^{\mathrm{h}}_t \mid h_t, u^{\mathrm{ai}}_t), \hat{\mu}(U^{\mathrm{h}}_t \mid \hat{\sigma}_t(h_t), u^{\mathrm{ai}}_t)) \le \varepsilon$, where, we use Remark 2, $\mathsf{P}(\tilde{x}_{t+1}|\hat{\pi}_t, u^{\mathrm{h}}_t) \le 1$, and (8). ■

Using Lemma 6, we establish that the recommendation strategy $\hat{\boldsymbol{g}}^{*\mathrm{ai}}_t = \hat{g}^{*\mathrm{ai}}_{0:t}$ from (9)–(10) is an approximation.

Theorem 2: Let $||\hat{V}||_\infty$ be an upper bound on $\hat{V}_t(\hat{\pi}_t)$ for all $\hat{\pi}_t$ and $t \in \mathcal{T}$. Then, $\hat{\boldsymbol{g}}^{*ai}_t$ computed using (9)–(10) is an approximate recommendation strategy in Problem 1 with an optimality gap of $4\varepsilon \cdot (r^{\max} + \sum_{t=1}^{T} \gamma^t \cdot (||\hat{V}||_\infty + r^{\max}))$.

TABLE I
STATE TRANSITION PROBABILITIES

Action	State 0	1	2	3	4
A	[0.9, 0.1, 0.0, 0.0, 0.0]	[0.2, 0.7, 0.1, 0.0, 0.0]	[0.0, 0.2, 0.7, 0.1, 0.0]	[0.0, 0.0, 0.2, 0.7, 0.1]	[0.0, 0.0, 0.0, 0.0, 1.0]
B	[0.7, 0.2, 0.1, 0.0, 0.0]	[0.1, 0.6, 0.2, 0.1, 0.0]	[0.0, 0.1, 0.6, 0.2, 0.1]	[0.0, 0.0, 0.1, 0.6, 0.3]	[0.0, 0.0, 0.0, 0.0, 1.0]
C	[0.5, 0.3, 0.1, 0.1, 0.0]	[0.0, 0.4, 0.4, 0.1, 0.1]	[0.0, 0.0, 0.4, 0.4, 0.2]	[0.0, 0.0, 0.0, 0.4, 0.6]	[0.0, 0.0, 0.0, 0.0, 1.0]
D	[1.0, 0.0, 0.0, 0.0, 0.0]	[0.5, 0.5, 0.0, 0.0, 0.0]	[0.1, 0.6, 0.3, 0.0, 0.0]	[0.0, 0.2, 0.6, 0.2, 0.0]	[0.0, 0.0, 0.2, 0.6, 0.2]
E	[1.0, 0.0, 0.0, 0.0, 0.0]	[0.8, 0.2, 0.0, 0.0, 0.0]	[0.4, 0.5, 0.1, 0.0, 0.0]	[0.1, 0.4, 0.4, 0.1, 0.0]	[0.0, 0.1, 0.4, 0.4, 0.1]

Proof: Lemma 6 establishes that the random variable $\hat{\Pi}_t = (\hat{S}_t, X_t)$ is sufficient to approximately evaluate the expected cost in (11) and is sufficient to approximately predict future observations in (12) for all $t \in \mathcal{T}$. Furthermore, from (7) in Definition 1 and system dynamics, we conclude that $\hat{\Pi}_t$ evolves in a state-like manner. Hence, it satisfies the conditions reported in [22, Definition 2] to qualify as an (ϵ, δ)-approximate information state for the human-AI POMDP, with $\epsilon = 2r^{\max} \cdot \varepsilon$ and $\delta = \varepsilon$. The result follows by substituting ϵ and δ into the performance bounds in [22, Th. 3]. ■

Remark 5: Consider the partially observed system described in Remark 1. Using arguments analogous to Theorem 2, an AHM similar to that in Definition 1 can be utilized in this scenario and that the corresponding approximate information state is $\hat{\Pi}_t := (\hat{S}_t, Q_t)$, where $Q_t = \mathsf{P}(X_t \mid H_t) \in \Delta(\mathcal{X})$. This result is rigorously proved in our online preprint [24].

C. Constructing an Approximate Human Model

We use supervised learning to learn the AHM in Definition 1. We assume that we can access multiple realized trajectories $(x_{t+1}, u_t^{\mathrm{h}}, u_t^{\mathrm{ai}} : t \in \mathcal{T})$ generated using an exploratory AI strategy and used for training. Then, we select two function approximators as follows: (1) The encoder is a recurrent neural network (e.g., LSTM or GRU) denoted by $\phi : \hat{\mathcal{S}} \times \mathcal{X} \times \mathcal{U}^2 \to \hat{\mathcal{S}}$ whose hidden state will be treated as $\hat{S}_t$ at each $t \in \mathcal{T}$. Thus, the inputs to ϕ are $(\hat{S}_{t-1}, X_t, U_{t-1}^{\mathrm{h}}, U_{t-1}^{\mathrm{ai}})$ and its output is $\hat{S}_t$. Note that typically, a larger $\hat{\mathcal{S}}$ would lead to greater computational complexity but a better ability to model human behavior. Thus, this selection should be made in cognizance of the available training data, computational resources, and task complexity. (2) The decoder is a feed-forward neural network $\rho : \hat{\mathcal{S}} \times \mathcal{U} \to \Delta(\mathcal{U})$, whose inputs at each $t \in \mathcal{T}$ are $(\hat{S}_t, U_t^{\mathrm{ai}})$ and whose output is the conditional distribution $\hat{\mu}$, represented conveniently as a vector in the probability simplex $\Delta(\mathcal{U})$. During training, we pass the realizations $(\hat{s}_{t-1}, x_t, u_{t-1}^{\mathrm{h}}, u_{t-1}^{\mathrm{ai}})$ into the encoder and then the realizations $(\hat{s}_t, u_t^{\mathrm{ai}})$ into the decoder. Then, the "target" at time t is the realized human action u_t^{h} from the corresponding datapoint. Thus, we select a training loss $L = -\sum_{t=0}^{B} \log(\hat{\mu}_t(u_t^{\mathrm{h}}))$, where $\hat{\mu}(u_t^{\mathrm{h}})$ is the probability of the specific realization u_t^{h} in the distribution $\hat{\mu}$. This loss function approximates the Kullback–Leibler divergence between the true distribution and $\hat{\mu}$, which forms an upper bound on the total variation distance in (8) by Pinker's inequality. Then, we have the following approaches to construct and train an AHM:

1) Combining empirical models with learning: The main idea is to *empirically select* an AHM space $\hat{\mathcal{S}}$ and evolution equation $\hat{\psi}$. The choice of $\hat{\mathcal{S}}$ is based on factors affecting human behavior within a specific application and ensures $\hat{S}_{t+1} = \hat{\psi}(\hat{S}_t, U_t^{\mathrm{ai}}, X_{t+1})$ for all $t \in \mathcal{T}$. For example, in the partial adherence model [16], [17], $\hat{S}_t$ is the human's unchanging probability of adherence at each t, or the opinion aggregation model [12], where $\hat{S}_t$ is the human's self-confidence. In the previous two examples, $\hat{S}_{t+1} = \hat{S}_t$. To learn an AHM, we feed $\hat{S}_t$ from the empirical model and U_t^{ai} to the decoder ρ at each t and train ρ with loss L.

2) Using only supervised learning: When we cannot use domain knowledge, we learn an AHM from data by assuming an encoder-decoder architecture. We consider the encoder ϕ and feed its internal state $\hat{S}_t$ with U_t^{ai} to the decoder ρ at each $t \in \mathcal{T}$. We train the complete network assembly with loss L.

TABLE II
REWARDS

Action	State 0	1	2	3	4
A	0.0	0.7	1.0	-0.5	-1.0
B	0.0	0.5	1.0	-0.2	-0.4
C	0.0	0.3	0.7	0.1	-0.1
D	0.0	0.1	0.5	0.1	0.2
E	0.0	0.0	0.0	0.2	0.5

IV. NUMERICAL EXAMPLE

In this section, we illustrate our results with a simple example. We consider a machine operation and maintenance problem with a human operator who receives suggestions from an AI platform. The state $X_t \in \{0, 1, 2, 3, 4\}$ represents the status of the machine at each $t \in \mathcal{T}$. The possible actions are $\mathcal{U} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}\}$, where A is "rest", B is "minor operate", C is "major operate", D is "minor fix" and E is "major fix". At each $t \in \mathcal{T}$, the status evolves using the transition probabilities in Table I, and the rewards are in Table II.

We consider a human operator, whose internal state is a tuple $S_t = (\Theta_t, A_t)$, where $\Theta_t \in \{0, 0.1, \ldots, 1\}$ denotes their *trust* at any $t \in \mathcal{T}$ and $A_t \in \{0, 1\}$ denotes their adherence. Trust determines the probability of adherence and adherence determines whether the human implements AI recommendations. If $U_t^{\mathrm{ai}} = \mathrm{A}$, the human recovers trust by resting, i.e., $\Theta_{t+1} = \Theta_t + 0.3$. If recommendation is a "minor" work, i.e., $U_t^{\mathrm{ai}} = \mathrm{B}$ or D, then the trust goes down by $\Theta_{t+1} = \Theta_t - 0.02$. Similarly, if the recommendation is a "major" work, i.e., $U_t^{\mathrm{ai}} = \mathrm{C}$ or E, then $\Theta_{t+1} = \Theta_t - 0.05$. Meanwhile, the trust is also affected by the state of the machine, e.g., the trust drops by $\Theta_{t+1} = \Theta_t - 0.05$ if $X_t = 3$ or $\Theta_{t+1} = \Theta_t - 0.1$ if $X_t = 4$, i.e., when the machine is functioning poorly. Then, the adherence A_t takes values with the distribution $\mathsf{P}(A_t =$

TABLE III
AVERAGE REWARDS OVER 500 SIMULATIONS

Initial Trust	$T = 20$		$T = 30$	
	Naive DP	Recommendation	Naive DP	Recommendation
0.2	0.416	**0.608**	0.447	**0.604**
0.4	1.380	**1.939**	1.384	**1.992**
0.6	2.401	**3.244**	2.401	**3.610**
0.8	3.203	**4.422**	3.203	**4.924**
1.0	4.288	**5.316**	4.197	**6.293**

$1 \mid \Theta_t) = \Theta_t$. The human implements the action $U^{\mathrm{h}}_t = U^{\mathrm{ai}}_t$ if $A_t = 1$ and $U^{\mathrm{h}}_t = \mathrm{A}$ if $A_t = 0$.

We construct an AHM using the first approach in Section III-C. The encoder ϕ has 3 linear layers of sizes $(3, 30), (30, 20), (20, 2)$ with `ReLU` activation in between, and GRU with hidden size of 1. At each $t \in \mathcal{T}$, the first layer receives $(X_t, U^{\mathrm{h}}_{t-1}, U^{\mathrm{ai}}_{t-1})$ as an input, and with the hidden state C_{t-1}, the GRU yields hidden state $C_t \in [-1, 1]$ with `tanh` activation. Then, the hidden state C_t and new recommendation U^{ai}_t becomes an input of the decoder ρ, which has 3 linear layers of sizes $(2, 12), (12, 12), (12, 5)$ with `ReLU` activation. We train this AHM over $10,000$ trajectories generated by randomizing the human's initial state and system evolution, with $T = 20$ and a learning rate 0.005 with loss L from Section III-C. Then, to simplify the DP, we discretize the output of the GRU to after training it to ensure that $\hat{S}_t = \nu(C_t) \in \hat{\mathcal{S}} = \{-1, -, 0.9, \ldots, 1\}$, where $\nu(\cdot)$ is a mapping from $[-1, 1]$ to the nearest point in $\hat{\mathcal{S}}$. Using this model, we solve the DP in (9)–(10) for only 5 time steps to reduce computational complexity and obtain an approximate recommendation strategy $\hat{\boldsymbol{g}}^{*\mathrm{ai}}$. As a baseline, we also compute a naive strategy by solving the optimal DP for both $T = 20$ and $T = 30$ time horizons, without including a human in the system. Our results are obtained by running 500 simulations for time horizon $T = 20$ and $T = 30$ using the "naively optimal" strategy as well as the approximate strategy implemented with a receding horizon of 5 steps. These results are summarized in Table III. For all different initial trust and time horizon, our approximate recommendation performed at least 24% and at most 50% better than the baseline method, which highlights the utility of the learned AHM.

Remark 6: A numerical example for the AHM with the partially observed systems in Remark 1 is in our preprint [24].

V. CONCLUDING REMARKS

In this letter, we developed a framework for CPHS with partially observed data with the human-AI POMDP. We established the structural form of optimal recommendations and provided an AHM for approximate recommendations. Finally, we presented an approach to constructing AHMs from data and illustrated its utility in a numerical example. Some limitations of this letter to be addressed in future work include scaling the approach with reinforcement learning, extending it to continuous domains, and applying it to practical applications involving complex human tasks.

REFERENCES

[1] M. Y. Uzun, E. Inanc, and Y. Yildiz, "Enhancing human operator performance with long short-term memory networks in adaptively controlled systems," *IEEE Control Syst. Lett.*, vol. 7, pp. 3507–3512, 2023.

[2] N. Venkatesh, V.-A. Le, A. Dave, and A. A. Malikopoulos, "Connected and automated vehicles in mixed-traffic: Learning human driver Behavior for effective on-ramp merging," in *Proc. 62nd IEEE Conf. Decis. Control (CDC)*, 2023, pp. 92–97.

[3] A. Dave, I. V. Chremos, and A. A. Malikopoulos, "Social media and misleading information in a democracy: A mechanism design approach," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2633–2639, May 2022.

[4] H. Bang, A. Dave, and A. A. Malikopoulos, "Routing in mixed transportation systems for mobility equity," in *Proc. Amer. Control Conf.*, 2023, pp. 1–6.

[5] B. Green and Y. Chen, "The principles and limits of algorithm-in-the-loop decision making," in *Proc. ACM Human-Comput. Interact.*, vol. 3, 2019, pp. 1–24.

[6] T. Samad, "Human-in-the-loop control and cyber–physical–human systems: Applications and categorization," in *Cyber-Physical-Human Systems: Fundamentals and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2023, pp. 1–23.

[7] A. A. Malikopoulos, "Separation of learning and control for cyber-physical systems," *Automatica*, vol. 151, May 2023, Art. no. 110912.

[8] M. Carroll et al., "On the utility of learning about humans for human-AI coordination," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–19.

[9] A. M. Annaswamy and V. J. Nair, "Human behavioral models using utility theory and prospect theory," in *Cyber-Physical-Human Systems: Fundamentals and Applications*, Hoboken, NJ, USA: Wiley-IEEE Press, 2023, pp. 25–41.

[10] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Manag. Sci.*, vol. 64, no. 3, pp. 1155–1170, 2018.

[11] J. Sun, D. J. Zhang, H. Hu, and J. A. Van Mieghem, "Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations," *Manag. Sci.*, vol. 68, no. 2, pp. 846–865, 2022.

[12] M. Balakrishnan, K. Ferreira, and J. Tong, "Improving human-algorithm collaboration: Causes and mitigation of over-and under-adherence." 2022. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4298669

[13] E. Sabaté, *Adherence to Long-Term Therapies: Evidence for Action*, World Health Org., Geneva, Switzerland, 2003.

[14] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manag. Ann.*, vol. 14, no. 2, pp. 627–660, 2020.

[15] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them Err," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, p. 114, 2015.

[16] J. Grand-Clément and J. Pauphilet, "The best decisions are not the best advice: Making adherence-aware recommendations," 2022, *arXiv:2209.01874*.

[17] I. Faros, A. Dave, and A. A. Malikopoulos, "A Q-learning approach for adherence-aware recommendations," *IEEE Control Syst. Lett.*, vol. 7, pp. 3645–3650, 2023.

[18] G. Chen, X. Li, C. Sun, and H. Wang, "Learning to make adherence-aware advice," 2023, *arXiv:2310.00817*.

[19] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human–robot collaboration," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2018, pp. 307–315.

[20] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in partially observed systems," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 483–565, 2022.

[21] A. Dave, I. Faros, N. Venkatesh, and A. A. Malikopoulos, "Worst-case control and learning using partial observations over an infinite time horizon," in *Proc. 62nd IEEE Conf. Decis. Control (CDC)*, 2023, pp. 6014–6019.

[22] J. Subramanian and A. Mahajan, "Approximate information state for partially observed systems," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, 2019, pp. 1629–1636.

[23] A. A. Malikopoulos, "On team decision problems with nonclassical information structures," *IEEE Trans. Autom. Control*, vol. 68, no. 7, pp. 3915–3930, Jul. 2023.

[24] A. Dave, H. Bang, and A. A. Malikopoulos, "A framework for effective AI recommendations in cyber-physical-human systems," 2024, *arXiv:2403.05715*.