Precise or Broad? Designing Algorithmic Advice for Learning in Sequential Decision Making

Philippe Blaettchen
Lee Kong Chian School of Business, Singapore Management University, pblaettchen@smu.edu.sg

Wichinpong Park Sinchaisri
Haas School of Business, University of California, Berkeley, parksinchaisri@berkeley.edu

Abstract. *Problem definition:* Organizations increasingly deploy algorithmic tools to support complex operational decisions, raising a practical design question: how should these tools be built when designers care not only about immediate performance, but also about preserving and building human skill that remains valuable when advice is unavailable, imperfect, or requires genuine oversight? We study how the *precision* of algorithmic advice shapes this trade-off. *Methodology/results:* We develop a stylized model of advice-taking and learning. The model characterizes a reward-learning frontier: precise, action-level advice is easier to implement and improves payoffs while available through higher compliance, whereas broad, strategic advice requires interpretation, induces greater exploration, and generates knowledge that is portable, even when decision environments differ. We test the model's predictions in two online experiments in an electric-vehicle routing and charging task, representing typical characteristics of sequential decision tasks. Consistent with the theory, precise numerical advice delivers the strongest gains during the advice phase, whereas broader advice can yield more robust performance after advice is removed, specifically if the new environment differs substantially, but not completely. We use inverse reinforcement learning to recover interpretable latent objective components from action traces, distinguishing transient compliance from persistent internalization. *Managerial implications:* Our results provide design guidance for advice systems that balance short-run operational efficiency with the development of long-run human capability. They also help validate inverse reinforcement learning as an effective tool for estimating human behaviors in complex sequential tasks.

Key words: human-ai collaboration, algorithmic advice, sequential decision making, inverse reinforcement learning

1. Introduction

Making good decisions on the job is difficult and costly. It takes time to build intuition about trade-offs, interpret noisy feedback, and develop a policy that performs well across different situations arising in practice. Even with experience, making effective decisions requires time and effort. Algorithmic tools promise to reduce these costs by guiding choices even when the ultimate decision maker is human, in domains such as fulfillment and warehouse operations (Sun et al. 2022), clinical scheduling (Ibanez et al. 2018), and other operational settings (e.g., Balakrishnan et al. 2026, Bastani et al. 2026). Yet, algorithmic tools may not perform as desired, not only because humans do not always adhere to their advice, but also because they alter how users engage with the underlying decision problem. A system that makes execution easy can improve short-run outcomes, but it may also reduce the practice that builds robust human capability.

As decision support becomes more pervasive this creates an increasingly salient design tension. On the one hand, organizations desire the inherent performance gains of improved decision making. On the other hand, over-reliance on automation can erode skills that matter when tools are unavailable, imperfect, or require genuine oversight, particularly in complex tasks such as sequential decision making. Evidence from

aviation, for example, links extensive autopilot use to declines in manual flying ability and slower recovery from errors (Casner et al. 2014). Similar concerns are emerging in education and knowledge work as new tools change when and how people practice core skills (Macnamara et al. 2024, Bastani et al. 2025). In many high-stakes operational settings, stakeholders also require humans to remain accountable for decisions, emphasizing the need for meaningful supervision rather than passive approval (British Medical Association 2024). The practical question is therefore not only whether a tool improves performance today, but also whether it helps or hinders the development of decision-making skills that remain valuable tomorrow.

Many tools provide precise, action-level prescriptions, such as an explicit quantity to implement. Other tools provide broader guidance, such as a rule that emphasizes robustness or encourages planning over a longer horizon. Precise advice can reduce cognitive costs and increase compliance while it is available. Broad advice requires interpretation and may initially appear less effective, but it can induce deeper reasoning about trade-offs and foster learning that transfers when the environment changes or advice is unavailable. We ask under which circumstances a specific degree of precision is desirable for a system designer.

We pursue this question in three steps. First, we develop a stylized model of advice-taking and learning across environments. The model yields testable predictions that structure the empirical analysis: while advice is available, precise advice improves contemporaneous performance largely through higher compliance. Broad advice, on the other hand, confers greater learning advantages in new environments that are sufficiently different, as it supports the decision maker abstracting away from a specific decision to obtain a higher-level understanding. However, the new environment cannot be too different, or transfer would be too complex to be successful. When the task is a Markov decision process, this learning advantage tends to grow with the decision horizon, motivating our study of *sequential* decision tasks. Here, human-in-the-loop concerns are especially acute. In fact, many operational tasks require a policy rather than a single choice: ordering decisions affect downstream stockouts and holding costs; dispatch decisions affect future congestion; clinical treatment decisions affect diagnostic possibilities and, thus, further treatments.

Second, we test our model's predictions in two online experiments using an electric-vehicle routing and charging task. Participants repeatedly decide when and how much to charge while driving along a route with stochastic traffic and nonlinear charging costs. This environment captures central operational features of sequential decision making, while being salient to experimental participants. Across studies, we find a consistent pattern: *precise* numerical advice delivers the strongest gains during the advice phase, whereas *broad* advice can yield more robust performance after advice is removed in a different environment, particularly when participants must infer the underlying logic to successfully transfer learnings.

Third, to open the black box of *how* advice changes sequential strategies, we introduce inverse reinforcement learning (IRL) as a practical measurement tool for behavioral operations research on sequential

decision making. IRL provides a way to recover participants' latent objectives from their observable actions, such as preferences for simplicity or risk aversion, and to distinguish transient compliance from persistent internalization. This addresses an important aspect of the human-in-the-loop problem in complex settings: to design decision environments with specific features (such as supporting learning), we must understand the broader strategies and behaviors they induce, not just measure contemporary performance.

1.1. Literature Review

Operational decision support combines prediction with guidance about action. Field evidence shows that workers systematically deviate from algorithmic prescriptions in ways that matter for performance, including in warehouse operations (Sun et al. 2022), radiology scheduling (Ibanez et al. 2018), and settings where discretion and private information affect outcomes (Balakrishnan et al. 2026). Grand-Clément and Pauphilet (2024) formalize this adherence problem, characterizing optimal advice when planners anticipate selective compliance. A central design question remains underexplored: when organizations care about both near-term outcomes and long-run human capability, how should decision support be structured?

This question is especially consequential in sequential tasks that require anticipating uncertain consequence of current actions on future actions. Because decisions are interdependent, decision biases can accumulate. Difficulties in dealing with uncertainty are well documented, for instance, with an overweighting of salient experiences (Tversky and Kahneman 1973) or the gambler's fallacy (Rabin and Vayanos 2010). In operational settings, decision makers often rely on simple heuristics rather than optimal strategies (Schweitzer and Cachon 2000). The accumulation of such biases creates demand for algorithmic guidance.

This guidance can also affect learning: which situations a decision maker encounters determines what they learn. Performance improvements require accumulated practice and exposure to varied conditions (Argote and Miron-Spektor 2011). A decision maker who always follows an optimal policy visits a narrow band of states, accumulating little information about alternatives (March 1991). Exploration, through experimentation or errors, exposes the decision maker to a broader set of states, yielding more generalizable knowledge. Over long horizons, small differences in behavior can compound into large differences in what is learned, especially when knowledge must transfer to new environments (Lapré and Van Wassenhove 2001).

Advice precision is a natural design lever because it changes how much cognitive work the decision maker must do. Action-level prescriptions reduce the effort needed to map information into a choice, which can increase adherence and improve short-run execution. Broader, strategy-level guidance leaves more of that mapping to the user. A large learning-sciences literature suggests that, when learners generate steps or explanations themselves, they may acquire knowledge that transfers more robustly, even if performance during practice is noisier. In contrast, when guidance fully specifies actions, it can reduce cognitive load and speed initial progress, especially for novices (Sweller 1994, Salden et al. 2010, Bjork et al. 2011). These

forces mirror our setting: more precise advice can improve performance while it is available, whereas broader advice may induce more interpretation and experimentation that supports retention and transfer once advice is removed or the environment changes. Despite extensive research on algorithm aversion and overreliance (e.g., Dietvorst et al. 2015, Buçinca et al. 2021), little is known about how advice precision shapes learning in sequential operational tasks or how this trade-off varies with task horizon and environmental change.

Operations management research has begun to model how decision makers respond to algorithmic guidance (Snyder et al. 2025, Bastani et al. 2026), but this work largely emphasizes performance while advice is present. In parallel, concerns about skill erosion under automation have grown (Casner et al. 2014, Macnamara et al. 2024). What remains missing is a formal and empirical characterization of how advice design shapes the performance-learning trade-off in operational environments, especially for sequential decision problems. Addressing this gap requires measurement beyond observed actions or realized performance: we must also infer what participants appear to optimize. In sequential settings, inverse reinforcement learning (IRL) provides a principled way to recover latent reward functions from behavioral traces (Arora and Doshi 2021), yet it remains underutilized in behavioral operations.

2. A Model for Decision Making under Algorithmic Advice

In our work, we focus on a decision task executed by a human decision maker, possibly supported by algorithmic advice, e.g., in the form of AI tools. The advice can be either *broad*, that is, it can inform the decision maker about the correct way of making an optimal decision. Alternatively, the advice can be *precise*, meaning it provides the decision maker with a concrete action.[1]

As a running example, we consider an inventory manager for a single-item periodic-review system under random demand. We initially assume that demand is stationary and the manager follows a base-stock policy, that is, they choose a base-stock level S, then all periods' ordering decisions are fully defined (ordering up to S). Let $u > 0$ and $o > 0$ denote underage and overage costs. It is well-known that, for demand with cumulative distribution function F, the optimal base stock $S^\star$ is the minimum quantity such that $F(S^\star) \geq \frac{u}{u+o}$. *Precise* advice could be to "order up to $S^\star$." *Broad* advice could communicate the rule mapping (u, o, F) to $S^\star$, for instance, "as $u/(u+o)$ rises, raise the target fractile; heavier upper tails push $S^\star$ upward."

Precise advice is easier to implement. However, broad advice may support learning by forcing the decision maker to engage with the task. This is particularly useful when algorithmic advice cannot always be provided, e.g., in emergency situations, or when workers need sufficient training to effectively supervise algorithms.

[1] We propose an intermediate alternative, as well as combinations of advice with deeper explanations, in Study 2 in Section 5.

2.1. A Simple Model of One-Shot Decision Making

While we aim to understand the decision maker's reaction to advice in the context of *sequential decision problems*—that is, tasks where they have to make multiple decisions in a row to achieve an over-arching goal—we start by modeling a one-shot decision-making task. This helps us to elucidate key trade-offs faced by the designer and define key behaviors to measure. In Section 2.2, we will show how these trade-offs directly extend to sequential decision making, with some important additional considerations.

Setup and Notation. Consider a one-shot decision task carried out by a human decision maker sequentially, in two different decision-making environments:

E_1: In this environment, the decision maker receives algorithmic advice.

E_2: In this environment, the decision maker does not have access to algorithmic advice.

In our experimental setup, we will also have an initial environment, E_0, without access to algorithmic advice, similar to E_2, but without having previously been exposed to the advice.

Rewards. If the decision maker implements the "best" action in a given environment, they receive (uncertain) reward R^+, while they receive R^- when implementing another action. We assume that $r := \mathbb{E}[R^+] - \mathbb{E}[R^-] > 0$, and that the decision maker maximizes their expected reward, minus the cost of effort.

Advice in E_1, effort, and compliance. The designer chooses advice type $a \in \{p, b\}$, where p is *precise* and b is *broad*. In E_1, the decision maker observes this advice, and chooses their effort $e_1 \geq 0$ to exert in decision making. The cost of effort is convex, corresponding to decreasing marginal benefits of exerting effort. That is, the decision maker incurs $c(e_1) = \frac{k}{2}e_1^2$ with $k > 0$.

The probability that the decision maker follows the advice and chooses the best action is

$$\pi_a^1(e_1) = \alpha_a + \beta_a e_1, \tag{1}$$

where we assume $0 < \alpha_b < \alpha_p < 1$ (precise advice is easier to follow), and $0 < \beta_p < \beta_b < 1$ (broad advice provides strategic insight, helping to convert effort into results). W.l.o.g, we further assume $\alpha_a + \beta_a < 1$.

Decision in E_2 (no advice). The decision maker faces a new environment, E_2, without access to advice. A larger value of $\delta \in [0, 1]$ indicates a bigger difference between the task in E_2 compared to E_1.

The decision maker again chooses to exert effort $e_2 \geq 0$ at cost $c(e_2) = \frac{k}{2}e_2^2$, but is not guided by any advice. The probability that they select the best action in this environment is:

$$\pi_a^2(e_1, e_2) = e_2\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\beta_a e_1\right]. \tag{2}$$

Here, $\lambda > 0$ measures how contemporaneous effort e_2 directly translates into choosing the best action. The effectiveness of exerting effort in the second environment is also influenced by learning from the first

environment: First, through the intercept α_a, which captures "effort-free implementability" of advice a in E_1. In E_2, this transfers only when the environments are *very similar*, hence the factor $(1-\delta)^2$. Second, through the slope β_a, which governs how E_1 efforts affect results. What transfers to E_2 is a higher-order understanding of the optimal action, enabled through deep engagement with the task. This is valuable whenever the new environment is *similar*, captured by the milder factor $(1-\delta)$. As $\delta \to 1$, both parts vanish, and $\pi_a^2 \to e_2\lambda$.[2]

Objectives. In each period, the decision maker maximizes their expected pay-off, consisting of a reward, obtained with probability π^1 (resp. π^2), minus the cost of the effort exerted:

$$\begin{aligned} U_1(e_1;a) &= \mathbb{E}[R^+]\pi_a^1(e_1) + \mathbb{E}[R^-](1-\pi_a^1(e_1)) - \tfrac{k}{2}e_1^2 = r(\alpha_a + \beta_a e_1) - \tfrac{k}{2}e_1^2 + \mathbb{E}[R^-] \\ U_2(e_2;a,e_1) &= \mathbb{E}[R^+]\pi_a^2(e_1,e_2) + \mathbb{E}[R^-](1-\pi_a^2(e_1,e_2)) - \tfrac{k}{2}e_2^2, \\ &= re_2\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\beta_a e_1\right] - \tfrac{k}{2}e_2^2 + \mathbb{E}[R^-]. \end{aligned}$$

We assume the decision maker maximize U_1 over e_1 without considering E_2, but takes e_1 (and induced learning) into account when maximizing U_2 over e_2. On the other hand, the designer chooses the advice $a \in \{p, b\}$ in order to maximize a reward that is proportional to a weighted sum of the expected pay-offs:

$$J(a) = \gamma r(\alpha_a + \beta_a e_1^\star(a)) + (1-\gamma)re_2^\star(a)\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\beta_a e_1^\star(a)\right] + \mathbb{E}[R^-],$$

where $e_t^\star(a)$ is the effort chosen by the decision maker, and $\gamma \in [0,1]$ represents the designers' weights on each environments' rewards. Notably, the decision maker's effort cost does not enter the designer's objective.

Parametric assumptions. We make the following additional assumptions:

(A_1) $\frac{r}{k} \le 1$ and $\frac{r}{k}\left[\lambda + \alpha_a + \frac{r\beta_a^2}{k}\right] \le 1$ for $a \in \{p, b\}$;

(A_2) $\alpha_p - \alpha_b > \frac{r}{k}(\beta_b^2 - \beta_p^2)$.

Assumption (A_1) ensures that all solutions are interior for any $\delta \in [0,1]$, given $\pi_a^t \le 1$. Assumption (A_2) allows us to focus on practically relevant cases by ensuring there is a short-term advantage of precise advice.

Effort and Compliance. The decision maker chooses e_1 to maximize $U_1(e_1;a)$. Using (A_1), we can derive the optimal effort: $e_1^\star(a) = \frac{r\beta_a}{k}$. We then observe the following (proofs are in Appendix A):

LEMMA 1 **(E_1 Behavior).** *Consider E_1. $e_1^\star(b) > e_1^\star(p)$, while $\pi_b^1\left(e_1^\star(b)\right) < \pi_p^1\left(e_1^\star(p)\right)$.*

In E_2, the decision maker chooses e_2 to maximize $U_2(e_2;a,e_1^\star)$. Under Assumption (A_2),

$$e_2^\star(a) = \frac{r}{k}\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\frac{r\beta_a^2}{k}\right]. \tag{3}$$

[2] We can rewrite (2) as $\pi_a^2(e_1,e_2) = e_2\lambda + e_2\left(\delta^2 \times 0 + \delta(1-\delta)\beta_a e_1 + (1-\delta)^2\pi_a^1(e_1)\right)$, lending an alternative interpretation: Environments differ in two dimensions, and either dimension is different with probability δ. Hence, with probability δ^2, both dimensions differ, and the second environment is too different to make use of learnings from the first. With probability $\delta(1-\delta)$, the environments are sufficiently similar so that effortful learning in the first environment can lead to sufficient understanding to better navigate the second environment. Finally, with probability $(1-\delta)^2$, the best action in the first environment is still optimal for the second one.

This depends on the environmental difference δ. It will be useful to define $\delta_0 := 1 - \frac{\frac{r}{k}(\beta_b^2 - \beta_p^2)}{\alpha_p - \alpha_b}$. By Assumption ($A_2$), $\delta_0 \in (0,1)$. We then have the following result:

LEMMA 2 (**E_2 Behavior**). *Consider E_2. If $0 \le \delta < \delta_0$, $e_2^\star(b) < e_2^\star(p)$ and $\pi_b^2\left(e_1^\star(b), e_2^\star(b)\right) < \pi_p^2\left(e_1^\star(p), e_2^\star(p)\right)$. If $\delta_0 < \delta < 1$, $e_2^\star(b) > e_2^\star(p)$ and $\pi_b^2\left(e_1^\star(b), e_2^\star(b)\right) > \pi_p^2\left(e_1^\star(p), e_2^\star(p)\right)$. Finally, if $\delta \in \{\delta_0, 1\}$, $e_2^\star(b) = e_2^\star(p)$ and $\pi_b^2\left(e_1^\star(b), e_2^\star(b)\right) = \pi_p^2\left(e_1^\star(p), e_2^\star(p)\right)$.*

While precise advice is easy to follow, broad advice requires exerting more effort to make use of. Hence, under broad advice, the decision maker is incentivized to exert more efforts in E_1. In the second environment, the decision maker benefits from the prior advice in two ways: either they can directly implement what they have observed, in which case precise advice in E_1 incentivizes increased efforts in E_2. Alternatively, when environments are sufficiently different, effort during E_1 increases in importance, because it induces higher-level, transferable understanding. In this case, broad advice in E_1 incentivizes increased efforts in E_2.

Key Trade-Offs for the Designer. Consider now the designer. We define the *reward gap* (the immediate benefit of precise over broad advice):

$$\Delta_1 := r\left[\pi_p^1\left(e_1^\star(p)\right)\right) - \pi_b^1\left(e_1^\star(b)\right)\right)\Big] = r(\alpha_p - \alpha_b) - \frac{r^2}{k}(\beta_b^2 - \beta_p^2), \tag{4}$$

where $\Delta_1 > 0$ by (A_3). We define the *learning gap* (the long-term benefit of broad over precise advice):

$$\Delta_2(\delta) := r\left[\pi_b^2\left(e_1^\star(b), e_2^\star(b)\right)\right) - \pi_p^2\left(e_1^\star(p), e_2^\star(p)\right)\right)\Big]. \tag{5}$$

Note that $J(p) - J(b) = \gamma\Delta_1 - (1-\gamma)\Delta_2$.

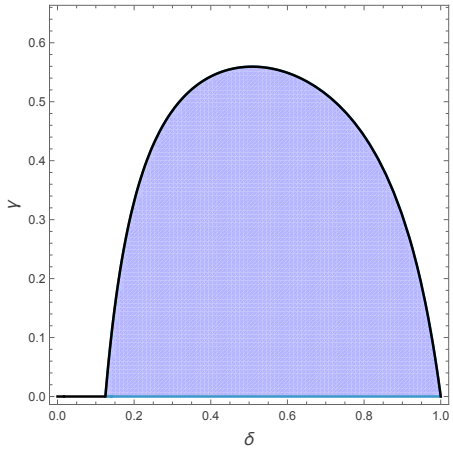
Based on the discussion above regarding decision maker's efforts in E_2, the learning gap is positive when environments are sufficiently dissimilar:

LEMMA 3 (**Learning Gap**). *The learning gap depends on the environmental difference δ:*

- *If $0 \le \delta < \delta_0$, $\Delta_2(\delta) < 0$ and $\Delta_2'(\delta) > 0$.*
- *If $\delta_0 < \delta < 1$, $\Delta_2(\delta) > 0$ and $\Delta_2'(\delta) > 0\ \forall \delta \in (\delta_0, \delta_1)$, $\Delta_2'(\delta) < 0\ \forall \delta \in (\delta_1, 1)$ for some $\delta_0 < \delta_1 \le 1$.*
- *If $\delta \in \{\delta_0, 1\}$, $\Delta_2(\delta) = 0$.*

Hence, while there is a direct benefit of precise advice (it can be more easily implemented, which is reflected in the *reward gap*), there can be an indirect benefit of broad advice, in that it leads to increased performance in the second environment. This is the *learning gap*. Overall, the designer selects the type of advice based on how much value can be generated in either of the environments:

Figure 1 Optimal design choice: precise versus broad.



Note. The designer optimally chooses *broad* advice within the purple region, and *precise* advice outside. Other parameters are $r = 1$, $k = 2$, $\alpha_p = 0.5$, $\alpha_b = 0.3$, $\beta_p = 0.1$, $\beta_b = 0.6$ and $\lambda = 0.7$.

PROPOSITION 1 **(Optimal Advice)**. *Define*

$$\gamma^\star(\delta) = \begin{cases} 0, & \text{if } \delta \in [0, \delta_0], \\ \frac{\Delta_2(\delta)}{\Delta_1 + \Delta_2(\delta)}, & \text{if } \delta \in (\delta_0, 1]. \end{cases}$$

Then the designer optimally chooses b if and only if $\gamma < \gamma^\star(\delta)$.

Figure 1 displays when the designer optimally chooses to provide broad advice. If the environments are very similar (δ low), the learning gap is negative and the designer never chooses broad. On the other hand, if the environments are somewhat dissimilar, the learning gap can easily outweigh the reward gap, as long as the designer places enough weight on rewards in the second environment (γ low). As the environments become too different, the learning gap shrinks again and the designer prefers precise advice.

Finally, we consider how the designer's threshold changes with the decision maker's characteristics. As would be expected, the region in which broad advice is optimal shrinks when the effectiveness of precise advice increases (either directly through α_p or indirectly through β_p) and expands when the effectiveness of broad advice increases. Moreover, the region shrinks when the cost of effort is increasing, because broad advice's learning advantage is channeled through the decision maker's effort:

PROPOSITION 2 **(Comparative Statics)**. *For any* $\delta \in (\delta_0, 1]$*:*

- $d\gamma^\star/d\alpha_p < 0$, $d\gamma^\star/d\beta_p < 0$*: higher effectiveness of precise advice shrinks the region for b.*
- $d\gamma^\star/d\alpha_b > 0$, $d\gamma^\star/d\beta_b > 0$*: higher effectiveness of broad advice expands the region for b.*
- $d\gamma^\star/dk < 0$*: higher effort cost shrinks the region for b.*

Inventory example. Recall the example of an inventory manager for a single-item periodic-review system. Under common assumptions, a deviation from the optimal base stock $S^\star$ only leads to a square-root deviation

in the expected costs. Hence, if E_2 only differs slightly from E_1, the decision-maker may perform better in E_2 after having observed precise advice, because simply using $S^\star$ from E_1 will provide a near-optimal solution in E_2. On the other hand, identifying the exact optimum may be harder in E_1 under broad advice. However, if the manager uses the advice to understand the right approach, they may be more successful at setting a sensible base-stock level in E_2 when the environment differs substantially.

2.2. Sequential Decision Making: Increasing the Importance of Broad Advice

We now consider a sequential decision-making task, showing that the learning gap favoring broad advice tends to increase compared to a one-shot task.

Modifications in Setup and Notation. The decision maker again faces two environments:

E_1: In this environment, the decision maker is faced with an MDP $\mathcal{M}_\theta = (\mathcal{S}, \mathcal{A}, P_\theta, r_s(\cdot), s_1)$ with horizon $T \in \mathbb{N}$. At $t = 1, \ldots, T$, the system is in state $s_t \in \mathcal{S}$, an action $\tilde{a}_t \in \mathcal{A}$ is chosen by the decision maker, the next state is drawn from $P_\theta(\cdot \mid s_t, \tilde{a}_t)$, and an immediate reward $r_{s_t}(\tilde{a}_t)$ accrues. The decision maker is supported by algorithmic advice, and the machine generating this advice has access to the environment parameter θ and the θ-optimal policy $\nu^*_\theta : \mathcal{S} \to \mathcal{A}$, such that $\nu^*_\theta \in \arg\max_{\nu_\theta} \mathbb{E}\left[\sum_{t=1}^{T} r_{s_t}(\nu_\theta(s_t))\right]$

E_2: In this environment, the decision maker faces a subsequent decision-making task without access to advice. This may be a similar MDP, or another type of task. Importantly, the decision maker's performance will depend on knowledge formed in E_1.

Recall that, in our experimental setup, we also have an initial environment E_0 similar to E_2, without access to advice, but no prior advice exposure. Moreover, each environment can contain multiple MDPs.

We can continue to rely on our inventory management example to illustrate E_1. However, we now assume that the inventory manager considers a system under non-stationary demand, aiming to find an optimal policy accounting for the current state and future demand predictions.

Advice, effort, and compliance in E_1. In each E_1 period, the decision maker has access to algorithmic advice of type $a \in \{p, b\}$, where p is *precise* and b is *broad*. The decision maker (possibly forward-looking within environment E_1) chooses to exert effort $e_t \in [0, 1]$ in period t at cost $\frac{k}{2}e_t^2$. Conditional on e_t, the probability of following the advice and choosing the best action is $\pi_a(e_t) = \alpha_a + \beta_a e_t$, where we again assume $0 < \alpha_b < \alpha_p < 1$ (precise advice is easier to follow) and $0 < \beta_p < \beta_b < 1$ (broad advice provides strategic insight, helping to better convert effort into results). Again, we assume $\alpha_a + \beta_a < 1$.

Let $e^\star_{t,a}$ denote the (possibly time-varying) effort choices of the human decision maker in E_1 under regime $a \in \{p, b\}$. We define the induced "compliance" as $c_{t,a} := \pi^t_a(e^\star_{1,a}, \ldots, e^\star_{t,a})$.

Executed action in E_1. If the decision maker complies at t, they execute $\nu^\star_\theta(s_t)$. We assume that, in the case of non-compliance, the decision maker uses a fixed fallback policy $\mu : \mathcal{S} \to \Delta(\mathcal{A})$, corresponding, e.g., to a habitual heuristic (see, e.g., Grand-Clément and Pauphilet 2024). Thus, the executed action at time t is

$$\tilde{a}_t = \begin{cases} \nu^\star_\theta(s_t), & \text{with prob. } c_{t,a}, \\ A \sim \mu(\cdot \mid s_t), & \text{with prob. } 1 - c_{t,a}. \end{cases}$$

Information in E_1 and learning measure. Let θ be a finite-dimensional parameter that governs the transition probabilities P_θ, and possibly also the reward function r_s. Assume the per-period observation (s_t, a_t, s_{t+1}) contributes *state-action-specific information* about θ, measured by a nonnegative function $\mathcal{I}(s_t, a_t)$. This could be, e.g., the Fisher information density. The decision maker's effort affects how much of this information is internalized: We define the *cumulative learning up to T* under regime a as $L_T(a) := \mathbb{E}_a\left[\sum_{t=1}^T \beta_a e^\star_{t,a} \mathcal{I}(s_t, \tilde{a}_t)\right]$. Finally, we denote with $\mathcal{S}_{\text{inf}} \subseteq \mathcal{S}$ the set of states for which there exists at least one action with strictly positive information: $\max_{\tilde{a} \in \mathcal{A}} \mathcal{I}(s, \tilde{a}) > 0$. We assume $\mathcal{S}_{\text{inf}} \neq \varnothing$

Parametric assumptions. We make the following assumptions:

(A'_1) $c_{t,p} \geq c_{t,b}$ for all $t \in \{1, 2, \ldots, T\}$;

(A'_2) $\beta_b\, e^*_{t,b} \geq \beta_p\, e^*_{t,p}$ for all $t \in \{1, 2, \ldots, T\}$;

(A'_3) The ν^*_θ-induced Markov chain starting from s_1 avoids $\mathcal{S}_{\text{inf}}$ almost surely;

(A'_4) Under μ, from any state, the chain reaches $\mathcal{S}_{\text{inf}}$ with positive probability in finite time;

(A'_5) There exists $\bar{I} > 0$ such that, for any s and any $a \in \{p, b\}$, $\mathbb{E}\big[\mathcal{I}(s_t, \tilde{a}_t) \mid s_t = s \in \mathcal{S}_{\text{inf}}\big] = \bar{I}$.

Assumption (A'_1) indicates that precise advice is more likely to be followed. This is the natural extension of (A_2) and means that the simplicity of precise advice outweighs the effort-inducing advantages of broad advice in terms of compliance. Meanwhile, Assumption (A'_2) implies a learning benefit from broad advice, directly extending the condition $\beta_b > \beta_p$ in the one-shot decision case.

Assumption (A'_3) means that the per-period observations (s_t, a_t, s_{t+1}), when following advice exactly, do not effectively contribute to information about θ. E.g., an optimal inventory policy usually keeps the system in a narrow, low-variance region. On the other hand, Assumption (A'_4) says that the human decision maker, if following their fallback policy, will enter informative states with positive probability. That is, because their policy does not rely on θ, it is bound to explore information about θ even if purely coincidentally.

While Assumptions (A'_1)–(A'_4) allow us to define a sequential decision-making task with the same key characteristics as the previously defined one-shot task, Assumption (A'_5) supports tractability, by equalizing the information obtained within informative states, independent of the action chosen.

Length of Horizon Amplifies Learning under Broad. Here, we compare the effect of learning from a system that always provides precise advice in E_1, with one that always provides broad advice.

THEOREM 1 **(Horizon Impact)**. *In E_1, for the stationary advice regimes* always-p *and* always-b*:*

(i) For every horizon $T \geq 1$, $L_T(b) \geq L_T(p)$.

(ii) The difference $\Delta L_T := L_T(b) - L_T(p)$ is increasing in T.

Theorem 1 shows that the learning gap is more pronounced in sequential tasks specifically when these tasks have a longer time horizon. This further sharpens the trade-off between immediate rewards and long-term effects of algorithmic advice in such a context.

Although the learning gap increases with the length of the sequential task, the total advantage of broad over precise might be decreasing in the length. This is because the immediate advantage of precise advice (reward gap) during E_1 is usually also increasing in the length of the sequential task. Appendix A.2 details the trade-off between precise and broad advice as a function of T.

2.3. Testable Hypotheses from our Model

Our analysis above immediately translates into several testable hypotheses. First, consider the behaviors and outcomes in E_1, where advice is available. Lemma 1 indicates that compliance with advice is higher when that advice is precise, than when it is broad. As a result, there is a positive *reward gap*, favoring precise advice. On the other hand, Theorem 1 stipulates that the decision maker in a sequential decision task tends to explore a broader set of outcomes under broad advice, possibly contributing to learning.

(H_{1a}) **Advice compliance.** Compliance with advice is lower under broad than under precise.

(H_{1b}) **Reward gap.** With-advice (E_1) performance is higher under precise than under broad advice.

(H_2) **Exploration.** Participants observing broad advice visit more different states.

Next, consider the behaviors and outcomes in E_2, when advice is no longer available. Lemma 2 indicates that decision makers perform better after having observed broad advice, *when E_2 is sufficiently different from E_1*. This results in a positive *learning gap*, favoring broad advice. However, as Lemma 3 specifies, the size of this learning gap will eventually decrease, when E_2 is too dissimilar. Hence, we expect the highest benefit from broad advice at an intermediate difference in environments.

(H_{3a}) **Post-advice strategy.** Participant's post-advice (E_2) strategies are closer to the optimal strategy after observing broad compared to precise advice in E_1, if and only if E_1 and E_2 are substantially different.

(H_{3b}) **Learning gap.** Post-advice (E_2), participants' performance is higher under broad than under precise advice, if and only if E_1 and E_2 are substantially different. This gap shrinks for very large differences.

Finally, we consider the design of the advice. Broad advice can support higher-order understanding of a task, which can then be transferred to different contexts. At the same time, it comes with a risk: In an effort to direct attention to the abstract, broad advice may be challenging to fully grasp, hurting the ability to learn. When increasing the accessibility of broad advice, without discouraging deeper engagement (thus, increasing β_b), we should observe higher efforts and performance across environments.

On the other hand, precise advice sacrifices learning for short-term gains from precision. It stands to reason that the designer can retain or improve these short-term gains while obtaining at least some of the learning effects of broad advice, by providing both the precise advice itself (which is easy to implement), and deeper explanations. By increasing β_p without sacrificing α_p, explanations should support decision makers in both environments. If environments are sufficiently similar, the improvement in E_2 from learning outweighs the improvement in E_1.[3] Explanations may also support broad advice, but they are likely to provide some of the same information as the advice itself, so we would expect them to be less beneficial.

(H_4) **Specificity of broad advice.** When broad advice's specificity is increased, participants' performance increases in both E_1 and E_2.

(H_{5a}) **Explanation comparison.** With explanations, participant's performance improves more under precise advice than under broad advice.

(H_{5b}) **Explanation impact.** Only when environments are sufficiently similar, can explanations improve participants' E_2 performance more than their E_1 performance.

3. Experimental Design: Sequential Decision-Making Game

We develop a behavioral experiment to examine human's sequential decision making under algorithmic advice. Participants complete a series of tasks with performance-based incentives, potentially supported by machine-generated advice. Our experiment proceeds in three phases:

E_0 (*pre-adivce phase*): Participants complete two to three rounds of a sequential decision-making task without advice. This baseline captures how individuals refine their strategies through trial and error.

E_1 (*with-advice phase*): Participants complete additional rounds with algorithmic advice designed to improve performance. In this phase, we implement our main treatments, varying advice characteristics (e.g., "precise" vs. "broad") and features of the environment (e.g., realized traffic patterns).

E_2 (*post-advice phase*): Without advice, participants complete additional rounds that remain similar or change substantially, to measure retention, as well as transfer to a new setting.

We implement the experiment in an electric vehicle (EV) charging task because it captures key features of operational decision making while remaining tractable in the laboratory. Route planning is a demanding executive function that requires integrating spatial knowledge, traffic conditions, and safety constraints under time pressure. Modern navigation applications provide real-time route and traffic information, yet users report recurring frustrations with overly detailed or poorly timed guidance that increases cognitive load (Nakhimovsky et al. 2010). As EV adoption grows, charging decisions introduce additional uncertainty

[3] We have $\frac{d\pi_p^2(e_1^\star(a),e_2^\star(a))}{d\beta_a} > \frac{d\pi_p^1(e_1^\star(a))}{d\beta_a} \Leftrightarrow \frac{2r}{k}(1-\delta)\left[\lambda + (1-\delta)^2\alpha_a + (1-\delta)\frac{r\beta_a^2}{k}\right] > 1$. When $\delta = 0$, the left-hand side may be smaller or greater 1. It is decreasing in δ and zero at $\delta = 1$.

about energy consumption and station availability, increasing reliance on algorithmic assistants. This makes EV routing and charging a useful testbed for studying how algorithmic advice can support short-run performance while fostering the situational awareness and strategy formation needed for robust performance when assistance is absent and to supervise algorithmic decisions.

The EV-charging context is particularly apt because it captures some of the key features of complex sequential decision making, described below, while being salient for typical experimental participant.

Sequentiality. A decision-making process is sequential when decisions made now enable or constrain decisions made later. EV charging is inherently sequential: choosing to exit and charge after a road segment changes future charging needs and may make later exits redundant.

Complexity. Complexity makes decision-support tools relevant. With EVs, the charging process introduces non-trivial trade-offs. Power delivery is non-uniform (Neubauer et al. 2012, Nicholas and Hall 2018), implying a "nonlinear cost" of charging. At the same time, exiting the highway, setting up the charger, and paying are largely independent of the amount of charge. With nonlinear variable costs and fixed overhead costs, participants need to weigh frequent small charges against infrequent large charges.

Decision Space. Many sequential decision problems are made more challenging by a large set of feasible actions. Much of the behavioral operations literature on decision support, however, focuses on a limited set of choices (e.g., assigning a task to a worker in Bastani et al. 2026 or chess players choosing their move in McIlroy-Young et al. 2022). In EV charging, the amount of charge is continuous: participants choose a charging amount ranging from 0% to 100%. Even when discretized into percentage points, the range is large enough that we can study how participants translate a qualitative recommendation into a numeric action.

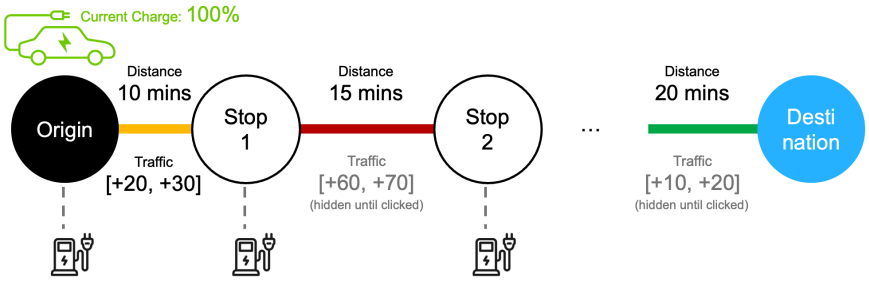
Uncertainty. Uncertainty adds an additional layer of difficulty and is a key driver of deviations from algorithmic recommendations (Dietvorst et al. 2015). In our task, drivers must anticipate uncertain traffic that affects the feasibility of a planned sequence of decisions. Because participants do not observe future traffic, they face a risk of running out of charge. The cost of this can be substantial in practice, as evidenced by the wide availability of EV roadside assistance coverage (Grand View Research 2024, Ford Motor Company 2024), and we capture it through a large penalty in the experiment.

3.1. Details of the Experimental Design

We begin by describing the experimental driving task, then explain how tasks are integrated into the overall game structure. Finally, we describe how we compute the optimal policy and how advice is generated.

Task: Driving an EV to the Destination. The main task requires participants to operate a virtual EV from an origin to a destination along a highway, aiming to minimize elapsed game time (measured in "in-game minutes"). Multiple exits ("stops") segment the highway, connected by road segments with known

Figure 2 Illustration of the experimental task.



Note. Each circle represents a highway exit ("stop") and a line represents a highway segment. Each segment is labeled with its length (in in-game minutes) and, possibly, the traffic range. The car symbol indicates the current location (in this example, the origin), and the current level of charge is displayed next to it.

base lengths. Distances are expressed in in-game minutes. The EV's charge is measured in percentage points, where one percentage point of charge enables one in-game minute of driving.

Traffic adds a random delay to each segment. Participants see a range of possible delays, and the actual delay is drawn from a uniform distribution over that range. For instance, if a segment indicates "[+5,+10]," the realized delay increases the segment's driving time by $\tau \sim U[5, 10]$ in-game minutes. Consequently, the EV needs enough charge for both the segment's base length and the realized delay.

Figure 2 illustrates the task layout. While the base length of every segment is always visible, only the next segment's traffic range is displayed automatically. Traffic ranges for later segments are revealed only if the participant clicks on them. This feature allows us to measure information acquisition efforts.

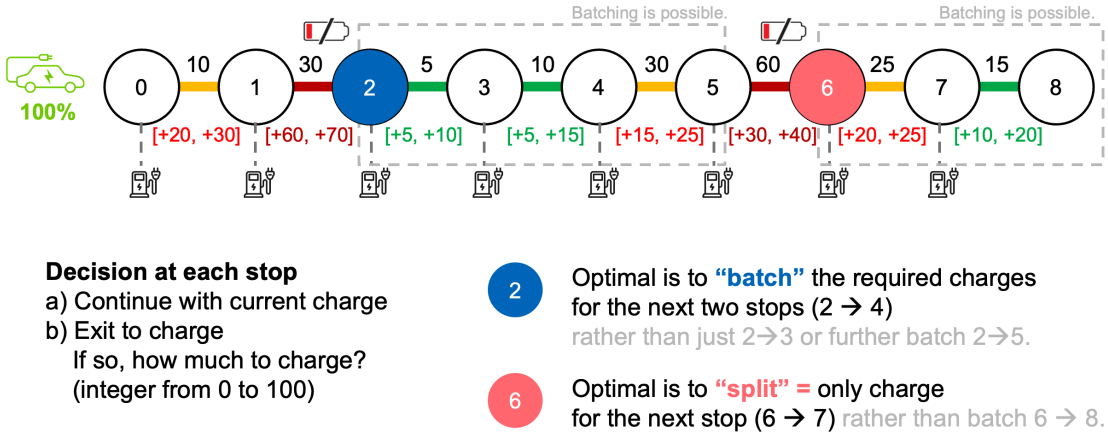
At each stop, a participant chooses whether to proceed to the next segment or to exit and recharge up to a maximum of 100%. Exiting incurs a fixed "exit time" of 30 in-game minutes, representing detour, setup, and payment overhead. If the participant exits, they select how many percentage points of charge to add (an integer from 0 to 100−current charge). Charging time reflects real-world nonlinearity, so adding charge from a low battery level requires less time per percentage point than from a high level. Let current charge be s and additional charge ℓ. Then, charging time is

$$Y(s,\ell) = \begin{cases} f(100) - f(s) + 30, & \text{if } s + \ell \geq 100, \\ f(s+\ell) - f(s) + 30, & \text{if } 0 < \ell < 100 - s, \\ 0, & \text{otherwise,} \end{cases} \quad \text{with} \quad f(x) = \lfloor 0.2 \cdot x^{1.55} \rceil .$$

After a participant commits to a decision, the segment's traffic delay is realized and added to the base length. We assume that exiting to charge does not change the traffic range or the realized delay. If the EV has enough charge to cover the segment's total travel time, the participant reaches the next stop with the starting charge minus the charge corresponding to the total travel time. If not, the participant incurs an *emergency charge penalty* of 300 in-game minutes and arrives at the next stop with 0% charge.

The task is successfully completed once the participant reaches the destination. Elapsed game time is the sum of travel time, exit time, charging time, and any emergency penalties.

Figure 3 Illustration of batching and splitting.



Maps. We refer to a tuple of origin (index 0), destination, intermediate stops, and connecting segments as a *map*. Each round contains a single map. Maps differ by: (i) the number of stops N, (ii) segment lengths $\{d_i\}$, (iii) traffic ranges $\{[+\underline{t}_i, +\bar{t}_i]\}$, and (iv) initial charge s_0. We design each map so that at least one recharge is required to avoid running out of charge and incurring the emergency penalty.

Traffic realizations vary across segments and rounds, even when the map is repeated. Within a given treatment, we hold fixed the sequence of maps and traffic realizations.

Optimal Charging Decisions. For each map, we compute the optimal policy via finite-horizon dynamic programming. The state is the current stop and charge level (0 to 100). At each stop, the decision is whether to exit and, if exiting, the amount of charge to add. The objective is to minimize expected elapsed game time, accounting for charging costs, exit overhead, traffic uncertainty, and the emergency penalty.

The game is designed so that the optimal strategy is simple: At each stop, it is optimal to charge up to the total required charge either for the next segment or for the next two segments, taking into account the worst-case traffic, and not to exit when the current charge exceeds the required charge.[4] This set-up allows us to effectively observe strategies and how they're shaped by advice, while retaining meaningful variation.

We call the decision to charge for one segment only *splitting* and the decision to charge for two (or multiple) consecutive segments *batching*. Each map is designed so that the optimal strategy includes both *splitting* and *batching* decisions. Figure 3 provides an illustration.

Advice Implementation. Across all studies, advice is generated offline from the optimization routine above, and does not incorporate future traffic realizations. Participants receive the same mapping from their state (current location, future segments with lengths and traffic ranges, and charge) to a recommended action—conditions only differ in how this action is communicated, not the underlying policy.

[4] This policy is reminiscent to the optimal policy in the classic dynamic lot-sizing model by Wagner and Whitin (1958), with the worst-case traffic replaced by known demand.

Advice Precision as Message Granularity. Our main manipulation varies advice *precision*, i.e., how directly the message relates to a concrete action. *Precise* advice provides an action-level prescription (e.g., whether to exit at the current stop and, if yes, a specific amount to charge). *Broad* advice provides a strategic guideline that participants must translate into a specific charge amount, given the map and traffic information. Study 1 implements a two-level version of this manipulation. Study 2 extends it by adding an intermediate level that increases strategic specificity without providing a numeric target.

Procedure and Study Flow. Participants first read instructions and play practice rounds to understand the task. They need to pass a comprehension check before proceeding. After the main task, they answer questions about their strategy, prior experience with EV charging, and demographics. Participants, recruited on Amazon Mechanical Turk and Prolific, receive a participation payment and earn performance bonuses.

Before running the main studies on learning and transfer, we conducted a pilot experiment that varied only the advice precision (precise vs. broad) in the same EV charging task, and did not include a post-advice phase. The results indicates a clear short-run benefit of precision: participants receiving precise advice improve completion time and their decisions concentrate tightly around the near-optimal strategy at a key batching stop. Meanwhile, broad advice yields substantially more dispersion in actions and outcomes. These patterns motivate our focus in the main studies on whether short-run efficiency comes at the expense of longer-run learning and transfer once advice is removed. Details of the pilot are reported in Appendix B.

4. Study 1: Learning and Transfer Across Environmental Similarity

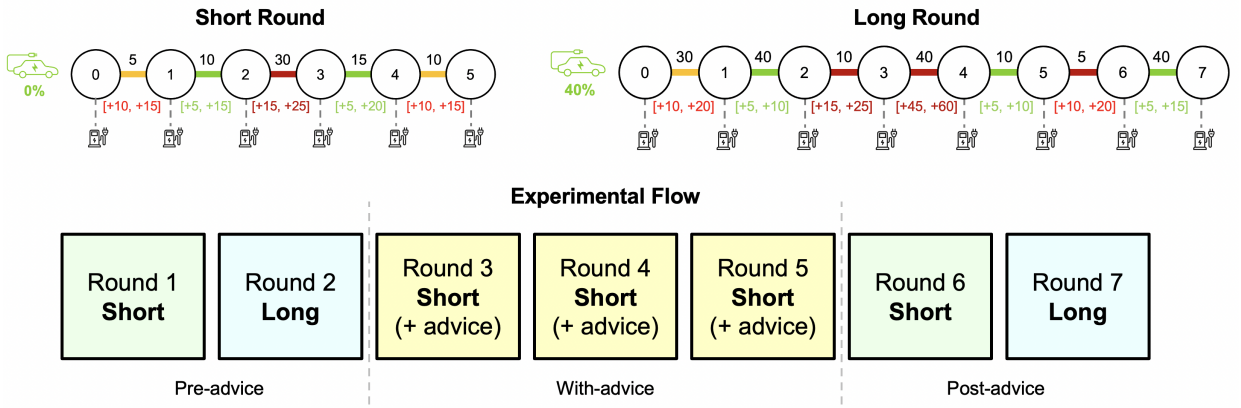
Study 1 tests our core theoretical predictions regarding the performance-learning trade-off. The design varies the precision of the advice (precise vs. broad) and the environmental similarity between the advice phase and the post-advice phase through traffic realization patterns. This allows us to test when broad advice's learning benefits materialize and when environmental differences become too large for effective transfer.

4.1. Study Design and Details

We adopt a 2×2 factorial design varying advice precision and task difficulty, while differentiating environmental similarity within-subject through two different post-advice rounds.

Advice Precision. This is manipulated through the content of the advice shown at each decision point during the with-advice phase. The *precise* advice provides an action-level recommendation, indicating whether the participant should exit at the current stop and, if so, the recommended charge level as a percentage of battery capacity (e.g., "Exit at this stop and charge 60%"). The *broad* advice provides a qualitative charging rule (e.g., "Charge enough to cover this segment and the next"). The timing and frequency of advice are held fixed across conditions; only message precision varies. This manipulation directly tests Hypotheses H_{1-2} regarding compliance, with-advice performance, and exploration.

Figure 4 Maps and study flow for Study 1.



Task Difficulty via Traffic Predictability. Real-world environments vary in complexity and margin for error. To test robustness across different operational complexities, we manipulate traffic realization patterns as a between-subject factor. In the *simple* traffic condition, realized traffic clusters toward the boundaries of the announced range (consistently extreme), making patterns more predictable and improving the performance of typical heuristic approaches. In the *complex* traffic condition, realized traffic centers around the middle of the range with greater variability, increasing decision difficulty. Importantly, both conditions present the same maps, including traffic ranges, so they do not affect environmental difference (δ) between phases.

Environmental Difference. To test the learning consequences of advice precision, participants complete two post-advice rounds, operationalizing different levels of environmental difference.

Round 6 presents the *short map*, the same five-exit map structure that participants experienced during the advice phase (same distances and announced traffic ranges), but without advice and with different traffic realizations. This represents a minimal environmental change, corresponding to low δ. Round 7 presents the *long map*, a structurally different map with seven exits instead of five, resulting in longer travel distances, more charging opportunities, and different batching trade-offs. Participants saw this map once in the pre-advice phase (Round 2) but never received advice for it. This represents substantial environmental change: not only does traffic vary, but the decision problem itself has different structure, corresponding to higher δ.

This within-subject design allows us to test Hypothesis H_{3b}: the learning gap favoring broad is positive only when environmental differences are sufficiently large, i.e., not in Round 6, where even memorized approaches may perform well, but in Round 7, where good performance requires complex transfer.

Maps and Study Flow. The study uses two distinct maps: *short* and *long* (Figure 4). The short map begins at 0% initial charge, forcing participants to charge at the origin, which is the only optimal batching location. The long map begins at 40% initial charge and features multiple charging opportunities, with an optimal strategy requiring batching at Exit 4. The study proceeds in three phases:

- *Pre-advice phase (E_0, Rounds 1–2):* Participants complete one short-map round and one long-map round without advice, establishing baseline behavior.
- *With-advice phase (E_1, Rounds 3–5):* Participants receive either precise or broad advice for three short-map rounds, providing a test for H_{1a-b} and H_2.
- *Post-advice phase (E_2, Rounds 6–7):* Advice is removed and participants complete one additional short-map round (small δ) and one long-map round (larger δ), providing a test for H_{3a-b}.

Within each traffic condition, we hold fixed the sequence of maps and traffic realizations to ensure that differences across advice conditions reflect learning rather than luck.

This study was administered via Amazon Mechanical Turk, with 90 participants completing the task (40% were female, 66.67% were aged 25–44, 95.56% held a valid driver's license, and 91.11% owned a car). The average payoff was $5.14, and the median (resp. mean) completion time was 39.5 (resp. 50.67) minutes.

4.2. Results

Our primary outcome is *normalized performance*: a participant's actual improvement in completion time from the worst case divided by the optimal improvement, where higher values indicate better performance.

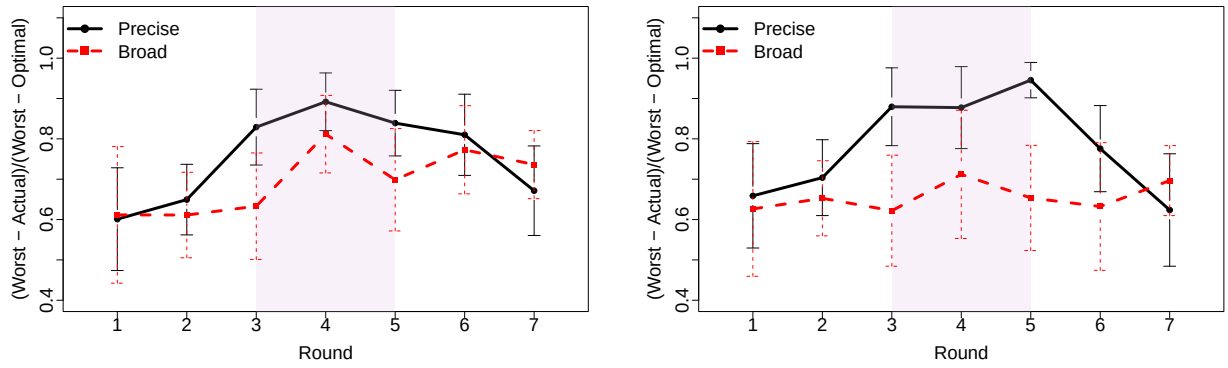
Performance: Precise Advice Helps Immediately, Broad Advice Helps Transfer. Figure 5 summarizes normalized performance. During the pre-advice phase (Rounds 1–2), there are no statistically significant differences between advice conditions, confirming successful randomization.

With-Advice Performance (Rounds 3–5). In E_1, precise advice substantially outperforms broad advice, supporting Hypothesis H_{1b}. The separation is clearest by the final round (Round 5), particularly in the complex traffic condition (Figure 5b): under precise advice, the normalized performance has mean = 0.95, $SD = 0.10$. Under broad advice, mean = 0.65, $SD = 0.26$ (Welch's $t(21.02) = 4.47$, $p < 0.001$, two-sided). This 30-percentage-point gap represents the reward gap at its peak. The pattern holds in the simple traffic condition as well (Figure 5a), though the gap is somewhat smaller.

This immediate performance advantage stems from precise advice providing explicit, executable actions. When told to "charge 60%," participants face minimal cognitive burden. When told to "charge enough for this segment and the next," participants must interpret current traffic ranges, estimate required charge, and decide on safety margins. The compliance and exploration mechanisms behind this gap are examined later.

Post-Advice Performance (Rounds 6–7). After advice is removed, precise advice participants still outperform broad advice participants in both traffic conditions on the short map without advice (Round 6). This indicates that the minimal change (i.e., same map with only different traffic realizations) represents such low environmental difference (δ) as to favor precise advice. On the long map (Round 7), for which participants never receive advice, the performance relationship changes, with participants previously exposed to broad

Figure 5 Normalized performance across rounds.



(a) *Simple* traffic condition (b) *Complex* traffic condition

Note. Normalized performance is the actual improvement in completion time from the worst case divided by the optimal improvement in completion time for that round (higher is better). Shaded regions represent the three rounds (3–5) in which participants received advice based on their treatment condition.

advice achieving higher performance. In the simple condition, the means are 0.74 versus 0.67 (Welch's $t(41.08) = 0.961$, $p = 0.342$, two-sided); in the complex condition, broad advice yields a mean of 0.70 versus 0.62 for precise (Welch's $t(36.48) = 0.923$, $p = 0.362$, two-sided).

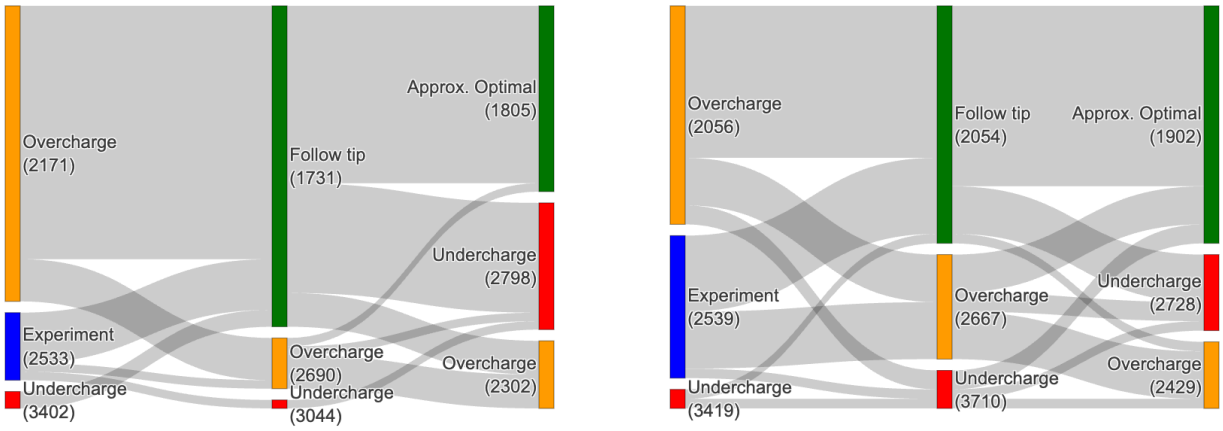
While these differences do not reach conventional statistical significance thresholds, their consistency across conditions, and the difference between Round 6 and Round 7, provide suggestive evidence for Hypothesis H_{3b}: broad advice can yield higher post-advice performance if environments differ sufficiently.

Behavioral Evidence: Compliance, Exploration, and Internalization. To understand the mechanisms behind these performance patterns, we examine how advice types shape behavior during and after the advice phase. We classify each charging decision relative to the time-minimizing strategy (see Appendix B.4 for methodology) and use sequence clustering to identify participants' strategy profiles in each phase.

Compliance and Exploration during the Advice Phase. Figure 6 tracks how participants move between strategy clusters. During the with-advice phase, precise advice induces a tight behavioral concentration around recommended actions. Participants cluster in strategies that we label "Follow tip," exhibiting near-optimal behavior that closely adheres to algorithmic guidance. This pattern supports Hypothesis H_{1a} (higher compliance under precise advice) and explains the strong contemporaneous performance advantages.

Under broad advice, behaviors remain substantially more dispersed. Participants fall into multiple clusters, including a large number who generally overcharge. This dispersion corresponds to greater state-space exploration, supporting Hypothesis H_2. In Appendix C.1, we quantify exploration using the coefficient of variation (CV) of *aftercharge* (i.e., battery level after charging but before the next segment). Figure EC.7 shows that, during Rounds 3–5, participants assigned to broad advice exhibit a systematically higher CV

Figure 6 Study 1 cluster transitions from pre-advice to with-advice by advice condition.



(a) Precise advice

(b) Broad advice

Note. The left / middle / right bar indicates the cluster to which a participant is assigned in the pre-advice / with-advice / post-advice phase. Numbers in parentheses report average in-game time within clusters, further validating the cluster labels.

than those assigned to precise advice, with the difference significant by Round 5 ($p < 0.05$, one-sided t-test). Greater exploration means that participants experience more diverse state-action pairs. This variation provides richer information about task structure, the trade-offs between charging frequency, overhead costs, and robustness to traffic uncertainty, which becomes valuable when advice is removed.

Internalization versus Transient Compliance. Clustering analysis reveals lasting differences in how participants internalize advice. Figure 6 shows distributions of the post-advice strategy conditional on with-advice behavior. Among participants who exhibited near-optimal behavior ("Follow tip" cluster) with advice, retention rates differ markedly by advice type: 76% of participants who received broad advice maintain near-optimal strategies in the post-advice phase, compared to only 55% of those who received precise advice.

This 21-percent retention gap provides behavioral support for Hypothesis H_{3a}: broad advice fosters internalization such that post-advice strategies can be closer to optimal. Participants who receive precise advice execute well *while the advice is available*, but many revert to inconsistent decision-making once the advice disappears. Participants who receive broad advice must engage more deeply during the advice phase to determine appropriate actions, and this engagement produces persistent strategic knowledge.

Residual charge analysis (Appendix C.2) clarifies that this post-advice advantage does not stem from uniform shifts in risk tolerance. We examine the median battery remaining upon arrival at exits, comparing pre-advice (Round 1) to post-advice (Round 6). We do not detect significant differences in how median residual charge changes. Instead, the transfer benefit appears to arise from participants developing a more flexible, context-sensitive understanding of when to batch charges versus when to split, strategic knowledge that generalizes when environments change moderately, but not when they change drastically.

5. Study 2: Advice Granularity, Explanations, and Transfer

Study 2 extends Study 1 in three ways. First, it varies environmental differences more explicitly, providing a stronger test of Hypothesis H_{3b}. Second, it introduces an intermediate advice condition that increases specificity within the qualitative format, allowing us to test Hypothesis H_4. Third, it explores adding explanations to the advice, in line with Hypotheses H_{5a-b}.

5.1. Study Design and Details

Study 2 adopts a $2 \times 3 \times 2$ between-subject factorial design along three dimensions: transfer distance, advice granularity, and explanation visibility. Participants complete seven rounds of the EV driving-and-charging game. Rounds 1–2 use a short map with five exits and no advice (pre-advice phase). Rounds 3–4 present the same short map with advice according to the assigned treatment (with-advice phase). Rounds 5–7 remove advice to measure retention and transfer (post-advice phase).

Transfer Distance. To examine whether participants can apply learned principles across environmental changes, we manipulate the degree of transfer required in the post-advice phase:

- *Familiar sequence:* Participants see the same short map again in Rounds 5–6, then face a modified version of the short map with different traffic ranges in Round 7. This sequence tests retention and adaptation under moderate environmental change.
- *Unseen sequence:* Participants switch to a longer map with seven exits for Rounds 5–7. This sequence tests transfer to a very different environment, with a larger state space.

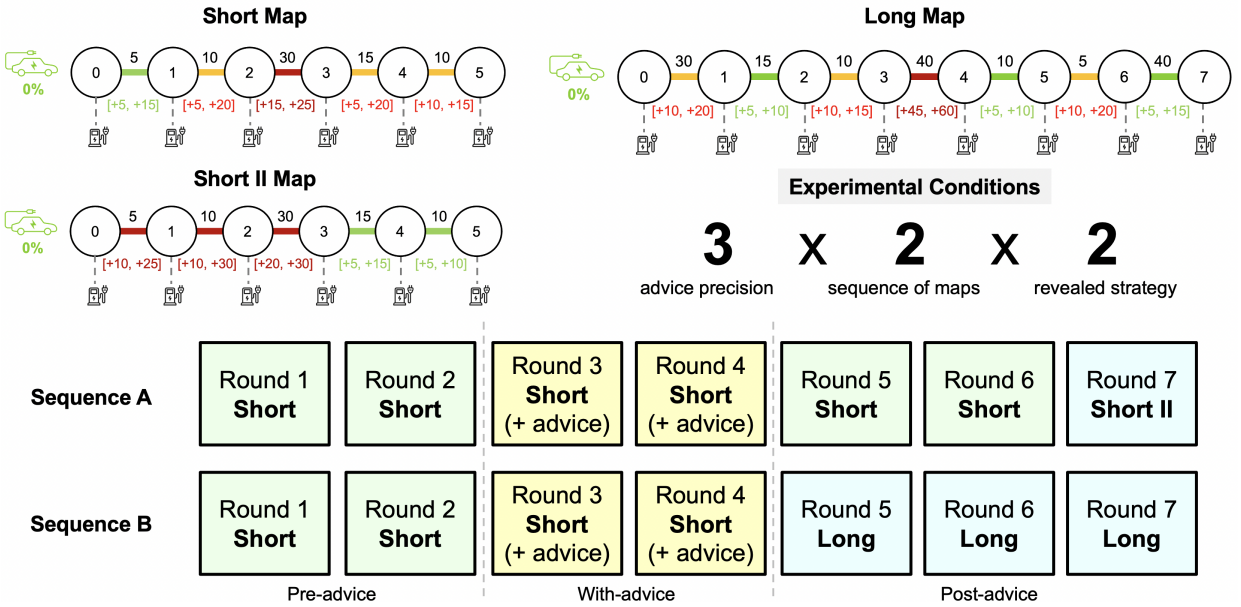
This manipulation directly operationalizes the environmental difference δ: the partially familiar sequence corresponds to intermediate δ, while the unseen sequence corresponds to higher δ. Hypothesis H_{3b} predicts that the learning gap favoring broad advice should be positive when differences are sufficiently large, but attenuated or absent when environmental differences become too large for transferring insights.

Advice Granularity. We expand the precision manipulation to three levels. Again, the conditions differ only in how recommendations are communicated, not in the underlying (optimal) policy.

- *Precise:* E.g., "Exit and charge [X]%." This is identical to the *precise* condition in Study 1.
- *Specific broad:* E.g., "Charge enough for this segment and the next, assuming worst-case traffic." This condition adds specificity, explicitly invoking the worst-case traffic contingency, while remaining qualitative. It represents an intermediate point on the precision continuum, increasing the implementability of broad advice without eliminating the need for participants to translate guidance into an action.
- *Broad:* E.g., "Charge enough for this segment and the next," identical to *broad* in Study 1.

This three-way distinction allows us to separate the effect of adding strategic guidance from the effect of providing an explicit numeric target. If increasing specificity within the qualitative format (specific broad vs. broad) yields performance gains, it suggests that performance improvements can be operationalized through

Figure 7 Maps and Study Flow for Study 2.



targeted clarifications rather than only via numeric precision (H_4). Otherwise, it could mean that removing the action-mapping burden is critical for short-term performance gains.

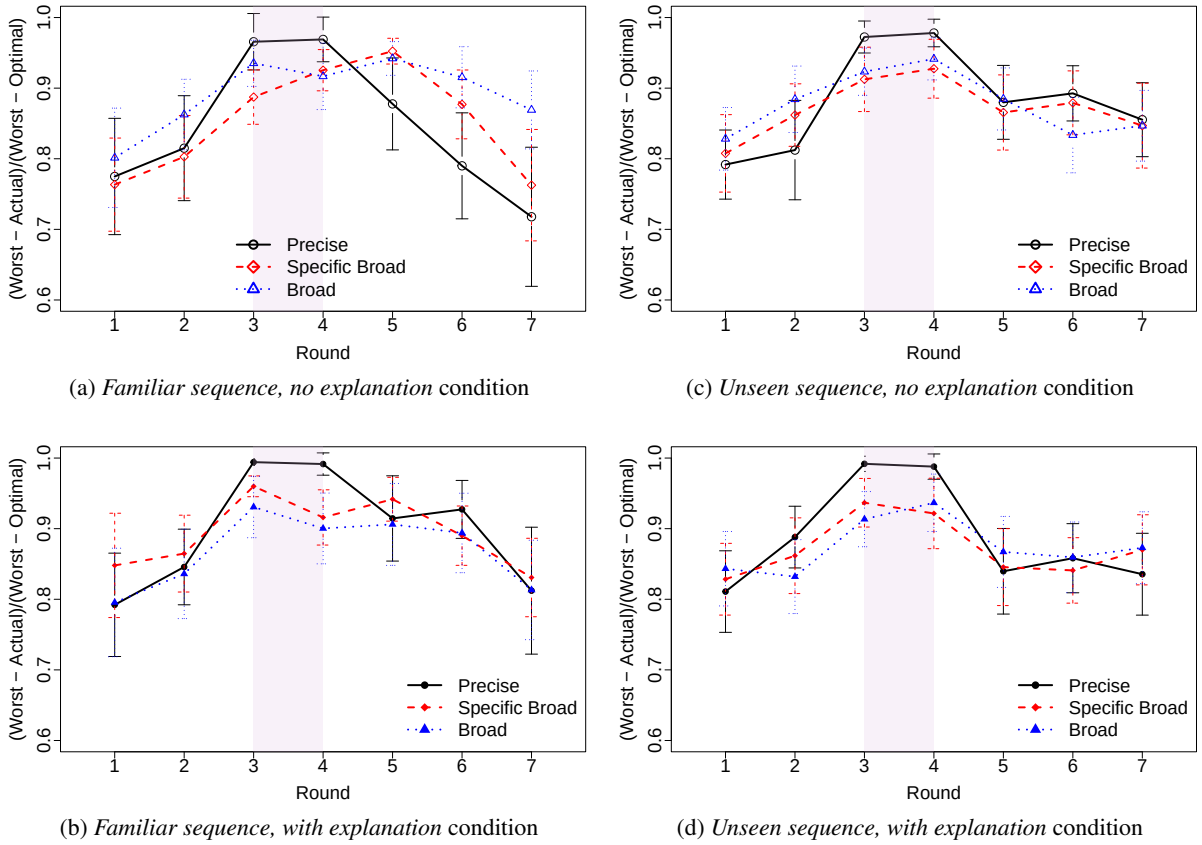
Explanation Visibility. Finally, we vary whether each tip is accompanied by a brief rationale:

- *Hidden:* Participants see only the tip (e.g., "Charge enough for this segment and the next").
- *Revealed:* Participants see the tip and a short explanation (e.g., "Charge enough for this segment and the next. This minimizes total charging time by reducing the number of stops, since exiting incurs a 30-minute overhead").

Prior work on algorithmic explainability points in competing directions. Explanations can improve engagement, calibration, and trust (Buçinca et al. 2021, Ghai et al. 2021), but they can also increase cognitive load or anchor users on reasoning they cannot evaluate (Rudin 2019). Hypothesis H_{5a} posits that explanations improve performance more under precise advice than under broad advice, because explanations provide some of the strategic context that broad advice already encourages participants to infer, thus compressing the difference across advice types. Hypothesis H_{5b} further predicts that explanations should boost post-advice (E_2) performance more than with-advice (E_1) performance, but only when environments are sufficiently similar for the explained principles to remain applicable.

Figure 7 illustrates the study design and flow. The study was conducted on Prolific, with 400 participants completing the task and passing attention and comprehension checks. Among these, 201 participants were assigned to the familiar sequence and 199 to the unseen sequence. The sample consisted of 49.5% female participants, 57% aged 25–44, 89% holding a valid driver's license, and 81.25% owning a car.

Figure 8 Normalized performance across rounds and treatment conditions.



(a) *Familiar sequence, no explanation* condition

(c) *Unseen sequence, no explanation* condition

(b) *Familiar sequence, with explanation* condition

(d) *Unseen sequence, with explanation* condition

Note. Normalized performance is defined as the ratio between the actual improvement in completion time compared to the worst case for the round and the optimal improvement in completion time (higher is better). Shaded regions represent the two rounds (3 and 4) in which participants received advice based on their treatment condition.

5.2. Results

Our primary outcome is again *normalized performance*, defined as completion time divided by optimal time, where higher values indicate better performance.

The Reward Gap: Numeric Precision versus Strategic Specificity. Figure 8 plots normalized performance across all experimental conditions, with shaded regions marking with-advice Rounds 3–4. During the advice phase, precise advice yields the strongest performance improvements across all conditions. By Round 4, precise advice achieves normalized performance of 0.98, substantially outperforming both specific broad (0.92) and broad (0.92), supporting Hypothesis H_{1b}.

The key test of Hypothesis H_4 is whether specific broad outperforms broad. We detect no statistically significant difference (mean difference < 0.001, $t = 0.02$, $p = 0.98$), indicating that H_4 is not supported in E_1. Explicitly invoking worst-case traffic contingencies does not improve immediate performance, suggesting that the reward gap between precise and broad advice stems from the translation burden inherent in any

qualitative guidance rather than from insufficient specificity. Analysis of charging decisions confirms the interpretation: both broad advice types yield similarly dispersed effective-charge distributions during the advice phase, indicating comparable exploration (H_2) and similar compliance patterns (H_{1a}).

Transfer and Environmental Similarity. The post-advice phase (Rounds 5–7) tests when broad advice's learning advantages materialize. We analyze the two transfer sequences separately.

In the familiar sequence with hidden explanations (Figure 8a), the post-advice phase reveals a striking reversal. While precise advice dominates during the advice phase, it exhibits the sharpest performance decline once advice is removed. We quantify this using *retention ratios*: each participant's normalized performance in Round 5 (first post-advice round) divided by their Round 4 performance (final with-advice round). Median retention ratios are 1.00 for broad, 1.01 for specific broad, and 0.97 for precise, with distributions differing significantly (Kruskal-Wallis $H = 16.5$, $p < 0.001$). Post-hoc pairwise tests (Wilcoxon rank-sum) confirm that precise differs significantly from both broad ($p = 0.009$) and specific broad ($p < 0.001$), while the two broad conditions do not differ from each other ($p = 0.23$).

By Round 7, when traffic ranges shift in addition to realizations, both broad conditions maintain their advantage. Mean normalized performance is 0.87 for broad, 0.76 for specific broad, and 0.72 for precise (Kruskal-Wallis $H = 6.2$, $p = 0.045$). Pairwise tests confirm that broad significantly outperforms precise (Wilcoxon $p = 0.015$), though specific broad does not differ significantly from precise ($p = 0.35$). These findings strongly support Hypothesis H_{3b}: at intermediate environmental differences, post-advice performance is higher under broad than under precise advice. They also provide evidence for H_{3a}: participants who received broad advice show strategies closer to optimal in the post-advice phase.

Regarding H_4, we do not find evidence that specific broad outperforms broad in E_2. Both conditions show superior performance relative to precise, but increased specificity does not provide additional learning advantages beyond what broad qualitative advice already provides. The similar retention ratios and Round 7 performance across the two broad conditions suggest that participants interpret qualitative guidance conservatively regardless of whether worst-case scenarios are explicitly invoked, and that the translation burden affects both qualitative formats similarly.

In the unseen sequence (Figures 8c–8d), participants switch to a longer map for Rounds 5–7. Across Rounds 5–7, mean performance is statistically indistinguishable across the three types of advice (0.86 for all conditions, Kruskal-Wallis $p > 0.70$ for each round and explanation condition). This null result highlights the boundary conditions of Hypothesis H_{3b}: When environmental differences are very large (e.g., introducing additional exits, different segment lengths, and more complex batching opportunities), even participants who internalized good strategies during the advice phase struggle to transfer them effectively. The environmental shift exceeds the zone where the learned strategies remain applicable, and the learning gap disappears.

The Role of Explanations. Comparing Figures 8a and 8b reveals how explanations reshape performance across advice types. In the familiar sequence without explanations, post-advice rounds (5–7) exhibit a sharp differentiation: broad advice yields a normalized performance of 0.89, while precise drops to 0.75. When explanations are revealed, all three advice types converge to 0.85–0.87.

This compression reflects asymmetric effects. For precise advice, explanations improve post-advice performance by 11.6 percentage points (from 0.75 to 0.87). For broad advice, explanations yield a modest 3.9-percentage-point *decline* (from 0.89 to 0.85). A linear model confirms this differential: the advice-type-by-explanation interaction is significant (coefficient = 0.155, $p = 0.006$), supporting H_{5a}. Participants receiving broad advice develop strategic understanding through exploration during the advice phase; explicit explanations offer little additional value and may introduce noise. For participants who follow precise numeric targets without building strategic foundations, on the other hand, explanations fill a critical gap.

Hypothesis H_{5b} examines whether explanations boost post-advice (E_2) performance more than with-advice (E_1) performance, and whether this depends on environmental similarity. We compute (E_2 revealed − E_2 hidden) − (E_1 revealed − E_1 hidden). For precise advice in the familiar sequence, this differential is 9.0 percentage points (bootstrap 95% CI [0.005, 0.173], $p = 0.018$), indicating that, when environmental changes are moderate, explanations are more valuable for learning than immediate execution. In the unseen sequence, however, the differential is −4.2 percentage points (95% CI [−0.094, 0.009]), and the contrast is significant (difference = 0.132, $p = 0.004$), confirming H_{5b}'s conditional prediction: explanations particularly support adaptation under moderate environmental change but not under substantial structural shifts.

Table 1 summarizes evidence for all hypotheses across studies. Meanwhile, Appendix D.2 clarifies behavioral mechanisms through residual charge analysis. Precise advice recipients reduce median safety margins more aggressively from pre- to post-advice (mean $\Delta = -9.9$ percentage points) than broad-advice participants (mean $\Delta = -3.0$ pp.; $t = -3.94$, $p < 0.001$). However, in the familiar sequence without explanations, broad advice achieves superior post-advice performance (0.89 vs. 0.75), despite similar median residual charge levels (51.5% vs. 46.1%, $p = 0.097$). This indicates that broad's advantage stems from context-sensitive judgment about when to batch versus split charges rather than from uniform shifts in conservatism.

6. Recovering Strategies using Inverse Reinforcement Learning

While our analysis indicates that decision makers facing sequential problems behave in line with our hypotheses in Section 2.3, we now turn to measuring (and predicting) their specific strategies. Our objective is two-fold: First, to validate our insights into decision makers' interactions with algorithmic advice (specifically, the trade-off between immediate benefits and learning). Second, to showcase inverse reinforcement learning (IRL) as a powerful method for quantifying human behaviors and strategies.

Table 1 Summary of hypotheses and experimental evidence.

Hyp.	Prediction (from Section 2.3)	Study 1	Study 2
H_{1a}	**Compliance.** Compliance with advice is lower under broad than under precise.	✓	✓
H_{1b}	**With-advice performance.** With advice available (E_1), performance is higher under precise than under broad.	✓	✓
H_2	**Exploration.** Participants observing broad advice visit more different states.	✓	✓
H_{3a}	**Strategy retention.** Post-advice strategies are closer to optimal after broad (if E_1, E_2 differ substantially).	✓	✓
H_{3b}	**Learning gap.** Post-advice performance is higher under broad, specifically at intermediate environmental differences.	(✓)	✓
H_4	**Specificity within broad advice.** Increasing broad advice specificity improves performance in E_1 and E_2.	–	✗
H_{5a}	**Explanations help precise more.** With explanations, performance improves more under precise than under broad.	–	✓
H_{5b}	**Explanations enable adaptation.** Explanations boost E_2 more than E_1, only if environments are similar.	–	✓

Note. ✓ = supported, (✓) = partially supported, ✗ = unsupported, – = not tested.

Regarding the second point, there are two crucial challenges for quantifying behaviors in our context. First is the challenge of estimating a decision maker's policy from limited data, in the absence of algorithmic advice. Because practical MDPs tend to have very large state spaces, we are usually only able to observe a few (and sometimes no) actions taken for a given state. Different decision makers may have different policies, and these policies may be probabilistic, so it can be challenging to reconstruct policies from such limited observations. Second is the challenge of disentangling the impact of advice from the decision maker's policy. A decision maker may choose a particular action because it aligns with their fallback policy or because the algorithmic advice indicates this action and the decision maker decides to follow the advice.

We first discuss how we estimate the fallback policy from limited observations absent advice, before detailing how we integrate advice and compliance estimation into our framework. Finally, we apply our estimation approach to our experimental data and describe key insights.

6.1. Estimating the Fallback Policy

Recall that a decision maker faces an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r_s(\cdot), s_1)$. States represent combinations of location and remaining charge, actions correspond to the charge added during a stop, while state transitions depend on charging decisions and traffic realizations. The state space $\mathcal{S}$ and the action space $\mathcal{A}$ are finite. We denote with $\mu : \mathcal{S} \to \Delta(\mathcal{A})$ the policy followed by a human decision maker absent algorithmic advice (the *fallback policy*). A policy is a (probabilistic) mapping from the state space to the action space.

The literature on learning policies from demonstrations provides two types of methods. First, direct methods, i.e., supervised learning techniques to estimate $\mu(\cdot|s_t)$ for states that are frequently observed in the data. However, the state space $\mathcal{S}$ may be large, making it challenging to generalize μ to little-visited states. Second, indirect methods aim at learning the MDP underlying the observed policies. In particular,

IRL assumes that the system dynamics are known, but that the decision maker optimizes their policy based on an unknown reward function (Hanawal et al. 2018).

Our setting is ideally suited for the application of IRL. The system dynamics are well defined, but the literature (e.g., Sun et al. 2022) and our experimental results indicate that decision makers do not decide based on the normative reward function alone. For example, an inventory manager may not maximize the firm's profits but their performance metrics, while also minimizing their (cognitive) effort. An electric vehicle driver may similarly aim at minimizing their cognitive effort and the risk of running out of charge.

In line with the IRL literature (Arora and Doshi 2021), we assume an unknown reward function $r_s : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and discount factor $\gamma \in [0, 1]$. For selected action $\tilde{a}_t$, the decision maker receives reward $\tilde{r}_t = r_{s_t}(\tilde{a}_t)$. The decision maker's policy μ leads to a trajectory $\tau = \langle (s_1, \tilde{a}_1), (s_2, \tilde{a}_2), \ldots, (s_T, \tilde{a}_T) \rangle$. Given the distribution of trajectories D^μ, we can denote the value of the policy by $V^\mu(s_1) = E_{\tau \sim D^\mu}\left[\sum_{t=1}^{T} \gamma^t \tilde{r}_t \mid s_1\right]$. The key assumption is that the policy fulfills $\mu^* \in \arg\sup_\mu V^\mu(s_1)$, given the decision maker's (unknown) reward function. Once the correct reward function has been identified, the policy can be computed. The challenge, then, is that many different reward functions can lead to the same observed behavior.

Arora and Doshi (2021) provide an overview of methods to overcome this challenge. We build on the ideas of Bayesian IRL to create a hierarchy of weights enabling us to capture similarities in idiosyncratic participants, and to derive posterior distributions rather than point estimates. Nevertheless, we also integrate the principle of maximizing the causal entropy (MaxCausalEnt, Ziebart et al. 2010), one of the most established approaches in the field. The intention of MaxCausalEnt is to choose the "least committed" probability distribution over trajectories that is consistent with the observed ones. For tractability, the reward function is usually decomposed as $r_{s_t}(\tilde{a}_t) = \sum_{j=1}^{m} \theta_j \phi_j(s_t, \tilde{a}_t)$, where the weights θ_j are unknown, but the potential individual reward components ϕ_j are known.[5]

To define $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$, we draw on participant feedback from a pilot study, in particular, participants' comments about their strategy and what advice they would have found helpful. Using topic modeling (see Appendix E.1), we extract $m = 7$ behavioral components that capture the main strategic considerations participants describe, including the elapsed game time (the *normative* reward function). Each component is a deterministic function of the current state and a specific action. Components may be weighted positive or negative, so it can be more appropriate to think of some as costs rather than rewards.

1. **Elapsed game time.** Expected time taken for charging, traveling and breakdowns, if any.
2. **Simplicity.** Charging zero or full (1), charging for one segment (0.2), otherwise 0.

[5] While some recent IRL algorithms forgo the linearity assumption (e.g., Baert et al. 2023), we retain it for two reasons. First, it enables interpreting the estimated reward functions. Second, it allows estimating reward functions from relatively few data points. This is important, because each decision maker has their own weights, and these may change as a result of observing advice.

3. **Risk exposure.** Level after charging, minus worst-case requirement for next segment. 0 if negative.
4. **Exposure after penalty.** Same as risk exposure, but only non-zero if participant was previously penalized for breakdown.
5. **Margin over worst.** Opposite of risk exposure.
6. **Batching preference.** Charge sufficient for two segments under worst-case traffic (1), charge sufficient for two segments under best-case traffic (0.5), otherwise 0.
7. **Splitting preference.** Charge no more than for one segment under worst-case traffic (1), otherwise 0.

These components are designed to be (i) behaviorally grounded, (ii) sufficiently rich to capture heterogeneity, and (iii) interpretable so that weights can be mapped to strategy differences.

We estimate the weights $\boldsymbol{\theta}_{sc,i} = (\theta^1_{sc,i}, \dots, \theta^m_{sc,i}) = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_{sc} + \boldsymbol{\Delta}_i$ for each decision maker $i \in \{1, \dots, n\}$ and each scenario $sc \in \mathcal{SC} := \{pre\} \cup \bigcup_{tip \in \{p,pr,s,sr,b,br\}} \{with\text{-}tip, post\text{-}tip\}$. Here, "tip" indicates the type of algorithmic advice displayed to participants, such as "**s**pecific broad" or "**b**road-**r**eveal". Hence, scenarios present combinations of the phase (pre-advice, with-advice, post-advice) and the type of advice that participants have access to in the current phase or in a previous phase.[6] Then, $\boldsymbol{\theta}_0$ is the average weight vector, $\boldsymbol{\Delta}_{sc}$ denotes the *scenario-specific shift* and $\boldsymbol{\Delta}_i$ the *individual shift* for a given decision maker.

6.2. Estimating Compliance with Advice

In line with our stylized model and prior literature (e.g., Grand-Clément and Pauphilet 2024), we assume that, when advice is available, decision makers follow that advice with a certain probability. Otherwise, they rely on their fallback policy. The normative objective is to minimize elapsed game time. Taking this as the (negative) reward function induces an optimal deterministic policy, denoted $\nu^\star$. The algorithmic advice is designed to recommend actions consistent with $\nu^\star$. Hence, given scenario sc, the decision maker chooses

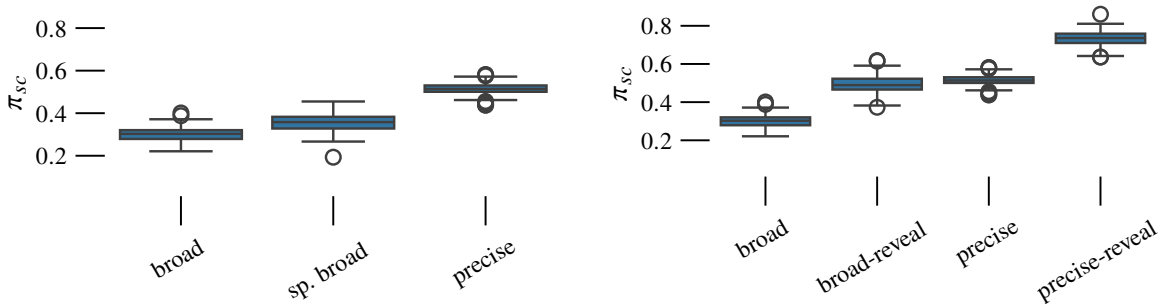
$$\tilde{a}_t = \begin{cases} \nu^\star(s_t), & \text{with prob. } \pi_{sc,i}, \\ A \sim \mu(\cdot \mid s_t), & \text{with prob. } 1 - \pi_{sc,i}. \end{cases}$$

Similar to before, we assume that $\pi_{sc,i} = \frac{1}{1+e^{-(\eta_{sc}+\eta_i)}}$ for decision maker $i \in \{1, \dots, n\}$ and scenario $sc \in \bigcup_{tip \in \{p,pr,s,sr,b,br\}} \{with\text{-}tip\}$ (the compliance probability is irrelevant for no-advice scenarios, so we arbitrarily set it to zero). Here, we use the standard logit-transformation to turn a linear model into a probability. Note that, by including a scenario-specific component in the compliance probability, we allow for the fact that broad advice may be more complex to follow, in line with our stylized model.

Appendix E.2 details our approach to identifying parameter distributions from observations, including the underlying theoretical framework. In Appendix E.3, we validate the approach and address identifiability.

[6] We do not differentiate the pre-advice phase, as participants' experience until its end is indistinguishable by their advice type.

Figure 9 Compliance probability by advice type.



(a) Advice types without explanation

(b) Adivce types with and without explanation

6.3. How Advice Affects Strategies

Advice directly affects participants' strategies through compliance. Hence, we compare the baseline compliance probabilities for different scenarios, that is $\pi_{sc} = \frac{1}{1+e^{-\eta_{sc}}}$.

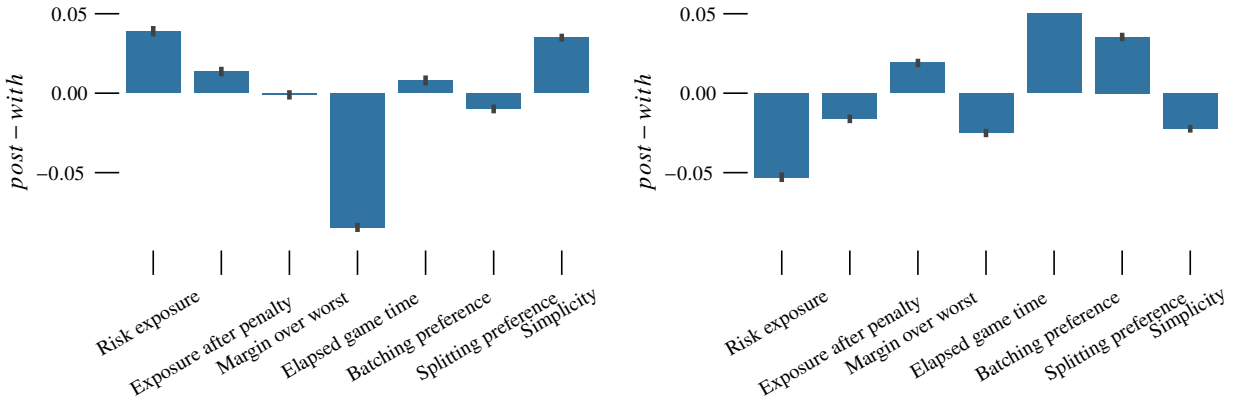
At the same time, algorithmic advice may indirectly affect behavior by inducing a change in participants' strategies. In particular, by observing advice, a participant may learn to improve their own decision making. Hence, the main objects of interest are the average weights for a given scenario, $\boldsymbol{\theta}_{sc} = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_{sc}$. We use these to see how advice changes the relative importance participants assign to certain reward components. To simplify exposition, we compute each component's relative feature importance within a scenario:

$$FI^j_{sc} = \frac{|\theta^j_{sc}|}{\sum_{k=1}^{m} |\theta^k_{sc}|}.$$

Findings. We run our approach on the complete set of experimental results, containing the Pilot Study and Studies 1–2. In total, there are 3,666 trajectories across 13 scenarios by 549 participants. In each trajectory, on average, there are 5.6 state-action pairs. We split trajectories into training (80%) and testing (20%) datasets to evaluate model fit also on real data (see Appendix E.3 for details). We then sample 300 times from each of the scenario-specific posteriors in order to derive a distribution for the metrics defined above.

Figure 9 shows sampled compliance probabilities in the different with-advice scenarios. In line with our prior analysis, *precise* advice leads to substantially higher compliance than *broad* advice, further supporting H_{1a}. The additional specificity of *specific broad* advice, however, only improves compliance slightly (Figure 9a), strengthening the evidence against H_4. Providing a strategic rationale is more powerful: In Figure 9b, we observe that the compliance probability under *broad-reveal* is as high as under *precise*. This positive impact of providing a rationale is not limited to *broad*, however. In fact, the compliance probability under *precise-reveal* increases by a similar number of percentage points. This points to a potential explanation for the higher boost that *precise* recipients receive from the *revealed* condition, besides the added learning benefits: an explanation behind a purely numeric advice may convince some participants of its validity.

Figure 10 Change in relative feature importance FI^j_{sc} from the with- to the post-advice phase.



(a) Precise advice

(b) Broad advice

Figure 10 displays the change of feature importance from pre- to post-advice phase. In line with our hypotheses, participants exposed to precise advice return to a more heuristic decision-making process. In particular, they focus on their risk exposure and show a preference for simple strategies (Figure 10a). On the other hand, participants exposed to broad advice retain a stronger preference for strategic considerations ("batching" and "splitting"), while showing less attention to the risk of running out of charge (Figure 10b).

7. Concluding Remarks

As algorithmic tools become embedded in operational decision making, their designers increasingly face a choice between optimizing for short-run performance and cultivating human capability. This paper shows that the precision of algorithmic advice is a key lever for managing this trade-off in sequential tasks.

On the theory side, we model a system designer, choosing between precise, action-level advice and broad, strategic advice for a human decision maker who first operates with assistance and later faces a related decision environment without. Precise advice is easier to follow, leading to higher immediate rewards, whereas broad advice induces greater effort and more portable learning, improving performance when advice is absent and environments are sufficiently different. Extending the analysis to finite-horizon Markov decision processes, we show that precise advice tends to steer behavior toward a narrow band of states, while broad advice encourages more explorative patterns. Thus, the cumulative stock of learning is higher under broad advice, and this learning advantage grows with the length of the decision horizon.

Our experimental results reinforce and refine these insights. In two studies using an EV charging task, precise advice generates the largest performance improvements during the advice phase, consistent with higher compliance and reduced cognitive effort. However, when advice is removed, participants exposed to broad advice retain more effective batching strategies, particularly in settings requiring substantial transfer, but with some degree of similarity. Hence, precise advice is attractive when environments are stable and

failures of human supervision are limited, whereas broad advice is more appropriate when designers value adaptability, transfer, and the ability of humans to effectively supervise algorithmic decisions. We also find that, while making broad advice more specific does not overcome its short-run disadvantages, adding explanations to precise advice can attenuate its learning-related disadvantages.

The IRL analysis sheds additional light on the mechanisms behind these patterns. By inferring participants' reward weights over interpretable components, we find that broad advice shifts latent objectives toward more strategic criteria that support complex charging plans. In contrast, precise advice mainly drives transient alignment with the system's intended policy, after which participants revert toward simple heuristics and risk-avoidance motives. This distinction underscores that the design of decision support systems influences not only actions, but also what users learn to value, with implications for long-term organizational capability.

Our study opens up avenues for future research. The experimental setting captures core features of operational sequential decisions, while being relatable to lab participants. However, field work is needed to validate the observed patterns in real operational work environments. Moreover, future models could incorporate richer forms of heterogeneity in users' cognitive costs and learning styles, as well as dynamic designers who update advice policies based on observed behavior. Empirically, applying IRL and related methods to large-scale behavioral data from deployed systems would allow researchers to study how algorithmic advice shapes human learning over weeks or months, and to test adaptive advice policies in practice.

References

Argote L, Miron-Spektor E (2011) Organizational learning: From experience to knowledge. *Organization Science* 22(5):1123–1137.

Arora S, Doshi P (2021) A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297:103500.

Baert M, Mazzaglia P, Leroux S, Simoens P (2023) Maximum causal entropy inverse constrained reinforcement learning. *arXiv preprint arXiv:2305.02857* .

Balakrishnan M, Ferreira KJ, Tong J (2026) Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science* 72(1):265–284.

Bastani H, Bastani O, Sinchaisri WP (2026) Improving human sequential decision making with reinforcement learning. *Management Science* 72(1):733–755.

Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R (2025) Generative ai without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences* 122(26):e2422633122.

Bjork EL, Bjork RA, et al. (2011) Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* 2(59-68):56–64.

British Medical Association (2024) Principles for Artificial Intelligence (AI) and its application in healthcare. URL https://tinyurl.com/3amkzw5n.

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5(CSCW1):1–21.

Casner SM, Geven RW, Recker MP, Schooler JW (2014) The retention of manual flying skills in the automated cockpit. *Human Factors* 56(8):1506–1516.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114–126.

Ford Motor Company (2024) Electric vehicle battery roadside assistance. URL https://tinyurl.com/327eunrh.

Ghai B, Liao QV, Zhang Y, Bellamy R, Mueller K (2021) Explainable active learning (XAL) toward AI explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW3):1–28.

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* Articles in Advance.

Grand View Research (2024) Vehicle roadside assistance market (2025–2030): Size, share & trends analysis report. URL https://tinyurl.com/whfzdecd.

Hanawal MK, Liu H, Zhu H, Paschalidis IC (2018) Learning policies for markov decision processes from data. *IEEE Transactions on Automatic Control* 64(6):2298–2309.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Lapré MA, Van Wassenhove LN (2001) Creating and transferring knowledge for productivity improvement in factories. *Management Science* 47(10):1311–1325.

Macnamara BN, Berber I, Çavuşoğlu MC, Krupinski EA, Nallapareddy N, Nelson NE, Smith PJ, Wilson-Delfosse AL, Ray S (2024) Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications* 9(1):46.

March JG (1991) Exploration and exploitation in organizational learning. *Organization Science* 2(1):71–87.

McIlroy-Young R, Wang R, Sen S, Kleinberg J, Anderson A (2022) Learning models of individual behavior in chess. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1253–1263.

Nakhimovsky Y, Miller AT, Dimopoulos T, Siliski M (2010) Behind the scenes of Google Maps navigation: Enabling actionable user feedback at scale. *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 3763–3768 (ACM).

Neubauer J, Brooker A, Wood E (2012) Sensitivity of battery electric vehicle economics to drive patterns, vehicle range, and charge strategies. *Journal of Power Sources* 209:269–277.

Nicholas M, Hall D (2018) Lessons learned on early electric vehicle fast-charging deployments. URL https://tinyurl.com/32vcek5s.

Rabin M, Vayanos D (2010) The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies* 77(2):730–778.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215.

Salden RJ, Koedinger KR, Renkl A, Aleven V, McLaren BM (2010) Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review* 22(4):379–392.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404–420.

Snyder C, Keppler S, Leider S (2025) Algorithm reliance: Fast and slow. *Management Science* Articles in Advance.

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

Sweller J (1994) Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 4(4):295–312.

Tversky A, Kahneman D (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5(2):207–232.

Wagner HM, Whitin TM (1958) Dynamic version of the economic lot size model. *Management Science* 5(1):89–96.

Ziebart BD, Bagnell JA, Dey AK (2010) Modeling interaction via the principle of maximum causal entropy. *International Conference on Machine Learning*, 1255–1262.

E-Companion to *"Designing Algorithmic Advice for Learning in Sequential Decision Making"*

Appendix A: Proofs and Additional Results for Section 2

A.1. Proofs of Mathematical Statements

Proof of Lemma 1. $e_1^\star(b) > e_1^\star(p)$ by $\beta_b > \beta_p$ and $e_1^\star(a) = \frac{r\beta_a}{k}$. Next, $\pi_1^p(e_1^\star(p)) - \pi_1^b(e_1^\star(b)) = (\alpha_p - \alpha_b) - \frac{r}{k}\left(\beta_b^2 - \beta_p^2\right)$, which is non-negative by Assumption (A_2). □

Proof of Lemma 2. From (3), $e_2^\star(b) - e_2^\star(p) = \frac{(1-\delta)r}{k}\left(\frac{r}{k}\left(\beta_b^2 - \beta_p^2\right) - (1-\delta)\left(\alpha_p - \alpha_b\right)\right)$. When $\delta = 1$, this is zero. In the outer parentheses, the first part is positive, while the second part is negative, because $\beta_b > \beta_p$ and $\alpha_p > \alpha_b$. Hence, for $\delta < 1$, the total term is positive if and only if δ is sufficiently large. In particular, it can easily be seen that the sign changes at δ_0. The result for the compliance ordering follows by noting that $\pi_a^2(e_2)$ is linear in e_2. □

Proof of Lemma 3. The sign of $\Delta_2(\delta)$ follows directly from Lemma 2, noting that $\Delta_2(\delta) = \frac{r^2}{k}\left[\left(e_2^\star(b)\right)^2 - \left(e_2^\star(p)\right)^2\right]$. For the derivatives, note first that $\Delta_2^{(4)}(\delta) = -\frac{24r^2}{k}(\alpha_p^2 - \alpha_b^2) < 0$. Moreover,

$$\Delta_2^{(3)}(\delta)\,|_{\delta=\delta_0} = \frac{12r^3}{k^2}\left[\alpha_b(\beta_b^2 - \beta_p^2) + \beta_p^2(\alpha_p - \alpha_b) + 2\alpha_p(\beta_b^2 - \beta_p^2)\right] > 0,$$

so $\Delta_2^{(3)}(\delta) > 0$ for all $\delta < \delta_0$. Next, consider

$$\Delta_2''(\delta)\,|_{\delta=\delta_0} = -\frac{2r^4(\beta_b^2 - \beta_p^2)}{k^3(\alpha_p - \alpha_b)}\left[4(\alpha_p - \alpha_b)\beta_b^2 + 4(\beta_b^2 - \beta_p^2)\alpha_b + (\alpha_p + \alpha_b)(\beta_b^2 - \beta_p^2)\right] - \frac{4r^2\lambda(\alpha_p - \alpha_b)}{k} < 0,$$

so $\Delta_2(\delta)$ is concave for all $\delta < \delta_0$. Thus, because

$$\Delta_2'(\delta)\,|_{\delta=\delta_0} = \frac{2r^3(\beta_b^2 - \beta_p^2)}{k^4}\left[\frac{r^2(\beta_b^2 - \beta_p^2)}{(\alpha_p - \alpha_b)^2}\left(\alpha_p(\beta_b^2 - \beta_p^2) + (\alpha_p - \alpha_b)\beta_p^2\right) + k^2\lambda\right] > 0,$$

$\Delta_2(\delta)$ is increasing everywhere on $\delta \in [0, \delta_0]$.

To show unimodularity on $[\delta_0, 1]$, note that, because $\Delta_2^{(4)}(\delta) < 0$, $\Delta_2''(\delta)$ is concave. Hence, there are two cases to consider: either $\Delta_2''(\delta) < 0$ for all $\delta \in (\delta_0, 1)$. In this case, $\Delta_2'(\delta)$ is monotonically decreasing. Because $\Delta_2'(\delta)\,|_{\delta=\delta_0} > 0$, unimodularity follows directly. Second, assume that $\Delta_2''(\delta) > 0$ for some $\delta \in (\delta_0, 1]$. Because of concavity, this implies that there are values δ_a, δ_b with $\delta_0 < \delta_a < \delta_b$ such that $\Delta_2''(\delta) > 0$ iff $\delta \in (\delta_a, \delta_b)$. In this case, $\Delta_2'(\delta)$ is decreasing at first (by continuity), then increasing again and, finally, it may be decreasing again. To establish that $\Delta_2(\delta)$ is, nevertheless, unimodular, we need to show that $\Delta_2'(\delta)$ never turns positive after turning negative. In particular, if $\Delta_2'(\delta)$ does turn negative, it reaches its maximum at $\min\{1, \delta_b\}$. Hence, we show the sufficient condition $\Delta_2'(\min\{1, \delta_b\}) < 0$.

First, it is easy to see that $\Delta_2'(\delta)\,|_{\delta=1} = -\frac{2r^3\lambda}{k^2}(\beta_b^2 - \beta_p^2) < 0$. Next, we derive δ_b with standard algebra. In particular, we find that $\delta_b = 1 - \frac{3r(\alpha_b\beta_b^2 - \alpha_p\beta_p^2)^2 - \sqrt{K}}{6k(\alpha_p^2 - \alpha_b^2)}$, with $K := 3r^2\left(2(\alpha_p\beta_b^2 - \alpha_b\beta_p^2)^2 + (\alpha_p\beta_p^2 - \alpha_b\beta_b^2)^2\right) - 12k^2\lambda(\alpha_p^2 - \alpha_b^2)(\alpha_p - \alpha_b) > 0$ iff $\Delta_2''(\delta)$ ever turns positive. We can then show that $\Delta_2'(\delta_b) = \frac{r^2}{9k^3(\alpha_p^2 - \alpha_b^2)}\Big[-(1-\delta_0)\left(3r(\alpha_p\beta_b^2 - \alpha_b\beta_p^2) - \sqrt{K}\right) - 3r(1-\delta_b)(\alpha_p\beta_b^2 - \alpha_b\beta_p^2)\left(3r(\alpha_b\beta_b^2 - \alpha_p\beta_p^2) + \sqrt{K}\right) - 2K(1-\delta_b)\Big]$. The final term in the square parentheses is clearly negative when $\delta_b < 1$. In that case, the second term also has to be negative: Because $\delta_b < 1 \Leftrightarrow 3r(\alpha_b\beta_b^2 - \alpha_p\beta_p^2) > \sqrt{K}$, the third of the inner three parentheses is greater $2\sqrt{K} > 0$. The other two inner parentheses are clearly also positive. Finally, the first term is also negative, because $\alpha_p\beta_b^2 - \alpha_b\beta_p^2 > \alpha_p\beta_p^2 - \alpha_b\beta_b^2$, so we have $\Delta_2'(\delta_b) < 0$ if $\delta_b < 1$, and the result follows. □

Proof of Proposition 1. $J(p) - J(b) = \gamma\Delta_1 - (1-\gamma)\Delta_2(\delta)$. When Δ_2 is non-positive ($\delta \le \delta_0$), then the entire term is positive, because $\Delta_1 > 0$. Otherwise (when $\delta > \delta_0$), the sign flips at $\gamma^\star(\delta)$. □

Proof of Proposition 2. With $\gamma^\star(\delta) = \dfrac{\Delta_2}{\Delta_1 + \Delta_2}$,

$$\frac{d\gamma^\star}{dq} = \frac{(\partial_q\Delta_2)(\Delta_1+\Delta_2) - \Delta_2(\partial_k\Delta_1 + \partial_q\Delta_2)}{(\Delta_1+\Delta_2)^2} = \frac{(\partial_q\Delta_2)\Delta_1 - \Delta_2(\partial_q\Delta_1)}{(\Delta_1+\Delta_2)^2},$$

for any parameter q. Note that $\Delta_1 > 0$ and $\Delta_2 \ge 0$ for $\delta \in (\delta_0, 1]$. Hence, when $\partial_q\Delta_2 < 0$ and $\partial_q\Delta_1 > 0$ (resp. $\partial_q\Delta_2 > 0$ and $\partial_q\Delta_1 < 0$) for all $\delta \in (\delta_0, 1]$, then $\frac{d\gamma^\star}{dq} < 0$ (resp. $\frac{d\gamma^\star}{dq} > 0$).

Consider first $\partial_q\Delta_1$ for all relevant parameters. We can directly compute:

$$\partial_{\alpha_p}\Delta_1 = r > 0, \partial_{\alpha_b}\Delta_1 = -r < 0, \partial_{\beta_p}\Delta_1 = \frac{2r^2\beta_p}{k} > 0, \partial_{\beta_b}\Delta_1 = -\frac{2r^2\beta_b}{k} < 0, \partial_k\Delta_1 = \frac{r^2}{k^2}(\beta_b^2 - \beta_p^2) > 0.$$

Consider now $\partial q\Delta_2$. We can directly obtain the first four derivatives:

$$\begin{aligned}
\partial_{\alpha_p}\Delta_2 &= -\frac{2r^2(1-\delta)^2}{k^2}\left[r\beta_p^2(1-\delta) + k\left(\alpha_p(1-\delta)^2 + \lambda\right)\right] < 0,\\
\partial_{\alpha_b}\Delta_2 &= \frac{2r^2(1-\delta)^2}{k^2}\left[r\beta_b^2(1-\delta) + k\left(\alpha_b(1-\delta)^2 + \lambda\right)\right] > 0,\\
\partial_{\beta_p}\Delta_2 &= -\frac{4r^3\beta_p(1-\delta)}{k^3}\left[r\beta_p^2(1-\delta) + k\left(\alpha_p(1-\delta)^2 + \lambda\right)]\right] < 0,\\
\partial_{\beta_b}\Delta_2 &= \frac{4r^3\beta_b(1-\delta)}{k^3}\left[r\beta_b^2(1-\delta) + k\left(\alpha_b(1-\delta)^2 + \lambda\right)]\right] > 0.
\end{aligned}$$

To see that $\partial_k\Delta_2 < 0$, we take the derivative to λ:

$$\frac{d(\partial_k\Delta_2)}{d\lambda} = -\frac{2r^2(1-\delta)}{k^2}(\alpha_p - \alpha_b)\left[(1-\delta_0) + (\delta - \delta_0)\right] < 0.$$

Hence, a sufficient condition is $\lim_{\lambda\to 0}\partial_k\Delta_2 < 0$. In fact,

$$\lim_{\lambda\to 0}\partial_k\Delta_2 = \frac{r^3(1-\delta)^2}{k^3}\left[-3(1-\delta_0)(\alpha_p - \alpha_b)(\beta_b^2 + \beta_p^2) - 4(1-\delta)(\alpha_b\beta_b^2 - \alpha_p\beta_p^2) + (\beta_b^2 - \beta_p^2)(\alpha_p + \alpha_b)\frac{(1-\delta)^2}{1-\delta_0}\right].$$

Because the first term is negative and the final term is positive, this can be upper-bounded by replacing δ_0 with δ. Then, the inner part simplifies to $-2(1-\delta)(\alpha_p\beta_b^2 - \alpha_b\beta_p^2) < 0$. □

Proof of Theorem 1. We show the result in four steps.

1. Coupling. Let $(U_t)_{t\ge 1} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$ be random variables. At time t and state s_t, both regimes draw the same U_t, indicating the decision maker's compliance with the advice: under $a \in \{p, b\}$, follow $\nu_\theta^*(s_t)$ iff $U_t \le c_a$, otherwise, follow $A \sim \mu(\cdot \mid s_t)$. Because $c_p \ge c_b$, we have $\{U_t \le c_b\} \subseteq \{U_t \le c_p\}$ almost surely for all $t \in \{1, 2, \ldots, T\}$. That is, the set of time steps in which the decision maker does not comply under b is almost surely a *superset* of that under p.

2. Visit dominance. Let $I_t^{(a)} := \mathbf{1}\{s_t^{(a)} \in \mathcal{S}_{\text{inf}}\}$. Under (A_3')–(A_4') and the coupling above, regime b has weakly more deviation episodes and hence weakly more opportunities to reach $\mathcal{S}_{\text{inf}}$. Therefore, for every T,

$$\mathbb{E}\Big[\sum_{t=1}^{T} I_t^{(b)}\Big] \ge \mathbb{E}\Big[\sum_{t=1}^{T} I_t^{(p)}\Big], \tag{EC.1}$$

with strict inequality for large T when returns to $\mathcal{S}_{\text{inf}}$ occur with positive probability.

3. Information decomposition. By Assumptions (A_3')–(A_5'),

$$\mathbb{E}\big[\mathcal{I}(s_t,\tilde{a}_t)\mid s_t=s\big]=c_{t,a}\mathcal{I}(s_t,\nu_\theta^\star(s_t))+(1-c_{t,a})\sum_{\tilde{a}\in\mathcal{A}}\mu(\tilde{a}|s_t)\mathcal{I}(s_t,\tilde{a})=(1-c_{t,a})\bar{I}\,\mathbf{1}\{s_t\in\mathcal{S}_{\text{inf}}\}.$$

Applying the law of iterated expectation, we have $\mathbb{E}\Big[\sum_{t=1}^T\mathcal{I}(s_t^{(a)},\tilde{a}_t^{(a)})\Big]=\bar{I}\sum_{t=1}^T(1-c_{t,a})\,\mathbb{E}\big[I_t^{(a)}\big]$.

4. Comparison of L_T. By definition, $L_T(a)=\sum_{t=1}^T\beta_a e_{t,a}^\star\,\mathbb{E}_a[\mathcal{I}(s_t^{(a)},\tilde{a}_t^{(a)})]=\bar{I}\sum_{t=1}^T\beta_a e_{t,a}^\star\,(1-c_{t,a})\,\mathbb{E}_a\big[I_t^{(a)}\big]$. Thus, $L_T(b)-L_T(p)=\bar{I}\sum_{t=1}^T\Big(\beta_b e_{t,b}^\star(1-c_{t,b})\,\mathbb{E}[I_t^{(b)}]-\beta_p e_{t,p}^\star(1-c_{t,p})\,\mathbb{E}[I_t^{(p)}]\Big)$. Add and subtract $\beta_p e_{t,p}^\star(1-c_{t,p})\,\mathbb{E}[I_t^{(b)}]$ inside the sum to get

$$=\bar{I}\sum_{t=1}^T\underbrace{\big(\beta_b e_{t,b}^\star(1-c_{t,b})-\beta_p e_{t,p}^\star(1-c_{t,p})\big)}_{\geq 0\text{ by }(A_1')\text{–}(A_2')}\mathbb{E}[I_t^{(b)}]\;+\;\bar{I}\sum_{t=1}^T\beta_p e_{t,p}^\star(1-c_{t,p})\underbrace{\big(\mathbb{E}[I_t^{(b)}]-\mathbb{E}[I_t^{(p)}]\big)}_{\geq 0\text{ by (EC.1)}}.$$

(i) Both sums are ≥ 0, hence $L_T(b)\geq L_T(p)$ for all T. (ii) Because each summand above is positive, ΔL_T is weakly increasing. A strict increase follows if, at some step, the second sum has a strictly positive summand (alternatively, if the first sum is strictly positive e.g., $\beta_b e_{t,b}^\star>\beta_p e_{t,p}^\star$ or $c_{t,p}>c_{t,b}$ and $\mathbb{E}[I_t^{(b)}]>0$). □

A.2. The Precise/Broad Trade-Off as a Function of T.

While learning benefits from increased time horizons in E_1, the designer's trade-off is more complex. To illustrate, assume the same one-shot decision task during E_2 as before. This is purely for illustrative purposes—a similar trade-off between immediate rewards and long-term learning occurs also when E_2 contains another sequential decision-making task (We highlight that the tasks in E_2 in our experiment, and the pre-advice environment E_0, also are MDPs similar to the one in E_1). Let $\pi_a^2\big(e_2,L_T(a)\big)=e_2\Big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi\big(L_T(a)\big)\Big]$, where, $\delta\in[0,1]$ measures the difference of E_2 from E_1, $\lambda>0$ is the channel through which efforts in E_2 affect compliance with the best strategy directly, $(1-\delta)^2\alpha_a$ captures the direct applicability of E_1's advice in E_2, and $(1-\delta)\Psi\big(L_T(a)\big)$ determines the impact of a higher-order understanding of optimal actions, enabled by engagement with the task during E_1. In particular, $\Psi:\mathbb{R}_+\to\mathbb{R}_+$ is assumed nondecreasing. Optimizing the level of effort e_2 gives $e_2^\star=(r/k)\big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi(L_T(a))\big]$ and

$$U_2\big(e_2^\star;L_T(a)\big)=\frac{r^2}{k}\Big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi\big(L_T(a)\big)\Big]^2+\mathbb{E}[R^-]-\frac{k}{2}\frac{r}{k}\big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi(L_T(a))\big].$$

Recall that the final term (the cost of effort for the decision maker) does not enter the system designer's objective function. Hence the learning gap of b over p is

$$\Delta_2(T,\delta)\;=\;\frac{r^2}{k}\left(\Big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi\big(L_T(b)\big)\Big]^2-\Big[\lambda+(1-\delta)^2\alpha_a+(1-\delta)\Psi\big(L_T(p)\big)\Big]^2\right).$$

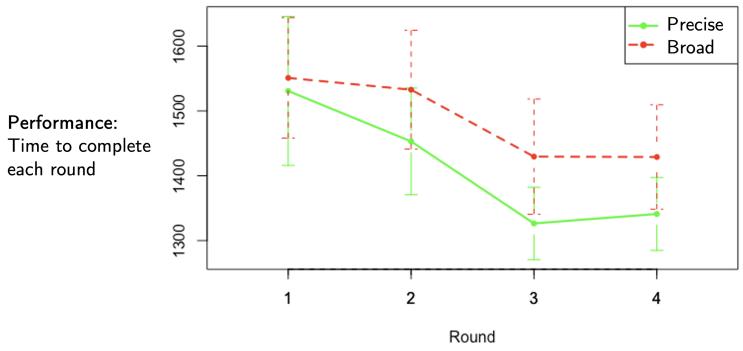
As Ψ is nondecreasing, and $L_T(b)-L_T(p)\geq 0$ is weakly increasing in T (Theorem 1), we have that $\Delta_2(T,\delta)$ is weakly increasing in T, as long as $\Delta_2(T,\delta)\geq 0$.

The reward gap during E_1, $\Delta_1(T)\;=\;r\sum_{t=1}^T(c_{t,p}-c_{t,b})$ is also increasing in T (A_1'). Hence, the decision boundary

$$\gamma^\star(T,\delta)\;=\;\frac{\Delta_2(T,\delta)}{\Delta_1(T)+\Delta_2(T,\delta)}$$

may be either increasing or decreasing in T. In particular, the decision boundary is increasing if and only if $\Delta_2\geq 0$ and $(\partial_T\Delta_2)\,\Delta_1>(\partial_T\Delta_1)\,\Delta_2$, that is, if the elasticity of Δ_2 to changes in T exceeds the elasticity of Δ_1 to changes in T.

Figure EC.1 Pilot study performance (elapsed game time) across rounds by advice condition.



Appendix B: Pilot Study: Immediate Effects of Advice Precision

This appendix reports a pilot study that isolates the immediate behavioral effects of advice precision, focusing on short-run compliance and performance (without traffic variation or advice removal).

B.1. Design and Sample

The pilot consists of two phases. In the pre-advice phase (E_0), participants complete two rounds without advice, letting us observe baseline behavior and early learning. In the with-advice phase (E_1), participants complete two additional rounds while receiving either *precise* or *broad* advice. The map used is the short map described in Study 1 (Section 4). We recruited 60 participants from Prolific (3,360 decision points). Average payoff was \$5.43.

B.2. Performance

Figure EC.1 plots EGT by round. Precise advice improves performance from E_0 to E_1 (mean improvement 13.4%).

B.3. Action-Level Evidence of Compliance at a Batching Decision.

To visualize how advice changes actions, we focus on Exit 2, where the time-minimizing strategy is to batch-charge to cover the two segments connecting Exits 2 through 4. Figures EC.2–EC.3 plot the distribution of *aftercharge* (the post-charging battery level) at Exit 2 across rounds. The shaded region marks the near-optimal range for batching.

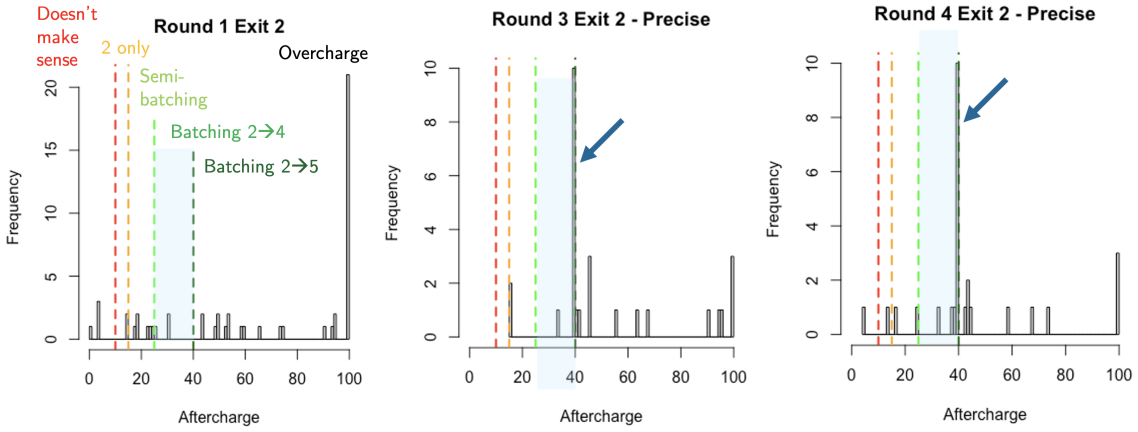
Figures EC.2–EC.3 show that precise advice sharply concentrates aftercharge within the near-optimal batching range, while broad advice shifts behavior toward batching but with substantially greater dispersion.

B.4. Decision Classification and Sequence Clustering

To summarize participants' strategies over the full task rather than a single exit, we classify each state-action pair $(s_i, \tilde{a}_i)$ by comparing the participant's charging decision to the charging decision minimizing the expected (in-game) time given uncertain traffic. Recall that the latter always corresponds to charging sufficiently for the next one or two segments, taking into account worst-case traffic. When the current level is above that value, the optimal charge is zero.

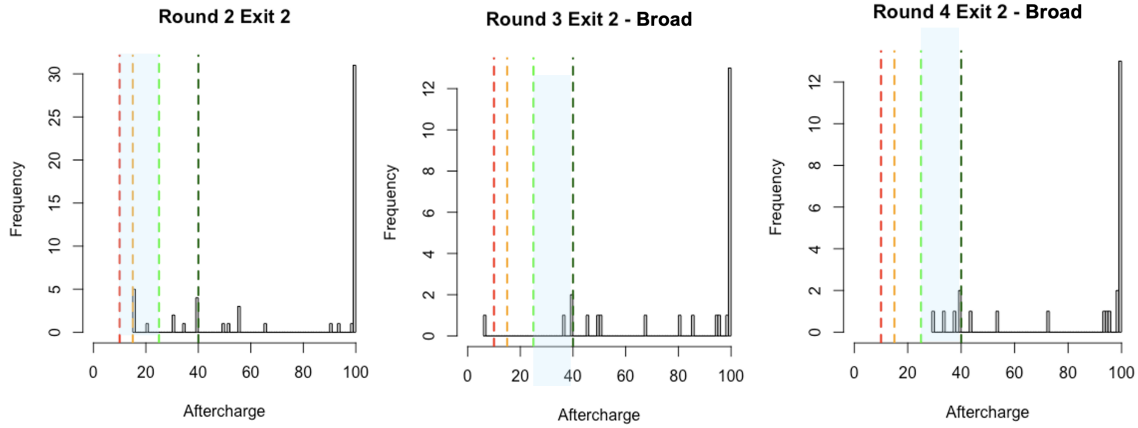
We classify each charging decision by how many segments the participant's post-charge battery can cover under worst-case traffic, relative to the time-minimizing benchmark. This yields four labels: *out* (insufficient for the next segment), *below* (covers fewer segments than the benchmark), *in* (matches the benchmark, allowing a small tolerance), and *above* (covers more segments than the benchmark). Figure EC.4 illustrates this mapping. We then apply sequence clustering (TraMineR; Gabadinho et al. 2011) to participants' label sequences within each phase, to summarize strategy profiles.

Figure EC.2 Pilot study action distributions at Exit 2, Part 1.



(a) Round 1, Exit 2 (no advice) (b) Round 3, Exit 2 (precise) (c) Round 4, Exit 2 (precise)

Figure EC.3 Pilot study action distributions at Exit 2, Part 2.



(a) Round 2, Exit 2 (no advice) (b) Round 3, Exit 2 (broad) (c) Round 4, Exit 2 (broad)

Note. At exit 2, batching is near-optimal. The shaded region indicates the near-optimal aftercharge range.

Figure EC.4 Pilot study classification of charging decisions relative to the time-minimizing decision.

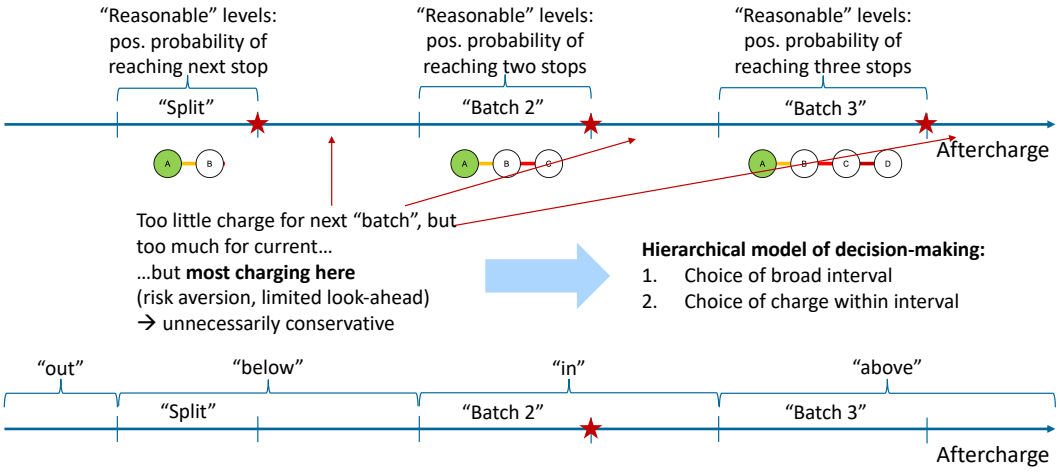
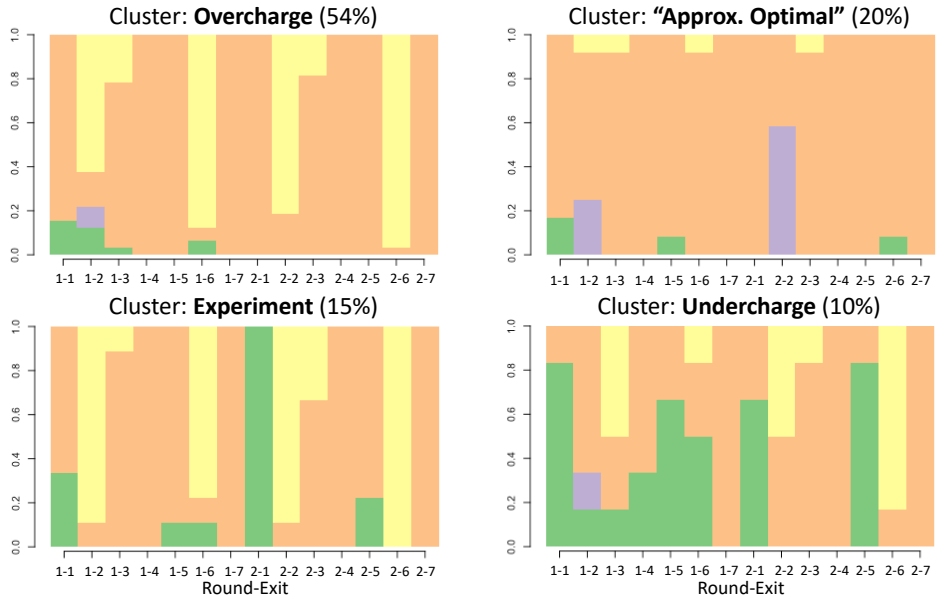


Figure EC.5 Pilot study clusters of baseline behavior (pre-tip).



Note. Colors indicate decision labels, with green = *out*, purple = *below*, orange = *in*, yellow = *above*.

Figure EC.5 displays clusters for the pre-advice phase.[7] In particular, for each round-exit combination, we observe the percentage of participants in a cluster whose decision corresponds to one of the decision labels, as indicated by their color. As seen here, the largest group is a cluster we label *Overcharge*: participants here tend to behave conservatively, either charging in line with the optimal number of segments or more. A smaller group exhibits behavior we label *Approx. Optimal*: participants' behavior is similar to the optimal charging strategy, even without advice.

B.5. Strategy Transitions with Advice.

After conducting clustering analyses for both pre-advice and the post-advice phases,[8] we consider the impact of advice on how participants' strategies change. Figure EC.6 tracks participants' movement between pre-advice and with-advice clusters, based on the type of advice they observe. Precise advice induces larger and faster shifts toward near-optimal strategies, though a meaningful subset of baseline overchargers remain conservative even when precise (numerical) advice is available. Broad advice produces a weaker and noisier transition. Interestingly, participants that are already behaving near optimally prior to observing advice all are part of the *Follow tip* cluster in the with-advice phase.

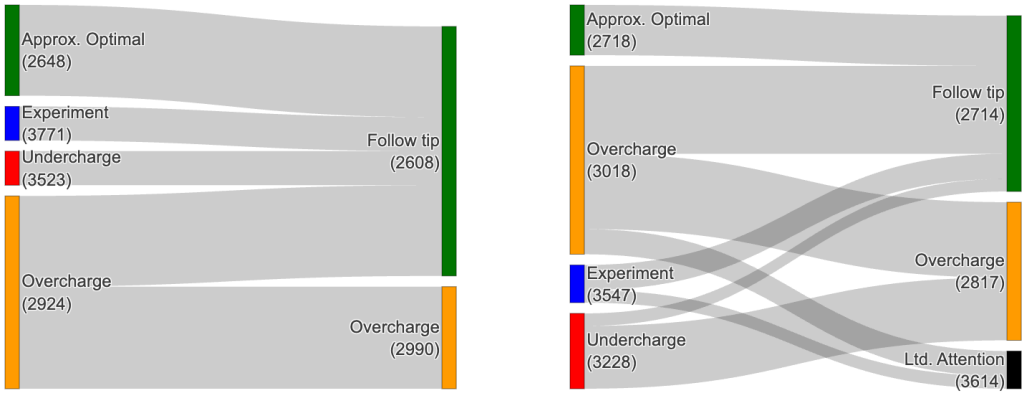
B.6. Takeaway for the Main Studies

Overall, the pilot provides evidence that precise advice increases compliance and improves short-run performance in our task, while broad advice leads to more dispersed actions. These patterns motivate the main studies, which explicitly test whether the short-run benefits of precision persist once advice is removed, and whether broader advice supports learning and transfer across environments.

[7] We exclude one participant for whom the majority of decisions were classified as *out*, indicating click-through without charging.

[8] We omit the latter due to space constraints, but figures are available from the authors. In the post-advice phase, we observe three clusters: participants that largely behave in line with the optimal strategy (*Follow tip*), those that continue to charge more than optimal (*Overcharge*), and those those who both over- and undercharge (*Ltd. Attention*).

Figure EC.6 Pilot study cluster transitions from pre-advice to with-advice by advice condition.



(a) Precise advice (b) Broad advice

Note. The left-hand side (resp. right-hand side) indicates the cluster to which a participant is assigned in the pre-advice (resp. with-advice) phase. Numbers in parentheses report average in-game time within clusters, further validating the cluster labels.

Appendix C: Additional Analyses for Study 1

This appendix reports behavioral diagnostics that complement the performance results in Section 4.2. In Appendix C.1, we discuss the role of exploration in supporting the learning advantage of *broad* compared to *precise* advice. Then, in Appendix C.2, we provide evidence that this learning advantage can be ascribed to a higher-level strategic understanding of the task rather than a simple shift in safety margins.

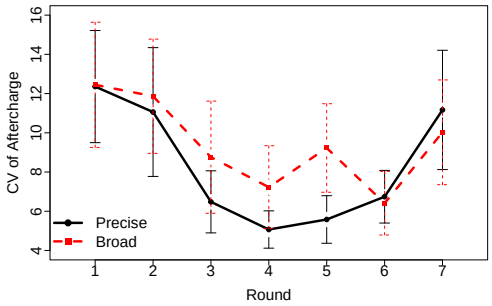
C.1. Learning through Exploration

We quantify how tightly participants' charging choices concentrate around a single implementation rule. For this, we compute the coefficient of variation (CV) of *aftercharge* (the total amount of charge after a charging decision and before the next segment's realization) across rounds. Figure EC.7 shows that dispersion declines over rounds for both conditions, consistent with participants converging toward more stable charging behavior as they gain experience. However, during the advice window (Rounds 3–5), participants assigned to *broad* advice exhibit systematically greater variability in aftercharge than those assigned to *precise* advice, with the difference peaking in Round 5 (the final with-advice round), where broad advice yields a significantly higher CV (one-sided t-test: $p < 0.05$). This higher dispersion is consistent with broad advice leaving more of the mapping from principle to action to the user, leading them to explore different actions and, thus, experience a larger number of different states, as posited in Hypothesis H_2. While CV alone cannot fully distinguish deliberate exploration, the fact that the separation emerges most strongly at the end of the with-advice phase supports the interpretation that broad advice sustains greater strategy heterogeneity. This mechanism can plausibly contribute to broad advice's advantage once advice is withdrawn (the learning gap).

C.2. Residual Charge Analysis

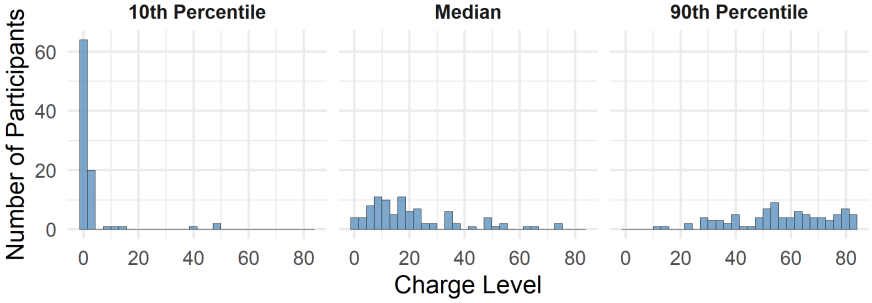
C.2.1. Residual Charge Distributions as a Proxy for Conservatism. We use residual charge as a proxy for a participant's implicit "safety margin," defined as the charge remaining upon arrival at each exit on the short map (excluding the initial state). As charging costs are convex, while the penalty for running out is high, a higher residual charge reflects more conservative planning, whereas lower residual charge reflects more aggressive planning.

Figure EC.7 Behavioral dispersion during the advice phase: CV of aftercharge.



Note. Points plot the round-level mean CV within each advice condition; vertical bars indicate uncertainty around the mean.

Figure EC.8 Participant-level summaries of residual charge across exits.



Note. Residual charge is a proxy for conservatism (larger buffers imply more conservative charging).

For each participant, we compute the 10th percentile, median, and 90th percentile of residual charge across exits (excluding Exit 0). Figure EC.8 visualizes these participant-level summaries. Two patterns are worth noting. First, participants exhibit substantial heterogeneity in baseline safety margins. Second, the lower tail can be constrained by segments where near-zero residual charge is effectively forced by distance and realized traffic; accordingly, the median and upper-tail summaries are typically more informative about conservative charging tendencies.

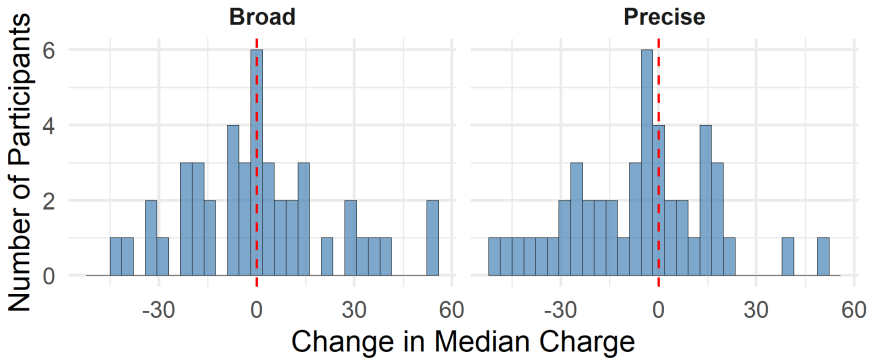
C.2.2. Pre- to Post-advice Changes in Safety Margins. To test whether advice exposure induces a persistent shift in conservatism on the short map, we compare each participant's median residual charge before versus after the advice phase. Specifically, we compute the change in median residual charge between the pre-advice short-map round (Round 1) and the post-advice short-map round without advice (Round 6):

$$\Delta_i^{\text{med}} \equiv \text{MedianResidualCharge}_{i,\text{Round 6}} - \text{MedianResidualCharge}_{i,\text{Round 1}}.$$

Negative values indicate carrying less residual charge after advice.

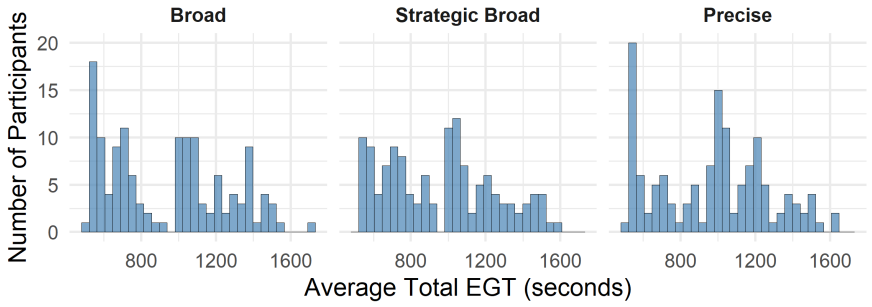
Figure EC.9 plots the distribution of Δ_i^{med} by advice condition. The distribution exhibits improvements for some participants and regressions for others, with substantial mass near zero. We complement Figure EC.9 with a Welch two-sample difference-in-means test comparing Δ_i^{med} across advice conditions. The estimated broad-minus-precise difference is 7.10 (95% CI $[-2.23, 16.4]$; $t = 1.51$, df $= 87.2$, $p = 0.134$), indicating that advice type does not induce a uniform post-advice shift in median safety margins on the short map.

Figure EC.9 Distribution of participant-level changes in median residual charge from Round 1 to Round 6.



Note. Negative values indicate reduced safety margins.

Figure EC.10 Final-round Elapsed Game Time by advice condition.



C.2.3. Interpretation Relative to Transfer. Section 4.2 documents a transfer advantage of broad advice on the long map, where advice is never provided. The residual-charge analyses above focus on the short map and provide two takeaways: (i) participants are heterogeneous in baseline safety margins, and (ii) advice type does not produce a uniform, persistent shift in safety buffers from Round 1 to Round 6. Together, these findings suggest that broad advice's transfer advantage is not well explained by a simple "be less conservative" shift in charging, and is more consistent with advice shaping higher-level, generalizable strategic understanding.

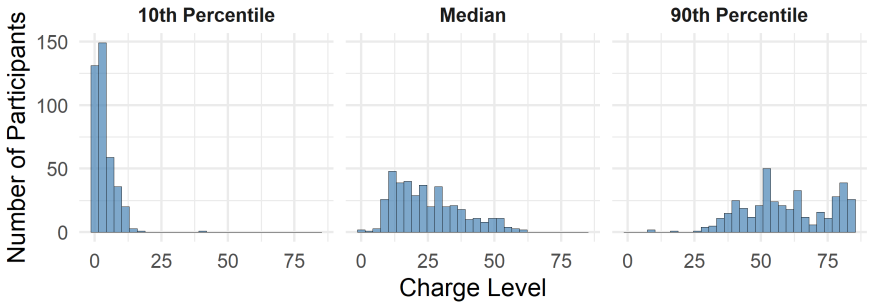
Appendix D: Additional Analyses for Study 2

This appendix reports behavioral diagnostics that complement the performance patterns in Section 5.2. The goal is to examine behavioral mechanisms underlying the performance patterns: (i) compliance while advice is available, (ii) exploration during the advice phase, and (iii) strategic understanding reflected in safety margins and batching behavior after the advice is removed. These diagnostics use existing plots generated from the experimental logs and are intended to inform mechanism interpretation. Figure EC.10 summarizes the final-round outcome using EGT, providing a single scalar view of end-of-task performance by advice condition.

D.1. Residual charge distributions as a proxy for safety margins.

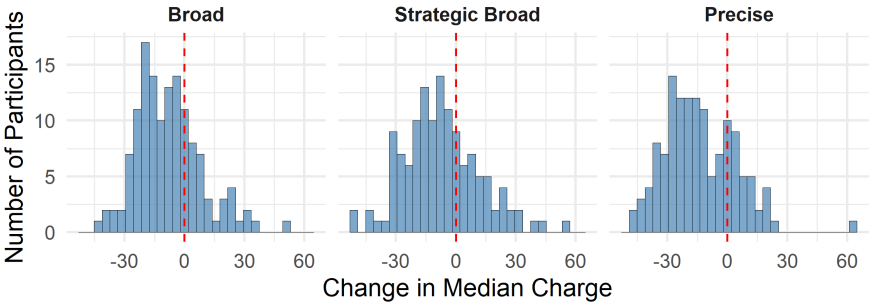
We use residual battery charge as a proxy for a participant's implicit safety margin. For each participant, we compute the 10th, 50th, and 90th percentile of residual charge across exits, excluding the initial state (see Figure EC.11). Participants exhibit meaningful heterogeneity in safety margins. Second, relative to earlier studies, dispersion appears lower, consistent with participants converging toward a narrower range of safety margins as they gain experience.

Figure EC.11 Participant-level summaries of residual charge across exits.



Note. Larger buffers indicate more conservative charging.

Figure EC.12 Distribution of participant-level changes in median residual charge from pre- to post-advice.



Note. Negative values indicate less conservative charging after advice.

D.2. Pre- to post-advice changes in median residual charge.

To assess whether advice exposure is associated with persistent changes in safety margins, we compare participant-level median residual charge after and before the advice rounds. Define $\Delta\widetilde{B}_i = \widetilde{B}_i^{\text{post}} - \widetilde{B}_i^{\text{pre}}$, where negative values indicate that a participant has smaller buffers after advice. Figure EC.12 plots the distribution of $\Delta\widetilde{B}_i$ by advice condition.

The distribution is consistent with many participants reducing safety buffers over time. Pairwise Welch tests indicate that precise differs from both broad ($p = 0.007$) and specific broad ($p = 0.004$), while specific broad and broad are not distinguishable ($p = 0.711$). These comparisons suggest that precise advice is associated with a stronger reduction in safety margins, whereas the two qualitative advice types yield similar shifts on this residual-charge measure.

D.3. Experienced traffic and subsequent conservatism.

Finally, we examine whether realized traffic relates to subsequent conservatism. We compare the traffic experienced by participants in the lower quartile of residual-charge behavior (more aggressive) versus the upper quartile (more conservative), using a traffic ratio defined as actual traffic divided by the midpoint of the segment's traffic range. One third (35 out of 105) of participants with more aggressive experienced a slightly higher average traffic ratio, compared to 59% of those who were more conservative. A Welch test rejects equality of means ($p < 0.01$), with the more conservative group experiencing a higher average traffic ratio. This pattern is consistent with adverse experienced traffic being associated with more conservative subsequent play, though the magnitude of the difference is small.

Table EC.1 Building reward components from identified topics.

Top words	Exemplary document	Reward component
charge, avoid	"Avoid overcharging to keep the charge times as low as possible"	Elapsed game time
attention, pay	"I tried to make estimates for the range ... looking ahead to see the traffic status for the next leg"	Elapsed game time
round, load	"... When I was extremely low I would load fully hoping to not have to load the next round"	Simplicity
safe, destination	"Be safe and charge for the maximum distance"	Risk exposure
penalty, 300	"Play it safe! That 300 minute penalty when you run out is killer"	Exposure after penalty
stop, possible	"Try and stop as little as possible"	Margin over worst
tried, make sure	"... try to charge enough to make it to that stop so you don't have to charge again the next time"	Batching preference
time, stop	"... after including worst traffic scenario and charge only up to what was needed ..."	Splitting preference
left, minutes	"I counted the max amount of minutes it would take me and charged to that amount"	Splitting preference

There are two additional topics similar to the last two and implying the *Splitting preference* component. The final three topics relate to advice usefulness and usage. As we deal with compliance separately from the reward components, we omit those topics.

Appendix E: Inverse Reinforcement Learning (IRL) Approach in Detail

E.1. Deriving the Reward Components

At the end of our pilot study (see Appendix B), we ask the participants, among others, to comment on two questions:

- "Please describe your strategy to perform well in this game?"
- "Please provide your best simple tip/advice on the gameplay/strategy to help future players. In other words, what tip should we display instead of what was shown in the last two rounds of the game?"

We make answering these questions obligatory, and most answers are comprehensive with 94 (resp. 64 – 155) words at the median (resp. IQR). Each answer forms a "document" ($n = 122$—we also include participants' answers if they were excluded from the main analysis). We then use BERTopic to derive topics relevant to these documents, where we set the minimum topic size to eight. We identify 14 topics in this way, from which we derive the seven reward components as indicated in Table EC.1.

E.2. Estimation Process

E.2.1. Theoretical Background. Under Maximum Causal Entropy (MaxCausalEnt) framework (Ziebart et al. 2010), an agent's policy $\mu(\tilde{a}_t \mid s_t)$ is defined as the distribution maximizing the causal entropy of the trajectory distribution, subject to matching the expected feature counts of the demonstrated behavior. We provide an overview of some key aspects that are relevant for introducing our own approach. Details can be found in Gleave and Toyer (2022).

MaxCausalEnt can be formulated as a constrained optimization problem,

$$\begin{aligned} \max_{\mu} \quad & H_{\text{causal}}(\mu) \;=\; -\mathbb{E}_{p_\mu}\Big[\sum_{t=1}^{T} \gamma^t \log \mu(\tilde{a}_t \mid s_t)\Big] \\ \text{s.t.} \quad & \mathbb{E}_{p_\mu}[f_j(\tau)] = \mathbb{E}_{\mathcal{D}}[f_j(\tau)], \quad j = 1, \ldots, m, \end{aligned}$$

where policy μ induces a trajectory distribution $p_\mu(\tau)$, $\mathcal{D}$ denotes the (empirical) distribution of trajectories, and $f_j(\tau) = \sum_{t=1}^{T} \gamma^t \phi_j(s_t, \tilde{a}_t)$ are feature counts, where $\boldsymbol{f} = (f_1, \ldots, f_m)$.

Introducing Lagrange multipliers $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ for the feature constraints, the Lagrangian is

$$L(\mu, \boldsymbol{\theta}) = H_{\text{causal}}(\mu) + \mathbb{E}_{p_\mu}[\boldsymbol{\theta}^T \boldsymbol{f}(\tau)] - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\theta}^T \boldsymbol{f}(\tau)].$$

Thus, using standard duality results, we can rewrite MaxCausalEnt as

$$\min_{\boldsymbol{\theta}} \max_{\mu} L(\mu, \boldsymbol{\theta}). \tag{EC.2}$$

This is usually solved iteratively in two steps: first, for a given $\boldsymbol{\theta}$, the optimal policy $\mu^*(\boldsymbol{\theta})$ is identified. Then, $\boldsymbol{\theta}$ is updated through gradient ascent. The solution to the inner problem is the "soft" Bellman policy

$$\mu(\tilde{a}_t \mid s_t) = \exp\big(Q_{\text{soft}}(s_t, \tilde{a}_t) - V_{\text{soft}}(s_t)\big), \tag{EC.3}$$

$$\text{where} \quad Q_{\text{soft}}(s_t, \tilde{a}_t) = \boldsymbol{\theta}^T \boldsymbol{\phi}(s_t, \tilde{a}_t) + \gamma \mathbb{E}_P\big[V_{\text{soft}}(s_{t+1})\big], \quad \text{and} \quad V_{\text{soft}}(s_t) = \log \sum_{\tilde{a}'_t \in \mathcal{A}} \exp\big(Q_{\text{soft}}(s_t, \tilde{a}'_t)\big).$$

For our study, we will assume that participants fully discount future rewards, that is, $\gamma = 0$. This assumption is in line with the observations that the vast majority of participants do not exhibit explicit forward-looking behavior by clicking on future segments to reveal traffic ranges Note that our formulation still allows for heuristic forward-looking behavior, e.g., through the component that measures participants' *Batching preference*. We also perform a robustness check with $\gamma > 0$, showing that our results are qualitatively unchanged (omitted due to space constraints).

When $\gamma = 0$, the future value term in (EC.3) vanishes and $Q_{\text{soft}}(s_t, \tilde{a}_t)$ reduces directly to the instantaneous reward $r_{s_t}(\tilde{a}_t)$. The optimal policy therefore collapses to the standard softmax policy

$$\mu(\tilde{a}_t \mid s_t) = \frac{\exp\big(\sum_{j=1}^m \theta_j \phi_j(s_t, \tilde{a}_t)\big)}{\sum_{\tilde{a}'_t \in \mathcal{A}} \exp\big(\sum_{j=1}^m \theta_j \phi_j(s_t, \tilde{a}'_t)\big)}. \tag{EC.4}$$

E.2.2. Hierarchical Model. A central challenge in IRL is that the problem is under-constrained: for any set of demonstrations, there exist infinitely many reward functions that could explain the behavior. Classical approaches often resolve this ambiguity via point-estimate constraints. Meanwhile, Ramachandran and Amir (2007) propose resolving this ambiguity probabilistically. Instead of seeking a single "best" reward vector, the authors' approach infers the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ conditional on the observed trajectories, where they equally assume that trajectories are obtained from a softmax policy.

This Bayesian IRL approach enables the prior $p(\boldsymbol{\theta})$ to constrain the search space. By selecting appropriate priors, the model penalizes complexity and explicitly discourages trivial or degenerate solutions, effectively acting as a regularizer that favors parsimonious explanations. In principle, domain knowledge can also be integrated, although we omit this aspect. Besides, applying a Bayesian approach enables us to capture the uncertainty in our parameter estimates explicitly (see, also, e.g., Brown and Niekum 2018).

Finally, the Bayesian framework naturally extends to hierarchical modeling, which is helpful when data is aggregated from multiple participants who share structural similarities but still exhibit heterogeneity. While the original formulation in Ramachandran and Amir (2007) assumes a single agent, subsequent work in Bayesian IRL (e.g., Choi and Kim 2012) has demonstrated that modeling the joint distribution of population-level parameters can yield superior generalization compared to training separate models.

Recall that our model assumes $\boldsymbol{\theta}_{sc,i} = (\theta^1_{sc,i}, \ldots, \theta^m_{sc,i}) = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_{sc} + \boldsymbol{\Delta}_i$ for each decision maker $i \in \{1, \ldots, n\}$ and each scenario $sc \in \mathcal{SC}$. At the top level we posit a global baseline $\boldsymbol{\theta}_0$ and scenario-specific deviations:

$$\boldsymbol{\theta}_0 \sim \mathcal{N}(0, \boldsymbol{I}_m)\,,$$
$$\boldsymbol{\Delta}_{sc} \mid \tau_{sc} \sim \mathcal{N}\Big(0, \tau^2_{sc}\boldsymbol{I}_m\Big)\,, \text{ with } \tau_s \sim \text{HalfNormal}(\sigma_1),$$

where we add a centering constraint $\sum_{sc\in\mathcal{SC}} \boldsymbol{\Delta}_{sc} = 0$, so that the $\boldsymbol{\Delta}_{sc}$ represent shifts around a common baseline.

To capture individual heterogeneity without a full m-dimensional random effect per participant, we use a low-rank factorization: $\boldsymbol{\Delta}_i = \boldsymbol{U}\boldsymbol{z}_i$, where $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ is a learned matrix and $z_i \in \mathbb{R}^d$ is a low-dimensional latent vector for participant i. We assume $d = 2$, but our results are not substantially altered with different values (we also try $d \in \{3, 4\}$).

We place Gaussian priors on these quantities: $z_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $\boldsymbol{u}_\ell \sim \mathcal{N}(0, \boldsymbol{I}_K)$, $\ell = 1, \ldots, d$, where $\boldsymbol{u}_\ell$ denotes the ℓ-th column of $\boldsymbol{U}$. In the implementation, each column $\boldsymbol{u}_\ell$ is normalized to unit length and scaled by a learned magnitude $\tau_\ell \sim \text{HalfNormal}(\sigma_2)$, which improves identifiability and numerical conditioning. In addition, as with the scenario-specific shifts, the values z_i are centered around the mean.

We then assume that participant i's fallback policy in scenario sc is a softmax policy (EC.4) with the corresponding weight $\boldsymbol{\theta}_{sc,i}$. We place an analogous hierarchical structure on the compliance probabilities to derive $\pi_{sc,i}$ and, finally, assume that the participant chooses action $\tilde{a}_t$ with probability

$$P_{sc,i}(\tilde{a}_t \mid s_t) = \pi_{sc,i}\nu^*(\tilde{a}_t \mid s_t) + (1 - \pi_{sc,i})\mu(\tilde{a}_t \mid s_t), \tag{EC.5}$$

where we recall that $\nu^*(\tilde{a} \mid s)$ is the policy of the algorithmic advice system.

E.2.3. Inference. While early Bayesian IRL approaches rely on Markov Chain Monte Carlo methods (Ramachandran and Amir 2007), these often suffer from poor scaling when applied to high-dimensional hierarchical models or large datasets. Thus, to overcome computational bottlenecks, we employ *Stochastic Variational Inference* (SVI, Hoffman et al. 2013). SVI enables scaling to large datasets while maintaining the benefits of uncertainty quantification (Blei et al. 2017). It recasts the inference problem as an optimization task, seeking a tractable distribution q_ψ that minimizes the Kullback-Leibler (KL) divergence to the true posterior. We first present the basic approach, then show how our formulation also allows us to integrate the causal entropy maximization ideas behind MaxCausalEnt.

Let $\mathcal{D}$ again denote the observed trajectories, and let Z denote the set of all parameters, including latent ones. Denote with $p(Z)$ the joint prior given by the product of prior distributions outlined previously, and with $p(\mathcal{D} \mid Z)$ the behavioral likelihood built from participants' mixed policies:

$$p(\mathcal{D} \mid Z) = \prod_{sc\in\mathcal{SC}} \prod_{i=1}^{n} \prod_{\tau\in\mathcal{D}_{sc,i}} \prod_{(s_t,\tilde{a}_t)\in\tau} P_{sc,i}\big(\tilde{a}_t \mid s_t; Z\big),$$

where $\mathcal{D}_{sc,i}$ are the trajectories specific to scenario sc and participant i. The corresponding posterior is

$$p(Z \mid \mathcal{D}) = \frac{p(Z)\,p(\mathcal{D} \mid Z)}{p(\mathcal{D})}, \qquad p(\mathcal{D}) = \int p(Z)\,p(\mathcal{D} \mid Z)\,dZ.$$

SVI chooses a tractable family $q_\psi(Z)$ and optimizes it to be close to the true posterior $p(Z \mid \mathcal{D})$ (Blei et al. 2017). Starting from the log marginal likelihood,

$$\log p(\mathcal{D}) = \log \int p(Z, \mathcal{D})\,dZ = \log \int q_\psi(Z)\,\frac{p(Z, \mathcal{D})}{q_\psi(Z)}\,dZ \geq \mathbb{E}_{q_\psi}\Big[\log p(Z, \mathcal{D}) - \log q_\psi(Z)\Big] = \mathcal{L}(\psi),$$

where the we use Jensen's inequality. The quantity $\mathcal{L}(\psi) = \mathbb{E}_{q_\psi}\left[\log p(Z,\mathcal{D})\right] - \mathbb{E}_{q_\psi}\left[\log q_\psi(Z)\right]$ is the *evidence lower bound* (ELBO). This can be re-expressed in terms of the Kullback–Leibler (KL) divergence. Using $p(Z \mid \mathcal{D}) = \frac{p(Z,\mathcal{D})}{p(\mathcal{D})}$,

$$\mathrm{KL}\big(q_\psi(Z) \,\|\, p(Z \mid \mathcal{D})\big) = \mathbb{E}_{q_\psi}\left[\log p(Z)\right] - \mathbb{E}_{q_\psi}\left[\log p(Z \mid \mathcal{D})\right] = \mathbb{E}_{q_\psi}\left[\log q_\psi(Z)\right] - \mathbb{E}_{q_\psi}\left[\log p(Z,\mathcal{D})\right] + \log p(\mathcal{D}),$$

so we obtain the identity $\mathcal{L}(\psi) = \log p(\mathcal{D}) - \mathrm{KL}\big(q_\psi(Z) \,\|\, p(Z \mid \mathcal{D})\big)$. Since $\log p(\mathcal{D})$ does not depend on ψ, maximizing the ELBO is equivalent to minimizing the KL divergence.

So far, we have described the general idea behind SVI, applied to our IRL task. However, we make an important modification by adding an entropy regularization term. Recall the inner maximization of the MaxCausalEnt problem in (EC.2). For fixed $\boldsymbol{\theta}$, the last term of the Lagrangian, $-\mathbb{E}_{\mathcal{D}}[\boldsymbol{\theta}^T \boldsymbol{f}(\tau)]$ is constant in μ, so maximizing $L(\mu,\theta)$ over μ is equivalent (up to an additive constant) to solving the problem

$$\max_{\mu}\ \mathbb{E}_{p_\mu}\left[\boldsymbol{\theta}^\top \boldsymbol{f}(\tau)\right] + H_{\text{causal}}(\mu).$$

We explicitly integrate entropy maximization through a regularization term. We use an annealing schedule to scale it, which encourages the optimization trajectory to explore high-entropy policies early in training before converging to sharper preferences. In particular, for each scenario and individual, we define the normalized entropy as

$$H_{sc,i}(Z) = \sum_{t=1}^{T} \frac{-\sum_{\tilde{a}'_t} P_{sc,i}(\tilde{a}'_t \mid s_t; Z)\, \log P_{sc,i}(\tilde{a}'_t \mid s_t; Z)}{\log |\mathcal{A}_{sc,i,t}|},$$

where $\mathcal{A}_{sc,i,t}$ indicates the set of actions available to agent i in scenario sc at time t. We then let $H(Z) = \sum_{sc\in\mathcal{SC}} \sum_{i=1}^{n} H_{sc,i}(Z)$. Meanwhile, we introduce the scaling term $\lambda > 0$, where $\lambda \to 0$ throughout the training process.

Then, instead of the baseline joint $p(Z,\mathcal{D}) = p(Z)\, p(\mathcal{D} \mid Z)$, we use an *entropy-augmented* log joint:

$$\log p_\lambda(Z,\mathcal{D}) = \log p(Z) + \log p(\mathcal{D} \mid Z) + \lambda\, H(Z).$$

The corresponding *entropy-regularized posterior* is $p_\lambda(Z \mid \mathcal{D}) \propto p(Z)\, p(\mathcal{D} \mid Z)\, \exp\big(\lambda H(Z)\big)$. Hence, the entropy bonus changes the target distribution from the baseline posterior $p(Z \mid \mathcal{D})$ to the entropy-biased posterior $p_\lambda(Z \mid \mathcal{D})$. A large λ favors parameters whose induced policies are more stochastic (higher entropy).

Replacing p by p_λ in the derivation above, the ELBO becomes

$$\mathcal{L}_\lambda(\psi) = \mathbb{E}_{q_\psi}\left[\log \tilde{p}_\lambda(Z,\mathcal{D})\right] - \mathbb{E}_{q_\psi}\left[\log q_\psi(Z)\right] = \log p(\mathcal{D}) - \mathrm{KL}\big(q_\psi(Z) \,\|\, p(Z \mid \mathcal{D})\big) + \lambda\, \mathbb{E}_{q_\psi}\left[H(Z)\right],$$

and we note that, again, $p(\mathcal{D})$ drops out in the optimization. Hence, maximizing $\mathcal{L}_\lambda(\psi)$ is equivalent to minimizing the KL divergence, regularized by the expected entropy.

We employ the SVI capabilities of the *NumPyro* package in Python (Phan et al. 2019), using a low-rank multivariate normal for q_ψ. To improve numerical stability, we apply contrast-whitening (specifically, ZCA whitening) to the features $\boldsymbol{\phi}$ prior to inference, which stabilizes the covariance structure of the inputs.

E.3. Validation

We validate our inference approach through several tests, both with synthetic and real data. Our focus here, especially in tests 1 and 3, lies on identifiability: are we able to accurately recover the true weights and compliance probabilities?

Table EC.2 Correlations between true and estimated (median) parameters for synthetic data.

	$\rho(\theta^j_{sc,i}, \hat{\theta}^j_{sc,i})$	$\rho(\theta^j_{sc}, \hat{\theta}^j_{sc})$	$\rho(\Delta^j_i, \hat{\Delta}^j_i)$		$\rho(\pi_{sc,i}, \hat{\pi}_{sc,i})$	$\rho(\pi_{sc}, \hat{\pi}_{sc})$	$\rho(\pi_i, \hat{\pi}_i)$
Avg. across $j = 1, \ldots, m$	0.96	0.96	0.88	Value	0.97	0.99	0.67
Std. across $j = 1, \ldots, m$	0.04	0.04	0.05				

Correlations are across all possible subscripts, for a given superscript (in the case of the reward weight-related parameters). For compliance probabilities, we only consider with-advice scenarios, that is, scenarios in which algorithmic advice is available. Note that $\theta^j_{sc} = \theta^j_0 + \Delta^j_{sc}$.

E.3.1. Synthetic Data: Identification. As a first validation step, we need to ensure that our approach is able to accurately recover model parameters when misspecification is not a concern. Hence, we create a synthetic data set according to the assumed (hierarchical) model. In order to generate data that is comparable with the real experimental data, we take the original trajectories and only replace chosen actions. In particular, we randomly draw reward weights and compliance probabilities, in line with the prior distributions outlined in Appendix E.2.2. Then, for a given scenario, individual, and state in the original trajectories, we generate the synthetic action based on the probabilities described in (EC.5). In total, we have 3,666 trajectories across 13 scenarios and 549 participants, with a total of 20,592 state-action pairs serving as observations. Each observation is assigned to the training set with 80% probability.

Next, we estimate the posterior distributions of model parameters. We then correlate each "true" parameter with the median of the corresponding posterior distribution. Table EC.2 provides an overview of the results. As shown here, we are able to accurately predict the main effects ($\theta^j_{i,sc} = \theta_0 + \Delta^j_{sc} + \Delta^j_i$, resp. $\pi_{sc,i}$), as well as the scenario-specific effects ($\theta^j_{sc} = \theta_0 + \Delta^j_{sc}$, resp. π_{sc}). While the prediction of the individual intercepts is imperfect, this is to be expected, given the relatively few observations for any one individual in each scenario. Moreover, for our analysis, the scenario-based effects are much more relevant, as they give us an idea of the "average" behavior across participants in a given setting.

E.3.2. Real Data: Log-Likelihoods and Posterior Predictive Checks. Next, we run our estimation approach on the real data, with the same structure. That is, we have 20,592 observations from 3,666 trajectories (80% in the training set), but the actions now are the real actions chosen by participants, rather than those generated by our model.

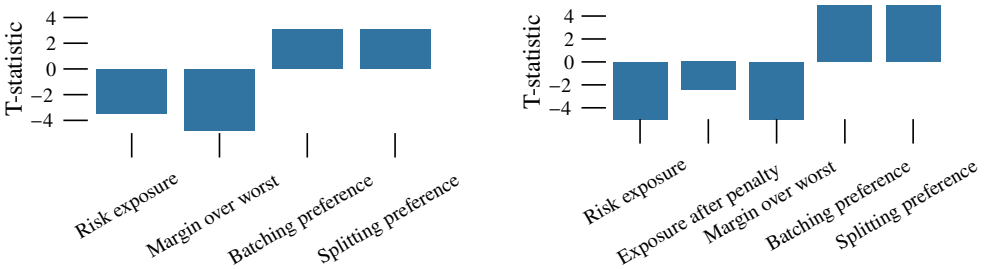
To derive the log-likelihoods of the predicted model, we draw ten times from the (estimated) posterior parameter distributions. For each draw h, and each scenario-participant combination (sc, i), we compute

$$\ell\ell^h_{sc,i} := \sum_{\tau \in \mathcal{D}^k_{sc,i}} \sum_{(s_t, \tilde{a}_t) \in \tau} \log P_{sc,i}(\tilde{a}_t \mid s_t),$$

where $\mathcal{D}^k_{sc,i}$ is the set of trajectories for (sc, i) in the $k \in \{train, test\}$-set. Next, we average the $\ell\ell^h_{sc,i}$ values across h and across combinations of (sc, i). This results in -13.03 for the training set and -13.41 for the test set. While the small difference ($\approx 3\%$) is already encouraging, we note that the same statistic on the test set in the synthetic experiment (Appendix E.3.1) is -11.81. Hence, although there is a risk of model-misspecification when using our approach on real data, the log-likelihoods are of the same order of magnitude as when we eliminate that risk through synthetic data.

Finally, we conduct a posterior predictive check (PPC), essentially the Bayesian counterpart to a goodness-of-fit test (Gelman et al. 1996). For each observation, the model returns a vector of predicted choice probabilities over the set of feasible actions. Concatenating those, we obtain a "probability matrix," with observations in the rows and actions in the columns. We then construct a corresponding "action matrix," in which each row is a one-hot vector indicating the action actually chosen (1 for the chosen action, 0 otherwise). The Brier score is computed as the mean squared difference

Figure EC.13 Comparison of Δ_i^j across clusters (t-test).



(a) Pre-advice phase (b) Post-advice phase

Note. T-statistic = $\frac{1}{|N_o|}\sum_{i\in N_o}\Delta_i^j - \frac{1}{|N\setminus N_o|}\sum_{i\in N\setminus N_o}\Delta_i^j$, where N is the set of participants and N_o are participants in the "Approx. Optimal" cluster in the corresponding scenario. We only display the weights for reward components with significant t-test ($p < 0.05$).

between these two matrices, i.e., we sum the squared differences across actions and average over observations. We choose the Brier score, because it is a well-established, strictly proper scoring rule (Gneiting and Raftery 2007).

We first compute the Brier score for the empirical data, yielding 36.60. We then draw 300 replicated datasets by generating model parameters from the posterior distribution, then simulating actions according to our model with the realized parameters. For each replication, we construct a new action matrix, and re-compute the Brier score against the original probability matrix. The mean score across the 300 replicated datasets is 38.35. Following standard practice, we summarize the comparison with a posterior predictive p-value, defined as the proportion of replicated datasets in which the simulated Brier score exceeds the observed one. In our case, 93% of the replicated datasets yield a larger Brier score than the observed data (i.e., $p \approx 0.93$).

E.3.3. Real Data: Consistency with Qualitative Insights. We also compare the IRL results with the qualitative observations from sequence clustering (Figures 6 and EC.8). Recall that we denoted clusters of participants as "Follow tip" or "Approx. Optimal" if their behavior was in line with the optimal strategy in the with-advice phase, respectively pre- or post-advice phases. We use this classification, comparing participants' (median) reward weight shifts and compliance probability shifts through t-tests. Figure EC.13 shows the t-statistic when comparing $(\Delta_i^j)_{i\in N_{sc}^\star}$ and $(\Delta_i^j)_{i\notin N^\star}$ for a given reward component j, where $N^\star$ are participants in the "Approx. Optimal" cluster of a given phase.

Participants assigned to the "Approx. Optimal" cluster in the pre-advice phase (Figure EC.13a) focus more on batching and splitting decisions, and are less concerned with risk exposure and simplicity. This is in line with a forward-looking approach to decision-making. We observe a similar difference when comparing participants in the "Approx. Optimal" cluster in the post-advice phase to those classified otherwise (Figure EC.13b).

Finally, we run a t-test on the compliance probability in the with-advice phase, differentiating $(\pi_i)_{i\in N^\star}$ and $(\pi_i)_{i\notin N^\star}$. Here, the t-statistic is 13.24 ($p < 0.01$), indicating that the compliance probability is substantially higher for participants in the "Follow tip" cluster. Overall, our model estimates are consistent with our earlier qualitative observations about participants' behaviors, further supporting the validity of our IRL approach.

References for the E-Companion

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518):859–877.

Brown D, Niekum S (2018) Efficient probabilistic performance bounds for inverse reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Choi J, Kim KE (2012) Nonparametric bayesian inverse reinforcement learning for multiple reward functions. *Advances in Neural Information Processing Systems* 25.

Gabadinho A, Ritschard G, Müller NS, Studer M (2011) Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software* 40(4):1–37.

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* 6:733–760.

Gleave A, Toyer S (2022) A primer on maximum causal entropy inverse reinforcement learning. *arXiv preprint arXiv:2203.11409* .

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. *The Journal of Machine Learning Research* 14(1):1303–1347.

Phan D, Pradhan N, Jankowiak M (2019) Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554* .

Ramachandran D, Amir E (2007) Bayesian inverse reinforcement learning. *IJCAI*, volume 7, 2586–2591.

Ziebart BD, Bagnell JA, Dey AK (2010) Modeling interaction via the principle of maximum causal entropy. *International Conference on Machine Learning*, 1255–1262.