Human-Centric Behavior-Aware Adaptive Off-Policy Selection

Ge Gao[1] , **Aishwarya Mandyam**[1] , **Joy He-Yueya**[1] , **Min Chi**[2] and **Emma Brunskill**[1]
[1]Stanford University
[2]North Carolina State University
{gegao, am2, heyueya, ebrun}@stanford.edu, mchi@ncsu.edu

Abstract

In many human-centric environments, such as education and healthcare, the unobservability of human underlying states has been recognized as a key obstacle for understanding individual needs, thus hindering our ability to provide personalized decision-making policies. Several reinforcement learning (RL)-related approaches have been used to facilitate sequential decision-making in these settings, including off-policy selection (OPS), which aids in safely evaluating and selecting optimal policies offline. However, existing OPS algorithms are unsuitable when both the state is unobserved and the setting requires a personalized policy. To address this challenge, we propose a behavior-aware adaptive policy selection framework (HBO) that first captures potentially unique characteristics of the state from human behaviors, and then estimates when and how to intervene with less uncertainty in a timely manner, with bounded error. HBO is evaluated over two real-world human-centric applications, intelligent tutoring and sepsis treatments, where it significantly enhanced participants' long-term course outcomes and survival rates. Broadly, our work enables improved policy personalization in high-stakes domains where extensive evaluation is not possible.

1 Introduction

There is significant interest in using reinforcement learning (RL) in human-centric systems (HCSs), such as healthcare and education, to improve downstream outcomes [Namkoong *et al.*, 2020; Gao *et al.*, 2020; Chi *et al.*, 2011; VanLehn, 2006; Abdelshiheed *et al.*, 2023; Mandel *et al.*, 2014; Ruan *et al.*, 2024; Gottesman *et al.*, 2019]. The ultimate goal is personalized, adaptive policies that support each person's needs, but achieving this faces several challenges in practice. Evaluating RL policies online in HCSs is risky and may require a long time horizon to observe results (*e.g.*, several years for clinical trials). Off-policy evaluation (OPE) and selection (OPS) mitigate this by using historical data to assess policies before deployment [Jiang and Li, 2016; Fu *et al.*, 2021; Thomas and Brunskill, 2016; Yang *et al.*, 2022; Gao *et al.*, 2022b]. However, collecting HCS data is usually time and resource intensive [Gao *et al.*, 2024a] and data is generally limited, which motivates the need for uncertainty-aware, data-efficient estimators. Moreover, HCS deployments often face external protocols and constraints that restrict which policies can be considered. Only a (small) subset of the candidate policies that are likely to be beneficial to the *overall* trial population are approved to be deployed, making careful off-policy selection necessary, rather than unrestricted policy learning.

Existing OPE/OPS techniques typically assume that the policy is Markov in the observed features. In HCSs, applying the same policy to different participants may not result in the same expected outcome, *e.g.*, patients diagnosed with the same type of cancer may have very different outcomes from the same treatment policy, given the patients' varied medical history and health conditions which may not be observable. While OPE methods (e.g., importance sampling or doubly-robust estimators [Precup, 2000; Jiang and Li, 2016]) can account for rewards and dynamics that depend on the full history of features, off-policy selection and optimization typically focus on Markov decision policies that condition only on the same immediate state, implicitly assuming identical decisions for all individuals with the same current features.

In HCSs, we are interested in supporting the best outcomes for a new individual. Unlike some other settings, in HCSs *each individual is exposed to a given policy only once*– there are no resets to rewind time and treat a patient with a different initial policy or teach a student fractions from scratch in a new way. But there can be an opportunity to use observations from interacting with a new individual to help inform and select the best policy, *from the set of approved policies overseen by related departments*, to maximize their expected outcomes. That is contrast with pure OPS, which commits to one policy before any interaction.

To address these challenges, we introduce Human-centric Behavior-aware adaptive Off-policy selection (**HBO**). HBO uses historical data and evaluates online observations with a new individual under a pre-approved acceptable policy, to decide *if* and *when* to switch to a new policy to optimize individual outcomes. We notice that if critical behavioral changes are observed about a new participant, that may enable the system to select a much better decision policy that will maximize expected outcomes for that individual, rather than

selecting a single decision policy to be used for all individuals. For example, in personalized medicine only a limited set of approved treatments (*e.g.*, chemotherapy drugs) exists for a given disease, and clinicians often adapt a patient's regimen based on their response to maximize outcomes.

Specifically, our approach first discovers critical temporal behavioral patterns (CBPs) in historical trial data that indicate when a policy switch could be beneficial (*e.g.*, early signs of disease progression). For a new individual under a default approved policy, we monitor their real-time behavior to detect any CBPs. When a CBP occurs, our algorithm estimates the expected outcomes of switching the individual to each of a finite set of alternate approved policies (or decides to wait for more data), and operates policy switching based on confidence. To the best of our knowledge, the most related work is first-glance OPS [Gao *et al.*, 2024b] which assigns policies to new participants according to the sub-grouping of their initial observations on the first time step. After the policy is assigned, it is used for that participant for the remainder of the planning horizon. In contrast, our approach is more flexible since it can actively select when and if to switch someone, during their trajectory.

The key contributions of this work are summarized as follows: (**i**) We introduce a novel framework, HBO, that addresses the challenges of partial observability and personalization simultaneously in OPS for HCSs, with the goal of improving outcomes for participants in ongoing trials. Our approach identifies CBPs from historical data, enabling real-time monitoring of participants to assess the potential benefits from revising their treatment plans (*i.e.*, switch to another pre-approved policy), while incorporating a confidence-informed decision mechanism to determine an appropriate timing for switching. (**ii**) We conduct extensive case studies to evaluate HBO in intelligent education and healthcare applications, using data collected from a real-world intelligent tutoring system and MIMIC-III [Johnson *et al.*, 2016] respectively. Our results show that HBO can effectively enhance long-term outcomes of participants compared to baseline approaches. (**iii**) The design and mechanism behind the HBO framework allows it to be readily available to real-world human trials, given the clear motivation and straightforward adaptation of building blocks available. Code is available at https://github.com/fay067/HBO.

2 Related Work

OPE/OPS. OPE and OPS refer to techniques used to estimate the cumulative returns of target policy candidates using historical experience (*i.e.*, offline data) collected from a distinct behavior policy. This can enable easier identification of a subset of (target) policies that can lead to promising returns once deployed. Standard techniques use importance sampling (IS) [Horvitz and Thompson, 1952; Precup *et al.*, 2000], direct method (DM) [Harutyunyan *et al.*, 2016; Li *et al.*, 2010], doubly robust (DR) [Jiang and Li, 2016; Thomas and Brunskill, 2016] and distributional correction estimation (DICE) [Yang *et al.*, 2020; Nachum *et al.*, 2019a; Zhang *et al.*, 2020; Nachum *et al.*, 2019b] approaches. Most existing OPE/OPS techniques are designed toward evaluating homogeneous agents that share largely similar specifications, without any online interactions. Our work leverages both offline dataset and the online observations to select policy for a new individual, aiming to maximize individual outcomes.

Partial observability in offline RL. Most RL algorithms assume that the state is fully known and base policy optimization process on this. However, in HCSs, this is unlikely to be true. For example, in patient treatment settings in clinical scenarios, we can only measure values such as vitals and lab values. However, this represents only a portion of the underlying patient state which is largely unobserved. Some works have attempted to account for this partial observability while learning optimal policies including Hidden Parameter MDPs (HiP-MDPs) [Doshi-Velez and Konidaris, 2013; Fu *et al.*, 2023], and partially observed MDPs (POMDP) [Kaelbling *et al.*, 1998], which assume that the underlying parameter representing the agent state is unknown. However, the HiP-MDP and POMDP frameworks in general maximize returns averaged over long horizons. In contrast, in HCSs, the trials are considered high-stake where immediate evaluations and actions are expected to be taken once needed, *e.g.*, emergency events in ICU. Moreover, most work along the line of HiP-MDP and POMDP are built with the goal of *learning a single optimal policy* that fits a population overall, and rely on sufficient historical experience for capturing environmental dynamics. In contrast, our goal is to adaptively assign pre-trained policies based on information collected from an individual in an ongoing trial. As a result, limited historical data collected from previous cohorts can be used, as well as little prior information given about the individuals in the current cohort, *i.e.*, mostly only the initial observations.

Personalization in policy learning. There are a few existing works that learn personalized policies by considering individual characteristics and assuming the state is fully known [Hallak *et al.*, 2015; Modi *et al.*, 2018; Sun *et al.*, 2024]. For instance, [Hallak *et al.*, 2015] propose Contextual MDP and Cluster-Explore-Classify-Exploit framework, of a multi-task setting where tasks are drawn from related MDPs, and the goal is to quickly learn to perform well in each new MDP. However, they generally assume that all exploration policies are acceptable in each new task– for example, [Hallak *et al.*, 2015] analyze using random exploration in the "Explore" step, and [Modi *et al.*, 2018] use a variant of tabular R-max when executing in a new task. However, in many HCSs, the underlying human state can be largely unobserved, and we often have important restrictions on the policies that can be deployed with each new student or patient, to ensure safety and performance. It is not acceptable to run any possible exploration policy, and pure exploration or the optimization-under-uncertainty policies (which may take risky actions due to their potential high performance) used by [Hallak *et al.*, 2015; Modi *et al.*, 2018; Sun *et al.*, 2024] are unlikely to be allowed by stakeholders.

3 Human-Centric Behavior-Aware Adaptive Off-Policy Selection

Due to the high-stakes nature of HCSs, policies used to interact with new participants typically require careful review and

approval by experts *a priori* [Gao *et al.*, 2024b]. Moreover, anchoring on a single policy over the entire horizon may sacrifice the optimality of the final outcome, as the policy cannot be personalized to participant characteristics [Balazadeh *et al.*, 2024; Chi *et al.*, 2011]. For example, a policy that improves the average learning outcomes of a group of students may not benefit every individual student equally. As a result, we introduce an approach, Human-centric Behavior-aware adaptive Off-policy selection (HBO), that can adapt the to the necessities of each participant given real-time observations over the task horizon.

3.1 Problem Formulation

In HCSs, agents often operate with incomplete or noisy observations of user states, such as intentions and cognitive load, which are not fully observable or directly measurable. This partial observability naturally motivates modeling the practical problem as a partially observable Markov decision process (POMDP), where the agent could infer hidden states from observation histories to make informed decisions. We consider a setting defined by a POMDP, represented as a 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{S}_0, \mathcal{R}, \mathcal{Z}, \gamma)$. Specifically, $\mathcal{S}$ is the state space which we assume is unknown, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ defines transition dynamics from the current state and action to the next state, $\mathcal{S}_0$ defines the initial state distribution, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathcal{Z}(o|s)$ is the observation model which is unknown, $\gamma \in [0, 1)$ is discount factor. All episodes have a finite horizon H.

A trajectory under policy π is denoted as $\tau^i_\pi = [\dots, (o^i_t, a^i_t, r^i_t, o^i_{t+1}), \dots]^H_{t=0}$. We have access to a historical (*i.e.*, offline) dataset collected under a behavioral policy β, $\mathcal{D}_\beta = \{..., \tau^i_\beta, ...\}^N_{i=1}$, which consist of N trajectories where each trajectory corresponds to a single participant.

Assumption 1 (Initial Policy Selection)**.** *Assume that when a new participant i' joins the HCS, there always exist a pre-trained policy $\pi_1 \in \Pi$ assigned to i' immediately upon the system receiving their initial observation $o^{i'}_1$; here, Π is the set of policies, pre-trained on historical data $\mathcal{D}_\beta$, that have been approved by experts a priori. This initial policy selection is based on a pre-given mapping $\mathcal{D}_\beta \times \mathcal{O} \to \Pi$ to ensure a smooth start of the new participant (e.g., determined by the experts).*

Then, to maximize the outcome of each participant over the horizon, we aim to select the best sequence of policies $(\pi^*_1, ..., \pi^*_H)$ such that each policy $\pi^*_t \in \Pi$ is optimal at each step t, for each of the new participants $i' \in \mathcal{I}' = \{N + 1, N + 2, \dots\}$ arriving at the system with an initial observation $o^{i'}_1$ (while the rest of the trajectory remains *unobservable*), that maximizes the participant's expected cumulative return, $V^{(\pi^*_0, ..., \pi^*_H)}$, $\max_{\pi^*_t \in \Pi, t \in [0,H]} V^{(\pi^*_0, ..., \pi^*_H)}$, over the full horizon H. Here $V^{(\pi^*_0, ..., \pi^*_H)} = \mathbb{E}_{(o_{t+1}, a_t) \sim \rho^{\pi^*_t}}[\sum^H_{t=0} \gamma^{t-1} r_t | (\pi^*_0, ..., \pi^*_H)]$, and $\rho^{\pi^*_t}$ is the observation-action visitation distribution under π^*_t at step t, *i.e.*, the marginal distribution over o_{t+1} and a_t induced by π^*_t and the POMDP dynamics.

However, motivated by the general guidance of *minimal trials*[1] in HCSs [Nie *et al.*, 2021; Gao *et al.*, 2024b], as well as to ensure sufficient bandwidth during online testing (*e.g.*, in randomized trials), in this work we consider identifying a single time point $h \in [1, H]$ at which a one-time policy assignment switch is made for each new participant $i' \in \mathcal{I}'$. The formal problem statement is below.

Problem 1 (Human-Centric Adaptive Off-Policy Selection Problem)**.** *We focus on the conservative setting in which the policy can be switched at most once over the horizon for each participant, to minimize participant burden and operational risk in HCSs. Specifically, the goal is that, given the fraction of the observed trajectory, $\tau^{i'}_{\pi_{t<h}}(0 : h) = [\dots, (o^i_t, a^i_t, r^i_t, o^i_{t+1}), \dots]^h_{t=0}$, pertaining to a new participant i' treated with the expert-selected policy $\pi_{t<h} \in \Pi$ (see Assumption 1) up until the current step $h \in [1, H]$, determine whether switching to another policy $\pi_{t \geq h} \in \Pi \backslash \pi_{t<h}$ can maximize the total discounted sum of rewards, i.e., $\max_{h, \pi_{t \geq h}} \Big(V^{\pi_{t<h}}(h) + V^{\pi_{t \geq h}}(h) \Big)$; here, $V^{\pi_{t<h}}(h) = \mathbb{E}[\sum^{h-1}_{t=0} \gamma^t r_t | \pi_{t<h}, s_0]$ and $V^{\pi_{t \geq h}}(h) = \mathbb{E}[\sum^H_{t=h} \gamma^t r_t | \pi_{t \geq h}, \tau_{\pi_{t<h}}(0 : h), s_h]$ are the expected cumulative return before and after the policy switch at the time step h, respectively.*

3.2 Critical Behavioral Patterns (CBPs)

Historical data collected from HCSs in general provides limited coverage of the state/observation space, given the high-stake nature [Mandel *et al.*, 2014]. This limited coverage can make the data highly implicit, and hence unidentifiable for decision-making in its raw form; *e.g.*, patients appeared with similar current symptoms could be caused by different underlying diseases, where the clinicians will rely on their knowledge and past experience to carry out diagnoses and treatment plans onwards.

To tackle this challenge, in this section, we introduce a critical behavior patterns (CBPs) mining approach, to map observations into a discrete space; in what follows, sub-grouping can be leveraged to identify critical behavioral changes from the *historical* trajectories, which could be pivotal for HBO (see Algorithm 1) to leverage and decide if the policy assignments should be changed for any *new participants* at the current time step.

***Step I* – Discrete Representation Mapping.** Discrete representation mapping has proven effective in capturing general abstractions from complex trajectories and capturing similar behaviors shared across samples [Yang *et al.*, 2021; Gao *et al.*, 2022a; Gao *et al.*, 2024a]. This is important for HCSs as often the observation space is large while offline coverage is low [Gao *et al.*, 2023b]. Specifically, each trajectory τ^i can be mapped to a temporal sequence χ^i of H low-dimensional discrete representations, *i.e.*, $\chi^i = (z^i_0, z^i_1, ..., z^i_H)$, where z^i_t is the representation at step t. This mapping could be facilitated by existing pattern mining algorithms, such as Toeplitz Inverse Covariance-Based Clustering (TICC) [Hallac *et al.*, 2017] or Multi-series Time-aware

[1] Refer to studies scoped to gather preliminary data on a new treatment or intervention with minimal burden on participants.

TICC (MT-TICC) [Yang *et al.*, 2021] which considers both cross-trajectory and temporal dependencies to apply discrete representation mapping.

***Step II* – Identify CBPs.** Then, sub-sequences of discrete representations $\chi^i_{t_1:t_2}$, with $0 \leq t_1 < t_2 \leq H$, are extracted from each χ^i. CBPs are the sub-sequences of discrete representations that appear frequently in the corresponding trajectories with undesirable final outcome (*e.g.*, progression of pathology or failed final exams), which can indicate *critical behavioral changes* where more aggressive treatments may be necessary. The set of CBPs is denoted as $\tilde{\chi} = \{\chi^i_{t_1:t_2} | i \in [1, N]\}$.[2]

3.3 Confidence-Informed Policy Switch

Many human-centric practices in the real world benefit from the increased volume of acquired observations to boost confidence when making decisions [Dann *et al.*, 2019; Nie *et al.*, 2021], due to the partial observability in HCSs. Motivated by this pattern, here we introduce a confidence estimation approach that balances the benefits against opportunity costs of switching the policy at the current time step h, *i.e.*, if one should collect the return of switching to a (seemingly) more beneficial policy earlier (according to existing observations) or switch the policy later when more observations are available to be more confident.

For a new participant, i', determining whether they will benefit from a policy switch at the current step h depends on two factors, *i.e.*, (*i*) if CBPs are identified from the past observations (following Section 3.2); and (*ii*) if switching to $\pi_{t\geq h} = \arg\max_{\pi\in\Pi} \mathbb{E}[\sum_{t=h}^{H} \gamma^t r_t | \pi, \tau_{\pi_{t<h}}(0:h), s_h]$ is estimated to result in more gains than sticking with $\pi_{t<h}$ and switching later once more observations are available.

Who would benefit from policy switching. Consider the trajectory $\tau^{i'}_{\pi_{t<h}}$ until step h, pertaining to the new participant i' who is still treated by the expert-selected policy $\pi_{t<h}$. There exists a *switching advantage* by switching from $\pi_{t<h}$ to $\pi_{t\geq h} \in \Pi \backslash \pi_{t<h}$ when $V^{\pi_{t\geq h}}(h) > V^{\pi_{t<h}}(h)$. And we aim to determine who is likely to gain the switching advantage during the online deployment. As described in Section 3.2, CBPs are the patterns significantly presented in the *historical data* that had led to undesirable final outcomes. If the fraction of the observed trajectory, $\tau^{i'}_{\pi_{t<h}}(0:h)$, of the new participant is identified as containing the CBPs, the participant has a risk to achieve undesirable final outcomes and may gain switching advantage.

To detect whether $\tau^{i'}_{\pi_{t\leq h}}(0:h)$ contains CBPs, we first encode $\tau^{i'}_{\pi_{t\leq h}}(0:h)$ into $\chi^{i'}_{1:h}$ using TICC-based methods, *i.e.*, by applying *Step I* of Section 3.2. Then if $\chi^{i'}_{1:h}$ contains any pre-identified CBPs from historical data (*i.e.*, in $\tilde{\chi}$), we flag this participant and keep tracking if switching to $\pi_{t\geq h}$ at current step h or switching later once more observations are available, as introduced below.

[2]Without loss of generality, $\chi^i_{t_1:t_2}$'s only refers to CBPs from this paragraph onwards. We slightly abuse the use this notation, as all other non-critical patterns are discarded.

When to switch policy. Given the interaction is performed in real time where observations onward from current step h are unseen, it's not practical to wait until receiving entire trajectory of the participant and then decide the best step to switch policies. To address this, one can adopt off-policy selection to estimate the switching advantage by comparing estimated returns achieved by switching or not, *i.e.*, $\hat{V}^{\pi_{t\geq h}}(h)$ and $\hat{V}^{\pi_{t<h}}(h)$ respectively. That is also reminiscent of the decision-making process of human experts, by monitoring participants' behaviors and regularly re-evaluating whether to update interventions based on observations accumulated so far and if any abnormalities have already been detected (*i.e.*, the presence of CBPs). Specifically, the estimated value of switching to another policy, $\pi_{t\geq h} \in \Pi \backslash \pi_{t<h}$, is a variant of weighted importance sampling (WIS) [Precup *et al.*, 2000] where only historical sub-trajectories containing the CBPs of interest are used. Specifically,

$$\hat{V}^{\pi_{t\geq h}}(h) = \frac{\sum_{i\in\Phi^{i'}} w(\tau^i, \pi_{t\geq h})(\sum_{t=h}^{H} \gamma^t r^i_t)}{\sum_{i\in\Phi^{i'}} w(\tau^i, \pi_{t\geq h})}, \quad (1)$$

where $\Phi^{i'} \subseteq \{1, 2, ..., N\}$ is the subset of historical trajectory indices, whose corresponding discrete sequences contain the same identified CBPs from the current observations $\tau^{i'}_{\pi_{t<h}}(0:h)$ of the new participant i', $w(\tau^i, \pi_{(\cdot)}) = \prod_{t=h}^{H} \frac{\pi_{(\cdot)}(a_t|\tau^i_\beta(0:t)}{\beta(a_t|\tau^i_\beta(0:t)}$ is the importance ratio for the trajectory τ^i in the offline dataset. Similarly, $\hat{V}^{\pi_{t<h}}(h)$ can be calculated following Equation 1 by plugging $\pi_{t<h}$ into $w(\tau^i, \pi_{(\cdot)})$. Therefore, the optimal target policy for the new participant to switch from step h can be obtained by

$$\pi^*_{t\geq h} = \underset{\pi\in\Pi\backslash\pi_{t<h}}{\arg\max}\ \hat{V}^{\pi}(h). \quad (2)$$

If $\hat{V}^{\pi^*_{t\geq h}}(h) > \hat{V}^{\pi_{t<h}}(h)$, there may exist an advantage to deploy $\pi^*_{t\geq h}$ to the participant i' starting from the current step h. To better balance the gain achieved from switching immediately against the confidence from holding until more observations are collected, we introduce the *look ahead mechanism* to bound our confidence on switching at the current step. Specifically, once a CBP is detected and $\hat{V}^{\pi^*_{t\geq h}}(h) > \hat{V}^{\pi_{t<h}}(h)$, the look ahead mechanism steps in as follows.

Definition 1 (Look Ahead Advantage)**.** *The look ahead advantage $\Delta(h)$ is defined as the difference of returns between deferring one more step followed by switching to the estimated best policy $\pi^*_{t\geq h+1}$ from step $h+1$, and switching to estimated best policy $\pi^*_{t\geq h}$ from current step h, given $\tau^{i'}_{\pi_{t<h}}(0:h)$, i.e.,*

$$\Delta(h) = \hat{V}^{\pi_{t<h}, \pi^*_{t\geq h+1}}(h+1) - \hat{V}^{\pi^*_{t\geq h}}(h), \quad (3)$$

where $\hat{V}^{\pi_{t<h}, \pi^*_{t\geq h+1}}(h+1) = \hat{r}_h + \gamma \sum_s p(s^{i'}_{h+1} | \tau^{i'}_{\pi_{t<h}}(0:h), \hat{o}^{i'}_{h+1}) \hat{V}^{\pi^*_{t\geq h+1}}(h+1)$. Specifically, $\hat{r}_h$ is the estimated reward received at current step h by remaining on $\pi_{t<h}$, and $\hat{o}^{i'}_{h+1}$ is estimated observation by switching to policy $\pi^*_{t\geq h+1}$ at step $h+1$. Both parts can be estimated through a model-based approach [Hafner et al.*, 2020; Gao* et al.*, 2022b]. Moreover, $p(s^{i'}_{h+1}|\cdot,\cdot)$ is the POMDP belief state model. As a result, if*

Algorithm 1 HBO.

Input: A set of target policies Π, offline dataset $\mathcal{D}_\beta$. A pre-trained initial policy π_1.
// Offline Phase.
Get discrete sequences χ from $\mathcal{D}_\beta$.
Get the set of critical behavioral patterns (CBPs) $\tilde{\chi}$ from χ following Section 3.2.
// Deployment Phase.
while the HCS receives the initial observation o_1 from a new participant **do**
  Initialize $\pi_o = \pi_1$.
  for each step $h \in [1, H]$ **do**
    Obtain the discrete representation z_h of o_h.
    if $\tau_{\pi_{t\leq h}}(0:h)$ contains CBPs **then**
      Get estimated optimal switching policy $\pi^*_{t\geq h}(h)$ from Equation 2.
      if $\hat{V}^{\pi^*_{t\geq h}}(h) > \hat{V}^{\pi^*_{t<h}}(h)$ **then**
        Get look ahead advantage $\Delta(h)$ from Equation 3.
        if $\Delta(h) \leq 0$ **then**
          Switch to $\pi^*_{t\geq h}(h)$. Update $\pi_o = \pi^*_{t\geq h}(h)$.
        end if
      end if
    end if
    Operate policy π_o.
  end for
end while

$\Delta(h) > 0$*, the policy switch should be on-hold and deferred to the next step when there is no looking ahead advantage.*

The HBO algorithm is summarized in Algorithm 1. Below we derive the upper-bound error of HBO-selected policies.

3.4 Theory

Under a few assumptions, it is straightforward to see that our proposed algorithm will at least achieve the same value as the behavior policy minus epsilon.

Assumption 2 (Full coverage)**.** *For all possible target policies* $\{\pi_i\}_{i=1}^{i=E}$, $\pi_i(a|o) > 0 \rightarrow \pi_b(a|o) > 0, \forall o \in \mathcal{O}, a \in \mathcal{A}$ *where* E *is the cardinality of* Π.

Assumption 3 (In-distribution population)**.** *The distribution of trajectories in the original dataset* $\tau \sim D$ *is the same as any new trajectory* τ'.

Assumption 4 (ϵ-accurate policy value estimation)**.** $|V^\pi - \hat{V}^\pi| \leq \epsilon$

Given 2, 3, 4, in each individual's MDP, the value of the selected policy $V^{HBO} \geq V^\beta - 2 * \epsilon$.

Proof. Note that for all time steps at which HBO follows the behavior policy, the bound trivially holds. We therefore consider a subjtrajectory t_h at which HBO switches to another policy π_i, which it then follows for the remaining horizon. Consider the worst case setting in which HBO overestimates the value $\hat{V}^{\pi_i}(\tau_h)$ of following a given target policy π_i, and underestimates the value of the behavior policy $\hat{V}^{\pi_i}(\tau_h)$. π_i is selected by HBO because $\hat{V}^{\pi_i}(\tau_h) > \hat{V}^\beta(\tau_h)$. By 4, this implies the following:

$$V^{\pi_i}(\tau_h) + \epsilon > \hat{V}^\beta(\tau_h) > \hat{V}^\beta(\tau_h) > \hat{V}^\beta(\tau_h) - \epsilon$$
$$V^{\pi_i}(\tau_h) > V^\beta(\tau_h) - 2 * \epsilon$$

□

4 Experiments

HBO is evaluated with two types of HCSs: interactive education and healthcare. Specifically, the interactive education experiment is based on a semi-synthetic simulator capturing the behavior of 1,342 students who participated in real-world college entry-level probability course over 7 academic semesters [Gao *et al.*, 2022b; Gao *et al.*, 2024b; Gao *et al.*, 2023b]. The goal is to use the data collected from half of the students under an expert-designed policy, to assign pre-trained RL policies to the rest of students, in order to maximize their final learning outcomes. The healthcare experiment targets selecting pre-configured policies that can best treat patients with sepsis, over a simulated environment widely adopted in existing works [Namkoong *et al.*, 2020; Tang and Wiens, 2021; Lorberbom *et al.*, 2021; Gao *et al.*, 2023a].

4.1 Baselines

We now describe an additional variant of the HBO method we introduced, four baselines, and an oracle setting.

(i) Vanilla one-policy-fits-all with existing OPE/OPS. This baseline chooses a single target policy, $\pi \in \Pi$, to be deployed to all participants, which achieves the maximum expected return over the *entire offline dataset* as estimated using by existing OPE/OPS methods, *e.g.*, [Precup, 2000; Thomas and Brunskill, 2016; Le *et al.*, 2019; Nachum *et al.*, 2019a]. However, as different OPE/OPS algorithms may converge to varied policy selections [Fu *et al.*, 2021], we consider deploying each single target policy as baselines. **(ii) Combinatorial one-policy-fits-all with existing OPE/OPS.** Firstly, we exhaustively list all possible combinations of initial policies, target policies to switch to, over all steps. This leads to the space, $(\Pi \times \Pi)^H$, which is effectively the same space searched by HBO. Then, this baseline uses an OPE method to select from this (much larger) set of policies. We consider two popular OPE estimators, weighted importance sampling (WIS) [Precup, 2000] and doubly robust (DR) [Jiang and Li, 2016], that have been widely used in OPE/OPS works, denoted as WIS* and DR*, respectively. Moreover, one could view WIS* and DR* as providing insights into the impact of pattern matching, since they excluded CBPs and performed OPS and switching by comparing the OPE estimates using all data (rather than only that which match in the pattern). **(iii) First-glance off-policy selection (FPS).** FPS assigns target policies to each new participant, at $t = 0$ when only o_0 is available, based on partitioning participants into sub-groups using both historical dataset and initial state of the new participant before online interactions [Gao *et al.*, 2024b]. **(iv) HBO with immediate policy switching (HBO-IS).** One variation of HBO is deciding policy switching once critical behavioral patterns are detected given a new participant, without the look-ahead step

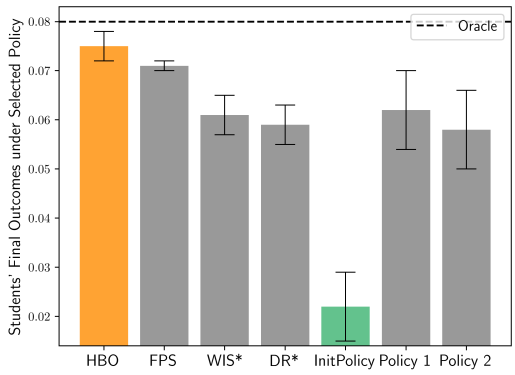

Figure 1: Final normalized learning gains of students under policies selected by HBO and baselines. WIS* and DR* use exhaustive switching combinations; InitPolicy is the lecturer-designed policy; Policy 1 & 2 are RL-induced one-policy-fits-all policies.

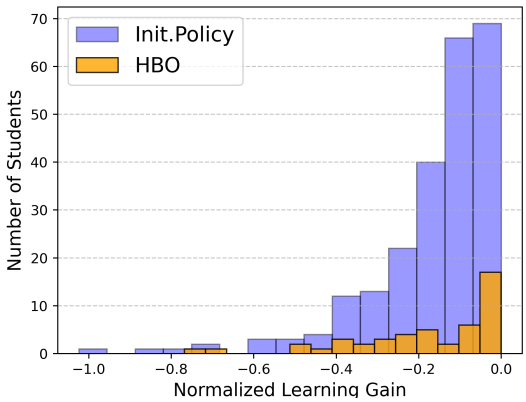

Figure 2: Distribution of final outcomes for *lower-performing* students under the initial policy and HBO. HBO substantially reduces the number of students with negative normalized learning gains.

introduced in Section 3.3. **(v) Oracle.** This approach computes the optimal selection for each participant in hindsight, *i.e.*, assuming at the beginning ($t = 0$) we could have picked for each participant the best policy as if the trajectory rollouts under all policies were available, what the final outcome would be. Although this approach is practically impossible as it would depend on access to future observations, it can inform approximately the best returns can be achieved by HBO (*i.e.*, the best policy is selected at $t = 1$).

4.2 Intelligent Tutoring

The intelligent tutoring system (ITS) has been incorporated into an undergraduate-level course on probability and statistics across 7 semesters, involving a total of 1,342 students. The study was conducted with approval from the Institutional Review Board (IRB) at the institution, ensuring adherence to ethical standards. Additionally, a departmental committee oversees the process to safeguard participants' academic performance and privacy. In this educational setting, each learning session is structured with a sequence of 12 problems, referred to as an "episode" (horizon $H = 12$). During each step, the ITS policy selects one of three tutoring actions: working independently, the students and the tutor co-construct the solution, or viewing the problem-solving demonstration from the ITS (mainly for studying purposes). The observation space per time steps consists of 142 log interaction features carefully designed by domain experts. These features capture various aspects of student activity, such as time spent on a problem and solution accuracy. There is zero reward for all intermediary steps. At the end of each episode, the reward is the normalized learning gain (NLG) derived from scores on two exams: a pre-test before using the system and a post-test afterward. The scores are normalized. Training data ($N = 628$) is collected under a lecturer-designed behavioral policy and used to train HBO for capturing critical behavioral patterns and estimating values of target policies. The initial behavior policy is the lecturer-ITS policy designed to ensure a smooth start and safety even without any policy switch. The target policies included two pre-trained alternate ITS RL policies.

Main results. Figure 1 presents students' final outcomes under policies selected by different methods. Overall, HBO and its variations were the most effective policy selection methods, having the highest average students' final outcomes compared to other baselines. There was no obvious difference on the rounded averaged final returns across HBO and its variations, but we observed that a small set of individuals benefited from HBO compared to HBO-IS. This is discussed further in later sections. Moreover, HBO surpassed the state-of-the-art approach FPS, underscoring the efficacy of HBO in capturing students' underlying characteristics and switching policies with estimating confidence in real time. Though the traditional OPE/OPS methods can be adapted into the huge policy space (*i.e.*, overall 3^{12} combinations) as our algorithm searched in, both methods, WIS* and DR*, yielded lower final outcomes, further indicating the need for HBO. It is also noted that the performance of the initial policy was worse than other target policies (*i.e.*, RL-induced Policy 1 & 2), while HBO can significantly enhance students' performance beginning with the initial policies, highlighting HBO's significant advantage in correctly switching to the right policy at proper timing. Last but not least, our approach is getting quite close to the oracle's performance, further justifying the advantages of HBO.

Individual gains from looking ahead. Figure 2 illustrates the distribution of final outcomes over lower-performers whose normalized learning gains were negative, under initial policy and HBO, respectively. It was observed that under the initial policy, a substantial portion of the cohort, specifically 229 out of 628 students, experienced negative learning gains. This high incidence of unfavorable outcomes highlights the limitations of the initial policy in adequately addressing the diverse needs of the student population. Conversely, under HBO, the number of students with negative learning gains significantly decreased to only 47 out of 628. This stark reduction by approximately 79% underscores HBO's capability to customize to support those who are failing to thrive under a one-size-fits-all policy.

Although HBO and HBO-IS have similar rounded average returns, a subset of students benefited from HBO's delayed lookahead. To examine this, we simulated participants' continuations under candidate switching times using the variational-

autoencoder-based simulator from prior work [Gao *et al.*, 2024a; Gao *et al.*, 2023b; Gao *et al.*, 2024b]. In one case, switching to the estimated optimal policy at $t = 3$ yielded a predicted return of 0.135, while HBO delayed the switch to $t = 4$ and achieved 0.23. In another case, immediate switching at $t = 3$ yielded 0.15, whereas delaying to $t = 4$ and $t = 5$ increased the returns to 0.233 and 0.30, respectively. These examples suggest that HBO's lookahead mechanism can improve individual outcomes by avoiding premature policy switches and waiting until additional observations support a more confident intervention.

4.3 Sepsis Treatment

We consider selecting the policy that can best treat sepsis for each patient in the ICU, leveraging the simulated environment introduced by [Oberst and Sontag, 2019], which has been widely adopted in existing works [Hao *et al.*, 2021; Nie *et al.*, 2022; Tang and Wiens, 2021; Lorberbom *et al.*, 2021; Gao *et al.*, 2023a; Namkoong *et al.*, 2020]. Specifically, the state space is constituted by a binary indicator for diabetes, and four vital signs {heart rate, blood pressure, oxygen concentration, glucose level} that take values in a subset of {very high, high, normal, low, very low}; size of the state space is $|\mathcal{S}| = 1440$. Actions are captured by combinations of three binary treatment options, {antibiotics, vasopressors, mechanical ventilation}, which lead to $|\mathcal{A}| = 2^3$. Seven candidate target policies are considered [Namkoong *et al.*, 2020], *i.e.*, (i) initial policy which is a soft physician policy, with 0.95 probability takes the action recommended by an RL policy trained following policy iteration, and with 0.05 probability picks randomly from other actions; (ii) with antibiotics on the initial step which register antibiotics right after the patient is admitted; (iii) without antibiotics on the initial step which does not administer antibiotics right after the patient is admitted; (iv) a mixture of two policies: 80% of the initial policy and 20% of a policy that is similar to the initial policy but the vasopressors action is flipped; (v) a mixture of two policies similar to (iv) but 60% of the initial policy; (vi) a policy that always administers antibiotics once the patient is admitted; (vii) a policy that never administer antibiotics. Moreover, simulated unrecorded comorbidities are introduced into the cohort to capture uncertainty from patients' underlying diseases or other latent characteristics, which may reduce the effectiveness of administered antibiotics.

Main results. HBO consistently improved outcomes in the longer-horizon setting ($H = 50$) and outperformed FPS and other baselines across horizon and training-size settings. HBO-IS, which switches immediately once CBPs are detected, performed worse than HBO in all settings. The gap is smaller for $H = 20$ but larger for $H = 50$, suggesting that lookahead is especially important in longer, evolving healthcare trajectories.

		H=20	H=50
One-policy-fits-all	Init. Policy	0.068 (.007)	0.025 (.007)
	With Antibiotics Init	0.068 (.000)	0.021 (.000)
	Without Antibiotics Init	0.037 (.000)	-0.02 (.000)
	80% Mix w. Expert	0.005 (.000)	-0.06 (.000)
	60% Mix w. Expert	-0.022 (.000)	-0.1 (.000)
	With Antibiotics All-way	0.044 (.000)	0.022 (.000)
	Without Antibiotics All-way	0.012 (.000)	-0.015 (.000)
N=1,000	FPS	0.08 (.003)	0.026 (.004)
	HBO-IS	0.081 (.007)	0.034 (.007)
	HBO	**0.086** (.007)	**0.039** (.002)
N=5,000	FPS	0.07 (.003)	0.032 (.003)
	HBO-IS	0.079 (.002)	0.033 (.002)
	HBO	**0.081** (.002)	**0.039** (.001)
	Oracle	0.166 (.008)	0.238 (.007)

Table 1: The final outcomes from deploying to each patient the corresponding candidate policy selected by HBO against baselines, on $H = \{20, 50\}$, with varied training size $N = \{1,000, 5,000\}$, respectively. Results are averaged over 5 different runs. Standard errors are rounded. WIS* and DR* are omitted because their rounded average returns match the listed baselines.

5 Conclusion, Limitation, and Future Work

In this work, we introduced HBO which provides adaptive personalized policy selection by capturing potentially unique characteristics from human behaviors and estimates when and how to intervene with less uncertainty in a timely manner. HBO was validated with extensive real-world human-centric applications, intelligent tutoring and sepsis treatment, where HBO achieved superior performance compared to baselines.

While this study primarily focuses on the off-policy selection task to identify improvements from capturing underlying unique human characteristics and adaptive policy assignments with minimal trials, further work could include extending HBO for full policy optimization beyond policy selection. Also, our assumption 2 may hold in many human-centric domains (such as our educational tutoring data and sepsis treatment) the deployed behavior policies are relatively consistent (e.g., due to standardized curricula and treatment protocols). If overlap is highly limited, one could use more conservative evaluation techniques. Relaxing or adapting this assumption is an important direction for future work. Future work may involve developing policy estimators that manage the bias-variance trade-off, such as incorporating weighted doubly robust or MAGIC methods [Thomas and Brunskill, 2016] instead of importance sampling weights. Incorporating a budgeted multi-switch model could increase flexibility while still managing overall risk in lower-risk scenarios. Moreover, our current switching rules based on CBPs are deliberately conservative to ensure safety: we only switch when there is strong evidence of a problematic pattern (as this is necessary in high-stake human-centric domains). A learned scoring model (for example, predicting expected improvement) could allow more nuanced decisions, but it would require additional data and might risk overfitting or losing transparency. High-stakes environments, such as healthcare, can restrict the range of available policy options. Our results suggest HBO is a promising approach for such settings, as it is an effective and scalable method that can work well even with only relatively small offline datasets.

Acknowledgments

We sincerely thank the anonymous reviewers, Hyunji (Alex) Nam, and Stephane Hatgis-Kessell for their valuable and constructive feedback. This research was also supported by the

NSF Grants: 1651909, 2013502, and 2507143.

References

[Abdelshiheed *et al.*, 2023] Mark Abdelshiheed, John Wesley Hostetter, Tiffany Barnes, and Min Chi. Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 291–303. Springer, 2023.

[Balazadeh *et al.*, 2024] Vahid Balazadeh, Keertana Chidambaram, Viet Nguyen, Rahul Krishnan, and Vasilis Syrgkanis. Sequential decision making with expert demonstrations under unobserved heterogeneity. In *Automated Reinforcement Learning: Exploring Meta-Learning, AutoML, and LLMs*, 2024.

[Chi *et al.*, 2011] Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1):137–180, 2011.

[Dann *et al.*, 2019] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

[Doshi-Velez and Konidaris, 2013] Finale Doshi-Velez and George Konidaris. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations, 2013.

[Fu *et al.*, 2021] Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

[Fu *et al.*, 2023] Haotian Fu, Jiayu Yao, Omer Gottesman, Finale Doshi-Velez, and GD Konidaris. Performance bounds for model and policy transfer in hidden-parameter mdps. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[Gao *et al.*, 2020] Qitong Gao, Michael Naumann, Ilija Jovanov, Vuk Lesi, Karthik Kamaravelu, Warren M Grill, and Miroslav Pajic. Model-based design of closed loop deep brain stimulation controller using reinforcement learning. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 108–118. IEEE, 2020.

[Gao *et al.*, 2022a] Ge Gao, Qitong Gao, Xi Yang, Miroslav Pajic, and Min Chi. A reinforcement learning-informed pattern mining framework for multivariate time series classification. In *31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[Gao *et al.*, 2022b] Qitong Gao, Ge Gao, Min Chi, and Miroslav Pajic. Variational latent branching model for off-policy evaluation. In *The Eleventh International Conference on Learning Representations*, 2022.

[Gao *et al.*, 2023a] Ge Gao, Song Ju, Markel Sanz Ausin, and Min Chi. Hope: Human-centric off-policy evaluation for e-learning and healthcare. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1504–1513, 2023.

[Gao *et al.*, 2023b] Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. Off-policy evaluation for human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 9065–9091, 2023.

[Gao *et al.*, 2024a] Ge Gao, Qitong Gao, Xi Yang, Song Ju, Miroslav Pajic, and Min Chi. On trajectory augmentations for off-policy evaluation. In *The Twelfth International Conference on Learning Representations*, 2024.

[Gao *et al.*, 2024b] Ge Gao, Xi Yang, Qitong Gao, Song Ju, Miroslav Pajic, and Min Chi. Off-policy selection for initiating human-centric experimental design. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Gottesman *et al.*, 2019] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

[Hafner *et al.*, 2020] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[Hallac *et al.*, 2017] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 215–223, 2017.

[Hallak *et al.*, 2015] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

[Hao *et al.*, 2021] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[Harutyunyan *et al.*, 2016] Anna Harutyunyan, Marc G. Bellemare, Tom Stepleton, and Remi Munos. Q(λ) with off-policy corrections, 2016.

[Horvitz and Thompson, 1952] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[Jiang and Li, 2016] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning, 2016.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony

Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Kaelbling *et al.*, 1998] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[Le *et al.*, 2019] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery.

[Lorberbom *et al.*, 2021] Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34:26792–26803, 2021.

[Mandel *et al.*, 2014] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

[Modi *et al.*, 2018] Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic learning theory*, pages 597–618. PMLR, 2018.

[Nachum *et al.*, 2019a] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

[Nachum *et al.*, 2019b] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections, 2019.

[Namkoong *et al.*, 2020] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

[Nie *et al.*, 2021] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

[Nie *et al.*, 2022] Allen Nie, Yannis Flet-Berliac, Deon Jordan, William Steenbergen, and Emma Brunskill. Data-efficient pipeline for offline reinforcement learning with limited data. *Advances in Neural Information Processing Systems*, 35:14810–14823, 2022.

[Oberst and Sontag, 2019] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[Precup *et al.*, 2000] Doina Precup, Richard Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 06 2000.

[Precup, 2000] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[Ruan *et al.*, 2024] Sherry Ruan, Allen Nie, William Steenbergen, Jiayu He, JQ Zhang, Meng Guo, Yao Liu, Kyle Dang Nguyen, Catherine Y Wang, Rui Ying, et al. Reinforcement learning tutor better supported lower performers in a math task. *Machine Learning*, pages 1–26, 2024.

[Sun *et al.*, 2024] Yuewen Sun, Biwei Huang, Yu Yao, Donghuo Zeng, Xinshuai Dong, Songyao Jin, Boyang Sun, Roberto Legaspi, Kazushi Ikeda, Peter Spirtes, et al. Identifying latent state-transition processes for individualized reinforcement learning. *Advances in Neural Information Processing Systems*, 37:124887–124924, 2024.

[Tang and Wiens, 2021] Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

[Thomas and Brunskill, 2016] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[VanLehn, 2006] Kurt VanLehn. The behavior of tutoring systems. *International Journal Artificial Intelligence in Education*, 16(3):227–265, 2006.

[Yang *et al.*, 2020] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

[Yang *et al.*, 2021] Xi Yang, Yuan Zhang, and Min Chi. Multi-series time-aware sequence partitioning for disease progression modeling. In *IJCAI*, 2021.

[Yang *et al.*, 2022] Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 4376–4396. PMLR, 2022.

[Zhang *et al.*, 2020] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pages 11194–11203. PMLR, 2020.