OPTIMIZING HUMAN-IN-THE-LOOP SYSTEMS UNDER DYNAMIC TRUST: AN ADHERENCE-AWARE REINFORCEMENT LEARNING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world decision systems adopt a human-in-the-loop architecture, where AI provides recommendations while human decision makers retain final authority. As a result, the implemented action may differ from the recommendation. Existing adherence-aware methods account for this mismatch, but typically model human adherence as a fixed probability. In practice, however, human reliance can evolve with previous interaction outcomes, making adherence endogenous to the sequential decision process. We propose an adherence-aware reinforcement learning framework that models dynamic adherence via a Belief Markov Decision Process (MDP). By augmenting the state space with a belief state over human trust, our approach simultaneously tracks dynamic adherence probabilities and adaptively calibrates recommendations to maximize expected long-term rewards despite human deviations. Based on the resulting Belief MDP, we derive an adherence-aware Bellman operator that marginalizes over human acceptance and fallback behavior, together with a scalable reinforcement learning implementation. Theoretically, we characterize the suboptimality induced by belief and human-model estimation errors. Evaluations in sepsis treatment, highway lane changing and supply chain inventory management tasks demonstrate that proactively modeling dynamic trust and making adherence-aware recommendations enhance overall system performance while simultaneously providing interpretable insights into how recommendations should adapt to varying levels of human trust.

1 INTRODUCTION

AI-based decision-support tools have become increasingly capable and widely deployed across a range of domains. In many real-world settings, particularly those involving high-stakes decisions, these tools operate in a *human-in-the-loop* paradigm, providing recommendations to human decision makers who retain authority over the action ultimately implemented (Sun et al., 2022; Kesavan & Kushwaha, 2020; Snyder et al., 2026; Cohen et al., 2026). Consequently, the implemented action may differ from the algorithmic recommendation, creating a recommendation-implementation mismatch that has recently received increasing attention. For example, Grand-Clément & Pauphilet (2026) show that deploying recommendation policies under the assumption of perfect compliance can lead to substantial losses in effective performance.

A natural response to this recommendation-implementation mismatch is to make the recommendation policy adherence-aware. The key insight is that the best action under perfect compliance need not be the best advice when the human may deviate from it (Grand-Clément & Pauphilet, 2026). Existing work formalizes human compliance through an *adherence rate*, which characterizes the likelihood that an algorithmic recommendation is implemented. Introducing this quantity makes it possible to explicitly distinguish the recommendation policy from the effective policy induced by human compliance. Instead of assuming that each recommended action will be implemented, the decision maker can evaluate recommendations based on their expected realized outcomes induced by both compliance and non-compliance. Building on this idea, a growing body of work has incorporated partial adherence into sequential decision-making frameworks, in particular reinforcement learning, demonstrating that adherence-aware recommendation policies can improve the realized performance (Faros et al., 2023; Chen et al., 2024; Sun et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, existing adherence-aware methods typically treat human adherence as fixed or exogenous during the decision process, independent of the recommendations made and their realized outcomes. This assumption is impractical because the success and failure of algorithmic advice systematically changes subsequent reliance, often with persistent and path-dependent effects, as demonstrated in empirical studies (Gill et al., 2024; Prahl & Van Swol, 2017; Chacon et al., 2022). Consequently, adherence should not be treated as a fixed or exogenous parameter that determines how recommendations are implemented, but as an evolving state affected by previous recommendation outcomes. Following this insight, we consider adherence as endogenous in human-AI interactions, with its evolution shaped by previous recommendations and their realized outcomes.

We propose an adherence-aware decision framework that explicitly accounts for the endogenous evolution of human adherence. We model the evolution of human adherence through a latent trust state, which captures the human's underlying propensity to follow algorithmic recommendations. The framework maintains a Bayesian belief over the latent trust state and incorporates this belief into sequential recommendation decisions. Within the resulting belief MDP, we derive an adherence-aware Bellman operator that explicitly marginalizes over whether the human accepts the recommendation or falls back to their own behavior, thereby aligning value learning with the actual dynamics of human-in-the-loop deployment. For environments with unknown or high-dimensional dynamics, we further develop a scalable reinforcement-learning implementation based on this formulation. We also provide a theoretical characterization of the framework's decision and learning properties, quantifying how modeling and estimation of dynamic adherence affect policy performance. In light of the above discussion, we summarize the main contributions of this work as follows:

- We formulate adherence-aware sequential decision making with endogenous human trust as a belief MDP. The formulation explicitly introduces an adherence-aware Bellman operator that marginalizes over the human's response.
- We provide a theoretical analysis of the resulting framework. We establish contraction of the adherence-aware Bellman operator and characterize the sub-optimality of the proposed method.
- We develop scalable reinforcement learning implementations for continuous or high-dimensional spaces. Experiments in sepsis treatment, highway lane changing, and supply chain inventory management demonstrate that modeling dynamic adherence can improve decision performance.

2 Related Work

Adherence-aware Recommendations. In human-in-the-loop systems, the action recommended by an algorithm may differ from the action ultimately implemented by the human decision-maker. Grand-Clément and Pauphilet (Grand-Clément & Pauphilet, 2026) formalize this mismatch by distinguishing the recommendation policy from the effective policy induced by partial adherence, and show that ignoring non-adherence can lead to arbitrarily poor outcomes. They introduce the adherence-aware MDP as a principled framework for modeling partial adherence. Building on this line of research, Faros et al. (Faros et al., 2023) extend the framework to settings with unknown environmental dynamics. They propose an adherence-aware Q-learning algorithm and establish its convergence. Chen et al. (Chen et al., 2024) further develop an adherence-aware UCB algorithm for sequential recommendation problems. On the application side, Sun et al. (Sun et al., 2025) apply adherence-aware reinforcement learning to lane-changing recommendations.

Although these methods explicitly account for the mismatch between recommended and executed actions, they typically model adherence using a fixed rate, which can be restrictive in real systems.

Dynamics of Human Reliance on AI. A complementary literature suggests that reliance on AI is shaped by the history and realized outcomes of previous interactions, rather than being well represented by a fixed behavioral parameter. Prahl & Van Swol (2017) show that receiving poor automated advice can substantially reduce subsequent advice utilization, with a stronger effect than for equally poor human advice, highlighting the sensitivity of reliance to observed algorithm performance. Chacon et al. (2022) provide longitudinal evidence that such effects persist over repeated interactions. More recently, Gill et al. (2024) find that early errors reduce later reliance, whereas successful early

predictions increase subsequent reliance. Thus, human reliance becomes endogenous to the decision process, motivating sequential models that explicitly account for its evolution over time.

Sequential Decision Making with Latent Human States. One natural way to model such evolving human factors is through latent-state and belief-based sequential decision models(Pineau et al., 2003; Golowich et al., 2023). Dynamically inferring latent human factors is common in human-AI and human-robot collaboration (Hoffman et al., 2024; Zahedi et al., 2023; Yu et al., 2024). Chen et al. (Chen et al., 2020) model trust as a latent state in a POMDP, learn trust dynamics and a human decision model from interaction data and plan for optimal robot actions. More broadly, POMDP and belief-based formulations have been widely used to reason about latent human factors, including intentions and goals (Cramer et al., 2021; Yow et al., 2023), as well as preferences and behavioral tendencies (Mangin et al., 2022; van der Spaa et al., 2024)

These works provide the general machinery for sequential decision making with latent, evolving human states. Building on this perspective, We formulate adherence-aware sequential decision making when human adherence is endogenous and dynamically evolves with prior interactions. This dynamic structure leads to a belief-based decision framework and an adherence-aware Bellman formulation that explicitly accounts for acceptance and fallback behavior.

3 PRELIMINARIES

3.1 PARTIALLY OBSERVED MARKOV DECISION PROCESS

A partially observed Markov decision process (POMDP) is defined by the tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, Z, R, \gamma),$$

where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ denote the state, action, and observation spaces, respectively. The transition kernel $P(s_{t+1} \mid s_t, a_t)$ specifies the probability of moving from state s_t to state s_{t+1} after taking action a_t, while the observation kernel $Z(o_{t+1} \mid s_{t+1}, a_t)$ specifies the probability of receiving observation o_{t+1} after the transition. The function $R(s_t, a_t)$ denotes the immediate reward at time step t. The within-step interaction follows

$$s_t \xrightarrow{a_t,\, P} s_{t+1} \xrightarrow{Z} o_{t+1}.$$

Since the latent state is not directly observed, decisions generally depend on the interaction history $H_t = (o_0, a_0, \ldots, a_{t-1}, o_t)$. This history can be summarized by the belief state

$$b_t(s) := \Pr(s_t = s \mid H_t),$$

which is a sufficient statistic for sequential decision making. Thus, a policy can condition on b_t rather than the full history.

Given action a_t, the belief first undergoes a prediction step and is then updated after observing o_{t+1}:

$$b^-_{t+1}(s') = \sum_{s \in \mathcal{S}} P(s' \mid s, a_t) b_t(s), \qquad b_{t+1}(s') = \frac{Z(o_{t+1} \mid s', a_t) b^-_{t+1}(s')}{\sum_{\bar{s} \in \mathcal{S}} Z(o_{t+1} \mid \bar{s}, a_t) b^-_{t+1}(\bar{s})}.$$

Equivalently,

$$b_t \xrightarrow[\text{prediction}]{a_t} b^-_{t+1} \xrightarrow[\text{correction}]{o_{t+1}} b_{t+1}.$$

3.2 HUMAN-IN-THE-LOOP DECISION PROCESS

The standard POMDP formulation above assumes that an action is selected before the state transition and the subsequent observation. In human-in-the-loop decision support, however, the algorithm first issues a recommendation u^r_t and then observes the human's response o_t representing whether they accept algorithm's recommendations. The executed action u_t drives the system transition.

We denote by $\theta_t \in \Theta$ the latent human trust state at time t, which governs the human's propensity to accept the algorithmic recommendation. At the beginning of time step t, the trust state has already

been propagated through the transition dynamics, yielding the predictive belief b_t^- which is

$$b_{t+1}^-(\theta') = \sum_{\theta \in \Theta} P_\beta(\theta' \mid \theta, x_t, u_t) b_t(\theta). \tag{1}$$

The recommendation u_t^r must therefore be selected based on b_t^-, before the human response o_t is observed. This yields the following Bayesian filtering cycle within each decision step.

$$(x_t, b_t^-) \xrightarrow{\text{recommend } u_t^r} o_t \xrightarrow{\text{Bayes update}} b_t \xrightarrow[\text{Bayes prediction}]{\text{execute } u_t} (x_{t+1}, b_{t+1}^-).$$

4 Adherence-aware Reinforcement Learning with Dynamic Trust

4.1 Adherence-Aware Bellman Operator

Building on the human-in-the-loop interaction, at each time step, the agent recommends u_t^r, which the human either accepts or rejects in favor of a fallback action $g^b(x_t)$. The acceptance probability under the predictive belief b_t^- is

$$\hat{p}_t = \sum_{\theta \in \Theta} P(o_t = 1 \mid x_t, \theta, u_t^r) b_t^-(\theta). \tag{2}$$

The executed action then drives the environmental transition $x_{t+1} \sim P_X(\cdot \mid x_t, u_t)$, while the latent trust state evolves according to $\theta_{t+1} \sim P(\cdot \mid \theta_t, x_t, u_t)$.

We now derive the Bellman equation induced by the human-in-the-loop interaction described above. After the human response o_t is observed, the posterior belief b_t and the executed action u_t are known. We therefore define the executed-action value function $Q(x_t, b_t, u_t)$, as the expected discounted return from this post-response information state. We use q to denote an arbitrary bounded candidate executed-action value function. For a candidate recommendation u^r, let $b^{\mathrm{acc}}(u^r)$ and $b^{\mathrm{rej}}(u^r)$ denote the posterior beliefs obtained after observing acceptance and rejection, respectively. We define the adherence-aware recommendation value

$$\Psi_q(x_t, b_t^-, u_t^r) = \hat{p}_t \, q(x_t, b_t^{\mathrm{acc}}(u_t^r), u_t^r) + (1 - \hat{p}_t) \, q\Big(x_t, b_t^{\mathrm{rej}}(u^r), g^b(x_t)\Big). \tag{3}$$

The resulting adherence-aware Bellman optimality operator is

$$(\mathcal{J}q)(x_t, b_t, u_t) = R(x_t, u_t) + \gamma \, \mathbb{E}_{x_{t+1} \sim P_X(\cdot \mid x_t, u_t)} \left[\max_{u_{t+1}^r \in \mathcal{U}} \Psi_q(x_{t+1}, b_{t+1}^-, u_{t+1}^r) \right]. \tag{4}$$

Accordingly, the recommendation policy is

$$\pi^{\mathrm{rec}}(x_t, b_t^-) = \arg\max_{u_t^r \in \mathcal{U}} \Psi_Q(x_t, b_t^-, u_t^r). \tag{5}$$

The key design of our adherence-aware Bellman operator is to learn the value of the action actually executed, $Q(x_t, b_t, u_t)$, thereby keeping accepted and rejected transitions structurally separated during value learning. This separation facilitates credit assignment by avoiding the conflation of heterogeneous outcomes induced by the same recommendation under different human responses. Candidate recommendations are then evaluated by explicitly marginalizing over their acceptance and fallback branches.

4.2 Joint Belief over Trust and Human Model Parameters

The previous formulation assumes that the transition function of the trust state θ is known. In practice, these components may be uncertain and must be inferred from interaction data. This setting is closely related to system identification in adaptive control, where the agent must simultaneously estimate latent states and unknown model parameters.

We introduce a trust transition model $P_\beta(\theta_{t+1} \mid \theta_t, x_t, u_t)$ parameterized by β and maintain a joint belief over (θ_t, β) defined as $b_t(\theta, \beta) = P(\theta_t = \theta, \beta \mid H_t)$.

After observing the adherence response o_{t+1}, the joint belief is updated according to Bayes' rule:

$$b_{t+1}(\theta', \beta) = \frac{P(o_{t+1} \mid x_{t+1}, \theta', u_{t+1}^r) \sum_{\theta \in \Theta} P_\beta(\theta' \mid \theta, x_t, u_t) b_t(\theta, \beta)}{\sum_{\bar{\beta}} \sum_{\bar{\theta} \in \Theta} P(o_{t+1} \mid x_{t+1}, \bar{\theta}, u_{t+1}^r) \sum_{\theta \in \Theta} P_{\bar{\beta}}(\bar{\theta} \mid \theta, x_t, u_t) b_t(\theta, \bar{\beta})}. \tag{6}$$

4.3 Scalable Adherence-aware Deep Reinforcement Learning

The Bellman operator in Section 4.1 assumes that the expectation over the environmental transition P_X can be evaluated exactly. In many applications, however, the environment dynamics are unknown or too complex to model explicitly. We therefore construct a model-free realization of the operator using sampled environmental transitions, while retaining the analytical marginalization over the human response. This preserves the structural separation between accepted and rejected outcomes rather than requiring their effects to be learned only through mixed recommendation-level samples.

Given a transition $(x_t, b_t, u_t, r_t, x_{t+1})$, we replace the expectation over x_{t+1} by the observed next state while retaining the analytical marginalization over the human response. This yields the sample-based Bellman target

$$y_t(q) = r_t + \gamma \max_{u^r_{t+1} \in \mathcal{U}} \Big[\hat{p}_{t+1}\, q\big(x_{t+1}, b^{\mathrm{acc}}_{t+1}(u^r_{t+1}), u^r_{t+1}\big) + (1-\hat{p}_{t+1})\, q\Big(x_{t+1}, b^{\mathrm{rej}}_{t+1}(u^r_{t+1}), g^b(x_{t+1})\Big) \Big]. \tag{7}$$

Here, $\hat{p}_{t+1}$ and the corresponding posterior beliefs are computed from the predictive belief b^-_{t+1} defined by the belief update in Section 4.1.

In large or continuous state spaces, however, representing the executed-action value $Q(x, b, u)$ exactly is generally infeasible. We therefore use a parameterized function approximator $Q_\phi(x, b, u)$ and fit it to the sample-based Bellman target by minimizing the temporal-difference loss

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{(x_t, b_t, u_t, r_t, x_{t+1}) \sim \mathcal{D}} \Big[\big(Q_\phi(x_t, b_t, u_t) - y_t(Q_\phi)\big)^2 \Big], \tag{8}$$

where $\mathcal{D}$ denotes the replay data.

The same formulation supports different reinforcement-learning architectures. For small recommendation spaces, the maximization over u^r can be evaluated directly, yielding a DQN-style implementation. For larger or continuous recommendation spaces, an actor can instead approximate the maximizing recommendation, leading to an actor–critic implementation. In both cases, the critic evaluates the executed action, while recommendation selection retains the adherence-aware marginalization defined in Section 4.1.

5 Theoretical Results

We now analyze the theoretical properties of the proposed adherence-aware framework. We first establish that the Bellman operator introduced in Section 4.1 defines a well-posed dynamic programming recursion. We then characterize how errors in belief estimation and the human model affect decision performance, and finally examine the structural limitations of static-adherence models. For conciseness, we present streamlined versions of the theoretical results in the main text. Complete variable definitions are provided in the corresponding proof sections in Appendix C.

5.1 Properties of Adherence-aware Bellman Operator

We begin by establishing the basic contraction property of the adherence-aware Bellman operator. Although the continuation value explicitly marginalizes over the human's acceptance and fallback branches, this additional structure preserves the standard contraction property of the Bellman operators.

Lemma 1 (Contraction of the adherence-aware Bellman operator)**.** *Let q_1 and q_2 be any bounded executed-action value functions. The adherence-aware Bellman optimality operator $\mathcal{J}$ defined in Section 4.1 satisfies*

$$\|\mathcal{J}q_1 - \mathcal{J}q_2\|_\infty \le \gamma \|q_1 - q_2\|_\infty.$$

Hence, $\mathcal{J}$ is a γ-contraction. Consequently, it admits a unique fixed point $Q^\star$, and for any bounded initial function q_0,

$$\lim_{k\to\infty} \|\mathcal{J}^k q_0 - Q^\star\|_\infty = 0.$$

The lemma shows that explicitly accounting for the acceptance and fallback branches does not alter the fundamental convergence properties of Bellman iteration. The proof is shown in Appendix C.1.

5.2 Error Decomposition for Dynamic-Trust Belief MDPs

We next quantify how estimation errors in the latent trust state and human model propagate to decision performance.

Theorem 1 (Expected Sub-optimality Gap Induced by State Estimation)**.** *Let τ_T denote the historical training trajectory up to time T, after which the estimated parameter $\hat{\beta}$ is fixed. During the execution phase at time step t, the expected sub-optimality gap of the policy $\hat{\pi}$ evaluated is as follows*

$$\mathbb{E}_{\tau_T}\left[G_M(x_t, b_t^{-,\star}, u_t^{r,\star}) - G_M(x_t, b_t^{-,\star}, \hat{u}_t^r)\right] \leq O(\sigma^t) + O(e^{-\tilde{\Delta}T}), \tag{9}$$

where $G_M(x, b^-, u^r) := \Psi_{Q_M^\star}(x, b^-, u^r)$, $\sigma \in (0,1)$ is the filter stability coefficient and $\tilde{\Delta} > 0$ is the effective decay rate of the parameter estimation error.

Theorem 1 shows that the effects of belief initialization and parameter estimation diminish as filtering and learning proceed. Thus, errors in initially understanding the human and in learning their trust dynamics do not persist indefinitely. With sufficient interaction and training, the resulting recommendation quality approaches that under the true human model.

As discussed in Section 4.3, a scalable function-approximation method is necessary in large-scale problems. To characterize how the function approximation interacts with the belief and parameter estimation errors, we use DQN as a representative deep reinforcement learning architecture and follow Fan et al. (2020) to analyze its neural fitted Q-iteration (FQI) surrogate. This leads to the following extension.

Corollary 1 (Total Expected Sub-optimality Gap)**.** *Let q_K denote the value-function approximation obtained after K iterations of Neural Fitted Q-Iteration (FQI). Define the induced recommendation-level approximation $G_K(x, b^-, u^r) := \Psi_{q_K}(x, b^-, u^r)$. Let $\mathcal{D}$ denote the replay data generated along the training trajectory τ_T. We have*

$$\mathbb{E}_{\tau_T, \mathcal{D}}\left[G_M(x_t, b_t^{-,\star}, u_t^{r,\star}) - G_M(x_t, b_t^{-,\star}, u_{t,K}^r)\right] \leq O(\sigma^t) + O(e^{-\tilde{\Delta}T}) + \mathcal{E}_{\mathrm{FQI}}(n, K),$$

where $\mathcal{E}_{\mathrm{FQI}}(n, K)$ denotes the statistical, approximation, and finite-iteration error of Neural FQI on the induced adherence-aware belief MDP.

Taking DQN as a representative deep-RL implementation, Corollary 1 shows that belief and parameter estimation do not introduce additional learning instability. Since the adherence-aware Bellman operator satisfies the standard properties required by the FQI analysis as stated in Lemma 1 to **??**, controlling the usual statistical and approximation errors of deep RL can retain the decision quality guarantees.

The above guarantees are established under correct model specification. This leaves a separate issue of model mis-specification, where the true trust dynamics may not be representable by the assumed model class. We next quantify the resulting performance loss.

Proposition 1 (Model mis-specification)**.** *Suppose the true model $M^\star$ need not belong to the model class $\{M_\beta : \beta \in \Omega\}$, and let $M^\dagger \in \arg\min_{M_\beta} D(M^\star, M_\beta)$ denote the pseudo-true model under the one-step Bellman mismatch D defined in Appendix C.2.*

Let $\delta_K^\dagger$ denote the maximum total-variation discrepancy between the one-step information-state transition kernels induced by $M^\star$ and $M^\dagger$. Then, for any fixed recommendation policy π,

$$\|G_{M^\star}^\pi - G_{M^\dagger}^\pi\|_\infty \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2}\delta_K^\dagger. \tag{10}$$

Proposition 1 highlights that the key quantity to control under model mis-specification is the one-step Bellman mismatch. If this mismatch remains small, model mis-specification need not lead to a large degradation in decision performance.

5.3 Structural Limitations of Static-Adherence Models

A static-adherence model replaces the evolving human response with a single fixed acceptance probability. The resulting error contains structural components that need not vanish even with unlimited data.

Table 1: Sepsis-treatment results and representative recommendation policies.

(a) Performance			
Method	Survival ↑	Mortality ↓	Recovery Steps ↓
PA-VI	83.9%	15.0%	11.90
SA-VI	83.9%	15.0%	11.90
DA-PBVI	**86.9%**	**12.3%**	11.92

(b) Representative policy in severe states			
Method	Low trust	Medium trust	High trust
PA-VI / SA-VI	Aggressive	Aggressive	Aggressive
DA-PBVI	Targeted	Targeted	Aggressive

Proposition 2 (Structural error of static-adherence models)**.** *Let* $\hat{p}_T := \frac{1}{T}\sum_{k=1}^{T} O_k,\ p_T^\dagger := \frac{1}{T}\sum_{k=1}^{T}\mathbb{P}(O_k = 1 \mid I_k)$*, and define* $\hat{M}_{\text{stat}} := M_{\text{stat}}(\hat{p}_T),\ M_{\text{stat}}^\dagger := M_{\text{stat}}(p_T^\dagger)$.

Let $b_t^{-,\star}$ *denote the true predictive belief and* $b_{t,\text{stat}}^{-,\dagger}$ *the oracle static predictive representation. Then, for any fixed recommendation policy* π*,*

$$\begin{aligned}&\left|G_{M^\star}^{\pi}(x_t, b_t^{-,\star}, u_t^r) - G_{\hat{M}_{\text{stat}}}^{\pi}(x_t, b_{t,\text{stat}}^{-,\dagger}, u_t^r)\right| \\ &\le \underbrace{\frac{R_{\max}}{1-\gamma}\left\|b_t^{-,\star} - b_{t,\text{stat}}^{-,\dagger}\right\|_1}_{\textit{belief approximation}} + \underbrace{\frac{D(M^\star, M_{\text{stat}}^\dagger)}{1-\gamma}}_{\textit{model approximation}} + \underbrace{\frac{2R_{\max}}{(1-\gamma)^2}|p_T^\dagger - \hat{p}_T|}_{\textit{estimation error}}.\end{aligned} \tag{11}$$

Proposition 2 shows that the first two error terms are structural and do not vanish with additional data, which is analyzed in detail in the proof in Appendix C.3. Thus, more accurate estimation of a static adherence rate alone cannot eliminate the performance loss caused by ignoring dynamic trust.

6 EXPERIMENTS

We evaluate our method on three human-in-the-loop sequential decision-making tasks with different environmental dynamics and safety considerations: sepsis treatment, highway lane-changing and supply chain inventory management.

6.1 CONTROLLED STUDY: WHEN DOES DYNAMIC ADHERENCE MATTER?

Baselines. We compare three planning models: **PA-VI**, which assumes perfect adherence; **SA-VI**, which uses a fixed estimated adherence rate; and **DA-PBVI**, which maintains a posterior over the latent trust state and plans in the corresponding belief MDP.

Table 1 shows that PA-VI and SA-VI learn essentially the same trust-independent policy, whereas explicitly modeling dynamic adherence changes the recommendation policy. In severe states, DA-PBVI recommends targeted treatment at low or medium trust and switches to aggressive treatment at high trust. This reflects a short-term reward versus long-term adherence trade-off: when trust is low, the policy favors the more reliable action to improve future reliance; once trust is high, it switches to the higher-value aggressive action A.4. The corresponding trust trajectories are reported in Appendix A.4, which shows DA-PBVI maintains a higher long-run adherence level.

6.2 SCALABLE ADHERENCE-AWARE REINFORCEMENT LEARNING

Baselines. We compare against five baselines: **PA-RL**, which assumes perfect adherence; **SA-RL**, which models adherence using a single fixed acceptance probability; **POMCP**, an online Monte Carlo POMDP planner; **PBVI**, point-based value iteration which is an offline belief-space POMDP planner; and **DRQN**, a recurrent policy that conditions on recent interaction histories without explicitly maintaining a belief state.

Table 2: Performance in the lane-changing task under two trust-dynamics models. Higher reward and speed are better; larger TTC indicates a greater safety margin.

	(a) Standard trust dynamics				(b) Conservative-preference dynamics			
Method	Reward	Speed	$\text{TTC}_{5\%}$	TTC_{med}	Reward	Speed	$\text{TTC}_{5\%}$	TTC_{med}
PA-RL	179.34 ± 6.37	26.09 ± 2.23	4.47	14.37	182.78 ± 6.91	24.38 ± 2.71	5.11	17.11
SA-RL	187.87 ± 5.88	26.52 ± 2.87	6.27	16.91	187.16 ± 5.49	25.45 ± 2.70	8.30	23.54
POMCP	182.52 ± 3.16	23.96 ± 2.04	12.07	21.63	184.88 ± 3.23	24.12 ± 1.96	10.50	19.60
PBVI	166.85 ± 2.06	24.05 ± 0.30	5.61	19.92	178.43 ± 1.88	20.34 ± 0.25	7.92	22.76
DRQN	183.77 ± 6.21	24.97 ± 2.16	6.23	20.58	182.96 ± 6.38	24.88 ± 2.15	5.02	17.83
Ours	$\mathbf{189.35 \pm 5.42}$	$\mathbf{27.34 \pm 2.36}$	6.14	15.59	$\mathbf{188.87 \pm 5.52}$	$\mathbf{26.35 \pm 2.59}$	7.47	20.75

Table 3: Performance comparison in the supply chain inventory management task. Higher net profit, revenue, and served demand are better.

Method	Net Profit ↑	Revenue ↑	Served Demand ↑
PA-RL	3643.68 ± 1255.28	6169.02 ± 2407.29	411.27 ± 160.49
SA-RL	3651.64 ± 1240.70	6157.27 ± 2341.61	410.48 ± 156.11
POMCP	4058.45 ± 884.53	6812.92 ± 1587.85	454.19 ± 105.86
PBVI	3965.46 ± 868.68	7019.35 ± 1568.78	467.96 ± 104.59
DRQN	4001.37 ± 893.07	6649.18 ± 1600.93	443.28 ± 106.73
Ours	$\mathbf{4298.05 \pm 751.23}$	$\mathbf{7771.41 \pm 1255.44}$	$\mathbf{518.09 \pm 83.70}$

6.2.1 Lane Changing

Table 2 evaluates the DQN-style implementation in the more complex lane-changing environment. Across both trust-dynamics settings, our method achieves the highest reward and average speed while maintaining competitive lower-tail TTC. The improvement over PA-RL and SA-RL highlights the value of explicitly tracking evolving adherence. Trust-conditioned recommendation behavior is shown in Appendix A.3. Our method also outperforms the history-based DRQN baseline, supporting explicit latent-state modeling over relying solely on recent interaction histories. It further avoids time-consuming online tree search and explicit discretization of the continuous traffic state.

6.2.2 Supply Chain Inventory Management

Table 3 reports the results in the supply chain inventory management task. Our method achieves the highest net profit, revenue, and served demand among all compared methods. We observe relatively large across-episode variation for all methods. This variance is largely driven by differences in the realized trust trajectories across episodes. To reduce uncertainty, we conducted paired comparisons in the same environment setting and random seeds and showed that our method outperforms the baselines for all cases with different initial trust levels. The relevant results are shown in the Appendix A.4.

6.3 Estimation, Learning and Robustness Analysis

In this section, we further examine whether the proposed framework works as intended in practice and assess its robustness to key sources of uncertainty and model mismatch.

6.3.1 Trust Tracking and Learning Efficiency

Tracking the latent trust state. We first evaluate whether observations contain sufficient information to track the latent adherence state. We compare the inferred adherence rate with the ground-truth process as shown in Figure 1.

Learning efficiency with trust tracking. We examine whether tracking dynamic trust makes policy learning more difficult. Figure 2 shows the return per step over training. Despite operating on an

Table 4: Effect of the adherence-aware Bellman structure

Bellman formulation	Return ↑	Served Demand ↑
Unstructured recommendation-level target	4117.04 ± 996.97	461.74 ± 124.29
Adherence-aware marginalized target	**4298.05 ± 751.23**	**518.09 ± 83.70**

augmented belief state, our method reaches the high-return regime earlier and converges to a higher performance level than the baselines. This result highlights the value of correctly modeling the human-in-the-loop dynamics. Although belief tracking increases the state dimension, representing the evolving adherence process makes the effective decision problem easier to learn.

6.3.2 Effect of the Adherence-Aware Bellman Structure

Our formulation proposed in Section 4.1 learns the value of the action that is actually executed and evaluates each recommendation by explicitly marginalizing over the acceptance and fallback branches. We compare this design with an unstructured alternative that directly learns a value for the recommendation from observed transitions. The distinction is important for credit assignment. Under the unstructured formulation, the same recommendation can be followed by substantially different rewards and next states. In contrast, the adherence-aware operator provides a cleaner learning signal. Table 4 shows that, under the same training budget, the adherence-aware formulation learns substantially better policies than the unstructured alternative.

6.3.3 Robustness Analysis.

Parameter non-identifiability We present a setting in which the underlying model parameters are not uniquely identifiable from the observed interaction trajectories. The cumulative conditional KL divergence between the true model and incorrect model grows slowly. Despite this ambiguity, the observationally equivalent parameterizations induce nearly identical predictive trust levels. This suggests that parameter recovery is a sufficient, but not necessary, condition for accurate downstream decision making when different parameterizations are behaviorally equivalent. By contrast, when different parameter values induce clearly distinguishable interaction trajectories, the true β can be accurately identified. The relevant results are shown in Figure 5 and 6 in Appendix A.4.

Model mis-specification Finally, we perturb the assumed trust-transition during evaluation. We measure the resulting degradation in task performance to assess whether the proposed method remains effective. This analysis tests whether a misspecified but sufficiently predictive surrogate model can remain useful for decision making. The results are shown in Table 9 in Appendix A.4

7 Conclusion

We study human-in-the-loop sequential decision making when human adherence evolves endogenously with previous interactions. To capture this dynamic dependence, we formulate the problem as a belief MDP over latent human trust and develop an adherence-aware reinforcement learning framework that explicitly distinguishes between recommended and executed actions. The resulting Bellman formulation accounts for both acceptance and fallback behavior, and supports scalable reinforcement learning. Our theoretical analysis establishes the contraction of the adherence-aware Bellman operator and characterizes how belief estimation, human-model estimation, and model mis-specification affect decision performance. Across sepsis treatment, highway lane changing, and supply chain inventory management, our experiments show that modeling dynamic adherence can improve realized system performance and lead to recommendations that adapt meaningfully to the evolving human state.

Future work may apply our framework in real-world human-in-the-loop settings. Existing empirical studies have provided valuable case-specific managerial insights into human reliance on algorithms(Caro et al., 2026). Our goal is to complement this literature by implementing an adherence-aware decision framework that can directly address partial adherence and assess whether these behavioral insights translate into measurable improvements in operational outcomes.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript editing. The literature search was primarily conducted by the authors during the early stages of the project, with generative AI used subsequently to check for potentially relevant studies that may have been overlooked. Generative AI was also used to double-check mathematical derivations and proofs after they had first been developed and manually verified by the authors. For the experimental component, generative AI was used to assist with the implementation of experimental code, while the experimental design, methodological choices, and implementation of the core formulation were completed by the authors. The experiments execution, iterative modification, and evaluation, were also conducted by the authors. We have reviewed and verified all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete formulation of the proposed adherence-aware framework in Section 4. The experimental implementation follows this formulation directly and does not rely on specialized algorithmic choices or extensive hyperparameter tuning. Implementation details and experimental settings are provided in Appendix A.2, and the corresponding code is included in the supplementary materials.

REFERENCES

Felipe Caro, Jean-Edouard Colliard, Elena Katok, Axel Ockenfels, Nicolas Stier-Moses, Catherine Tucker, and DJ Wu. Introduction to the special issue on the human-algorithm connection. *Management Science*, 72(1):1–13, 2026.

Alvaro Chacon, Edgar E Kausel, and Tomas Reyes. A longitudinal approach for understanding algorithm use. *Journal of Behavioral Decision Making*, 35(4):e2275, 2022.

Guanting Chen, Xiaocheng Li, Chunlin Sun, and Hanzhao Wang. Learning to make adherence-aware advice. In *International Conference on Learning Representations*, volume 2024, pp. 25294–25305, 2024.

Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(2):1–23, 2020.

Maxime C Cohen, Tinglong Dai, Georgia Perakis, Narendra Agrawal, Gad Allon, Robert N Boute, Gérard P Cachon, Zhe Chen, Morris Cohen, Rares Cristian, et al. Om forum—supply chain management in the ai era: A vision statement from the operations management community. *Manufacturing & Service Operations Management*, 28(3):687–705, 2026.

Martijn Cramer, Karel Kellens, and Eric Demeester. Probabilistic decision model for adaptive task planning in human-robot collaborative assembly based on designer and operator intents. *IEEE Robotics and Automation Letters*, 6(4):7325–7332, 2021.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pp. 486–489. PMLR, 2020.

Ioannis Faros, Aditya Dave, and Andreas A Malikopoulos. A q-learning approach for adherence-aware recommendations. *IEEE Control Systems Letters*, 7:3645–3650, 2023.

Andrej Gill, Robert M Gillenkirch, Julia Ortner, and Louis Velthuis. Dynamics of reliance on algorithmic advice. *Journal of Behavioral Decision Making*, 37(4):e2414, 2024.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 349–362, 2023.

Julien Grand-Clément and Jean Pauphilet. The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science*, 72(1):667–692, 2026.

Guy Hoffman, Tapomayukh Bhattacharjee, and Stefanos Nikolaidis. Inferring human intent and predicting human action in human–robot collaboration. *Annual Review of Control, Robotics, and Autonomous Systems*, 7(1):73–95, 2024.

Saravanan Kesavan and Tarun Kushwaha. Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science*, 66(11): 5182–5190, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Olivier Mangin, Alessandro Roncone, and Brian Scassellati. How to be helpful? supportive behaviors and personalization for human-robot collaboration. *Frontiers in Robotics and AI*, 8:725780, 2022.

Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for pomdps. In *Ijcai*, volume 3, pp. 1025–1032, 2003.

Andrew Prahl and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017.

Benjamin Rolf, Ilya Jackson, Marcel Müller, Sebastian Lang, Tobias Reggelin, and Dmitry Ivanov. A review on reinforcement learning algorithms and applications in supply chain management. *International Journal of Production Research*, 61(20):7151–7179, 2023.

Clare Snyder, Samantha Keppler, and Stephen Leider. Algorithm reliance: Fast and slow. *Management Science*, 72(1):368–385, 2026.

Jiankun Sun, Dennis J Zhang, Haoyuan Hu, and Jan A Van Mieghem. Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865, 2022.

Weihao Sun, Heeseung Bang, and Andreas A Malikopoulos. Ai recommendation systems for lane-changing using adherence-aware reinforcement learning. *arXiv preprint arXiv:2504.20187*, 2025.

Linda van der Spaa, Jens Kober, and Michael Gienger. Simultaneously learning intentions and preferences during physical human-robot cooperation. *Autonomous Robots*, 48(4):11, 2024.

Yuexuan Wang, Jingyuan Zhou, and Kaidi Yang. Hierarchical decision making with structured policies: A principled design via inverse optimization. *arXiv preprint arXiv:2606.28764*, 2026.

J-Anne Yow, Neha Priyadarshini Garg, and Wei Tech Ang. Shared autonomy of a robotic manipulator for grasping under human intent uncertainty using pomdps. *IEEE Transactions on Robotics*, 40:332–350, 2023.

Pian Yu, Shuyang Dong, Shili Sheng, Lu Feng, and Marta Kwiatkowska. Trust-aware motion planning for human-robot collaboration under distribution temporal logic specifications. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12949–12955. IEEE, 2024.

Zahra Zahedi, Mudit Verma, Sarath Sreedharan, and Subbarao Kambhampati. Trust-aware planning: Modeling trust evolution in iterated human-robot interaction. In *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*, pp. 281–289, 2023.

A EXPERIMENTS

A.1 PROBLEM DESCRIPTIONS

Sepsis treatment. We construct a stylized sequential treatment environment in which an AI system recommends treatment intensity to a clinician who retains final control over the executed action(Komorowski et al., 2018). The patient state consists of three severity levels $\{\text{moderate}, \text{severe}, \text{critical}\}$ and three short-term trends $\{\text{improving}, \text{stable}, \text{worsening}\}$, together with terminal `recovery` and `death` states. The recommendation space is

$$\mathcal{A} = \{\text{conservative}, \text{targeted}, \text{aggressive}\}.$$

Our purpose in this task is to test whether modeling time-varying adherence improves decision making, without confounding effects from neural function approximation.

Lane changing. We further evaluate our reinforcement-learning method in `highway-v0`, using a three-lane highway and five discrete recommendations,

$$\mathcal{A} = \{\text{left}, \text{right}, \text{stay}, \text{slower}, \text{faster}\}.$$

Unlike the discrete sepsis environment, the traffic state is high-dimensional and continuous, making explicit belief-space planning much less practical.

Supply chain inventory management. We also consider a supply chain inventory management problem in which a warehouse serves multiple retail stores under stochastic customer demand(Rolf et al., 2023; Wang et al., 2026).At each decision epoch, the AI recommends the replenishment quantity for the warehouse, while the downstream allocation of inventory to individual stores is determined by a fixed heuristic policy. The recommendation space is

$$\mathcal{A} = \{0, 1, \ldots, C\},$$

where C denotes the warehouse ordering capacity. The human decision maker may accept the recommendation or reject it and follow an S-type inventory policy as the fallback decision.

A.2 RL METHODOLOGY, BASELINES, AND EXPERIMENTAL SETUP

A.2.1 ADHERENCE-AWARE RL.

Lane changing. For lane changing, we use a DQN-style implementation for the five discrete recommendations. The Q-network evaluates executed actions, while recommendation values are obtained by marginalizing over the acceptance and fallback branches.

Supply chain inventory management. For supply-chain inventory management, we use an actor–critic implementation. A GNN encodes the physical state, the posterior mean of trust is used as the belief feature, and the critic evaluates executed replenishment quantities.

A.2.2 BASELINES

The baselines we use are as follows

- **Perfect-Adherence VI (PA-VI):** ignores adherence and plans as if every recommendation were executed.
- **Static-Adherence VI (SA-VI):** models the accept-or-fallback structure but assumes a single fixed adherence probability $\hat{\theta}$ estimated from interaction data.
- **Dynamic-Adherence PBVI (DA-PBVI):** maintains a posterior over the latent trust state and approximately plans in the corresponding belief MDP.
- **Perfect-Adherence RL (PA-RL):** a reinforcement-learning baseline that assumes every recommendation is accepted, corresponding to an adherence probability of one.

Table 5: Main reinforcement-learning hyperparameters.

Hyperparameter	Lane changing	Supply chain
Discount factor γ	0.99	0.99
Main architecture	DQN	Actor–critic
Hidden dimension	$128, 128$	GNN: 64
Q/critic learning rate	10^{-3}	3×10^{-4}
Actor learning rate	–	10^{-4}
Replay capacity	50,000	–
Batch size	64	–
Target update	500 steps	$\tau = 0.01$
Gradient clipping	10	200
DRQN history length	10	10
Recurrent hidden size	LSTM: 128	GRU: 32

Table 6: Planning-baseline hyperparameters.

Hyperparameter	Lane changing	Supply chain
POMCP simulations	100	300
POMCP search depth	8	8
POMCP rollout depth	4	8
UCT coefficient	2.0	2.0
Time budget per decision	–	30 ms
PBVI model trajectories	300 episodes	1,500 episodes
PBVI belief generation	300 traces, horizon 12	400 episodes, up to 700 points
PBVI stopping rule	8 backup sweeps	convergence or 200 iterations

- **Static-Adherence RL (SA-RL):** an adherence-aware RL baseline that models acceptance using a single fixed probability throughout the interaction, without tracking temporal changes in trust.
- **POMCP:** performs online Monte Carlo tree search over the latent trust state using the same generative human-response and environment model. UCT is used for recommendation selection, and the belief is updated after each observed human response.
- **PBVI:** serves as an offline approximate belief-space planner. Since the physical state spaces are continuous, we first construct finite state abstractions from simulated trajectories, collect reachable belief points, and then perform point-based Bellman backups.
- **DRQN**: a recurrent baseline that directly conditions the recommendation policy on recent sequences, without explicitly maintaining a Bayesian belief state. We keep the recurrent representations of the most recent 10 human–AI interactions. We use an LSTM for lane changing and a GRU together with the physical-state GNN for inventory management.

A.2.3 Experimental Setup

The main hyperparameters for reinforcement learning and planning methods are listed in Table 5 and 6

A.3 Interpretations of Experimental Results

To verify that the learned policy actively uses the inferred trust state, we conduct a matched-state counterfactual analysis. We first collect physical traffic states from real evaluation trajectories. For each collected state, we keep the traffic observation fixed, and intervene only on the trust belief by setting $\theta = 0.3$ (low trust) or $\theta = 0.9$ (high trust). We then compare the corresponding greedy recommendations. The results in Table 7 confirm that the learned recommendation policy does not merely respond to the physical traffic state, but systematically adapts its decisions to the inferred trust level. Especially, the dominant transition is from Stay under low trust to Faster under high trust.

Table 7: Trust-conditioned recommendation behavior under matched traffic states. The physical traffic state is held fixed and only the trust belief is varied.

(a) Policy disagreement across traffic scenarios

Traffic scenario	States	Disagreement rate
Front clear	19,107	90.0%
Dangerous following	793	42.0%
Left-lane opportunity	130	23.1%
Right-lane opportunity	83	28.9%

(b) Representative policy in front-clear states

	Low trust ($\theta = 0.3$)	High trust ($\theta = 0.9$)
Most frequent recommendation	STAY	FASTER
Frequency	81.8%	90.3%

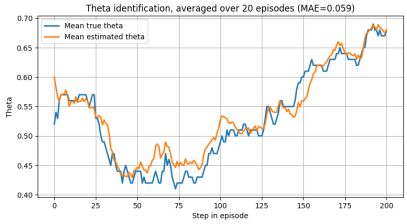


Figure 1: Comparison between inferred adherence rates and the ground-truth process.

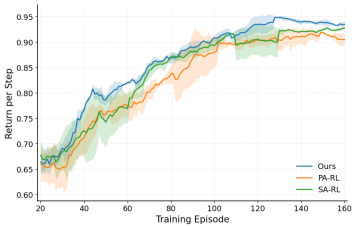


Figure 2: Training return per step across episodes.

A.4 Complementary Experimental Results

Table 8 illustrates the mechanism underlying the trust-dependent policy in a representative `severe-stable` state. Although aggressive treatment provides a larger immediate expected reward, targeted treatment has a higher probability of clinical improvement. This creates a trade-off between short-term reward and the longer-term value of maintaining human trust.

In the following, Figure 1 and 2 show trust estimation and learning efficiency. Figure 5 and 6 explore parameter non-identifiability issues. And Table 9 show the performance under model misspecification.

Table 8: One-step treatment statistics in the `severe-stable` state.

Treatment	Immediate Expected Reward	Improvement Probability
Targeted	0.032	**0.395**
Aggressive	**0.690**	0.255

B Assumptions

Notation. Let

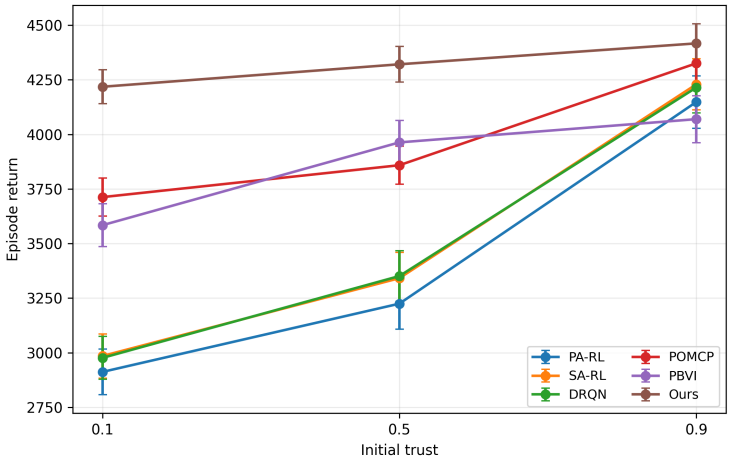
$$\mathcal{B}_\Theta := \Delta(\Theta)$$

denote the space of probability distributions over the latent trust state $\theta \in \Theta$. Let $\beta^\star$ denote the true trust-transition parameter.

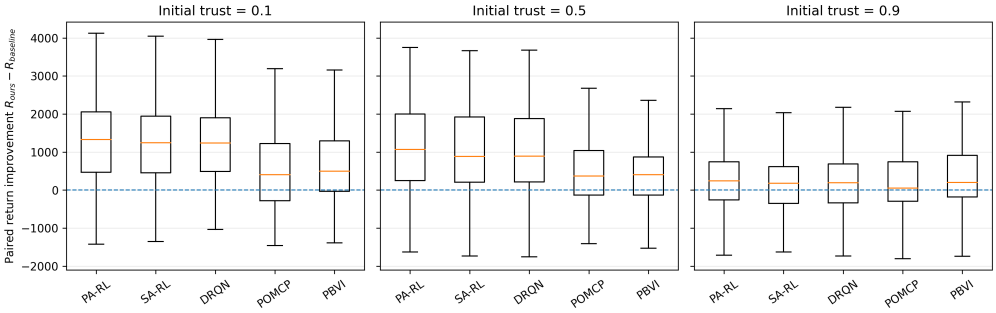
Throughout the analysis, the observation model $P(o \mid x, \theta, u^r)$ and the fallback policy g^b are assumed known and fixed. The trust-transition model is parameterized by

$$P_\beta(\theta' \mid \theta, x, u),$$

where u denotes the executed action.



(a) Performance under different initial trust levels.



(b) Paired return improvement of our method over the baselines.

Figure 3: Performance across heterogeneous trust conditions in the supply chain inventory management task. Top: average return under low, medium, and high initial trust. Bottom: distributions of paired return differences between our method and each baseline under matched evaluation conditions.

Table 9: Robustness to human-model mis-specification in the lane-changing task.

Model	Return ↑	θ MAE ↓
β-scale	187.47 ± 0.91	0.0506
Strict human	187.19 ± 0.95	0.0593
Reward-driven	188.03 ± 0.84	0.0794

For parameter estimation, inference is performed over a finite candidate set Ω with $\beta^\star \in \Omega$. We assume that the transition family P_β is defined for every $\beta \in \mathrm{conv}(\Omega)$, so that the posterior mean estimator used below is a valid model parameter.

Assumption 1 (Uniform Lipschitz trust-transition model)**.** *There exists a constant $L_\beta > 0$ such that, for every $\theta \in \Theta$, environmental state x, executed action u, and parameter $\widehat{\beta} \in \mathrm{conv}(\Omega)$,*

$$\left\| P_{\beta^\star}(\cdot \mid \theta, x, u) - P_{\widehat{\beta}}(\cdot \mid \theta, x, u) \right\|_1 \le L_\beta \|\beta^\star - \widehat{\beta}\|_1.$$

Thus, small perturbations of the trust-transition parameter induce uniformly small perturbations of the conditional distribution of the next latent trust state.

Assumption 2 (Exponential Bayesian filter forgetting)**.** *For an admissible observed context $c_k = (x_k, u_k^r, o_k, u_k)$, let $\Phi_{\beta^\star}^{c_k}$ denote the predictive-belief update defined above. For $0 \le s < t$, define*

$$\Phi_{\beta^\star}^{c_{s:t-1}} := \Phi_{\beta^\star}^{c_{t-1}} \circ \cdots \circ \Phi_{\beta^\star}^{c_s}.$$

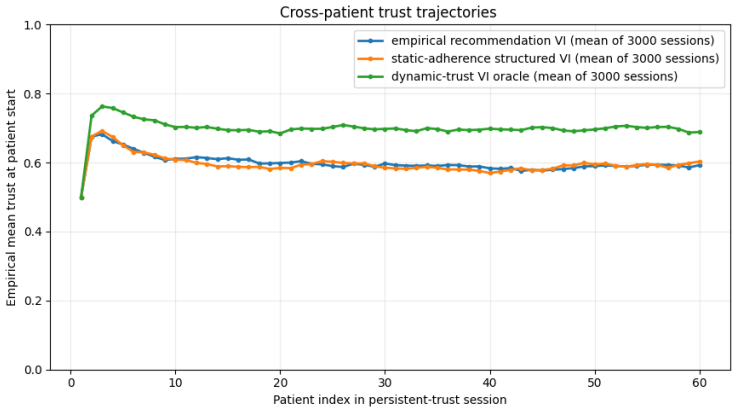
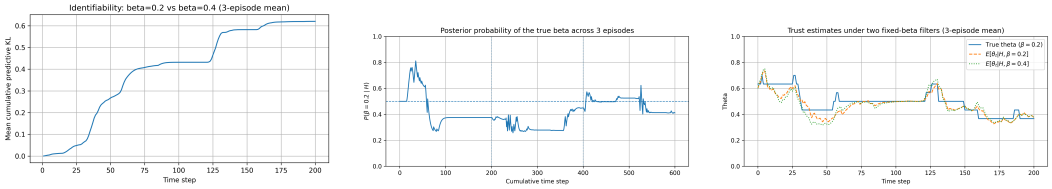


Figure 4: Empirical trust evolution across consecutive patients from the same initial trust condition. Dynamic-adherence planning maintains a higher long-run adherence level by adapting recommendations to the current trust state.



(a) Cumulative predictive KL divergence.

(b) Posterior probability of the true β.

(c) Trust estimates under two fixed-β filters.

Figure 5: Weak parameter identifiability.

We assume that the predictive trust filter exponentially forgets its initialization. Specifically, there exist constants $C_f \geq 1$ and $\rho \in (0,1)$ such that, for every $b^-, \tilde{b}^- \in \mathcal{B}_\Theta$, every $0 \leq s < t$, and every admissible observed context sequence $c_s, \ldots, c_{t-1}$,

$$\left\| \Phi_{\beta^\star}^{c_{s:t-1}}(b^-) - \Phi_{\beta^\star}^{c_{s:t-1}}(\tilde{b}^-) \right\|_1 \leq C_f \rho^{t-s} \|b^- - \tilde{b}^-\|_1.$$

Assumption 3 (Trajectory-level identifiability and bounded likelihood ratios)**.** *The parameter space Ω is finite, $\beta^\star \in \Omega$, and the prior assigns positive probability to every candidate parameter:*

$$\rho_0(\beta) > 0, \qquad \forall \beta \in \Omega.$$

At time step k, let

$$\mathcal{I}_k := \sigma(H_{k-1}, X_k, U_k^r)$$

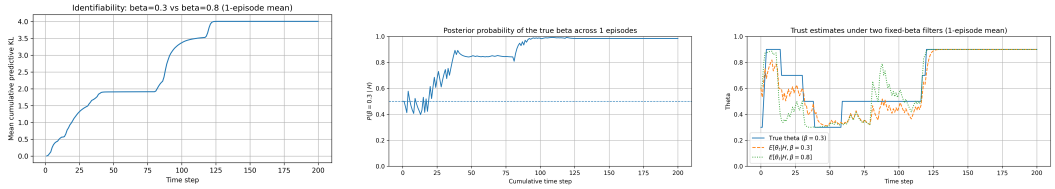
denote the information available after observing the current environmental state and selecting the recommendation, but before observing the adherence response O_k. Under a candidate parameter β, the posterior-predictive likelihood of the response is

$$P_\beta(O_k \mid \mathcal{I}_k) = \sum_{\theta \in \Theta} P(O_k \mid X_k, \theta, U_k^r) b_{k,\beta}^-(\theta),$$

where $b_{k,\beta}^-$ is the predictive trust belief obtained under β. For every incorrect parameter $\beta \neq \beta^\star$, define the conditional log-likelihood ratio

$$\Lambda_k(\beta) := \log \frac{P_{\beta^\star}(O_k \mid \mathcal{I}_k)}{P_\beta(O_k \mid \mathcal{I}_k)}.$$

We assume that the training trajectory is persistently informative enough to distinguish the true parameter from every incorrect candidate. Specifically, for every $\beta \neq \beta^\star$, there exists $\Delta_\beta > 0$ such



(a) Cumulative predictive KL divergence.

(b) Posterior probability of the true β.

(c) Trust estimates under two fixed-β filters.

Figure 6: Strong parameter identifiability.

that

$$\sum_{k=1}^{T} \mathbb{E}_{\beta^\star}[\Lambda_k(\beta) \mid \mathcal{I}_k] = \sum_{k=1}^{T} D_{\mathrm{KL}}(P_{\beta^\star}(\cdot \mid \mathcal{I}_k) \,\|\, P_\beta(\cdot \mid \mathcal{I}_k)) \geq T\Delta_\beta,$$

almost surely under the true model. Since Ω is finite, this implies

$$\Delta_{\min} := \min_{\beta \neq \beta^\star} \Delta_\beta > 0.$$

Thus, the cumulative conditional KL divergence between the true parameter and every incorrect candidate grows at least linearly with the length of the training trajectory.

In addition, the conditional log-likelihood ratios are uniformly bounded: there exists $B < \infty$ such that

$$|\Lambda_k(\beta)| \leq B$$

almost surely for every k and every $\beta \neq \beta^\star$.

C PROOFS

C.1 LEMMA

Contraction of the adherence-aware Bellman operator.

Let $\mathcal{Q}$ denote the space of bounded real-valued functions on $\mathcal{Z} \times \mathcal{U}$ equipped with the supremum norm $\|\cdot\|_\infty$. For any $q_1, q_2 \in \mathcal{Q}$, the adherence-aware Bellman optimality operator J defined in equation 4 satisfies

$$\|Jq_1 - Jq_2\|_\infty \leq \gamma \|q_1 - q_2\|_\infty.$$

Hence, J is a γ-contraction. Consequently, J admits a unique fixed point $Q^\star$, and for any initial bounded function q_0,

$$\lim_{k \to \infty} \|J^k q_0 - Q^\star\|_\infty = 0.$$

Proof. Let

$$q_1, q_2 : \mathcal{X} \times \mathcal{B} \times \mathcal{U} \to \mathbb{R}$$

be two bounded executed-action value functions.

For a candidate next recommendation u^r_{t+1}, define the corresponding continuation value as

$$\mathcal{C}_q\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) := \hat{p}_{t+1} q\big(x_{t+1}, b^{\mathrm{acc}}_{t+1}(u^r_{t+1}), u^r_{t+1}\big) + (1-\hat{p}_{t+1}) q\Big(x_{t+1}, b^{\mathrm{rej}}_{t+1}(u^r_{t+1}), g^b(x_{t+1})\Big),$$

where b^-_{t+1} is the predictive joint belief over (θ_{t+1}, β), and b^{acc}_{t+1} and b^{rej}_{t+1} are the posterior joint beliefs following acceptance and rejection, respectively.

The adherence-aware Bellman operator can then be written as

$$(Jq)(x_t, b_t, u_t) = R(x_t, u_t) + \gamma \mathbb{E}_{x_{t+1} \sim P_X(\cdot \mid x_t, u_t)} \left[\max_{u^r_{t+1} \in \mathcal{U}} \mathcal{C}_q\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) \right].$$

Therefore,

$$
\begin{aligned}
\|Jq_1 - Jq_2\|_\infty &= \sup_{x_t,b_t,u_t} |(Jq_1)(x_t,b_t,u_t) - (Jq_2)(x_t,b_t,u_t)| \\
&= \gamma \sup_{x_t,b_t,u_t} \Bigg| \mathbb{E}_{x_{t+1}\sim P_X(\cdot|x_t,u_t)} \Bigg[\max_{u^r_{t+1}\in\mathcal{U}} \mathcal{C}_{q_1}\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) \\
&\qquad\qquad - \max_{u^r_{t+1}\in\mathcal{U}} \mathcal{C}_{q_2}\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) \Bigg] \Bigg| \\
&\le \gamma \sup_{x_t,b_t,u_t} \mathbb{E}_{x_{t+1}\sim P_X(\cdot|x_t,u_t)} \Bigg[\max_{u^r_{t+1}\in\mathcal{U}} \Big| \mathcal{C}_{q_1}\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) \\
&\qquad\qquad - \mathcal{C}_{q_2}\left(x_{t+1}, b^-_{t+1}, u^r_{t+1}\right) \Big| \Bigg] \\
&\le \gamma \sup_{x_t,b_t,u_t} \mathbb{E}_{x_{t+1}\sim P_X(\cdot|x_t,u_t)} \Bigg[\max_{u^r_{t+1}\in\mathcal{U}} \Big\{ \hat{p}_{t+1} \Big| q_1\big(x_{t+1}, b^{\mathrm{acc}}_{t+1}(u^r_{t+1}), u^r_{t+1}\big) \\
&\qquad\qquad - q_2\big(x_{t+1}, b^{\mathrm{acc}}_{t+1}(u^r_{t+1}), u^r_{t+1}\big) \Big| \\
&\qquad + (1-\hat{p}_{t+1}) \Big| q_1\Big(x_{t+1}, b^{\mathrm{rej}}_{t+1}(u^r_{t+1}), g^b(x_{t+1})\Big) \\
&\qquad\qquad - q_2\Big(x_{t+1}, b^{\mathrm{rej}}_{t+1}(u^r_{t+1}), g^b(x_{t+1})\Big) \Big| \Big\} \Bigg] \\
&\le \gamma \sup_{x_t,b_t,u_t} \mathbb{E}_{x_{t+1}\sim P_X(\cdot|x_t,u_t)} \Bigg[\max_{u^r_{t+1}\in\mathcal{U}} \Big\{ \hat{p}_{t+1}\|q_1 - q_2\|_\infty + (1-\hat{p}_{t+1})\|q_1 - q_2\|_\infty \Big\} \Bigg] \\
&= \gamma \|q_1 - q_2\|_\infty. \qquad (12)
\end{aligned}
$$

Hence,

$$\|Jq_1 - Jq_2\|_\infty \le \gamma \|q_1 - q_2\|_\infty.$$

Since $\gamma \in [0,1)$, J is a γ-contraction under the supremum norm. Therefore, by the Banach fixed-point theorem, J admits a unique fixed point $Q^\star$, and repeated exact Bellman iteration converges to $Q^\star$. □

Lemma 2 (Greedy consistency)**.** *Let $\mathcal{J}$ be the adherence-aware Bellman optimality operator and $\mathcal{J}^\pi$ its evaluation operator under a recommendation policy π. For any bounded executed-action value function q, let π_q be greedy with respect to Ψ_q. Then*

$$\mathcal{J}^{\pi_q} q = \mathcal{J} q \ge \mathcal{J}^\pi q$$

pointwise for every recommendation policy π.

Proof. For any next physical state x' and predictive belief b'^-, the definition of π_q gives

$$\Psi_q\big(x', b'^-, \pi_q(x', b'^-)\big) = \max_{u^r\in\mathcal{U}} \Psi_q(x', b'^-, u^r) \ge \Psi_q\big(x', b'^-, \pi(x', b'^-)\big).$$

Taking expectation with respect to $x' \sim P_X(\cdot \mid x, u)$, multiplying by γ, and adding the common immediate reward $R(x,u)$ yields

$$(\mathcal{J}^{\pi_q} q)(x,b,u) = (\mathcal{J} q)(x,b,u) \ge (\mathcal{J}^\pi q)(x,b,u).$$

Since the argument holds for every (x,b,u), the result follows. □

Lemma 3 (Policy-transition representation)**.** *For any recommendation policy π and bounded executed-action value functions q_1, q_2,*

$$\mathcal{J}^\pi q_1 - \mathcal{J}^\pi q_2 = \gamma \widetilde{P}^\pi (q_1 - q_2),$$

where $\widetilde{P}^\pi$ is the Markov transition operator induced by the environmental transition, recommendation policy, and human accept-or-fallback response.

Proof. For any bounded function h, define

$$(\widetilde{P}^{\pi}h)(x,b,u) := \mathbb{E}_{x'\sim P_X(\cdot|x,u)}\left[\hat{p}^{\pi}h(x',b^{\mathrm{acc},\pi},u^{r,\pi}) + (1-\hat{p}^{\pi})h(x',b^{\mathrm{rej},\pi},g^{b}(x'))\right],$$

where b'^{-} is the predictive belief generated from (x,b,u),

$$u^{r,\pi} := \pi(x',b'^{-}), \qquad \hat{p}^{\pi} := \hat{p}(x',b'^{-},u^{r,\pi}),$$

and $b^{\mathrm{acc},\pi}$ and $b^{\mathrm{rej},\pi}$ are the corresponding posterior beliefs after acceptance and rejection.

By the definition of Ψ_q,

$$(\mathcal{J}^{\pi}q)(x,b,u) = R(x,u) + \gamma(\widetilde{P}^{\pi}q)(x,b,u).$$

Therefore,

$$\mathcal{J}^{\pi}q_1 - \mathcal{J}^{\pi}q_2 = \gamma\widetilde{P}^{\pi}(q_1-q_2).$$

Moreover, $\widetilde{P}^{\pi}$ is linear because expectation and the acceptance–fallback mixture are linear in h. Since $\hat{p}^{\pi} \in [0,1]$, the two branch weights are nonnegative and sum to one, so $\widetilde{P}^{\pi}$ is a Markov operator. □

Lemma 4 (Conditional unbiasedness of the sampled Bellman target)**.** *Fix a bounded executed-action value function q, and define the sampled target*

$$Y_i(q) := R_i + \gamma \max_{u^r\in\mathcal{U}} \Psi_q(X_i', B_{i+1}^{-}, u^r),$$

where B_{i+1}^{-} is the predictive joint belief generated from (X_i,B_i,U_i). If $\mathbb{E}[R_i \mid X_i = x, B_i = b, U_i = u] = R(x,u)$ and $X_i' \sim P_X(\cdot \mid x,u)$, then

$$\mathbb{E}[Y_i(q) \mid X_i = x, B_i = b, U_i = u] = (\mathcal{J}q)(x,b,u).$$

Proof. Conditioning on $(X_i,B_i,U_i) = (x,b,u)$ fixes the predictive belief B_{i+1}^{-}. Therefore,

$$\begin{aligned}\mathbb{E}[Y_i(q) \mid x,b,u] &= R(x,u) + \gamma\mathbb{E}_{X_i'\sim P_X(\cdot|x,u)}\left[\max_{u^r\in\mathcal{U}} \Psi_q(X_i', B_{i+1}^{-}, u^r)\right] \\ &= (\mathcal{J}q)(x,b,u),\end{aligned}$$

where the last equality follows directly from the definition of the adherence-aware Bellman operator. □

C.2 Proposition 1

Let $M^{\star}$ denote the true trust dynamics model, which need not belong to the candidate model class $\mathfrak{M} := \{M_\beta : \beta \in \Omega\}$.

Define

$$V_{\max} := \frac{R_{\max}}{1-\gamma},$$

and the one-step Bellman mismatch

$$D(M^{\star}, M) := \sup_{\pi} \sup_{q:\,\|q\|_\infty \le V_{\max}} \left\|\mathcal{J}_{M^{\star}}^{\pi}q - \mathcal{J}_{M}^{\pi}q\right\|_\infty.$$

Let

$$M^{\dagger} = M_{\beta^{\dagger}} \in \arg\min_{M_\beta\in\mathfrak{M}} D(M^{\star}, M_\beta)$$

denote a pseudo-true model.

Let $\mathsf{K}_M(\cdot \mid x,b,u)$ denote the induced one-step transition kernel from the current post-response information state (x,b,u) to the next pre-response information state (x',b'^{-}) under model M, and define

$$\delta_K^{\dagger} := \sup_{x,b,u} \left\|\mathsf{K}_{M^{\star}}(\cdot \mid x,b,u) - \mathsf{K}_{M^{\dagger}}(\cdot \mid x,b,u)\right\|_{\mathrm{TV}}.$$

Assume that the reward function, environmental dynamics, observation model, and fallback policy are correctly specified, so that mis-specification arises only from the trust-transition model. Then, for any fixed recommendation policy π,

$$|G_{M^{\star}}^{\pi} - G_{M^{\dagger}}^{\pi}| \le \frac{2\gamma R_{\max}}{(1-\gamma)^2}\delta_K^{\dagger}.$$

Proof. For a fixed recommendation policy π, define the adherence-aware Bellman evaluation operator under model M by

$$(\mathcal{J}_M^\pi q)(x,b,u) := R(x,u) + \gamma \mathbb{E}_{(x',b'^-)\sim \mathsf{K}_M(\cdot|x,b,u)} \left[\Psi_q\big(x', b'^-, \pi(x',b'^-)\big)\right].$$

Here u is the executed action, while $\pi(x', b'^-)$ is the recommendation issued at the next decision epoch. The acceptance and fallback branches are therefore incorporated through the adherence-aware recommendation value Ψ_q.

For notational convenience, define

$$V_q^\pi(x', b'^-) := \Psi_q\big(x', b'^-, \pi(x', b'^-)\big).$$

Since Ψ_q is a convex combination of the acceptance and fallback branches,

$$\|V_q^\pi\|_\infty \le \|q\|_\infty.$$

Let $Q_{M^\star}^\pi$ and $Q_{M^\dagger}^\pi$ denote the fixed points of $\mathcal{J}_{M^\star}^\pi$ and $\mathcal{J}_{M^\dagger}^\pi$, respectively. Since the reward is bounded by $R_{\max}$,

$$\|Q_{M^\star}^\pi\|_\infty \le V_{\max}, \qquad \|Q_{M^\dagger}^\pi\|_\infty \le V_{\max}, \qquad V_{\max} := \frac{R_{\max}}{1-\gamma}.$$

Using the fixed-point equations,

$$\begin{aligned}\|Q_{M^\star}^\pi - Q_{M^\dagger}^\pi\|_\infty &= \|\mathcal{J}_{M^\star}^\pi Q_{M^\star}^\pi - \mathcal{J}_{M^\dagger}^\pi Q_{M^\dagger}^\pi\|_\infty \\ &\le \|\mathcal{J}_{M^\star}^\pi Q_{M^\star}^\pi - \mathcal{J}_{M^\star}^\pi Q_{M^\dagger}^\pi\|_\infty \\ &\quad + \|\mathcal{J}_{M^\star}^\pi Q_{M^\dagger}^\pi - \mathcal{J}_{M^\dagger}^\pi Q_{M^\dagger}^\pi\|_\infty.\end{aligned}$$

The Bellman evaluation operator is a γ-contraction, so

$$\|\mathcal{J}_{M^\star}^\pi Q_{M^\star}^\pi - \mathcal{J}_{M^\star}^\pi Q_{M^\dagger}^\pi\|_\infty \le \gamma \|Q_{M^\star}^\pi - Q_{M^\dagger}^\pi\|_\infty.$$

Moreover, because $\|Q_{M^\dagger}^\pi\|_\infty \le V_{\max}$, the definition of the one-step Bellman mismatch gives

$$\|\mathcal{J}_{M^\star}^\pi Q_{M^\dagger}^\pi - \mathcal{J}_{M^\dagger}^\pi Q_{M^\dagger}^\pi\|_\infty \le D(M^\star, M^\dagger).$$

Therefore,

$$\|Q_{M^\star}^\pi - Q_{M^\dagger}^\pi\|_\infty \le \gamma \|Q_{M^\star}^\pi - Q_{M^\dagger}^\pi\|_\infty + D(M^\star, M^\dagger),$$

and rearranging yields

$$\|Q_{M^\star}^\pi - Q_{M^\dagger}^\pi\|_\infty \le \tfrac{D(M^\star, M^\dagger)}{1-\gamma}.$$

We next relate the one-step Bellman mismatch to the discrepancy between the induced transition kernels. For any q satisfying $\|q\|_\infty \le V_{\max}$, the reward terms cancel because the reward function is correctly specified. Hence,

$$\begin{aligned}&|(\mathcal{J}_{M^\star}^\pi q)(x,b,u) - (\mathcal{J}_{M^\dagger}^\pi q)(x,b,u)| \\ &= \gamma \left| \int V_q^\pi(x', b'^-) \left[\mathsf{K}_{M^\star}(dx', db'^- \mid x,b,u) - \mathsf{K}_{M^\dagger}(dx', db'^- \mid x,b,u)\right] \right|.\end{aligned}$$

Using

$$\|V_q^\pi\|_\infty \le \|q\|_\infty \le V_{\max}$$

and the standard total-variation inequality

$$\left| \int f \, d(P-Q) \right| \le 2\|f\|_\infty \|P - Q\|_{\mathrm{TV}},$$

we obtain

$$\begin{aligned}&|(\mathcal{J}_{M^\star}^\pi q)(x,b,u) - (\mathcal{J}_{M^\dagger}^\pi q)(x,b,u)| \\ &\le 2\gamma V_{\max} \, \|\mathsf{K}_{M^\star}(\cdot \mid x,b,u) - \mathsf{K}_{M^\dagger}(\cdot \mid x,b,u)\|_{\mathrm{TV}}.\end{aligned}$$

Taking the supremum over (x, b, u), π, and all $\|q\|_\infty \le V_{\max}$ gives

$$D(M^\star, M^\dagger) \le 2\gamma V_{\max} \delta_K^\dagger.$$

Substituting $V_{\max} = R_{\max}/(1-\gamma)$ into the previous fixed-point perturbation bound yields

$$\|Q^\pi_{M^\star} - Q^\pi_{M^\dagger}\|_\infty \le \tfrac{2\gamma R_{\max}}{(1-\gamma)^2} \delta_K^\dagger.$$

Finally, recall that the recommendation-level value under a fixed policy π is

$$G^\pi_M(x, b^-, u^r) := \Psi_{Q^\pi_M}(x, b^-, u^r).$$

Since the observation model and fallback policy are common to $M^\star$ and $M^\dagger$, for the same (x, b^-, u^r) the adherence probability and the acceptance/rejection posterior beliefs are identical under the two models. Therefore,

$$\begin{aligned} &\left|G^\pi_{M^\star}(x, b^-, u^r) - G^\pi_{M^\dagger}(x, b^-, u^r)\right| \\ &\le \|Q^\pi_{M^\star} - Q^\pi_{M^\dagger}\|_\infty. \end{aligned}$$

Taking the supremum over (x, b^-, u^r) gives

$$|G^\pi_{M^\star} - G^\pi_{M^\dagger}| \le \frac{2\gamma R_{\max}}{(1-\gamma)^2} \delta_K^\dagger.$$

This proves the result. □

C.3 Proposition 2

Define the estimated static adherence rate and its population counterpart by

$$\hat{p}_T := \frac{1}{T}\sum_{k=1}^{T} O_k, \qquad p_T^\dagger := \frac{1}{T}\sum_{k=1}^{T} \mathbb{P}(O_k = 1 \mid I_k),$$

and let

$$\hat{M}_{\text{stat}} := M_{\text{stat}}(\hat{p}_T), \qquad M^\dagger_{\text{stat}} := M_{\text{stat}}(p_T^\dagger).$$

Let

$$\mathcal{M}_{\text{stat}} := \{M_{\text{stat}}(p) : p \in [0, 1]\}$$

denote the static-adherence model class, and $b^{-,\dagger}_{t,\text{stat}}$ denote the predictive-belief representation associated with the oracle static model $M^\dagger_{\text{stat}}$ under this embedding.

Then, for any fixed recommendation policy π, environmental state x_t, and recommendation u_t^r,

$$\begin{aligned} &\left|G^\pi_{M^\star}(x_t, b_t^{-,\star}, u_t^r) - G^\pi_{\hat{M}_{\text{stat}}}(x_t, b^{-,\dagger}_{t,\text{stat}}, u_t^r)\right| \\ &\le \underbrace{\frac{R_{\max}}{1-\gamma}\left\|b_t^{-,\star} - b^{-,\dagger}_{t,\text{stat}}\right\|_1}_{\text{belief approximation error}} + \underbrace{\frac{D(M^\star, M^\dagger_{\text{stat}})}{1-\gamma}}_{\text{model approximation error}} \\ &\quad + \underbrace{\frac{2R_{\max}}{(1-\gamma)^2}|p_T^\dagger - \hat{p}_T|}_{\text{finite-sample estimation error}}. \end{aligned}$$

Moreover,

$$\mathbb{E}|p_T^\dagger - \hat{p}_T| \le \frac{1}{2\sqrt{T}}.$$

and therefore the finite-sample estimation term vanishes at rate $O(T^{-1/2})$. In contrast, the first two approximation terms need not vanish with increasing T.

Proof. We decompose the discrepancy by inserting the oracle static belief representation $b^{-,\dagger}_{\mathrm{stat}}$ and the oracle static model $M^{\dagger}_{\mathrm{stat}}$:

$$\begin{aligned}
&\left|G^{\pi}_{M^\star}(x_t, b_t^{-,\star}, u_t^r) - G^{\pi}_{\hat{M}_{\mathrm{stat}}}(x_t, \hat{b}^{-}_{\mathrm{stat}}, u_t^r)\right| \\
&\le \left|G^{\pi}_{M^\star}(x_t, b_t^{-,\star}, u_t^r) - G^{\pi}_{M^\star}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r)\right| \\
&\quad + \left|G^{\pi}_{M^\star}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r) - G^{\pi}_{M^\dagger_{\mathrm{stat}}}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r)\right| \\
&\quad + \left|G^{\pi}_{M^\dagger_{\mathrm{stat}}}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r) - G^{\pi}_{\hat{M}_{\mathrm{stat}}}(x_t, \hat{b}^{-}_{\mathrm{stat}}, u_t^r)\right|.
\end{aligned}$$

Belief-representation error. Fix the current recommendation u_t^r and continuation recommendation policy π. As in the proof of Theorem 1, define

$$\alpha^{\pi,u_t^r}_{M^\star}(x_t, \theta) := \mathbb{E}^{\pi}_{M^\star}\left[\sum_{j=0}^{\infty}\gamma^j R_{t+j} \,\middle|\, x_t, \theta_t = \theta, u_t^r\right].$$

The recommendation-level value is linear in the initial predictive belief:

$$G^{\pi}_{M^\star}(x_t, b^-, u_t^r) = \sum_{\theta\in\Theta} b^-(\theta)\alpha^{\pi,u_t^r}_{M^\star}(x_t, \theta).$$

Since

$$\left|\alpha^{\pi,u_t^r}_{M^\star}(x_t, \theta)\right| \le \frac{R_{\max}}{1-\gamma},$$

we obtain

$$\left|G^{\pi}_{M^\star}(x_t, b_t^{-,\star}, u_t^r) - G^{\pi}_{M^\star}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r)\right| \le \frac{R_{\max}}{1-\gamma}\left\|b_t^{-,\star} - b^{-,\dagger}_{\mathrm{stat}}\right\|_1.$$

Unlike the estimation error considered in Theorem 1, this term need not vanish with T, because the oracle static representation is restricted to a fixed-adherence model and need not reproduce the time-varying true predictive belief.

Static-model error. We next hold the predictive belief fixed at $b^{-,\dagger}_{\mathrm{stat}}$ and compare the true model with the oracle static model. By the fixed-point perturbation argument used in Proposition 1,

$$\left\|Q^{\pi}_{M^\star} - Q^{\pi}_{M^\dagger_{\mathrm{stat}}}\right\|_\infty \le \frac{D(M^\star, M^\dagger_{\mathrm{stat}})}{1-\gamma}.$$

Because $M^\dagger_{\mathrm{stat}}$ is represented as a member of the same belief-model class, the observation model and fallback policy are common when the two models are evaluated at the same predictive belief. Therefore,

$$\begin{aligned}
&\left|G^{\pi}_{M^\star}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r) - G^{\pi}_{M^\dagger_{\mathrm{stat}}}(x_t, b^{-,\dagger}_{\mathrm{stat}}, u_t^r)\right| \\
&\le \left\|Q^{\pi}_{M^\star} - Q^{\pi}_{M^\dagger_{\mathrm{stat}}}\right\|_\infty \le \frac{D(M^\star, M^\dagger_{\mathrm{stat}})}{1-\gamma}.
\end{aligned}$$

To interpret the relation between the static and dynamic model classes, we temporarily view M at the level of the induced adherence model. For any fixed adherence rate $p \in [0,1]$, consider the degenerate dynamics

$$P(\theta_{t+1} = p \mid \theta_t = p, x_t, u_t) = 1.$$

Starting from $\theta_0 = p$, this construction gives

$$\theta_t = p, \qquad P(O_t = 1 \mid I_t) = p, \qquad \forall t,$$

and is therefore equivalent to the static-adherence model with rate p. Thus, at the level of the induced adherence process, constant adherence is a degenerate special case of dynamic adherence. As a result, $\mathcal{M}_{\mathrm{stat}} \subseteq \mathcal{M}_{dyn}$ and it further implies

$$D(M^\star, M^\dagger_{dyn}) = \inf_{M\in\mathcal{M}_{dyn}} D(M^\star, M) \le \inf_{M\in\mathcal{M}_{\mathrm{stat}}} D(M^\star, M) \le D(M^\star, M^\dagger_{\mathrm{stat}}).$$

Thus, even if the dynamic model class is itself misspecified, the best achievable Bellman mismatch within that class is no larger than the best achievable mismatch within the static-adherence class.

Finite-sample estimation error. It remains to compare the oracle static rate $p_T^\dagger$ with its empirical estimate $\hat{p}_T$. Consider two static-adherence models with fixed rates p and p'. Let $\mathcal{J}_p^\pi$ and $\mathcal{J}_{p'}^\pi$ denote their Bellman evaluation operators. Since the two models differ only in the weights assigned to the acceptance and fallback branches,

$$\left\|\mathcal{J}_p^\pi q - \mathcal{J}_{p'}^\pi q\right\|_\infty \le 2\gamma\|q\|_\infty |p - p'|.$$

Using the γ-contraction property and $\|Q_{p'}^\pi\|_\infty \le R_{\max}/(1-\gamma)$ gives

$$\left\|Q_p^\pi - Q_{p'}^\pi\right\|_\infty \le \frac{2\gamma R_{\max}}{(1-\gamma)^2}|p - p'|.$$

Passing to the recommendation-level value yields

$$\begin{aligned}\left\|G_p^\pi - G_{p'}^\pi\right\|_\infty &\le \left\|Q_p^\pi - Q_{p'}^\pi\right\|_\infty + \frac{2R_{\max}}{1-\gamma}|p - p'| \\ &\le \frac{2R_{\max}}{(1-\gamma)^2}|p - p'|.\end{aligned}$$

Setting $p = p_T^\dagger$ and $p' = \hat{p}_T$ therefore gives

$$\left\|G_{M_{\text{stat}}^\dagger}^\pi - G_{\hat{M}_{\text{stat}}}^\pi\right\|_\infty \le \frac{2R_{\max}}{(1-\gamma)^2}|p_T^\dagger - \hat{p}_T|.$$

Finally, writing

$$p_k^\star := \mathbb{P}(O_k = 1 \mid I_k),$$

we have

$$\hat{p}_T - p_T^\dagger = \frac{1}{T}\sum_{k=1}^{T}(O_k - p_k^\star).$$

Since

$$\mathbb{E}[O_k - p_k^\star \mid I_k] = 0,$$

the summands form bounded martingale differences with conditional variance at most $1/4$. Hence

$$\mathbb{E}|\hat{p}_T - p_T^\dagger| \le \sqrt{\mathbb{E}(\hat{p}_T - p_T^\dagger)^2} \le \frac{1}{2\sqrt{T}}.$$

Combining the three bounds proves the result. □

C.4 THEOREM 1

Let τ_T denote the historical training trajectory up to time T, after which an estimated trust-transition parameter $\hat{\beta}$ is fixed. Let $M = M_{\beta^\star}$ denote the true model and $\hat{M} = M_{\hat{\beta}}$ the estimated model.

At execution time t, let $b_t^{-,\star}$ and $\hat{b}_t^-$ denote the true and estimated predictive beliefs over the latent trust state, respectively. Define

$$u_t^{r,\star} \in \arg\max_{u^r \in \mathcal{U}} G_M(x_t, b_t^{-,\star}, u^r),$$

and

$$\hat{u}_t^r \in \arg\max_{u^r \in \mathcal{U}} G_{\hat{M}}(x_t, \hat{b}_t^-, u^r).$$

Then the expected recommendation regret under the true model satisfies

$$\mathbb{E}_{\tau_T}\left[G_M(x_t, b_t^{-,\star}, u_t^{r,\star}) - G_M(x_t, b_t^{-,\star}, \hat{u}_t^r)\right] \le O(\sigma^t) + O(e^{-\widetilde{\Delta} T}),$$

where $\sigma \in (0,1)$ is the Bayesian-filter stability coefficient and $\widetilde{\Delta} > 0$ is the effective posterior-concentration rate for the trust-transition parameter.

Proof. For a model M and predictive belief b^-, define the optimal recommendation-level value

$$G_M(x, b^-, u^r) := \Psi_{Q_M^\star}(x, b^-, u^r),$$

where $Q_M^\star$ denotes the fixed point of the adherence-aware Bellman operator under model M. Accordingly, the optimal recommendation under M is

$$u_M^{r,\star}(x, b^-) \in \arg\max_{u^r \in \mathcal{U}} G_M(x, b^-, u^r).$$

Throughout the proof, the observation model $P(o_t \mid x_t, \theta_t, u_t^r)$ and the fallback policy g^b are assumed known and fixed; the trust-transition parameter β is the only estimated human-model parameter. Accordingly, for any predictive belief b^- and recommendation u^r, the adherence probability is

$$\hat{p}(x, b^-, u^r) = \sum_{\theta \in \Theta} P(o = 1 \mid x, \theta, u^r) b^-(\theta),$$

and is not identified directly with the latent trust state θ.

Step 1: Recommendation-regret decomposition. For notational simplicity, define

$$G_M^\star(b^-, u^r) := G_M(x_t, b^-, u^r), \qquad G_{\hat{M}}^\star(b^-, u^r) := G_{\hat{M}}(x_t, b^-, u^r).$$

The physical state x_t is fully observed and therefore suppressed below.

By definition,

$$u_t^{r,\star} \in \arg\max_{u^r} G_M^\star(b_t^{-,\star}, u^r), \qquad \hat{u}_t^r \in \arg\max_{u^r} G_{\hat{M}}^\star(\hat{b}_t^-, u^r).$$

Hence,

$$\begin{aligned} & G_M^\star(b_t^{-,\star}, u_t^{r,\star}) - G_M^\star(b_t^{-,\star}, \hat{u}_t^r) \\ &= \left[G_M^\star(b_t^{-,\star}, u_t^{r,\star}) - G_{\hat{M}}^\star(\hat{b}_t^-, u_t^{r,\star})\right] \\ &\quad + \left[G_{\hat{M}}^\star(\hat{b}_t^-, u_t^{r,\star}) - G_{\hat{M}}^\star(\hat{b}_t^-, \hat{u}_t^r)\right] \\ &\quad + \left[G_{\hat{M}}^\star(\hat{b}_t^-, \hat{u}_t^r) - G_M^\star(b_t^{-,\star}, \hat{u}_t^r)\right]. \end{aligned} \tag{13}$$

Since $\hat{u}_t^r$ maximizes $G_{\hat{M}}^\star(\hat{b}_t^-, \cdot)$, the middle term is non-positive. Therefore,

$$G_M^\star(b_t^{-,\star}, u_t^{r,\star}) - G_M^\star(b_t^{-,\star}, \hat{u}_t^r) \le 2 \sup_{u^r \in \mathcal{U}} \left| G_M^\star(b_t^{-,\star}, u^r) - G_{\hat{M}}^\star(\hat{b}_t^-, u^r) \right|.$$

For every fixed recommendation u^r,

$$\begin{aligned}&\left|G^\star_M(b_t^{-,\star}, u^r) - G^\star_{\hat{M}}(\hat{b}_t^-, u^r)\right| \\ &\leq \left|G^\star_M(b_t^{-,\star}, u^r) - G^\star_M(\hat{b}_t^-, u^r)\right| + \left|G^\star_M(\hat{b}_t^-, u^r) - G^\star_{\hat{M}}(\hat{b}_t^-, u^r)\right|. \end{aligned} \tag{14}$$

The two terms respectively capture predictive-belief mismatch and trust-model mismatch.

For the predictive-belief mismatch term, fix a current recommendation u^r and a continuation recommendation policy π, viewed as a mapping from observable interaction histories to future recommendations. Define

$$\alpha_M^{\pi,u^r}(x,\theta) := \mathbb{E}_M^\pi\left[\sum_{k=0}^\infty \gamma^k R_{t+k} \,\middle|\, x_t = x, \theta_t = \theta, \text{ current recommendation } u^r\right].$$

For fixed (π, u^r), the expected return is linear in the initial predictive belief:

$$G_M^\pi(x, b^-, u^r) = \sum_{\theta\in\Theta} b^-(\theta)\alpha_M^{\pi,u^r}(x,\theta).$$

Since $|\alpha_M^{\pi,u^r}(x,\theta)| \leq R_{\max}/(1-\gamma)$,

$$\left|G_M^\pi(x, b_1^-, u^r) - G_M^\pi(x, b_2^-, u^r)\right| \leq \frac{R_{\max}}{1-\gamma}\|b_1^- - b_2^-\|_1.$$

Taking the supremum over continuation policies preserves this bound, so

$$\left|G_M^\star(x, b_1^-, u^r) - G_M^\star(x, b_2^-, u^r)\right| \leq \frac{R_{\max}}{1-\gamma}\|b_1^- - b_2^-\|_1.$$

For the model-mismatch term in equation 14, fix the same predictive belief b^- and current recommendation u^r under $M = M_{\beta^\star}$ and $\widehat{M} = M_{\hat{\beta}}$.

Define $\epsilon_\beta := L_\beta\|\beta^\star - \hat{\beta}\|_1$. By Assumption 1,

$$\sup_{\theta,x,u}\left\|P_{\beta^\star}(\cdot \mid \theta, x, u) - P_{\hat{\beta}}(\cdot \mid \theta, x, u)\right\|_1 \leq \epsilon_\beta.$$

For any fixed continuation policy π, a standard one-step simulation argument on the latent process therefore gives, for any fixed continuation policy π,

$$\sup_{x,b^-,u^r}\left|G_M^\pi(x, b^-, u^r) - G_{\widehat{M}}^\pi(x, b^-, u^r)\right| \leq \gamma V_{\max}\epsilon_\beta + \gamma \sup_{x,b^-,u^r}\left|G_M^\pi(x, b^-, u^r) - G_{\widehat{M}}^\pi(x, b^-, u^r)\right|,$$

where $V_{\max} = R_{\max}/(1-\gamma)$. Rearranging yields

$$\sup_{x,b^-,u^r}\left|G_M^\pi(x, b^-, u^r) - G_{\widehat{M}}^\pi(x, b^-, u^r)\right| \leq \frac{\gamma R_{\max} L_\beta}{(1-\gamma)^2}\|\beta^\star - \hat{\beta}\|_1.$$

Finally, since $G_M^\star = \sup_\pi G_M^\pi$ and $G_{\widehat{M}}^\star = \sup_\pi G_{\widehat{M}}^\pi$, we have

$$\left|G_M^\star(x, b^-, u^r) - G_{\widehat{M}}^\star(x, b^-, u^r)\right| \leq \frac{\gamma R_{\max} L_\beta}{(1-\gamma)^2}\|\beta^\star - \hat{\beta}\|_1.$$

Combining the two bounds gives

$$\mathcal{R}_t^{\text{rec}} \leq 2\left(\frac{R_{\max}}{1-\gamma}\|b_t^{-,\star} - \hat{b}_t^-\|_1 + \frac{\gamma R_{\max} L_\beta}{(1-\gamma)^2}\|\beta^\star - \hat{\beta}\|_1\right). \tag{*}$$

Step 2: Predictive-belief estimation error. Since recommendations are selected before observing the human response, the relevant information state for decision making is the predictive belief b_t^-. We therefore directly analyze the error between the true and estimated predictive beliefs.

For a predictive belief b_t^-, after recommendation u_t^r and observed response O_t, define the Bayesian correction operator

$$\mathcal{B}_t(b_t^-)(\theta) := \frac{P(O_t \mid x_t, \theta, u_t^r) b_t^-(\theta)}{\sum_{\tilde{\theta}\in\Theta} P(O_t \mid x_t, \tilde{\theta}, u_t^r) b_t^-(\tilde{\theta})}.$$

Because the observation model is assumed known and fixed, $\mathcal{B}_t$ does not depend on the trust-transition parameter β.

After the response is observed, the executed action u_t is known. The next predictive belief is obtained by propagating the corrected belief through the trust-transition model:

$$\Phi_{\beta,t}(b_t^-)(\theta') := \sum_{\theta\in\Theta} P_\beta(\theta' \mid \theta, x_t, u_t)\mathcal{B}_t(b_t^-)(\theta).$$

Hence,

$$b_{t+1}^{-,\star} = \Phi_{\beta^\star,t}(b_t^{-,\star}), \qquad \widehat{b}_{t+1}^- = \Phi_{\hat{\beta},t}(\widehat{b}_t^-).$$

For $0 \le s < t$, define the multi-step filtering operator

$$\Phi_{\beta,s:t} := \Phi_{\beta,t-1} \circ \cdots \circ \Phi_{\beta,s}.$$

Hence,

$$b_t^{-,\star} = \Phi_{\beta^\star,0:t}(b_0^{-,\star}), \qquad \hat{b}_t^- = \Phi_{\hat{\beta},0:t}(\hat{b}_0^-).$$

The predictive-belief error can be decomposed as

$$\begin{aligned} e_t^- &= \left\|\Phi_{\beta^\star,0:t}(b_0^{-,\star}) - \Phi_{\hat{\beta},0:t}(\hat{b}_0^-)\right\|_1 \\ &\le \left\|\Phi_{\beta^\star,0:t}(b_0^{-,\star}) - \Phi_{\beta^\star,0:t}(\hat{b}_0^-)\right\|_1 \\ &\quad + \left\|\Phi_{\beta^\star,0:t}(\hat{b}_0^-) - \Phi_{\hat{\beta},0:t}(\hat{b}_0^-)\right\|_1. \end{aligned}$$

By Assumption 2, the first term satisfies

$$\left\|\Phi_{\beta^\star,0:t}(b_0^{-,\star}) - \Phi_{\beta^\star,0:t}(\hat{b}_0^-)\right\|_1 \le C_f\rho^t e_0^-.$$

For the parameter-mismatch term, since the Bayesian correction operator is common to both models, Assumption 1 gives, for any predictive belief b^-,

$$\left\|\Phi_{\beta^\star,k}(b^-) - \Phi_{\hat{\beta},k}(b^-)\right\|_1 \le L_\beta\|\beta^\star - \hat{\beta}\|_1.$$

A telescoping decomposition over the t filtering steps, followed by Assumption 2 for the remaining true-filter updates, therefore gives

$$\begin{aligned} &\left\|\Phi_{\beta^\star,0:t}(\hat{b}_0^-) - \Phi_{\hat{\beta},0:t}(\hat{b}_0^-)\right\|_1 \\ &\le C_f L_\beta \sum_{k=0}^{t-1} \rho^{t-k-1}\|\beta^\star - \hat{\beta}\|_1 \\ &= C_f L_\beta \frac{1-\rho^t}{1-\rho}\|\beta^\star - \hat{\beta}\|_1. \end{aligned}$$

Consequently,

$$e_t^- \le C_f\rho^t e_0^- + C_f L_\beta \frac{1-\rho^t}{1-\rho}\|\beta^\star - \hat{\beta}\|_1.$$

Step 3: Parameter estimation and completion of the proof. We next bound the estimation error of the trust-transition parameter. Let Ω be the finite candidate parameter space, with $\beta^\star \in \Omega$, and assume that the prior satisfies $\rho_0(\beta) > 0$ for every $\beta \in \Omega$. Let $\rho_T(\beta)$ denote the posterior distribution over β after the historical training trajectory τ_T.

We use the posterior mean estimator

$$\widehat{\beta} := \sum_{\beta \in \Omega} \beta \, \rho_T(\beta).$$

Its estimation error satisfies

$$\begin{aligned} \|\widehat{\beta} - \beta^\star\|_1 &= \left\| \sum_{\beta \in \Omega} (\beta - \beta^\star) \rho_T(\beta) \right\|_1 \\ &\le \sum_{\beta \neq \beta^\star} \|\beta - \beta^\star\|_1 \rho_T(\beta) \\ &\le D_\Omega \sum_{\beta \neq \beta^\star} \rho_T(\beta), \end{aligned}$$

where

$$D_\Omega := \max_{\beta, \beta' \in \Omega} \|\beta - \beta'\|_1.$$

It therefore remains to control the posterior mass assigned to incorrect parameters. Let

$$I_k := \sigma(H_{k-1}, X_k, U_k^r)$$

denote the information available immediately before observing the human response O_k. Under a candidate parameter β, the posterior-predictive response likelihood is

$$P_\beta(O_k \mid I_k) = \sum_{\theta \in \Theta} P(O_k \mid X_k, \theta, U_k^r) b_{k,\beta}^-(\theta),$$

where $b_{k,\beta}^-$ is the predictive trust belief obtained under parameter β.

For every $\beta \neq \beta^\star$, define the conditional log-likelihood ratio

$$\Lambda_k(\beta) := \log \frac{P_{\beta^\star}(O_k \mid I_k)}{P_\beta(O_k \mid I_k)},$$

and let

$$S_T(\beta) := \sum_{k=1}^{T} \Lambda_k(\beta).$$

Bayes' rule gives the posterior odds ratio

$$\frac{\rho_T(\beta)}{\rho_T(\beta^\star)} = \frac{\rho_0(\beta)}{\rho_0(\beta^\star)} \exp\{-S_T(\beta)\}.$$

Since $\rho_T(\beta) \le 1$ and $\rho_T(\beta) \le \rho_T(\beta)/\rho_T(\beta^\star)$,

$$\rho_T(\beta) \le \min\left\{ \frac{\rho_0(\beta)}{\rho_0(\beta^\star)} e^{-S_T(\beta)}, 1 \right\}.$$

For any fixed $\lambda \in (0,1)$, using $\min\{x,1\} \le x^\lambda$ for $x \ge 0$ gives

$$\rho_T(\beta) \le \left(\frac{\rho_0(\beta)}{\rho_0(\beta^\star)} \right)^\lambda \exp\{-\lambda S_T(\beta)\}.$$

Let

$$\mu_k(\beta) := \mathbb{E}_{\beta^\star}[\Lambda_k(\beta) \mid I_k].$$

Under the trajectory-level identifiability condition,

$$\sum_{k=1}^{T} \mu_k(\beta) \geq T\Delta_\beta$$

almost surely for every $\beta \neq \beta^\star$, where $\Delta_\beta > 0$.

Assume additionally that the conditional log-likelihood ratios are uniformly bounded,

$$|\Lambda_k(\beta)| \leq B$$

for all k and $\beta \neq \beta^\star$. By the conditional Hoeffding bound,

$$\mathbb{E}_{\beta^\star}\left[\exp\left\{-\lambda\big(\Lambda_k(\beta) - \mu_k(\beta)\big)\right\} \,\middle|\, I_k\right] \leq \exp\left\{\frac{\lambda^2 B^2}{2}\right\}.$$

Applying backward conditioning over $k = T, T-1, \ldots, 1$ yields

$$\mathbb{E}_{\tau_T}\left[\exp\left\{-\lambda \sum_{k=1}^{T}\big(\Lambda_k(\beta) - \mu_k(\beta)\big)\right\}\right] \leq \exp\left\{\frac{T\lambda^2 B^2}{2}\right\}.$$

Therefore,

$$\mathbb{E}_{\tau_T}\left[e^{-\lambda S_T(\beta)}\right] \leq \exp\left\{-T\lambda\Delta_\beta + \frac{T\lambda^2 B^2}{2}\right\}.$$

Since Ω is finite, let

$$\Delta_{\min} := \min_{\beta \neq \beta^\star} \Delta_\beta > 0.$$

Choose $\lambda \in (0,1)$ sufficiently small such that

$$\widetilde{\Delta} := \lambda\Delta_{\min} - \frac{\lambda^2 B^2}{2} > 0.$$

Then, for every $\beta \neq \beta^\star$,

$$\mathbb{E}_{\tau_T}[\rho_T(\beta)] \leq \left(\frac{\rho_0(\beta)}{\rho_0(\beta^\star)}\right)^{\lambda} e^{-\widetilde{\Delta}T}.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_{\tau_T}\left[\|\widehat{\beta} - \beta^\star\|_1\right] &\leq D_\Omega \sum_{\beta \neq \beta^\star} \mathbb{E}_{\tau_T}[\rho_T(\beta)] \\
&\leq D_\Omega \sum_{\beta \neq \beta^\star} \left(\frac{\rho_0(\beta)}{\rho_0(\beta^\star)}\right)^{\lambda} e^{-\widetilde{\Delta}T} \\
&= O(e^{-\widetilde{\Delta}T}).
\end{aligned}$$

We now combine this parameter-estimation bound with the predictive-belief bound from Step 2. Recall that

$$e_t^- \leq C_f \rho^t e_0^- + \frac{C_f L_\beta (1-\rho^t)}{1-\rho}\|\beta^\star - \hat{\beta}\|_1.$$

Taking expectation over the training trajectory τ_T gives

$$\mathbb{E}_{\tau_T}[e_t^-] \leq C_f \rho^t e_0^- + \frac{C_f L_\beta (1-\rho^t)}{1-\rho}\mathbb{E}_{\tau_T}\left[\|\beta^\star - \hat{\beta}\|_1\right] = O(\rho^t) + O(e^{-\widetilde{\Delta}T}).$$

Finally, recall from Step 1 that

$$R_t^{\mathrm{rec}} \leq 2\left(\frac{R_{\max}}{1-\gamma} e_t^- + \frac{\gamma R_{\max} L_\beta}{(1-\gamma)^2}\|\beta^\star - \hat{\beta}\|_1\right).$$

Therefore,

$$\mathbb{E}_{\tau_T}[R_t^{\mathrm{rec}}] \leq 2\left(\frac{R_{\max}}{1-\gamma}\mathbb{E}_{\tau_T}[e_t^-] + \frac{\gamma R_{\max} L_\beta}{(1-\gamma)^2}\mathbb{E}_{\tau_T}\left[\|\beta^\star - \hat{\beta}\|_1\right]\right) = O(\rho^t) + O(e^{-\widetilde{\Delta}T}).$$

This proves that the expected sub-optimality gap induced by belief initialization and trust-parameter estimation errors decays exponentially with the execution step t and the training length T.

□

C.5 COROLLARY 1

Let q_K denote the value-function approximation when K denotes number of iterations of Neural Fitted Q-Iteration (FQI) and n denotes the number of samples. Define the induced recommendation-level approximation

$$G_K(x, b^-, u^r) := \Psi_{q_K}(x, b^-, u^r),$$

and let

$$u_t^{r,\star} \in \arg\max_{u^r \in \mathcal{U}} G_M(x_t, b_t^{-,\star}, u^r), \qquad u_{t,K}^r \in \arg\max_{u^r \in \mathcal{U}} G_K(x_t, \hat{b}_t^-, u^r).$$

Let $\mathcal{D}$ denote the replay data generated along the training trajectory τ_T. We have

$$\mathbb{E}_{\tau_T, \mathcal{D}}\big[G_M(x_t, b_t^{-,\star}, u_t^{r,\star}) - G_M(x_t, b_t^{-,\star}, u_{t,K}^r)\big] \le O(\sigma^t) + O(e^{-\widetilde{\Delta} T}) + \mathcal{E}_{\mathrm{FQI}}(n, K), \quad (15)$$

where $\mathcal{E}_{\mathrm{FQI}}(n, K)$ denotes the statistical, approximation, and finite-iteration error of Neural FQI on the induced adherence-aware belief MDP.

Proof. For notational simplicity, suppress x_t and define

$$u^\star \in \arg\max_{u^r} G_M(b^{-,\star}, u^r), \qquad \hat{u} \in \arg\max_{u^r} G_{\hat{M}}(\hat{b}^-, u^r), \qquad u_K \in \arg\max_{u^r} G_K(\hat{b}^-, u^r).$$

Then

$$\begin{aligned} & G_M(b^{-,\star}, u^\star) - G_M(b^{-,\star}, u_K) \\ = \; & \big[G_M(b^{-,\star}, u^\star) - G_{\hat{M}}(\hat{b}^-, u^\star)\big] + \big[G_{\hat{M}}(\hat{b}^-, u^\star) - G_{\hat{M}}(\hat{b}^-, u_K)\big] \\ & + \big[G_{\hat{M}}(\hat{b}^-, u_K) - G_M(b^{-,\star}, u_K)\big]. \end{aligned}$$

Since $\hat{u}$ maximizes $G_{\hat{M}}(\hat{b}^-, \cdot)$,

$$G_M(b^{-,\star}, u^\star) - G_M(b^{-,\star}, u_K) \le 2 \sup_{u^r \in \mathcal{U}} \Big|G_M(b^{-,\star}, u^r) - G_{\hat{M}}(\hat{b}^-, u^r)\Big| + \Delta_{\mathrm{FQI}},$$

where

$$\Delta_{\mathrm{FQI}} := G_{\hat{M}}(\hat{b}^-, \hat{u}) - G_{\hat{M}}(\hat{b}^-, u_K).$$

The first term is exactly the estimation-induced discrepancy controlled in the proof of Theorem 1; therefore,

$$\mathbb{E}_{\tau_T}\left[2 \sup_{u^r} \Big|G_M(b^{-,\star}, u^r) - G_{\hat{M}}(\hat{b}^-, u^r)\Big|\right] \le O(\sigma^t) + O(e^{-\widetilde{\Delta} T}).$$

It remains to control Δ_{FQI}. Since the replay data $\mathcal{D}$ are generated from the same training process as τ_T, by the law of total expectation,

$$\mathbb{E}_{\tau_T, \mathcal{D}}[\Delta_{\mathrm{FQI}}] = \mathbb{E}_{\tau_T}\left[\mathbb{E}_{\mathcal{D}|\tau_T}[\Delta_{\mathrm{FQI}}]\right].$$

Let $\varepsilon_{\mathrm{FQI}}(n, K; \tau_T)$ denote the Neural-FQI iterate-error bound conditional on the realized training trajectory. Since Lemma 2 to Lemma **??** establish that the adherence-aware Bellman operator satisfies the properties required by the FQI error-propagation argument, our adherence-aware formulation introduces no additional instability. Therefore, Neural FQI remains a valid theoretical surrogate for analyzing the DQN implementation in our setting.

$$\mathbb{E}_{\mathcal{D}|\tau_T}[\Delta_{\mathrm{FQI}}] \le \mathcal{E}_{\mathrm{FQI}}(n, K).$$

Therefore,

$$\mathbb{E}_{\tau_T, \mathcal{D}}\big[G_M(x_t, b_t^{-,\star}, u_t^{r,\star}) - G_M(x_t, b_t^{-,\star}, u_{t,K}^r)\big] \le O(\sigma^t) + O(e^{-\widetilde{\Delta} T}) + \mathcal{E}_{\mathrm{FQI}}(n, K).$$

□