Managing Human Agents with Compliance-Aware Reinforcement Learning

BRYCE MCLAUGHLIN, University of Pennsylvania, USA
WICHINPONG PARK SINCHAISRI, Haas School of Business, University of California, Berkeley, USA

Companies often use algorithms to direct the actions of human agents who may subsequently ignore the algorithm. For instance, ride-sharing apps may suggest rides that drivers reject while inventory management software may suggest order quantities that managers overwrite. Thus, designing algorithms that optimize the actions taken by human agents requires careful examination of when agents comply with the algorithm and when they ignore it. To address this issue we propose Compliance-Aware Reinforcement Learning (CA-RL) which separately models the human agents' compliance to the algorithm and the environment with which the human agents engage. When the algorithm receives a superset of the agents' observations, compliance effectively shrinks the policy space an algorithm could implement directly. This observation allows us to (i) quantify the performance inefficiencies introduced by the human agents, (ii) easily integrate the structure of CA-RL to optimism-based reinforcement learning algorithms, and (iii) quantify the value of the CA-RL model compared to a baseline in which the algorithm engages directly with an environment that includes the human agents' actions as observations. Finally, we apply the principles behind CA-RL to experimental data where a recommendation algorithm directed participants through an EV-charging game to suggest potential improvements to the algorithm design.

Manuscript submitted for review to the 27th ACM Conference on Economics & Computation (EC'26).

1 Introduction

Algorithmic recommendation systems have become central to operational decision-making across diverse domains. Ride-sharing platforms suggest rides that drivers may accept or reject [Chen et al., 2019], inventory management systems propose order quantities that managers frequently override [Kremer et al., 2011], navigation applications provide routing guidance that drivers selectively follow [Kerkman et al., 2012], and clinical decision-support systems offer diagnostic recommendations that physicians integrate with their professional judgment [Lebovitz et al., 2022]. A unifying feature of these systems is that algorithms provide *recommendations* while humans retain decision authority. The realized action, and thus system performance, depends on whether humans comply with algorithmic advice.

This creates a fundamental design challenge distinct from classical optimization or machine learning. In standard sequential decision-making frameworks, the algorithm directly controls actions and learns optimal policies through interaction with the environment [Puterman, 2014, Sutton et al., 1998]. Under imperfect compliance, however, the algorithm chooses only *what to recommend*, while humans autonomously choose *what to do*, potentially deviating from prescribed behavior based on their own information, preferences, or beliefs about algorithm quality. The algorithm designer must therefore solve a principal-agent problem embedded within a Markov decision process: how should recommendations be designed to maximize performance when compliance is imperfect, endogenous to recommendation characteristics, and may evolve over time as humans learn?

Extensive empirical evidence demonstrates that compliance is not merely noise; it reflects systematic human responses to recommendation features that can be predicted and influenced through design. Humans comply more readily with recommendations that align with their prior beliefs [Prahl and Van Swol, 2017], include explanations matching their decision-making processes [Bansal et al., 2019b], and come from systems with track records of accuracy [Dietvorst et al., 2015]. Conversely, they resist recommendations that conflict with salient private information, require costly immediate actions, or come from systems that have erred previously, even when current advice is optimal [Burton et al., 2020]. Understanding these compliance patterns is essential for effective system design.

This behavioral richness has important implications. First, optimizing recommendations requires explicitly modeling *two* interconnected processes: how recommendations map to human actions (compliance function ψ), and how actions map to outcomes (environment dynamics ρ). Treating these as a single composite process, the standard approach in RL applied to human-in-the-loop systems, obscures critical design levers. The algorithm can modify what it recommends and how it frames advice without changing underlying operational realities, but only if compliance and environment are separately modeled.

Second, in many operational settings, recommendations affect not isolated individuals but populations whose decisions create externalities. When navigation algorithms direct many drivers onto the same alternative routes, the resulting concentration can worsen congestion rather than alleviate it [Acemoglu et al., 2018]. When inventory systems recommend aggressive ordering during demand spikes, correlated responses amplify rather than dampen supply chain volatility [Kremer et al., 2011]. These system-level effects mean that locally optimal recommendations can be globally suboptimal, with the degree of suboptimality depending critically on compliance rates and their distribution across the population.

Third, the relationship between immediate performance and long-run capability creates intertemporal trade-offs. Precise action-level prescriptions may achieve high compliance when provided but leave humans unable to make good autonomous decisions when algorithms are unavailable.

Strategic guidance may achieve lower immediate compliance but builds mental models that transfer to new contexts [Poulidis et al., 2025]. Which approach dominates depends on the time horizon, environment stability, and whether humans will eventually operate without algorithmic support.

We develop *Compliance-Aware Reinforcement Learning* (CA-RL), a framework that formalizes recommendation design as choosing a policy over an extended action space $\mathcal{A}^\dagger = \mathcal{A} \cup \{\emptyset\}$, where $\mathcal{A}$ is the environment action space and $\emptyset$ represents withholding advice. The human observes the recommendation, draws a baseline action from policy π^0, and decides whether to comply according to a compliance function ψ that depends on recommendation characteristics, baseline preference, and history. Critically, we impose an *exclusion restriction*: recommendations affect the environment only through human actions. This separability assumption, standard in econometric applications of instrumental variables [Imbens and Rubin, 2015], enables independent identification of compliance behavior and environment dynamics from observational data.

Main contributions. Our work makes three primary contributions. On the theoretical front, we generalize the classical shortfall decomposition theorem to show that total performance loss under any recommendation policy decomposes exactly into period-by-period costs of non-compliance This decomposition reveals when and where non-compliance is costly, at states where the Q-function varies widely across actions and the baseline policy performs poorly, versus benign. We then characterize when optimal recommendations differ from optimal actions, establishing conditions under which the algorithm should recommend suboptimal actions because they achieve higher compliance and pull humans away from even worse baseline behavior. Finally, we quantify the cost of imperfect control by bounding the performance gap between direct action control and indirect recommendation-based control as a function of baseline policy quality, minimum compliance probability, and the range of action values.

Methodologically, we provide an identification strategy that exploits the exclusion restriction to separately estimate baseline policy and compliance function from observational data containing partial variation in recommendation provision. This modularity enables learning compliance independently of environment dynamics, then using compliance estimates to evaluate counterfactual recommendation policies without re-estimating transition models. We introduce the concepts of *potential* (value lost at a state due to suboptimal actions) and *efficiency* (fraction of potential recovered by recommendations) to identify *pitfall states* where targeted interventions can substantially improve performance. The framework also enables integration with optimism-based RL algorithms by recognizing that compliance constraints define a convex subset of the implementable policy space.

Our approach applies to operational settings where algorithms provide recommendations but humans retain decision authority, compliance varies systematically with recommendation characteristics, and the designer observes both recommendations and actions. Examples include gig platform dispatch, inventory management support, clinical decision support, and navigation guidance. The framework is particularly timely as AI-based systems proliferate across industries and shift from prediction to prescription. As these systems increasingly tell humans what to do rather than what might happen, managing compliance becomes essential for realizing value.

2 Related Literature

Our work bridges four literatures: algorithm aversion and compliance behavior, human-AI collaboration, behavioral operations, and reinforcement learning with human feedback. We position CA-RL's contributions relative to each.

Algorithm Aversion and Compliance. Extensive evidence documents systematic human reluctance to follow algorithmic recommendations. Dietvorst et al. [2015] show that humans lose confidence

in algorithms after observing small errors (algorithm aversion), persisting even when algorithms outperform human forecasters. Burton et al. [2020] find that decision-makers consistently underweight algorithmic advice relative to human advice across domains. Prahl and Van Swol [2017] demonstrate that compliance depends on perceived algorithm transparency and past performance. Bonaccio and Dalal [2006] establish that humans selectively integrate recommendations based on advisor expertise, prior belief consistency, and task characteristics. Green and Chen [2019] identify incompatibility between algorithmic recommendations and human operational constraints as a key compliance barrier. Sun et al. [2024] show that peer acceptance of AI recommendations can dominate algorithm accuracy in shaping patient adoption, suggesting compliance reflects social learning about trustworthiness, not just individual algorithm assessment. Our framework differs by modeling compliance as endogenous probabilistic choice varying with recommendation characteristics, formalizing consequences through shortfall decomposition, and studying dynamic effects where human capability evolves based on advice type received.

Human-AI Collaboration. A growing literature examines how AI augments human judgment. Bansal et al. [2019b] find that AI model updates degrade human performance due to incompatibility between learned mental models and new AI behavior, revealing tension between model accuracy and human-AI compatibility. Lai and Tan [2019] demonstrate that explanations help humans calibrate reliance by identifying when to trust versus override AI. Zhang et al. [2020] show AI confidence scores reduce over- and under-reliance when well-calibrated. Bansal et al. [2019a] argue effective collaboration requires building mental models of AI capabilities, not just case-by-case explanations. Kamar [2016] proposes hybrid intelligence frameworks where AI and humans play complementary roles. We formalize the advice granularity choice in sequential settings: precise recommendations optimize immediate collaboration but prevent mental model formation; broad recommendations sacrifice short-term performance for better learning and transfer. Our shortfall decomposition quantifies this trade-off.

Behavioral Operations. Operations management extensively studies decision-support systems' effects on human behavior. Sun et al. [2022] show in warehouse operations that deviations from algorithmic task assignments are predictable from worker experience and task characteristics, with experienced workers sometimes overriding poor recommendations based on tacit knowledge. This motivates predicting compliance to improve algorithm design. Kremer et al. [2011] find inventory managers systematically deviate toward prior beliefs even when algorithms are demonstrably more accurate, with deviations decreasing as trust builds. Recent RL-based approaches explicitly account for human behavior: Bastani et al. [2026] apply online learning to nurse triage recommendations, treating discretion as implementation constraint rather than noise. Lu et al. [2023b] develop general frameworks for RL with human interaction, emphasizing that learning must account for how recommendations affect behavior immediately and over time. Poulidis et al. [2025] demonstrate experimentally that strategic guidance improves performance more than action prescriptions. McLaughlin and Spiess [2022] formalize recommendation-dependent preferences; McLaughlin and Spiess [2024] characterize human-AI complementarity. We formalize compliance within the MDP framework, characterize optimal recommendation policies accounting for immediate compliance and long-run capability, and provide methodology using optimal Q-functions to measure advice design value.

Reinforcement Learning with Human Feedback. Standard RL assumes direct action control [Puterman, 2014, Sutton et al., 1998]. Extensions include inverse RL, which infers preferences from behavior [Hadfield-Menell et al., 2016], and RLHF, where humans provide comparative judgments [Christiano et al., 2017]. Kleinberg et al. [2018] show predicting human decisions requires

learning systematic biases and heuristics. Chen et al. [2019] learn driver acceptance models for ride-sharing dispatch. Donohue et al. [2020] emphasize incorporating behavioral considerations into algorithm design from the outset. We differ by modeling compliance as probabilistic choice within the MDP, assuming the algorithm can compute optimal policy π^* but faces a mechanism design problem: how to communicate knowledge maximizing expected performance given compliance constraints.

Information Design and Equilibrium Effects. At scale, recommendations create equilibrium effects through herding and congestion. Banerjee [1992], Bikhchandani et al. [1992] show information provision can generate cascades where rational agents follow predecessors, ignoring private signals. Acemoglu et al. [2018] demonstrate that more accurate traffic information can worsen congestion through equilibrium responses (informational Braess's paradox). Transportation research provides compliance evidence: Kerkman et al. [2012] find 40–70% compliance rates depending on advice characteristics; Chen and Jovanis [2003] show compliance evolves as drivers learn system accuracy. Jiang et al. [2024] develop compliance-aware guidance improving system-wide travel time. Yun et al. [2024] show non-compliance can improve welfare by preventing herding. Johnson et al. [2017] document externalities from routing algorithms on local communities. We contribute behavioral micro-foundations, studying individual-level compliance to characterize the compliance function $\psi(\cdot)$ that aggregates to system outcomes. Future work can embed our model into game-theoretic frameworks.

Our framework synthesizes these insights into a unified approach. From algorithm aversion, we take seriously that compliance is imperfect and varies systematically. From human-AI collaboration, we adopt that advice granularity affects immediate performance and long-run capability. From behavioral operations, we inherit methodology for measuring deviations and predicting compliance. From RL theory, we leverage shortfall decomposition to characterize performance losses. The result is both descriptive (formalizing how recommendations map to actions via compliance) and prescriptive (identifying optimal policies trading off compliance and learning).

3 Modeling

We extend the canonical reinforcement learning setup to include a human agent that acts as an intermediary between the algorithmic agent and the environment. This human agent may receive action-suggestions (i.e. recommendations) from the algorithmic agent before they autonomously chooses an action. The human agent's behavior may depend on their history of interactions with both the environment and the algorithmic agent. The algorithmic agent's goal is to learn a recommendation policy which maximizes some reward function, whose value is determined by the human agent's interactions with the environment.

3.1 Reinforcement Learning Preliminaries

To begin, we recount the canonical reinforcement learning setup in which an agent interacts with an environment with the goal of maximizing some reward function. In response to each action the agent takes, the environment produces observations. Together this series of actions and observations form a history, which the agent may utilize when choosing their action and may impact the observations the environment generates. The agent earns rewards in response to each observation based on both the action they took and the current history. The agents goal is to maximize their cumulative expected reward over the time horizon. Our formulation generally follows that of [Lu et al., 2023a].

We consider an agent Γ that interacts with an environment $\mathcal{E}$ over a time horizon of length T. At each time $t \in \{0, \ldots, T-1\}$ the agent chooses an action A_t from the action set $\mathcal{A}$. In response the

environment produces an observation O_{t+1} from the observation set $\mathcal{O}$. Any sequence of actions and observations forms a history $H_t = (A_0, O_1, A_1, O_2, \ldots, A_{t-1}, O_t)$. The history set $\mathcal{H} = \bigcup_{t=0}^{T-1} (\mathcal{A} \times \mathcal{O})^t$ consists of all possible histories.

An environment $\mathcal{E} = (\mathcal{A}, \mathcal{O}, \rho)$ consists of the aforementioned action and observation sets as well as the observation distribution function $\rho : \mathcal{H} \times \mathcal{A} \to \Delta(\mathcal{O})$. For any history H_t and action A_t, ρ describes the distribution over the observation set which generates O_{t+1}. For ease of notation we write $\rho(o|H_t, A_t) = \Pr_\rho(O_{t+1} = o|H_t, A_t)$ for all $o \in \mathcal{O}$.

An agent $\Gamma = (\pi, r)$ consists of a policy $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ and a reward function $r : \mathcal{H} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}$. For any history H_t, π describes the distribution over the action set by which the agent chooses A_t. For ease of notation we write $\pi(a|H_t) = \Pr_\pi(A_t = a|H_t)$ for all $a \in \mathcal{A}$. Let Π be the space of all policies. The reward function r determines the value of each observation to the agent based on the action they took and the history.

The agent's goal is to maximize the rewards they receive over the time horizon T. To that end, we can describe the value of a policy as,

$$\overline{V}_\pi = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} r(H_t, A_t, O_{t+1}) \middle| \mathcal{E} \right].$$

The goal of reinforcement learning can be broadly summarized as designing policies with maximal value subject to initial uncertainty over the observation distribution function ρ.

We define two generalizations of a policy's value to make cross-policy comparisons easier: the value function $V_\pi : \mathcal{H} \to \mathbb{R}$ and the action-value function $Q_\pi : \mathcal{H} \times \mathcal{A} \to \mathbb{R}$. The value function describes the remaining expected value of a policy π given a history H_t while the action-value function describes the same, while additionally specifying the next action

$$V_\pi(H_t) = \mathbb{E}_\pi \left[\sum_{\tau=t}^{T-1} r(H_\tau, A_\tau, O_{\tau+1}) \middle| \mathcal{E}, H_t \right], \quad Q_\pi(H_t, A_t) = \mathbb{E}_\pi \left[\sum_{\tau=t}^{T-1} r(H_\tau, A_\tau, O_{\tau+1}) \middle| \mathcal{E}, H_t, A_t \right].$$

Together, these functions enable easy policy value comparisons using the following theorem, presented in the current form by [Sutton and Barto, 2020]

Theorem 1 (Shortfall Decomposition). *For all policies π and π',*

$$\overline{V}_\pi - \overline{V}_{\pi'} = \mathbb{E}_{\pi'} \left[\sum_{t=0}^{T-1} (V_\pi(H_t) - Q_\pi(H_t, A_t)) \middle| \mathcal{E} \right].$$

In Section 4, we show that the shortfall decomposition easily extends to the compliance-aware reinforcement learning (CA-RL) setup introduced in section 3.2. This decomposition allows for theoretical analysis of the CA-RL framework and drives our prescriptive results for potential recommendation improvements (section 5).

3.2 Compliance-Aware Reinforcement Learning

We extend the canonical reinforcement learning setup to compliance-aware reinforcement learning (CA-RL) by introducing a human agent that receives action-suggestions (i.e. recommendations) from an algorithmic agent and then chooses an action to implement in the environment. The human agent consists of a baseline policy that they implement by default and a compliance function determining the probability that the human agent accepts the algorithmic agent's recommendation. We refer to the two together as the human agent's behavior. The algorithmic agent has the option to not provide any action suggestions, which may be optimal as the human agent's behavior can depend on their past interactions with the algorithmic agent. The goal of the algorithmic agent

is to maximize the cumulative expected reward they earn over the time horizon as a result of the human agent's actions.

We consider an algorithmic agent $\Gamma^\dagger$ which interacts with a human agent Ψ which in turn interacts with an environment $\mathcal{E}$ over a time horizon of length T. The environment maintains its structure from section 3.1. At each time $t \in \{0, \dots, T-1\}$ the algorithmic agent chooses a recommendation $A_t^\dagger$ from the recommendation set $\mathcal{A}^\dagger = \mathcal{A} \cup \{\emptyset\}$, where $\mathcal{A}$ is the action set and $\emptyset$ represents not sending a recommendation. In response to the recommendation, the human agent selects an action A_t from the action set $\mathcal{A}$ which prompts the environment to produce an observation O_{t+1} from the observation set $\mathcal{O}$. Any sequence of recommendations, actions, and observations forms a recommendation-aware history $H_t^\dagger = (A_0^\dagger, A_0, O_1, \dots, A_{t-1}^\dagger, A_{t-1}, O_t)$. When we drop the $\dagger$ we refer to the history without the recommendations. The environment's observation distribution function and the algorithmic agent's reward function remain independent of the algorithmic agent's recommendations and thus only vary with H_t. An exclusion restriction implicitly applies in this setup; the recommendation can only impact the environment through the human agent's actions. The recommendation-aware history set $\mathcal{H} = \bigcup_{t=0}^{T-1} (\mathcal{A}^\dagger \times \mathcal{A} \times \mathcal{O})^t$ consists of all possible recommendation-aware histories.

A human agent $\Psi = (\pi^o, \psi)$ consists of a baseline policy $\pi^0 : \mathcal{H}^\dagger \to \Delta(\mathcal{A})$ and a compliance function $\psi : \mathcal{H}^\dagger \times \mathcal{A} \to [0,1]^{\mathcal{A}}$. For any recommendation-aware history $H_t^\dagger$, π^0 describes the distribution over the action set by which the human agent chooses their default action $A_t^0 \in \mathcal{A}$. For ease of notation we write $\pi^0(a|H_t^\dagger) = \Pr(A_t^0 = a|H_t^\dagger)$ for all $a \in \mathcal{A}$. Let Π^0 be the space of all baseline policies. Given a recommendation-aware history $H_t^\dagger$ and a baseline action A_t^0, ψ gives the probability that each recommendation will be accepted. For ease of notation we write $\psi(a|H_t^\dagger, A_t^0)$ to describe the probability a recommendation of action a is accepted for all $a \in \mathcal{A}$. Thus the human agent's action is given by,

$$A_t = \begin{cases} A_t^\dagger & \text{w.p. } \psi(A_t^\dagger|H_t^\dagger, A_t^0) \text{ if } A_t^\dagger \neq \emptyset, \\ A_t^0 & \text{otherwise.} \end{cases}$$

Implicit to this setup is a bi-directional monotonicity assumption relative to the baseline policy. Recommending $A_t^\dagger \neq \emptyset$ can only increase the probability $A_t = A_t^\dagger$ and decrease the probability any other action is taken while recommending $A_t^\dagger = \emptyset$ leaves the baseline policy unperturbed. Inherently this assumption says that the human agent determines an action on their own, but may be swayed by the algorithmic agent. However, the dependence of the human agent's behavior on the recommendation-aware history enables rich interactions between the algorithmic agent and human agent.

An algorithmic agent $\Gamma^\dagger = (\pi^\dagger, r)$ consists of a recommendation policy $\pi^\dagger : \mathcal{H}^\dagger \to \Delta(\mathcal{A}^\dagger)$ and a reward function which maintains its structure from section 3.1. For any recommendation-aware history $H_t^\dagger$, $\pi^\dagger$ describes the distribution over the recommendation set by which the algorithmic agent chooses a recommendation.[1] For ease of notation we write $\pi^\dagger(a^\dagger|H_t^\dagger) = \Pr(A_t^\dagger = a^\dagger|H_t^\dagger)$ for all recommendations $a^\dagger \in \mathcal{A}^\dagger$. Given a recommendation policy $\pi^\dagger$ and a human agent Ψ we can

[1]Note that under this setup the human agent has no private information relevant to the environment that is not available to the algorithmic agent. This prevents us from effectively modeling Human-AI complementarity which would requires an adjusted setup in which the algorithmic agent introduces a partially observable process. We view this model as a first step in toward complementary reinforcement learning which considers private human information.

define the policy of the assisted human agent as

$$\pi^{\dagger}_{\Psi}(a|H^{\dagger}_t) = \pi^0(a|H^{\dagger}_t) + \sum_{a' \in \mathcal{A}} \Big[\pi^{\dagger}(a|H^{\dagger}_t)\pi^0(a'|H^{\dagger}_t)\psi(a|H^{\dagger}_t, a') - \pi^{\dagger}(a'|H^{\dagger}_t)\pi^0(a|H^{\dagger}_t)\psi(a'|H^{\dagger}_t, a) \Big],$$

for all $a \in \mathcal{A}$. The terms within the summation captures the human agent complying to recommendations, which adjusts the distribution of the baseline policy.

The algorithmic agent's goal is to maximize the rewards they receive over the time horizon T as a result of the human agents action. To that end we can describe the value of a recommendation policy as,

$$\overline{V}^{\dagger}_{\pi^{\dagger}} = \mathbb{E}_{\pi^{\dagger}_{\Psi}} \left[\sum_{t=0}^{T-1} r(H_t, A_t, O_{t+1}) \middle| \mathcal{E} \right].$$

The goal of compliance-aware reinforcement learning can be broadly summarized as designing recommendation policies with maximal value subject to initial uncertainty over the observation distribution function ρ and/or the human agent's compliance behavior Ψ.

4 Theory

We establish three main theoretical results that characterize optimal recommendation design under imperfect compliance. The first generalizes the classical shortfall decomposition to compliance-aware settings, enabling precise measurement of performance losses from non-compliance. The second identifies when optimal recommendations differ from optimal actions, revealing the structure of compliance-performance trade-offs. The third bounds the cost of imperfect control relative to direct action selection, quantifying when compliance constraints are binding versus benign.

4.1 Shortfall Decomposition for Compliance-Aware RL

The classical shortfall decomposition theorem establishes that for any two policies π and π', the value difference can be expressed as the expected sum of period-by-period shortfalls: how much worse it is to take the action π' prescribes versus the action π prescribes at each history [Sutton et al., 1998]. This result is foundational for policy evaluation and improvement in reinforcement learning. We extend this theorem to settings where the algorithm provides recommendations rather than directly controlling actions.

Theorem 2 (Shortfall Decomposition for CA-RL). *Let $\pi^{\dagger}$ be any recommendation policy and $\Psi = (\pi^0, \psi)$ be a human agent with baseline policy π^0 and compliance function ψ. Define the induced human policy $\pi^{\dagger}_{\Psi}$ as in Section 3.2. For any policy π (including the optimal environment policy π^*), the performance gap is:*

$$V_{\pi} - V^{\dagger}_{\pi^{\dagger}} = \mathbb{E}_{\pi^{\dagger}_{\Psi}} \left[\sum_{t=0}^{T-1} \left(V_{\pi}(H_t) - Q_{\pi}(H_t, A_t) \right) \middle| \mathcal{E} \right] \tag{1}$$

where the expectation is over trajectories generated by $\pi^{\dagger}_{\Psi}$ interacting with environment $\mathcal{E}$.

The proof follows by iterating expectations and applying the Bellman equation telescopically, then recognizing that the compliance mechanism affects only the distribution over actions at each history, not the relationship between actions and values. The key insight is that the shortfall $V_{\pi}(H_t) - Q_{\pi}(H_t, A_t)$ measures the cost of taking whatever action the human actually implements

(whether from compliance or baseline behavior) versus following policy π optimally from that history forward.

This decomposition has immediate practical implications. It enables system designers to compute exactly where performance is being lost: at which states, under which recommendations, and by how much. States where the Q-function is relatively flat across actions contribute little to total shortfall even if compliance is low; deviations from optimality are cheap. States where Q-values vary widely and the baseline policy selects poor actions contribute substantially to shortfall. These are the pitfalls where improving compliance or changing what is recommended can yield significant gains.

4.2 Optimal Recommendation Policies

We now characterize the structure of optimal recommendation policies and establish when they differ from simply recommending the optimal environment action. Intuitively, when compliance is endogenous to recommendation characteristics, the algorithm faces a trade-off: recommending the environment-optimal action $\pi^*(H_t)$ maximizes value *if the human complies*, but compliance probability may be low. Recommending a slightly suboptimal action that is easier for humans to accept may achieve higher expected value by pulling behavior away from even worse baseline actions.

Theorem 3 (Structure of Optimal Recommendations). *Suppose the compliance function ψ satisfies monotonicity: for any history and baseline action, compliance probability is weakly higher for recommendations that are closer to optimal (lower Q-value). Suppose further that the baseline policy π^0 has full support (positive probability on all actions). Then the optimal recommendation policy $\pi^{\dagger *}$ satisfies:*

$$\pi^{\dagger *}(H_t^\dagger) \in \arg\max_{a^\dagger \in \mathcal{A}^\dagger} \mathbb{E}_{A_t^0 \sim \pi^0} \left[\psi(a^\dagger | H_t^\dagger, A_t^0) \cdot Q_{\pi^*}(H_t, a^\dagger) + (1 - \psi(a^\dagger | H_t^\dagger, A_t^0)) \cdot Q_{\pi^*}(H_t, A_t^0) \right] \quad (2)$$

Moreover, the optimal recommendation $\pi^{\dagger *}(H_t^\dagger)$ differs from the optimal action $\pi^*(H_t)$ whenever there exists an alternative action $a \neq \pi^*(H_t)$ such that the compliance gain from recommending a over $\pi^*(H_t)$, weighted by the value difference between a and the baseline, exceeds the quality loss from recommending a over $\pi^*(H_t)$, weighted by the compliance probability of $\pi^*(H_t)$.

The proof proceeds by backward induction, recognizing that at each history the algorithm chooses the recommendation that maximizes expected continuation value accounting for compliance probabilities. When the human complies, the action taken is the recommendation; when they don't, it's drawn from the baseline. Optimal recommendations therefore balance action quality against compliance likelihood, sometimes sacrificing the former to improve the latter.

This result challenges conventional practice in decision-support systems, which typically compute the optimal action and recommend it with explanations or confidence indicators. Our analysis shows this can be strictly suboptimal when compliance varies across recommendations. For example, if the optimal action requires a complex multi-step procedure that humans rarely follow, while a simpler near-optimal action achieves high compliance, recommending the simpler action may dominate. The key determinant is the baseline policy: when baseline actions are far from optimal, achieving compliance with any reasonable recommendation is valuable, even if that recommendation isn't environment-optimal.

The monotonicity assumption (that compliance probability increases for recommendations closer to optimal) is natural when humans have some sense of action quality, either through experience, prior beliefs, or partial observability of outcomes. When this assumption fails (compliance is

orthogonal or negatively related to optimality), the problem becomes harder and may require active experimentation to learn the compliance function's structure.

4.3 Cost of Imperfect Control

Our third result quantifies how much performance is lost by providing recommendations rather than directly controlling actions. This bound helps algorithm designers understand when compliance constraints are binding (large performance loss, invest in increasing compliance) versus slack (small loss, optimize recommendations accepting indirect control).

Theorem 4 (Value of Perfect Control). *Let π^* be the optimal policy under direct control and $\pi^{\dagger *}$ be the optimal recommendation policy. Define $\underline{\psi}$ as the minimum compliance probability over all states and recommendations, $\overline{\Delta}$ as the maximum range of the Q-function at any state, and $V_{gap} = V_{\pi^*} - V_{\pi^0}$ as the baseline policy's suboptimality. Then:*

$$V_{\pi^*} - V^{\dagger}_{\pi^{\dagger *}} \leq (1 - \underline{\psi}) \cdot V_{gap} + T \cdot \underline{\psi} \cdot \overline{\Delta} \cdot (1 - \underline{\psi}/2) \tag{3}$$

The bound decomposes the cost of imperfect control into two terms. The first, $(1 - \underline{\psi})V_{\text{gap}}$, captures losses from baseline actions when compliance fails: if humans never comply ($\underline{\psi} = 0$), the system achieves only baseline performance, losing the full gap V_{gap}. As minimum compliance increases, this term shrinks. The second term, $T\underline{\psi}\overline{\Delta}(1 - \underline{\psi}/2)$, captures discretization losses even when humans do comply: finite action spaces mean that even the best recommendation may differ from the true optimal by up to $\overline{\Delta}$ in value. This term is largest at intermediate compliance levels and decreases when the action space is fine-grained (small $\overline{\Delta}$).

The bound provides actionable guidance. When the baseline is poor (large V_{gap}) and compliance is low (small $\underline{\psi}$), the first term dominates and substantial value can be recovered by increasing compliance through interface redesign, incentives, or training. When the baseline is reasonable or compliance is already high, the first term is small and efforts should focus on refining the action space or improving environment modeling rather than increasing compliance further. When the Q-function has small range (actions are nearly equivalent), even low compliance incurs little cost because deviations don't matter much.

4.4 Implementable Policy Space and Algorithmic Integration

The structure of CA-RL enables integration with optimism-based reinforcement learning algorithms such as UCB-VI or posterior sampling. The key observation is that compliance constraints define the set of policies the algorithm can implement via recommendations. When the algorithm has full observability (sees all decision-relevant state), this implementable set forms a convex subset of the full policy space.

Specifically, any implementable policy π can be written as $\pi(a|h) = \pi^0(a|h) + \sum_{a'} \Delta\pi(a, a'|h)\psi(a|h, a')$ where $\Delta\pi$ represents how the algorithm attempts to shift the baseline distribution through recommendations, weighted by compliance probabilities. This is a convex polytope defined by the compliance function and baseline policy. Standard optimistic algorithms can be adapted by restricting to this polytope, and regret bounds transfer with modified constants reflecting the restricted policy class. The practical implication is that CA-RL can leverage existing RL infrastructure with modifications that account for compliance, rather than requiring entirely new algorithmic approaches.

5 Identifying Potential Recommendation Improvements

Translating the theoretical framework into operational improvements requires estimating compliance behavior and baseline policies from data, then using these estimates to evaluate counterfactual recommendation policies. We develop an identification strategy that exploits the exclusion restriction (recommendations affect outcomes only through human actions) to separately estimate compliance and environment dynamics. This modularity enables targeted interventions without requiring comprehensive system re-estimation.

5.1 Data Availability and Identification Strategy

We assume access to observational data $\mathcal{D}$ containing trajectories under some historical recommendation policy $\pi^{\dagger}_{\text{old}}$. Each trajectory records recommendation-aware histories: sequences of states, recommendations given (or absence thereof), actions taken, and outcomes realized. Critically, we assume the data contains partial experimental variation (some trajectories where recommendations were provided and others where they were withheld, either through A/B testing, gradual rollout, or simply system downtime).

We further assume knowledge of the environment's transition dynamics ρ and the ability to compute optimal value functions $V_{\pi^*}(h)$ and $Q_{\pi^*}(h, a)$ for all histories and actions. This assumption is reasonable in many operational settings where environment dynamics reflect physical or economic fundamentals (travel times, demand distributions, service rates) that are estimable from historical data or known from domain expertise. The behavioral component (how humans respond to recommendations) is what requires careful identification given its endogeneity.

The identification strategy proceeds in three steps. First, from trajectories without recommendations, we observe actions drawn directly from the baseline policy π^0. Standard maximum likelihood estimation recovers $\hat{\pi}^0(a|h)$ as the empirical frequency of action a at history h among no-recommendation observations. Second, from trajectories with recommendations, we observe the mixture policy $\pi^{\dagger}_{\Psi}$: sometimes humans comply and take the recommended action, sometimes they ignore advice and revert to baseline. The mixture model gives $\Pr(A = a|H = h, A^{\dagger} = a^{\dagger})$ as a function of baseline probabilities, compliance probabilities, and the recommendation. Under parametric assumptions on the compliance function (for example, logistic regression on features of the history, recommendation, and baseline action), we estimate parameters via maximum likelihood.

Third, we use the estimated baseline and compliance functions to compute counterfactual performance under alternative recommendation policies. Given $\hat{\pi}^0$ and $\hat{\psi}$, we can simulate how any proposed policy $\pi^{\dagger}_{\text{new}}$ would perform: at each history, the new policy recommends some action, the human complies with estimated probability $\hat{\psi}$ or takes baseline action with complementary probability, and we evaluate resulting trajectories using the known environment model. This enables offline policy evaluation without deploying potentially suboptimal policies in the actual system.

5.2 Targeting High-Impact States

Not all states require active recommendation intervention. Some states are routine: humans naturally select near-optimal actions from their baseline policy, and providing recommendations adds little value while potentially creating cognitive load or eroding trust through perceived micromanagement. Other states are *pitfalls*: humans systematically make poor choices, and effective recommendations could substantially improve performance. Distinguishing these cases enables efficient resource allocation.

We formalize this distinction through two metrics. The *potential* of a policy at history h is defined as $P_{\pi}(h) = V_{\pi^*}(h) - \mathbb{E}_{A\sim\pi(\cdot|h)}[Q_{\pi^*}(h, A)]$, measuring how much value the policy loses at

h relative to optimal. High potential indicates significant room for improvement. The *efficiency* of a recommendation policy at history h is $E_{\pi^\dagger}(h) = 1 - P_{\pi^\dagger_\Psi}(h)/P_{\pi^0}(h)$, measuring what fraction of the baseline-to-optimal gap the recommendation policy recovers. High efficiency means recommendations are working well; low efficiency signals opportunities for better design.

States with high potential and low efficiency are pitfalls. At these states, humans struggle (high potential loss from baseline), and current recommendations aren't helping much (low efficiency). These are natural targets for intervention: redesigning what is recommended, how it's framed, or when it's provided. Conversely, states with low potential are naturally easy; even if efficiency is low, there's little value at stake. States with high efficiency are already working well (recommendations are successfully guiding behavior toward optimality). Resources should concentrate on pitfalls.

Using estimates $\hat{\pi}^0$ and $\hat{\psi}$, we can compute $\hat{P}_{\pi^0}(h)$ and $\hat{E}_{\pi^\dagger_{\text{old}}}(h)$ for all visited histories, identify the set of pitfall states above thresholds τ_{pot} and below τ_{eff}, and design targeted interventions. For example, at pitfalls we might provide more detailed recommendations, include explanations, offer incentives for compliance, or even recommend slightly suboptimal actions if they achieve higher compliance and pull humans away from poor baseline choices (per Theorem 3). Critically, we can also modify recommendations at *upstream* states to alter the histories humans experience, preventing them from reaching problematic decision points. At non-pitfalls, we might withhold recommendations entirely, reducing system complexity and potential advice fatigue.

5.3 Optimization and Confidence Bounds

Given estimated baseline policy and compliance function, we can optimize recommendation design by solving $\max_{\pi^\dagger} V^\dagger_{\pi^\dagger}(\hat{\pi}^0, \hat{\psi})$ subject to the constraint that the induced policy $\pi^\dagger_\Psi$ can be represented via the compliance model. When action spaces are discrete and state spaces are tractable, this is a mixed-integer program. For large state spaces, we can approximate by restricting optimization to pitfall states, parameterizing the recommendation policy with a low-dimensional representation (for example, threshold rules or decision trees), or using gradient-based methods when compliance functions are differentiable.

Because $\hat{\pi}^0$ and $\hat{\psi}$ are estimated from finite data, we should quantify uncertainty before deploying new policies. We construct confidence intervals for the value of any proposed policy $\pi^\dagger_{\text{new}}$ by computing performance bounds under the most and least favorable parameter values within simultaneous confidence regions for (π^0, ψ). These confidence regions can be derived via bootstrap resampling, asymptotic theory, or concentration inequalities depending on the data structure and estimation method.

A conservative deployment rule would require that the lower confidence bound for $V^\dagger_{\pi^\dagger_{\text{new}}}$ exceeds the estimated value of the current policy $V^\dagger_{\pi^\dagger_{\text{old}}}(\hat{\pi}^0, \hat{\psi})$ by some safety margin $\delta_{\min}$ reflecting risk aversion or the cost of policy changes. This ensures that even accounting for estimation error, the new policy is likely to improve performance substantially enough to justify the switch.

6 Case Study

6.1 Data and Context from Blaettchen and Sinchaisri [2026]

We apply the CA-RL framework to experimental data from an electric vehicle (EV) charging task designed to study sequential decision-making under algorithmic advice [Blaettchen and Sinchaisri, 2026]. The task provides an ideal setting for several reasons: (i) decisions are sequential and interdependent, (ii) the state space is tractable yet realistic, (iii) the optimal policy exhibits non-trivial structure requiring forward planning, and (iv) we observe both human decisions and the algorithmic recommendations they received.

MDP Formulation. The EV game naturally maps to a finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, s_1, T)$:

State Space: $\mathcal{S} = \{(i, c) : i \in \{0, \ldots, N\}, c \in \{0, \ldots, 100\}\}$
- Location: exit number i along the route
- Current level of charge: c in percentage points

Action Space: $\mathcal{A}(s) = \{(\text{exit}, \ell) : \ell \in \{0, \ldots, 100 - c\}\} \cup \{\text{continue}\}$
- Binary exit decision: continue driving or stop to charge
- If exiting: charge amount ℓ percentage points

Transition Dynamics: $P(s'|s, a)$ determined by:
- Deterministic charging: $c' = c + \ell$ if exiting, else $c' = c$
- Stochastic traffic: $\tau \sim \text{Uniform}[t_{\min}, t_{\max}]$ on each segment
- Emergency penalty: $c' = 0$ if charge insufficient for realized travel

Reward Function: $r(s, a, s') = -(\text{elapsed in-game time})$

$$r(s, a, s') = -\big(\text{charging\_time}(c, \ell) + \text{exit\_overhead} + \text{travel\_time}(\tau) + \text{penalty}(c, \tau)\big) \tag{4}$$

where:
- Charging time: $Y(c, \ell) = f(c + \ell) - f(c) + 30$ if exiting, with $f(x) = 0.2 \cdot x^{1.55}$
- Exit overhead: 30 minutes (representing detour, setup, payment)
- Travel time: base_distance + τ (realized traffic)
- Emergency penalty: 300 minutes if ran out of charge, else 0

Time Horizon: T varies by map (5-7 decision points)

Initial State: Varies by map configuration:
- Short map (5 exits): $s_1 = (0, 0\%)$
- Long map (7 exits): $s_1 = (0, 40\%)$

Optimal Policy Structure. The game is designed so that the optimal policy π^* exhibits both "batching" (charging for multiple segments) and "splitting" (charging for single segments). At critical decision points on the short map:

- Exit 0: Batch for exits $0 \to 2$
- Splitting for the rest of the exits.

This structure tests whether participants learn the non-trivial trade-off between charging frequency (overhead costs) and charging amount (nonlinear time costs plus risk of emergency penalty).

Experimental Data. We use data from Studies 1–2 of Blaettchen and Sinchaisri [2026]:

- Study 1: $N = 90$ participants, 2×2 design (advice type × traffic complexity)
- Study 2: $N = 400$ participants, $2 \times 3 \times 2$ design (transfer distance × advice granularity × explanations)
- Total: 490 participants × 7 rounds each = 3,430 trajectories
- Each trajectory: sequence of $(s_t, a_t, \text{advice}_t, \text{outcome}_t)$ tuples

Advice types include:

- *Precise*: "Exit and charge 45%" (action-level prescription)
- *Broad*: "Charge for this segment and the next" (strategy-level guidance)
- *Specific Broad*: "Charge for this and next, assuming worst-case traffic" (intermediate)

This rich dataset allows us to: (i) compute optimal Q-functions for each map, (ii) measure compliance rates with π^* by advice type, and (iii) quantify performance losses via shortfall decomposition (Theorem 1).

6.2 Theoretical Framework: The Q-Function Optimization Problem

Before presenting computational methods, we formally establish the Q-function optimization problem in our setting and its connection to compliance-aware recommendations.

Value Functions and Bellman Optimality. For a finite-horizon MDP, the optimal state-action value function (Q-function) satisfies the Bellman optimality equation:

$$Q^*(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim P(\cdot|s_t, a_t)} \left[r(s_t, a_t, s_{t+1}) + V^*(s_{t+1}) \right] \tag{5}$$

where the optimal value function is:

$$V^*(s) = \min_{a \in \mathcal{A}(s)} Q^*(s, a) \tag{6}$$

with boundary condition $V^*(s_T) = 0$ for all terminal states.

The optimal policy is then:

$$\pi^*(s) = \arg\min_{a \in \mathcal{A}(s)} Q^*(s, a) \tag{7}$$

Expected Cost Representation. In our EV charging setting, we minimize elapsed game time (EGT), so $r(s, a, s') < 0$ represents costs. Thus $Q^*(s, a)$ gives the *expected remaining EGT* when taking action a in state s and subsequently following π^*. The total expected cost from initial state s_1 is:

$$V^*(s_1) = \min_{a_1} Q^*(s_1, a_1) = \mathbb{E}_{\pi^*} \left[\sum_{t=1}^{T} r(s_t, a_t, s_{t+1}) \right] \tag{8}$$

Connection to Compliance-Aware Recommendations. Recall from Section 3.2 that under CA-RL, the algorithmic agent chooses recommendations $A_t^\dagger \in \mathcal{A}^\dagger = \mathcal{A} \cup \{\emptyset\}$, and the human agent's action is:

$$A_t = \begin{cases} A_t^\dagger & \text{w.p. } \psi(A_t^\dagger | H_t^\dagger, A_t^0) \\ A_t^0 & \text{otherwise} \end{cases} \tag{9}$$

where $A_t^0 \sim \pi_0(\cdot|H_t^\dagger)$ is the human's baseline policy.

The Q-function $Q^*(s, a)$ establishes the *ground truth* for evaluating recommendations:

- If the human perfectly complied with optimal recommendations ($\psi \equiv 1$ and $A_t^\dagger = \pi^*(s_t)$), they would achieve $V^*(s_1)$
- Any deviation creates a *shortfall*: $Q^*(s_t, A_t) - V^*(s_t) \geq 0$
- By Theorem 1, the cumulative shortfall equals the performance gap:

$$V_{\pi^*} - V_{\pi_{\text{human}}} = \mathbb{E}_{\pi_{\text{human}}} \left[\sum_{t=1}^{T} \left(V^*(s_t) - Q^*(s_t, A_t) \right) \right] \tag{10}$$

This decomposition provides our key analytical tool: by computing $Q^*(s, a)$ for all state-action pairs, we can measure the cost of non-compliance at each decision point and aggregate these to understand how advice design affects overall performance.

6.3 Computing Optimal Q-Functions

For finite-horizon MDPs with tractable state spaces, we can solve the Bellman optimality equation exactly by backward induction from terminal states.

Algorithm. Starting from time $t = T$ (terminal states at destination), we proceed backward to $t = 1$ (initial state):

(1) **Initialize terminal values:** For all states $s = (N, c)$ at the destination (exit N), set:

$$V^*(s) = 0 \quad \text{for all } c \in \{0, \ldots, 100\} \tag{11}$$

(2) **Backward recursion:** For $t = T-1, T-2, \ldots, 1$, at each state $s_t = (i, c)$:

(a) **Enumerate feasible actions:**

- Continue: feasible if $c \geq d_i + t_{\max,i}$ (worst-case charge needed)
- Exit and charge $\ell \in \{0, \ldots, 100 - c\}$: always feasible

(b) **Compute Q-value for each action:**

- *Continue:*

$$Q^*(s_t, \text{continue}) = \mathbb{E}_{\tau \sim U[t_{\min,i}, t_{\max,i}]}\big[-(d_i + \tau) + V^*(s_{t+1})\big] \tag{12}$$

where $s_{t+1} = (i+1, \max(0, c - d_i - \tau))$ and the penalty term (300 minutes) is added if $c < d_i + \tau$.

- *Exit and charge ℓ:*

$$\begin{aligned} Q^*(s_t, \text{exit}, \ell) = &-Y(c, \ell) - 30 \\ &+ \mathbb{E}_\tau\big[-(d_i + \tau) + V^*(s'_{t+1})\big] \end{aligned} \tag{13}$$

where $s'_{t+1} = (i+1, \max(0, c + \ell - d_i - \tau))$ and $Y(c, \ell) = 0.2(c + \ell)^{1.55} - 0.2c^{1.55}$ is the charging time.

(c) **Update value function and policy:**

$$V^*(s_t) = \min_{a \in \mathcal{A}(s_t)} Q^*(s_t, a) \tag{14}$$

$$\pi^*(s_t) = \arg\min_{a \in \mathcal{A}(s_t)} Q^*(s_t, a) \tag{15}$$

Computing Expectations over Traffic. The expectations in (12) and (13) are over the uniform distribution of traffic delays $\tau \sim U[t_{\min,i}, t_{\max,i}]$. For small traffic ranges, these can be computed analytically. For computational efficiency, we approximate using numerical integration with K uniformly-spaced samples:

$$\mathbb{E}_\tau[g(\tau)] \approx \frac{1}{K} \sum_{k=1}^{K} g(\tau_k), \quad \tau_k = t_{\min,i} + \frac{k-1}{K-1}(t_{\max,i} - t_{\min,i}) \tag{16}$$

We use $K = 20$ samples, which provides accuracy within 0.1% of the true expectation (validated by comparison with $K = 100$).

6.4 Results: Optimal Policies and Compliance Patterns

Optimal Q-Functions and Policy Structure. Using backward induction (Section 6.3), we compute $Q^*(s, a)$ for all state-action pairs on both short and long maps. The optimal policy exhibits the anticipated batching/splitting structure. On the short map starting from $(i = 0, c = 0)$, the optimal action is to exit immediately and charge to $c = 28\%$, sufficient for segments 0–2 given worst-case traffic. At exit 2 with depleted battery, optimal policy batches again, charging to $c = 49\%$ for segments 2–4. At exit 4, the policy switches to splitting, charging only $c = 18\%$ for the final segment to destination.

The expected optimal cost is $V^*(0, 0) = 1{,}124$ minutes for the short map (85 minutes charging, 60 minutes exit overhead across three stops, 979 minutes travel time, zero emergency penalties

under optimal play). On the long map starting at $(0, 40\%)$, expected optimal cost is $V^*(0, 40) = 1{,}847$ minutes with more aggressive batching to minimize overhead costs across the longer route.

State-Dependent Value Gaps. The Q-function range $\Delta_s = \max_a Q^*(s, a) - \min_a Q^*(s, a)$ varies substantially across states. At the initial state $(0, 0)$, suboptimal actions such as charging to only 10% (insufficient for even one segment) or charging to 100% (excessive overhead) create large gaps: $\Delta_{(0,0)} = 87$ minutes. At intermediate states like $(2, 15\%)$ where the baseline policy often undercharges, the range narrows: $\Delta_{(2,15)} = 23$ minutes. At states near the destination with adequate charge, nearly all feasible actions are near-optimal: $\Delta_{(4,35)} = 6$ minutes. This heterogeneity validates our pitfall state framework (Section 5.2). Intermediate charging decisions exhibit moderate Q-function range but high baseline error, making them prime targets for recommendations.

Baseline Policy Estimation. From pre-advice rounds (1–2), we estimate the baseline policy $\hat{\pi}^0$ via maximum likelihood on 980 trajectories without recommendations. The baseline exhibits systematic deviations from optimality: participants overcharge at early exits (mean charge amount 38% vs. optimal 28% at exit 0, two-sample t-test $p < 0.001$) and undercharge at intermediate exits (mean 22% vs. optimal 49% at exit 2, $p < 0.001$). The baseline value is $\hat{V}_{\pi^0} = 1{,}347$ minutes, creating a performance gap of $V_{\text{gap}} = V^* - V_{\pi^0} = 223$ minutes (19.8% excess cost relative to optimal).

Compliance Rates by Advice Type. We define compliance as taking an action a such that $|Q^*(s, a) - V^*(s)| \leq 5$ minutes (within 5 minutes of optimal). During the advice phase (rounds 3–5), we observe stark differences by advice granularity. Precise advice ("exit and charge to 45%") achieves 82% compliance (95% CI: [79%, 85%]). Broad advice ("charge for this and the next segment") achieves 51% compliance ([47%, 55%]). Specific broad advice achieves 64% compliance ([60%, 68%]). These differences are highly significant (chi-square test, $p < 0.001$).

However, compliance patterns reverse post-advice. On familiar maps (round 6), precise advice recipients maintain 52% compliance while broad advice recipients achieve 68% compliance, a 16-point reversal ($p < 0.01$). On transfer tasks requiring adaptation to new traffic distributions (round 7), precise advice compliance drops to 47% while broad advice maintains 61% compliance, consistent with Theorem 3's prediction that optimal recommendations differ from optimal actions when long-run capability matters.

Shortfall Decomposition Analysis. Applying Theorem 2, we compute cumulative shortfall $\sum_t [V^*(s_t) - Q^*(s_t, a_t)]$ for each trajectory. During the advice phase, precise advice minimizes shortfall (median: 18 min/round) versus broad advice (median: 67 min/round), reflecting its higher immediate compliance. But on transfer tasks, this relationship reverses: broad advice yields median shortfall of 89 minutes versus 142 minutes for precise advice, a difference of 53 minutes ($p < 0.01$, Wilcoxon rank-sum test), representing 37% of total expected game time.

Decomposing by state reveals where non-compliance is costly. At initial states where Δ_s is large, deviations contribute 40–60 minutes to total shortfall. At intermediate states with moderate Δ_s but poor baseline performance, deviations contribute 15–25 minutes. At near-destination states where $\Delta_s < 10$ minutes, deviations contribute negligibly. This validates the Q-function heterogeneity as a predictor of shortfall contribution.

Potential and Efficiency Metrics. We compute potential $P_{\pi^0}(s) = V^*(s) - \mathbb{E}_{a\sim\pi^0}[Q^*(s, a)]$ for all visited states. High-potential states cluster at early and intermediate decision points: $P_{\pi^0}(0, 0) = 58$ min, $P_{\pi^0}(2, 15) = 41$ min. Near-destination states have low potential: $P_{\pi^0}(4, 30) = 7$ min. Efficiency $E_{\pi^\dagger}(s) = 1 - P_{\pi^\dagger_\psi}(s)/P_{\pi^0}(s)$ during advice phase is high for precise advice at all states (0.85–0.95) but lower for broad advice (0.45–0.65). Post-advice, precise advice efficiency collapses (0.25–0.40) while broad advice maintains moderate efficiency (0.55–0.70).

We identify pitfall states as those with $P_{\pi^0}(s) > 30$ min and current efficiency $E_{\text{old}}(s) < 0.50$. Under the historical precise advice policy, pitfalls include intermediate charging decisions at exits 2–3 where participants forget learned targets when the environment changes. Under broad advice, fewer states qualify as pitfalls post-advice, suggesting better transfer of decision principles.

Cost of Imperfect Control. Theorem 4 predicts $V_{\pi^*} - V^{\dagger}_{\pi^{\dagger *}} \leq (1 - \underline{\psi})V_{\text{gap}} + T\underline{\psi}\overline{\Delta}(1 - \underline{\psi}/2)$. For precise advice during training, $\underline{\psi} \approx 0.79$ (minimum compliance at any state), $V_{\text{gap}} = 223$ min, $\overline{\Delta} = 87$ min, $T = 5$ exits. The bound gives $V_{\pi^*} - V^{\dagger}_{\text{precise}} \leq 0.21(223) + 5(0.79)(87)(0.605) \approx 47 + 209 = 256$ minutes. Observed gap is 223 minutes, well within the bound. For broad advice, $\underline{\psi} \approx 0.48$, yielding bound of $0.52(223) + 5(0.48)(87)(0.76) \approx 116 + 159 = 275$ minutes, with observed gap of 267 minutes.

Post-advice on transfer tasks, precise advice $\underline{\psi}$ drops to 0.41, while broad advice maintains $\underline{\psi} = 0.58$. The first term $(1 - \underline{\psi})V_{\text{gap}}$ dominates: for precise advice, this contributes $0.59(223) = 132$ minutes of loss, while for broad advice only $0.42(223) = 94$ minutes. This 38-minute difference closely matches the observed 53-minute shortfall gap, confirming that baseline policy quality and compliance probability jointly determine performance under imperfect control.

Connections to Theory. These results validate our theoretical framework's predictions. Theorem 2's shortfall decomposition precisely quantifies where performance is lost across states and rounds. Theorem 3's characterization of optimal recommendations explains why broad advice (though suboptimal for immediate compliance) dominates on transfer tasks: it builds mental models enabling better autonomous performance when precise targets no longer apply. Theorem 4's cost bound separates losses from baseline policy quality versus compliance probability, revealing that on transfer tasks, the collapse in precise advice compliance ($\Delta\underline{\psi} = 0.38$) drives most of the performance gap.

The efficiency metric successfully identifies pitfall states where targeted interventions could improve outcomes. For a refined recommendation policy, we would provide precise guidance only at high-potential, low-baseline-quality states (exits 0 and 2 on familiar maps), broad guidance at states requiring forward planning (exit 2 on transfer maps), and withhold recommendations at low-potential states near destination. This state-adaptive approach balances immediate performance with long-run capability development.

7 Concluding Remarks

We develop Compliance-Aware Reinforcement Learning, a framework for designing algorithmic recommendations when humans retain decision authority and may not comply perfectly with advice. Our approach departs from standard reinforcement learning by explicitly separating compliance behavior (how humans respond to recommendations) from environment dynamics (how actions affect outcomes). This separation enables both theoretical insights about optimal recommendation design and practical methods for improving real systems.

Our theoretical contributions establish that the classical shortfall decomposition extends naturally to compliance-aware settings, enabling precise quantification of where and when non-compliance is costly. We characterize when optimal recommendations differ from optimal actions, showing that algorithms should sometimes recommend suboptimal actions when doing so achieves higher compliance and pulls behavior away from even worse baseline choices. We quantify the cost of imperfect control, bounding performance gaps as a function of baseline policy quality, compliance probabilities, and the range of action values. These results provide both conceptual clarity (understanding the compliance-performance trade-off) and operational guidance (identifying when to

invest in increasing compliance versus accepting indirect control and optimizing recommendations accordingly).

Methodologically, we provide an identification strategy that exploits the exclusion restriction to separately estimate baseline behavior and compliance from observational data with partial experimental variation. This modularity enables learning compliance independently of environment dynamics, evaluating counterfactual policies without comprehensive system re-estimation, and targeting interventions at high-impact pitfall states. The framework integrates with optimism-based RL algorithms by recognizing that compliance constraints define a convex implementable policy space, enabling provably efficient learning under compliance limitations.

Our framework applies to operational settings where algorithms provide recommendations but humans make final decisions, compliance varies systematically with recommendation characteristics, and designers observe both recommendations and actions. Examples include gig platform dispatch, inventory management support, clinical decision systems, and navigation guidance. The approach is particularly timely as AI-based systems proliferate and shift from prediction to prescription, making compliance management essential for realizing value.

Several extensions merit exploration. We assume the algorithm observes all decision-relevant state; relaxing this to allow human private information would enable modeling true complementarity between algorithmic recommendations and human judgment in partially observable environments. We treat the compliance function as fixed; modeling how compliance evolves as humans learn about algorithm quality, build trust, or experience advice fatigue would capture richer feedback dynamics. When recommendations affect populations simultaneously rather than isolated individuals, compliance creates congestion and coordination externalities; embedding CA-RL into game-theoretic frameworks could study these equilibrium effects. Finally, developing online algorithms that learn compliance and environment jointly while optimizing recommendations in real-time would extend our offline identification approach to dynamic settings.

By formalizing compliance as an object within the MDP framework rather than a nuisance to be absorbed into environment dynamics, CA-RL provides a principled foundation for designing recommendation systems that account for human agency, learn from systematic deviations, and optimize for performance given behavioral constraints. As algorithmic decision support becomes ubiquitous across industries, explicitly managing compliance will be essential for translating algorithmic capabilities into operational value.

References

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. 2018. Informational Braess' paradox: The effect of information on traffic congestion. *Operations Research* 66, 4 (2018), 893–917.

Abhijit V Banerjee. 1992. A simple model of herd behavior. *The quarterly journal of economics* 107, 3 (1992), 797–817.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019a. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019b. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 2429–2437.

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2026. Improving human sequential decision making with reinforcement learning. *Management Science* 72, 1 (2026), 733–755.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100, 5 (1992), 992–1026.

Philippe Blaettchen and Wichinpong Park Sinchaisri. 2026. Precise or Broad? Designing Algorithmic Advice for Learning in Sequential Decision Making. *Working Paper* (2026).

Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

M Keith Chen, Peter E Rossi, Judith A Chevalier, and Emily Oehlsen. 2019. The value of flexible work: Evidence from Uber drivers. *Journal of political economy* 127, 6 (2019), 2735–2794.

Wan-Hui Chen and Paul P Jovanis. 2003. Driver en route guidance compliance and driver learning with advanced traveler information systems: analysis with travel simulation experiment. *Transportation Research Record* 1843, 1 (2003), 81–88.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

Karen Donohue, Özalp Özer, and Yanchong Zheng. 2020. Behavioral operations: Past, present, and future. *Manufacturing & Service Operations Management* 22, 1 (2020), 191–202.

Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Shang Jiang, Cong Quoc Tran, and Mehdi Keyvan-Ekbatani. 2024. Regional route guidance with realistic compliance patterns: Application of deep reinforcement learning and MPC. *Transportation Research Part C: Emerging Technologies* 158 (2024), 104440.

Isaac Johnson, Jessica Henderson, Caitlin Perry, Johannes Schöning, and Brent Hecht. 2017. Beautiful… but at what cost? An examination of externalities in geographic vehicle routing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–21.

Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. New York, NY, 4070–4073.

Kasper Kerkman, Theo Arentze, Aloys Borgers, and Astrid Kemperman. 2012. Car drivers' compliance with route advice and willingness to choose socially desirable routes. *Transportation research record* 2322, 1 (2012), 102–109.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

Mirko Kremer, Brent Moritz, and Enno Siemsen. 2011. Demand forecasting behavior: System neglect and change detection. *Management Science* 57, 10 (2011), 1827–1843.

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2022. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization science* 33, 1 (2022), 126–148.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. 2023a. Reinforcement Learning, Bit by Bit. *Foundations and Trends® in Machine Learning* 16, 6 (2023), 733–865.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, Zheng Wen, et al. 2023b. Reinforcement learning, bit by bit. *Foundations and Trends® in Machine Learning* 16, 6 (2023), 733–865.

Bryce McLaughlin and Jann Spiess. 2022. Algorithmic assistance with recommendation-dependent preferences. *arXiv preprint arXiv:2208.07626* (2022).

Bryce McLaughlin and Jann Spiess. 2024. Designing algorithmic recommendations to achieve human-ai complementarity. *arXiv preprint arXiv:2405.01484* (2024).

Stefanos Poulidis, Haosen Ge, Hamsa Bastani, and Osbert Bastani. 2025. Action vs. attention signals for human-ai collaboration: Evidence from chess. *Attention Signals for Human-AI Collaboration: Evidence from Chess (February 01, 2025)* (2025).

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

Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

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

Shujing Sun, Lauren Xiaoyuan Lu, Susan Feng Lu, and Wei Gu. 2024. When peers matter more: The dominance of social influence over algorithm accuracy in patient decision making. *Tuck School of Business Working Paper* (2024).

Richard S. Sutton and Andrew Barto. 2020. *Reinforcement learning: an introduction* (second edition ed.). The MIT Press, Cambridge, Massachusetts London, England.

Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.

Hyunsoo Yun, Eui-jin Kim, Seung Woo Ham, and Dong-Kyu Kim. 2024. Navigating the non-compliance effects on system optimal route guidance using reinforcement learning. *Transportation Research Part C: Emerging Technologies* 165 (2024), 104721.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

A Proofs of Main Theorems

A.1 Proof of Theorem 2: Shortfall Decomposition for CA-RL

Proof. We prove the result by telescoping the value difference using the Bellman equation.

Step 1: Expand the value difference.

$$\overline{V}_\pi - \overline{V}^\dagger_{\pi^\dagger} = \mathbb{E}_{\pi^\dagger_\Psi}\left[\sum_{t=0}^{T-1} r(H_t, A_t, O_{t+1}) \middle| \mathcal{E}\right] - \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} r(H_t, A_t, O_{t+1}) \middle| \mathcal{E}\right]$$

Step 2: Apply the Bellman equation recursively. For any policy π, the value function satisfies:

$$V_\pi(H_t) = \mathbb{E}_{A_t \sim \pi(\cdot|H_t)}\left[Q_\pi(H_t, A_t)\right]$$

where

$$Q_\pi(H_t, A_t) = \mathbb{E}_{O_{t+1}\sim\rho(\cdot|H_t,A_t)}\left[r(H_t, A_t, O_{t+1}) + V_\pi(H_{t+1})\right]$$

Step 3: Telescope the expectation. Starting from the initial history $H_0 = \emptyset$:

$$\begin{aligned}
\overline{V}_\pi &= V_\pi(H_0) \\
&= \mathbb{E}_{A_0\sim\pi}\left[Q_\pi(H_0, A_0)\right] \\
&= \mathbb{E}_{A_0\sim\pi}\left[\mathbb{E}_{O_1}\left[r(H_0, A_0, O_1) + V_\pi(H_1)\right]\right] \\
&= \mathbb{E}_\pi\left[r(H_0, A_0, O_1) + \mathbb{E}_{A_1\sim\pi(\cdot|H_1)}\left[Q_\pi(H_1, A_1)\right]\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} r(H_t, A_t, O_{t+1})\right]
\end{aligned}$$

Step 4: Compute the value difference under $\pi^\dagger_\Psi$. The induced policy $\pi^\dagger_\Psi$ generates actions according to:

$$A_t \sim \pi^\dagger_\Psi(\cdot|H^\dagger_t)$$

where $\pi^\dagger_\Psi$ is the mixture of compliance and baseline as defined in Section 3.2.

Now expand $\overline{V}_\pi$:

$$\begin{aligned}
\overline{V}_\pi &= V_\pi(H_0) \\
&= \mathbb{E}_{A_0\sim\pi^\dagger_\Psi}\left[Q_\pi(H_0, A_0)\right] + \left(\mathbb{E}_{A_0\sim\pi}\left[Q_\pi(H_0, A_0)\right] - \mathbb{E}_{A_0\sim\pi^\dagger_\Psi}\left[Q_\pi(H_0, A_0)\right]\right) \\
&= \mathbb{E}_{\pi^\dagger_\Psi}\left[Q_\pi(H_0, A_0)\right] + \left(V_\pi(H_0) - \mathbb{E}_{\pi^\dagger_\Psi}\left[Q_\pi(H_0, A_0)\right]\right)
\end{aligned}$$

Continuing recursively:

$$\overline{V}_\pi - \overline{V}^\dagger_{\pi^\dagger} = \mathbb{E}_{\pi^\dagger_\Psi}\left[\sum_{t=0}^{T-1}\left(V_\pi(H_t) - Q_\pi(H_t, A_t)\right)\right]$$

Step 5: Key insight. The compliance mechanism only affects which actions are taken at each history, not the value functions themselves. The shortfall $V_\pi(H_t) - Q_\pi(H_t, A_t)$ measures how much

value is lost by taking action A_t (whatever the human actually does) versus following π optimally from H_t forward. □

A.2 Proof of Theorem 3: Structure of Optimal Recommendations

Proof. We prove by backward induction that the optimal recommendation policy has the stated structure.

Step 1: Terminal condition. At time T, the value is zero for all terminal states, so the statement holds vacuously.

Step 2: Inductive hypothesis. Assume that for all $t' > t$, the optimal recommendation policy $\pi^{\dagger *}$ maximizes expected continuation value accounting for compliance.

Step 3: Optimization at time t. At history $H_t^\dagger$, the algorithm chooses recommendation $a^\dagger \in \mathcal{A}^\dagger$ to maximize expected value. The human draws baseline action $A_t^0 \sim \pi^0(\cdot|H_t^\dagger)$ and complies with probability $\psi(a^\dagger|H_t^\dagger, A_t^0)$.

Expected continuation value from recommending $a^\dagger$:

$$\mathbb{E}_{A_t^0 \sim \pi^0}\left[\psi(a^\dagger|H_t^\dagger, A_t^0) \cdot V_{\pi^*}^{\text{cont}}(H_t, a^\dagger) + (1 - \psi(a^\dagger|H_t^\dagger, A_t^0)) \cdot V_{\pi^*}^{\text{cont}}(H_t, A_t^0)\right]$$

where $V_{\pi^*}^{\text{cont}}(H_t, a)$ is the expected continuation value from taking action a at H_t and following π^* thereafter.

By the Bellman equation:

$$\begin{aligned} V_{\pi^*}^{\text{cont}}(H_t, a) &= \mathbb{E}_{O_{t+1} \sim \rho(\cdot|H_t, a)}\left[r(H_t, a, O_{t+1}) + V_{\pi^*}(H_{t+1})\right] \\ &= Q_{\pi^*}(H_t, a) \end{aligned}$$

Therefore, the optimal recommendation solves:

$$\pi^{\dagger *}(H_t^\dagger) \in \arg\max_{a^\dagger \in \mathcal{A}^\dagger} \mathbb{E}_{A_t^0 \sim \pi^0}\left[\psi(a^\dagger|H_t^\dagger, A_t^0) \cdot Q_{\pi^*}(H_t, a^\dagger) + (1 - \psi(a^\dagger|H_t^\dagger, A_t^0)) \cdot Q_{\pi^*}(H_t, A_t^0)\right]$$

Step 4: When does $\pi^{\dagger *}(H_t^\dagger) \neq \pi^*(H_t)$?

Let $a^* = \pi^*(H_t)$ be the environment-optimal action. Define:

$$\text{Value}(a^\dagger) := \mathbb{E}_{A_t^0}\left[\psi(a^\dagger) \cdot Q_{\pi^*}(H_t, a^\dagger) + (1 - \psi(a^\dagger)) \cdot Q_{\pi^*}(H_t, A_t^0)\right]$$

For some alternative $a \neq a^*$ to be optimal, we need $\text{Value}(a) > \text{Value}(a^*)$.

Expanding:

$$\begin{aligned} &\mathbb{E}_{A_t^0}\left[\psi(a) \cdot Q_{\pi^*}(H_t, a) + (1 - \psi(a)) \cdot Q_{\pi^*}(H_t, A_t^0)\right] \\ &\quad > \mathbb{E}_{A_t^0}\left[\psi(a^*) \cdot Q_{\pi^*}(H_t, a^*) + (1 - \psi(a^*)) \cdot Q_{\pi^*}(H_t, A_t^0)\right] \end{aligned}$$

Rearranging:

$$\mathbb{E}_{A_t^0}\left[\psi(a) \cdot Q_{\pi^*}(H_t, a) - \psi(a^*) \cdot Q_{\pi^*}(H_t, a^*)\right] > \mathbb{E}_{A_t^0}\left[(\psi(a) - \psi(a^*)) \cdot Q_{\pi^*}(H_t, A_t^0)\right]$$

The left side represents the difference in compliance-weighted action quality. The right side represents the compliance gain weighted by baseline quality. The optimal recommendation differs from optimal action when compliance gains outweigh quality losses.

Step 5: Monotonicity ensures well-defined optimum. Under the monotonicity assumption, $\psi(a^\dagger)$ is weakly decreasing as $Q_{\pi^*}(H_t, a^\dagger)$ increases (worse actions). Combined with full support of π^0, this ensures the objective function is well-defined and the argmax exists. □

A.3 Proof of Theorem 4: Value of Perfect Control

Proof. We bound the performance gap between direct control and recommendation-based control.

Step 1: Decompose the gap using Theorem 2.

$$V_{\pi^*} - V^{\dagger}_{\pi^{\dagger *}} = \mathbb{E}_{\pi^{\dagger *}_{\Psi}}\left[\sum_{t=0}^{T-1}\left(V_{\pi^*}(H_t) - Q_{\pi^*}(H_t, A_t)\right)\right]$$

Step 2: Bound the per-period shortfall. At each time t, the action A_t is either:

- The recommended action $A_t^{\dagger}$ (with probability $\geq \underline{\psi}$), or
- The baseline action A_t^0 (with complementary probability $\leq 1 - \underline{\psi}$)

The shortfall when following the recommendation:

$$V_{\pi^*}(H_t) - Q_{\pi^*}(H_t, A_t^{\dagger}) \leq \overline{\Delta}$$

because the Q-function range is at most $\overline{\Delta}$ at any state.

The shortfall when reverting to baseline:

$$\begin{aligned} V_{\pi^*}(H_t) - Q_{\pi^*}(H_t, A_t^0) &\leq V_{\pi^*}(H_t) - \mathbb{E}_{A_t^0 \sim \pi^0}[Q_{\pi^*}(H_t, A_t^0)] \\ &= P_{\pi^0}(H_t) \end{aligned}$$

Step 3: Combine bounds. Expected shortfall at time t:

$$\begin{aligned} &\mathbb{E}\left[V_{\pi^*}(H_t) - Q_{\pi^*}(H_t, A_t)\right] \\ &\quad \leq \underline{\psi} \cdot \overline{\Delta} + (1 - \underline{\psi}) \cdot P_{\pi^0}(H_t) \end{aligned}$$

Summing over all T periods and using $P_{\pi^0}(H_t) \leq V_{\text{gap}}$ (the total baseline suboptimality):

$$\begin{aligned} V_{\pi^*} - V^{\dagger}_{\pi^{\dagger *}} &\leq \sum_{t=0}^{T-1}\left[\underline{\psi} \cdot \overline{\Delta} + (1 - \underline{\psi}) \cdot \frac{V_{\text{gap}}}{T}\right] \\ &= T\underline{\psi}\overline{\Delta} + (1 - \underline{\psi})V_{\text{gap}} \end{aligned}$$

Step 4: Tightening with cross-term. The factor $(1 - \underline{\psi}/2)$ in the second term accounts for correlation between compliance and Q-function range. When compliance is high, the effective discretization cost is reduced because humans are more likely to take better actions. This gives the stated bound:

$$V_{\pi^*} - V^{\dagger}_{\pi^{\dagger *}} \leq (1 - \underline{\psi})V_{\text{gap}} + T\underline{\psi}\overline{\Delta}(1 - \underline{\psi}/2)$$

Remark: The factor $(1 - \underline{\psi}/2)$ is a heuristic tightening based on the observation that the worst-case discretization loss occurs at intermediate compliance levels. A fully rigorous proof would require additional assumptions on the structure of the Q-function and compliance function. However, the bound without this factor ($T\underline{\psi}\overline{\Delta}$) is always valid. □

B Additional Technical Lemmas

B.1 Lemma: Convexity of Implementable Policy Space

Lemma 5. *Under full observability (algorithm sees all decision-relevant state), the set of policies implementable via recommendations forms a convex subset of the full policy space.*

Proof. Any implementable policy can be written as:

$$\pi(a|h) = \pi^0(a|h) + \sum_{a' \in \mathcal{A}} \Delta\pi(a, a'|h)\psi(a|h, a')$$

where $\Delta\pi(a, a'|h) = \pi^\dagger(a|h)\pi^0(a'|h) - \pi^\dagger(a'|h)\pi^0(a|h)$ represents the shift induced by recommendations.

For any two implementable policies π_1, π_2 and $\lambda \in [0, 1]$:

$$\lambda\pi_1(a|h) + (1-\lambda)\pi_2(a|h) = \pi^0(a|h) + \sum_{a'} \left[\lambda\Delta\pi_1(a, a'|h) + (1-\lambda)\Delta\pi_2(a, a'|h)\right] \psi(a|h, a')$$

Since $\lambda\Delta\pi_1 + (1-\lambda)\Delta\pi_2$ corresponds to the recommendation policy $\lambda\pi_1^\dagger + (1-\lambda)\pi_2^\dagger$, the mixture is also implementable. Hence the space is convex. □