Precise or Broad? The Reward-Learning Frontier in Algorithmic Advice Design

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:* Algorithmic tools increasingly guide operational decisions by telling users what action to take. Although precise recommendations can improve immediate execution, organizations may also care about what users can do when guidance is unavailable or conditions change. We study how *advice precision*, whether the same policy is communicated as an executable action or as a qualitative rule the user must apply, affects performance while advice is available and after it is withdrawn. *Methodology/results:* We develop a model of advice-taking and transfer across environments using rational inattention. Precise advice is easier to implement and therefore improves current performance. Broad advice requires users to translate a rule into an action, imposing an immediate implementation cost but preserving more portable decision strategies when the subsequent environment is related but not identical. We test these predictions in two preregistered online experiments in which participants decide when and how much to charge an electric vehicle on a route with uncertain traffic. All recommendations come from the same dynamic-programming policy; treatments vary only how the recommendation is presented. Precise advice produces the highest performance while advice is visible. Specific-broad advice adds the operating condition needed to apply the qualitative rule but withholds the numerical target. Its performance remains close to broad advice and below precise advice, which points to the action-mapping burden as a central source of the short-run gap. After advice is withdrawn, the qualitative formats yield higher unsupported performance relative to precise advice in several related environments when no rationale is shown, but not on an unfamiliar map. Rationales raise agreement with precise advice without reliably helping the qualitative formats, indicating that explanation does not substitute for format. Inverse reinforcement learning estimates are consistent with the precise-advice advantage operating largely through compliance. *Managerial implications:* Precise advice is attractive when immediate execution dominates. Broader advice may be preferable when users must later act without executable guidance in related situations, but its relative value diminishes as the subsequent environment becomes more distant.

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

1. Introduction

Algorithmic decision support increasingly tells people what to do next. Platforms direct drivers toward demand, inventory systems support replenishment decisions under uncertainty, warehouse systems recommend packing decisions (Sun et al. 2022), and clinical decision support guides discretionary prioritization (Ibanez et al. 2018). In each setting, the algorithm may identify an action that improves the immediate outcome, but the organization still chooses how to present that recommendation (Bastani et al. 2026). A system can prescribe a zone, route, quantity, or charge amount. However, it may also communicate a rule, such as remaining near a high-demand corridor or increasing the order-up-to level when underage costs are high. The informational source may be the same. The work left to the decision maker is not.

A user can perform well because the system supplies an action whenever a decision arises, or because the user has internalized enough of the task structure to act without the system. These channels are difficult to separate while advice is visible, but they diverge when guidance is removed. Even when the recommended action is usually optimal, connectivity can fail, model coverage can be incomplete, local conditions can move outside the training distribution, and governance may require humans to supervise rather than passively approve algorithmic output. Evidence from aviation links extensive automation to weaker retention of some cognitive skills needed for manual flight (Casner et al. 2014). Recent work on AI-enabled education, operational expertise, and decision support raises a parallel concern: assistance that improves immediate output can reduce later unassisted performance when it substitutes for users' own problem solving (Lehmann et al. 2024, Bastani et al. 2025, Poulidis et al. 2025a, Siderius et al. 2026). The relevant managerial objective around implementing systems for algorithmic advice is thus broader than adoption or immediate accuracy. It includes what users can do when the interface no longer supplies an executable action.

We study the precision of algorithmic advice as a design lever. *Precise advice* names an executable action. *Broad advice* gives a qualitative rule that narrows the relevant actions but leaves the final choice to the user. A rationale is different: it explains why a recommendation is sensible without changing whether that recommendation is precise or broad. The design choice is therefore how much of the mapping from state to action the system should complete for the user, trading off the simplicity of precise advice with the benefits of exploration and learning.

We formalize this trade-off with a model of advice-taking and transfer across environments based on the rational inattention framework (see Matějka and McKay 2015). Precise advice improves current performance by pointing to a single action. Broad advice leaves a larger set of actions to consider, requiring more effort, but exposes the user to a wider range of options. These channels matter differently after advice is removed. When the environment remains similar, the performance achieved under precise advice can remain useful through imitation. As the environment changes, effortful learning can become more valuable. When the new environment is too different, neither imitation nor learning are useful. The model therefore predicts that the relative value of broad advice is greatest at an intermediate level of environmental novelty.

We apply this model to the context of sequential decision making. Here, an action changes the feedback the user observes, and the future options that remain feasible. Good performance depends on a state-contingent policy, not only on a good one-time action (Kagan et al. 2026). The trade-off between immediate rewards and learning is even more relevant here, because, as our analysis shows, broad advice can lead to more experiential learning, in addition to increased deliberation.

We test our model in two online experiments using an electric vehicle charging task. Participants decide when and how much to charge while traveling along a route with stochastic traffic and nonlinear charging

times. Good performance requires a forward-looking charging policy. We solve the task by dynamic programming and generate every recommendation from the same optimal policy. The treatments change only how recommendations are presented.

In both studies, precise advice produces the highest performance while advice is visible. Study 2 then introduces specific-broad advice, which supplies the operating condition needed to apply a qualitative rule but does not state the numerical target. Specific-broad remains statistically indistinguishable from broad advice and below precise advice in assisted performance and action agreement. Both qualitative formats also produce longer decisions and broader experiences than precise advice in Study 2. Additional strategic detail is therefore not enough to reproduce the short-run performance of a recommendation that supplies the final action.

While Study 1 offers suggestive evidence of post-advice benefits, Study 2 provides a cleaner test because it randomizes the task participants face after advice is removed. Broad advice helps most when the later task preserves the same decision logic but changes the specific conditions. This advantage disappears when participants move to a completely novel environment. The pattern is consistent with the idea that broad advice can teach a rule that transfers across nearby situations, but not across environments that are too different.

Study 2 also introduces rationales that explain why a recommendation is sensible. Rationales improve later performance when paired with precise advice in a familiar post-advice task and increase agreement with precise recommendations while advice is visible. They do not reliably help recipients of broad advice, and the positive effect under precise advice does not carry over to a novel environment. Thus, explanations appear to add useful strategic context to an executable recommendation, but they are not a substitute for changing the action-level format of advice.

Finally, we introduce inverse reinforcement learning (IRL) to recover the decision criteria that best explain participants' action traces. In a sequential task, similar completion times can arise from different strategies: copying the recommendation, maintaining a conservative safety buffer, or learning when to batch and split charges. IRL helps separate compliance with advice from changes in the user's own decision rule. The estimates suggest that precise advice works largely by inducing compliance, whereas broad advice is associated with post-advice strategies that place more weight on the structure of the charging problem.

Overall, our paper shifts the design question from whether users follow algorithmic advice to how the format of that advice affects assisted execution and later unsupported decisions. The design question becomes how much of the mapping from state to action the interface should complete, and what that allocation does to the user's decision process.

1.1. Related literature

Our first contribution is to the operations literature on algorithmic decision support. Field studies show that decision makers often retain discretion even when algorithms recommend actions, and that deviations from prescriptions can materially affect operational performance. Examples include discretionary task ordering in radiology (Ibanez et al. 2018), worker decisions in warehouse operations (Sun et al. 2022), and human-algorithm collaboration when humans hold private information (Balakrishnan et al. 2026). Related theory characterizes recommendation design when a planner anticipates selective adherence (Grand-Clément and Pauphilet 2026), and recent work shows that coarsening an AI signal can improve assisted decisions by reducing precision in a calibrated way (Hoong and Dreyfuss 2025). We take the recommendation policy as fixed and study how the same policy should be communicated when the user may later act without advice.

The paper also relates to work on algorithm aversion, overreliance, and automation. Users may reject algorithmic forecasts after observing errors (Dietvorst et al. 2015, 2018), discount advice when they believe human judgment is more appropriate (Özer et al. 2011, Castelo et al. 2019), or overrely on algorithmic output when the system is easy to accept or appears authoritative (Buçinca et al. 2021, Passi and Vorvoreanu 2022, Vasconcelos et al. 2023). This literature studies whether users accept, reject, or weight advice. Our question is what advice format does to later unsupported behavior in a sequential task. High compliance is valuable while advice is visible, but it can also substitute for the cognitive work through which users form transferable decision strategies.

Sequential operations make this concern especially salient. Current actions shape future states, information, and feasible actions, so observed behavior must be interpreted as a state-contingent strategy rather than as a collection of isolated choices (Kagan et al. 2026). Decision makers in such environments often rely on heuristics rather than optimal policies (Schweitzer and Cachon 2000), and biases in processing feedback can accumulate over time (Tversky and Kahneman 1973, Rabin and Vayanos 2010). Learning depends on exposure to varied conditions and accumulated experience (Lapré and Van Wassenhove 2001, Lapré and Tsikriktsis 2006, Argote and Miron-Spektor 2011), even though exploration can be costly in the short run (March 1991). Related work uses reinforcement learning to identify high-value tips for improving human sequential decision making (Bastani et al. 2026). We study the interface layer that communicates such a policy: an executable action versus a rule the user must implement.

Learning-sciences and automation research provide a mechanism for the trade-off. Generating steps, explanations, or mappings oneself can produce more durable transfer than passively receiving a completed solution, although such productive effort may reduce performance during practice (Sweller 1994, Salden et al. 2010, Bjork et al. 2011). Work on automation and skill retention similarly shows that substituting tools for practice can weaken the capabilities needed when the tool is unavailable (Casner et al. 2014,

Kluge and Frank 2014, Macnamara et al. 2024). Recent evidence on generative AI in education echoes this point: assistance can raise output while weakening later understanding when it substitutes for user reasoning (Lehmann et al. 2024, Bastani et al. 2025). Our contribution is to translate these mechanisms into an operations setting in which the advice is generated from an optimal policy and the outcome of interest is policy transfer.

The closest paper to ours is Poulidis et al. (2025b), who compare action signals with attention signals in chess. Action signals prescribe a move, whereas attention signals mark a position deserving attention. Their results show that action signals have larger immediate effects but can weaken later decisions, while attention signals require more effort and create benefits that persist beyond the signaled state. Broad advice in our setting differs because it communicates a strategic rule that still implies a set of feasible actions, rather than only directing attention to a state. Study 2 further separates strategic detail from action-level implementability through the specific-broad treatment, and the post-advice design varies the novelty of the post-advice environment.

Our rationale treatment connects the paper to explainable AI and human-AI collaboration. Explanations can support transparency, engagement, and trust calibration (Bader et al. 2011, Ghai et al. 2021, Buçinca et al. 2021, Bauer et al. 2023, Vereschak et al. 2021). They can also add cognitive load, create an illusion of understanding, rationalize a black-box recommendation, or encourage automation bias (Rudin 2019, Chromik et al. 2021, Passi and Vorvoreanu 2022). We therefore treat rationales as an interface feature whose effect depends on advice format. A rationale can supply strategic context to a numerical prescription, but it is not equivalent to broad advice because it does not change the action-level format of the recommendation.

Finally, the IRL analysis contributes to measurement in behavioral operations management (BOM). Most studies of decision support systems evaluate actions, adherence, or performance. In sequential tasks, however, similar outcomes can be generated by different policies. IRL provides a principled way to infer participants' policies from action traces (for a survey, see Arora and Doshi 2021). To the best of our knowledge, we provide the first study using IRL in the context of BOM. We employ IRL as mechanism evidence, in particular, to distinguish compliance with recommendations from persistent differences (and changes) in participants' decision criteria.

2. Advice precision and the reward-learning frontier

We focus on a task executed by a human decision maker who may be supported by algorithmic advice, which is distinct in its *breadth*. The model is deliberately stylized: Its role is to isolate how the same design feature that makes advice easy to implement may also reduce the effort and state-space exposure through which decision makers learn. We use the model to organize the empirical design and generate testable predictions, rather than to characterize all forms of advice-taking.

As an example, 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: They choose a base-stock level S, then all ordering decisions are fully defined (ordering up to S). Let $u > 0$ and $o > 0$ denote underage and overage costs. For demand with cumulative distribution function F, the optimal base stock S^* is the minimum quantity such that $F(S^*) \geq \frac{u}{u+o}$. *Precise* advice, with minimal breadth, could be to "order up to S^*," pointing out one specific action within the large set of feasible actions available to the manager. *Broad* advice could communicate a rule mapping (u, o, F) to S^*, for instance, "if u is higher than o, order above the median," indicating that the optimal action lies within a specific subset of the feasible set.

Precise advice is easy to implement. However, broad advice may support learning by forcing the decision maker to engage with different alternatives. This is particularly useful when advice cannot always be provided, for example, when workers need training to effectively supervise algorithms.

2.1. A simple model of one-shot decision making

While we aim to understand the decision maker's reaction to advice in *sequential decision problems*, where they have to make multiple decisions in a row to achieve an overall goal, we start by modeling a one-shot decision task. This allows us to isolate key trade-offs faced by the designer and define measurable behaviors. In §2.2, we show how these trade-offs extend to sequential decision making.

Setup and notation. Consider a task carried out by a human decision maker, wherein one out of a finite (but large) set of feasible actions has to be chosen. The decision maker sequentially faces three decision-making environments, E_0–E_2, where $\mathcal{A}_\tau$ denotes the set of available actions in environment E_τ. In E_1 only, the decision maker is supported by algorithmic advice.

Rewards. If the decision maker chooses the "best" action, they receive reward R^+, while they receive R^- otherwise. We assume that $r := \mathbb{E}[R^+] - \mathbb{E}[R^-] > 0$, and the decision maker maximizes the expected reward, minus the cost of effort in the current environment. When decision maker i faces E_τ, they have probability $p_{i,\tau}$ of choosing the best action without any effort, due to some fallback heuristic (e.g., Grand-Clément and Pauphilet 2026). When talking about a single decision maker, we use p_τ for simplicity. The probabilities p_τ and $p_{\tau+1}$ are linked, as we explore below.

Information processing. In a given environment E_τ, the decision maker may consider a set of possible actions $\mathcal{B}_\tau \subseteq \mathcal{A}_\tau$ with $B_\tau = |\mathcal{B}_\tau|$. We will discuss below how $\mathcal{B}_\tau$ relates to $\mathcal{A}_\tau$. Let A^*_τ be the best action (leading to R^+). The decision maker believes that $P(A^*_\tau \in \mathcal{B}_\tau) = \rho_\tau \in (0, 1]$ (we assume the same ρ_τ for all decision makers for tractability). Moreover, their prior about which of the possible actions from $\mathcal{B}_\tau$ is the correct one is uniform: $P(A^*_\tau = a \mid A^*_\tau \in \mathcal{B}_\tau) = 1/B_\tau \ \forall a \in \mathcal{B}_\tau$.

The decision maker can explore, or process, the B_τ different actions to increase the probability of picking the best one. In determining how the decision maker processes this information, we follow the framework

of rational inattention (Matějka and McKay 2015, Boyacı et al. 2024). In particular, let A_τ be the action chosen by the decision maker and denote with q their accuracy, that is, for $B_\tau > 1$, $P(A_\tau = a_\tau^* \mid A_\tau^* = a_\tau^*, A_\tau^* \in \mathcal{B}_\tau) = q$ and $P(A_\tau = a \mid A_\tau^* = a_\tau^*, A_\tau^* \in \mathcal{B}_\tau) = \frac{1-q}{B_\tau - 1}$ $\forall a \in \mathcal{B}_\tau \setminus \{a_\tau^*\}$. Then, the decision maker needs to exert effort to achieve this level of accuracy proportional to the information required, $M_{B_\tau}(q) = \log B_\tau + q \log q + (1-q) \log\left(\frac{1-q}{B_\tau - 1}\right)$. We provide a full derivation in Appendix A.1. When $B_\tau = 1$, we assume that $M_1(q) = 0$, because the only considered action is the optimal one, conditional on $A_\tau^* \in \mathcal{B}_\tau$.

If the decision maker decides to engage with action set $\mathcal{B}_\tau$, they choose their precision by solving

$$\max_{q \in [1/B_\tau, 1]} r\rho_\tau q - \kappa M_{B_\tau}(q) + \mathbb{E}[R^-],$$

where $\kappa > 0$ is the cost of information processing. Denoting $K_\tau = \exp(r\rho_\tau/\kappa)$, we have $q^* = q^*(B_\tau, \rho_\tau) = \frac{K_\tau}{K_\tau + B_\tau - 1}$, with $q^* = 1$ when $B_\tau = 1$. Hence, the decision maker's reward from engaging with set $\mathcal{B}_\tau$ is $V(B_\tau, \rho_\tau) := r\rho_\tau q^* - \kappa M_{B_\tau}(q^*) + \mathbb{E}[R^-] = \kappa \log\left(\frac{K_\tau + B_\tau - 1}{B_\tau}\right) + \mathbb{E}[R^-]$, with $V(1, \rho_\tau) = r\rho_\tau + \mathbb{E}[R^-]$. Without engagement, the decision maker's reward is $p_\tau r + \mathbb{E}[R^-]$, as there is no effort. Thus, the decision maker engages with the action set $\mathcal{B}_\tau$ if and only if $\kappa \log\left(\frac{K_\tau + B_\tau - 1}{B_\tau}\right) \geq p_\tau r$. We let the indicator $u(p_\tau, B_\tau, \rho_\tau) = 1$ if the inequality holds, and zero otherwise.

Breadth of advice. Absent advice, the decision maker has no information about how they can focus their exploration efforts. Hence, $\mathcal{B}_\tau = \mathcal{A}_\tau$ (and, thus, $\rho_\tau = 1$). Advice, on the other hand, is intended to focus the decision maker's efforts, so $\mathcal{B}_\tau \subsetneq \mathcal{A}_\tau$ and $\rho_\tau \leq 1$. At the most extreme, *precise* advice focuses the decision maker on a single action, such as an exact base stock S^* in our example. We denote advice with $B > 1$ as *broad*. The decision maker may choose to explore $\mathcal{B}_\tau$ as indicated by the advice, or $\mathcal{A}_\tau \setminus \mathcal{B}_\tau$ if they do not trust the advice. However, we focus on a relatively high trust setting (high ρ_τ), so that, if any exploration takes place, it is of the set $\mathcal{B}_\tau$. Hence, the decision maker's accuracy in environment E_τ is $A(p_\tau, B_\tau, \rho_\tau) := p_\tau + u(p_\tau, B_\tau, \rho_\tau)\,[q^*(B_\tau, \rho_\tau) - p_\tau]$, and their expected reward is $rA(p_\tau, B_\tau, \rho_\tau) + \mathbb{E}[R^-]$.

Learning and imitation. Decision makers carry knowledge across environments by updating their fallback heuristic. We assume that they *learn* from (costly) exploration, and learning is a function of exploration effort: $L(p_\tau, B_\tau, \rho_\tau) := \Psi\left(u(p_\tau, B_\tau, \rho_\tau) \times M_{B_\tau}\left(q^*(B_\tau, \rho_\tau)\right)\right)$, where we assume some function $\Psi : [0, \infty) \to [0, 1]$ with $\Psi(0) = 0$, $\Psi'(\cdot) > 0$, and $\Psi''(\cdot) \leq 0$. Thus, exploration has increasing but diminishing returns to learning. Learning, in turn, can be applied if the next environment encountered is similar enough to the previous ones. In particular, we assume that learning L enters the decision maker's updated heuristic in environment $\tau + 1$ with a factor $H(\delta_{\tau+1})$, where $\delta_{\tau+1} \in [0, 1]$ presents the structural difference between $\mathrm{E}_{\tau+1}$ and the previous environments, or the *novelty* of $\mathrm{E}_{\tau+1}$. We assume that $H(0) > 0$, $H'(\delta_{\tau+1}) < 0$, and $\lim_{\delta \to 1} H(\delta) = 0$, meaning that, when $\mathrm{E}_{\tau+1}$ is completely novel, learning does not have any use.

The decision maker may also *imitate* previous actions if the new environment is sufficiently similar to the previous ones. This is reflected by adjusting $p_{\tau+1}$ towards $A(p_\tau, B_\tau, \rho_\tau)$. We use a factor $D(\delta_{\tau+1})$ to moderate the effect, where $D(0) > 0$, $D'(\delta_{\tau+1}) < 0$, and $\lim_{\delta\to 1} D(\delta) = 0$. Assuming similar task difficulty, we have the following update to the decision maker's heuristic:

$$p_{\tau+1} = p_\tau + D(\delta_{\tau+1})\left[A(p_\tau, B_\tau, \rho_\tau) - p_\tau\right] + H(\delta_{\tau+1})(1-p_\tau)L(p_\tau, B_\tau, \rho_\tau).$$

For this to be well-defined, we require $H(0) + D(0) \leq 1$. Based on evidence that observational learning improves immediate performance without producing readily articulable schemas, whereas analogical learning and experience across related tasks produces more abstract and transferable knowledge (Nadler et al. 2003, Schilling et al. 2003), we assume that the effectiveness of imitation decays more quickly with environmental novelty than that of effortful learning, i.e., $D(\delta)/H(\delta)$ is decreasing in δ whenever $H(\delta) > 0$.

Recall our example. Under common assumptions, a deviation from the optimal base stock S^* 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 effectively use the previous S^* to arrive at a near-optimal solution and higher performance in E_1 translates directly into higher performance 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 by evaluating different possibilities, they may be more successful at setting a sensible base-stock level in E_2 when the environment differs substantially.

In the analysis, we do not explicitly consider the pre-advice phase E_0 and the transition from $p_0 \to p_1$. Instead, we assume the decision maker enters E_1 with a fallback heuristic with p_1, a sufficient statistic. Consequently, we only consider the novelty of E_2 and let $\delta := \delta_2$.

Designer's objective. The designer determines the breadth of advice to provide in E_1, that is, the size $B := B_1$ of the set $\mathcal{B}_1$. We assume that the optimal action $a_1^* \in \mathcal{B}_1$, even though the decision maker may have doubts about the advice quality, i.e., we allow for $\rho_1 < 1$. The designer is risk neutral and cares about performance in environments E_1–E_2, but not directly about the decision maker's cognitive cost (in Appendix A.4, we show that results are qualitatively unchanged if they do). If $\gamma \in [0,1]$ is the designer's weight for each environment, and within-environment rewards are proportional to those of the decision maker, we arrive at the designer's objective function:

$$J(B) := \mathbb{E}[R^-] + r\left[\gamma A(p_1, B, \rho_1) + (1-\gamma)A(p_2, |\mathcal{A}_2|, 1)\right],$$

where we recall that $\mathcal{B}_2 = \mathcal{A}_2$ and, thus, $\rho_2 = 1$.

Expected behaviors. We derive several hypotheses about the reward-learning frontier. First, we consider the impact of advice breadth on decision makers' behaviors (proofs in Appendix A.2):

LEMMA 1. *Assume some environment E_τ:*

(i) Let $\bar{B}_\tau := \frac{\exp(r\rho_\tau/\kappa)-1}{\exp(r p_\tau/\kappa)-1}$. Then, $u(p_\tau, B_\tau, \rho_\tau) = 1 \Leftrightarrow B_\tau \le \bar{B}_\tau$. $\bar{B}_\tau \ge 1$ if and only if $p_\tau \le \rho_\tau$.

(ii) Let $p_{i,\tau} \ge p_{j,\tau}$ for two decision makers $i \ne j$. Then, $u(p_{i,\tau}, B_\tau, \rho_\tau) \le u(p_{j,\tau}, B_\tau, \rho_\tau)$.

(iii) $q^*(B_\tau, \rho_\tau)$ is strictly decreasing in B_τ.

(iv) $A(p_\tau, B_\tau, \rho_\tau)$ is weakly decreasing in B_τ. It is strictly decreasing for $B_\tau \le \bar{B}_\tau$.

(v) $L(p_\tau, 1, \rho_\tau) \ge 0$ with $L(p_\tau, 1, \rho_\tau) = 0$ and $L(p_\tau, B_\tau, \rho_\tau) = 0\ \forall B > \bar{B}_\tau$. If $\bar{B}_\tau \ge 2$, then L is unimodal on $B_\tau \in \{2, \dots, \lfloor \bar{B}_\tau \rfloor\}$, allowing for two consecutive points taking the maximal value. If we assume that B_τ is continuous, then $B^*_\tau = \min\left\{\bar{B}_\tau, \frac{(K_\tau-1)^2}{K_\tau(\log K_\tau - 1)+1}\right\}$ is the unique maximizer of $L(p_\tau, B_\tau, \rho_\tau)$.

Advice is used when B is sufficiently small (i). This threshold depends on individual characteristics, reflected in the baseline performance p_0, which impacts p_1. We hypothesize:

(H_{1a}) *Compliance.* Compliance with advice is lower, the broader the advice.

Meanwhile, broad advice leads to less accuracy when used (iii). Thus,

(H_{1b}) *Exploration.* The broader the advice, the more dispersed the actions of compliant participants.

This conditional reduction of accuracy translates into a performance difference (iv):

(H_{1c}) *Reward gap.* The broader the advice, the lower the E_1 performance of compliant participants.

There is another side of the reward-learning trade-off: learning that is enabled by access to broad advice and the corresponding incentive to exert effort and explore different alternatives (v). Thus,

(H_{1d}) *Effort gap.* The broader the advice, the more effort exerted by compliant participants.

Next, we consider the decision maker's learning process. That is, the change in the baseline probability of choosing the correct action from one environment to the next:

PROPOSITION 1. *We have $p_\tau \le A(p_\tau, B_\tau, \rho_\tau)$ and $p_\tau \le p_{\tau+1} \le 1$.*

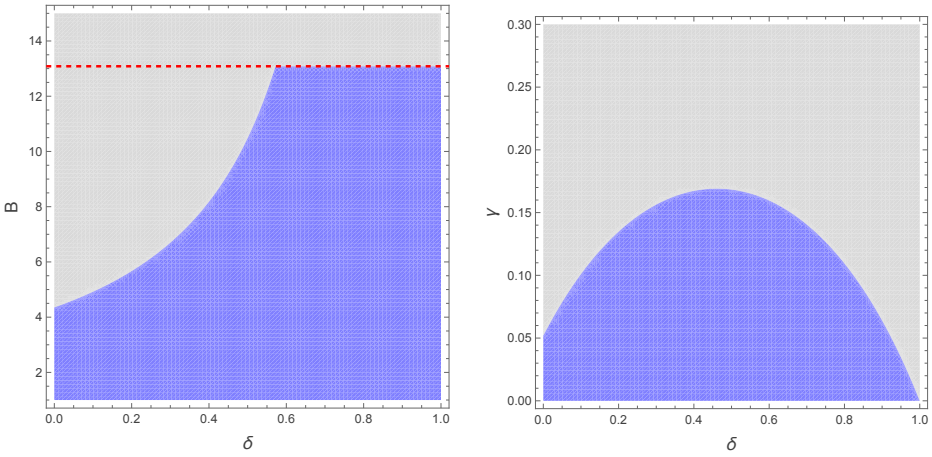
Hence, if tasks are of similar difficulty, we would expect performance absent advice to improve:

(H_2) *Performance trajectory.* Participants exposed to similar task difficulties in E_0 and E_2 have higher performance in the latter environment.

Finally, we consider the designer's choice between *precise* advice in E_1 ($B = 1$), and *broad* advice with $B > 1$ (including none, i.e., $B_1 = |\mathcal{A}_1|$). Our first result here makes explicit the accuracy advantage of precise advice that drives the reward gap but may also affect E_2 performance:

LEMMA 2. *We have $\alpha_B := A(p_1, 1, \rho_1) - A(p_1, B, \rho_1) \ge 0$. If $\bar{B}_1 > 1$, then $\alpha_B > 0$. Moreover, if $\bar{B}_1 \ge B$, then $\alpha_B = \frac{B-1}{\exp(r\rho_1/\kappa)+B-1}$.*

Figure 1 Broad compared to precise advice.



(a) E_2 performance (b) Total performance

Note. The designer prefers *precise* advice ($B = 1$) within the gray region and *broad* advice ($B > 1$) in the blue region, based on (a) E_2 performance, or (b) total weighted performance. In (a), there is no compliance with *broad* advice in E_1 above the red line. Other parameters and functions are $p_1 = 0.2$, $r = 1$, $\rho = 0.9$, $\kappa = 0.4$, $H(\delta) = 0.3(1-\delta)$, $D(\delta) = 0.7(1-\delta)^2$, $\Psi(\cdot) = \sqrt{\cdot}$, and $B = 5$ in (b).

We now turn to E_2 specifically. Denote with $p_2(B)$ the baseline probability of choosing the best action in environment E_2 after having observed advice with breadth B:

PROPOSITION 2. *Let* $\alpha_B := A(p_1, 1, \rho_1) - A(p_1, B, \rho_1) \geq 0$ *and* $\beta_B := (1-p_1)L(p_1, B, \rho_1) \geq 0$. *Then,* $p_2(B) - p_2(1) = H(\delta)\beta_B - D(\delta)\alpha_B$, *with* $\lim_{\delta \to 1} p_2(B) - p_2(1) = 0$.

From Lemma 1, $\beta_B > 0$ if and only if $B \leq \bar{B}_1$ and $p_1 < 1$. Then, $1 < B \leq \bar{B}_1$, and also $\alpha_B > 0$ (Lemma 2). Thus, if B is relatively large, precise advice may be followed, leading to an imitation effect, but broad advice not, so there is no learning. Now, assume $B \leq \bar{B}_1$. While α_B is strictly increasing in B, β_B is first increasing, then decreasing. Moreover, $D(\delta)/H(\delta)$ is decreasing in δ. Hence, $p_2(B) - p_2(1)$ tends to be largest when the environment is sufficiently novel and B is at an intermediate level, as shown in Figure 1a. Because $D(1) = H(1) = 0$, any difference will tend to zero as the novelty is too high. Note that, unless $\mathcal{A}_2$ is very small, we have $A(p_2, |\mathcal{A}_2|, 1) = p_2$, so E_2 performance is likely to behave in the same way as p_2. We have:

(H_{3a}) *Learning gap.* If E_2 is sufficiently novel and B not too large, participants complying with broad advice in E_1 have higher E_2 performance than those complying with precise advice.

(H_{3b}) *High novelty.* When E_2 is very novel, the previous performance difference disappears.

Figure 1b visualizes the trade-off between immediate performance and learning that the designer faces between broad and precise advice across environments E_1–E_2. We also state their optimal advice breadth under the assumption of a linear learning function Ψ:

PROPOSITION 3. *Assume* $\Psi(m) = m$, $p_1 < 1$, $|\mathcal{A}_2| > \bar{B}_2$, *and* $\gamma < 1$, *and let*

$$B^* = \frac{(1-\gamma)(1-p_1)(K_1-1)^2 H(\delta)}{(1-\gamma)(1-p_1)(K_1 \log K_1 - K_1 + 1)H(\delta) + K_1\left[\gamma + (1-\gamma)D(\delta)\right]}.$$

If $\lceil B^* \rceil \leq \bar{B}_1$, *the designer optimally chooses* $\lfloor B^* \rfloor$ *or* $\lceil B^* \rceil$. *If* $B^* \leq 1$, *precise advice is optimal. Moreover,* $dB^*/d\delta < 0$ *for all* $\delta > \delta_0$, *for some* $\delta_0 \in [0,1]$. *If* $\gamma \leq \gamma_0$ *for some* $\gamma_0 \in [0,1)$, *then* $\delta_0 > 0$.

The effect of explanations. Consider now what happens if the designer includes explanations in its advice. There are three channels through which explanations may affect the decision maker:

Perceived quality channel. In line with evidence that explanations can strengthen users' trust in algorithmic recommendations (Wang et al. 2018), explanations may increase ρ_1. This makes it more likely that advice is used, and also increases $q^*(B, \rho_1)$, conditional on use. Accuracy in E_2 can be improved through imitation of the higher E_1 accuracy. However, conditional on advice use, there is no additional learning.

Processing cost channel. Decision-support systems can alter the effort required to implement a decision (Todd and Benbasat 1991). We therefore consider that explanations may lower the effort cost κ. Whenever $B > 1$, this increases $q^*(B, \rho_1)$, but it does not affect learning under precise advice. Hence, conditional on advice uptake, both E_1 performance (directly) and E_2 performance (through learning) is improved.

Effective breadth channel. An explanation may change the effective breadth of advice, e.g., by making salient counterfactuals in order to justify an action. With precise advice, $B = 1$, so this is likely to increase breadth substantially. As long as effective B is not increased too much, advice continues to be used, but now enables learning, improving E_2 performance. For broad advice, because the set of actions highlighted by the explanation may overlap with the set of actions implied by the advice, the increase in effective B is attenuated. Moreover, $B > 1$ may already be in the decreasing part of $L(p_1, B, \rho_1)$ where a higher effective breadth reduces learning.

We propose the following empirical comparison to elicit the active channel(s) through which explanations act. In particular, let $\tilde{Y}$ denote E_2 performance after controlling for E_0 performance, E_1 performance, and compliance with advice during E_1. Moreover, let $\mathcal{T}_B = \mathbb{E}\left[\tilde{Y} \mid B, 1\right] - \mathbb{E}\left[\tilde{Y} \mid B, 0\right]$, where the former term indicates that advice of breadth B is available together with explanations, while the latter indicates that advice of breadth B is available without explanations.

(Oa) *Perceived quality.* If this channel dominates, we expect $\mathcal{T}_B \approx 0$ for any level of B.

(Ob) *Processing costs.* If this channel dominates, we expect $\mathcal{T}_B > 0$ for $B > 1$, but $\mathcal{T}_1 \approx 0$.

(Oc) *Effective breadth.* If this channel dominates, we expect $\mathcal{T}_1 > 0$ and $\mathcal{T}_B < \mathcal{T}_1$ for $B > 1$.

2.2. Sequential decision making: Adding experiential learning

We now extend the model to sequential decision-making tasks, highlighting that the trade-offs related to advice breadth remain and the role of learning becomes more pronounced.

Modifications in setup and notation. In E_1, the decision maker faces a time-homogeneous MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r_s(\cdot), s_1)$ with horizon $T \in \mathbb{N}$. At $t = 1, \ldots, T$, the system is in state $s_t \in \mathcal{S}$, the decision maker chooses action $\tilde{a}_t \in \mathcal{A}$, the next state is drawn from $P(\cdot \mid s_t, \tilde{a}_t)$, and reward $r_{s_t}(\tilde{a}_t)$ accrues. In our example, this could be the case of non-stationary demand, where the manager aims to find an optimal policy accounting for the current state and future demand predictions.

The decision maker is supported by algorithmic advice, and the algorithm has access to an optimal policy. Let $\tilde{a}^*(s)$ denote an optimal action at state s. We consider stationary advice regimes: Regime a induces at each state s a consideration set $\mathcal{B}_a(s) \subseteq \mathcal{A}$ containing $\tilde{a}^*(s)$ and with breadth $B_a(s) := |\mathcal{B}_a(s)|$. Precise advice ($a = p$) has breadth $B_p(s) = 1$. Broad advice ($a = b$) has $B_b(s) \geq 1$ with $B_b(s) > 1$ for some states. Let $\mu_i(\cdot \mid s)$ be the fallback policy of decision maker i (we again drop the subscript i when we talk of a single decision maker) and define its state-specific accuracy by $p_\mu(s) := \mu(\tilde{a}^*(s) \mid s)$. Using the operators from the one-shot model, the optimal accuracy when the decision maker engages with the advice-induced set is $q_a(s) := q^*(B_a(s), \rho_1)$, and the indicator of engagement is $u_a(s) := u(p_\mu(s), B_a(s), \rho_1)$.

For $B_a(s) > 1$, the rational inattention problem induces the conditional choice distribution

$$\nu_a(\tilde{a} \mid s) = \begin{cases} q_a(s), & \tilde{a} = \tilde{a}^*(s), \\ \frac{1-q_a(s)}{B_a(s)-1}, & \tilde{a} \in \mathcal{B}_a(s) \setminus \{\tilde{a}^*(s)\}, \\ 0, & \tilde{a} \notin \mathcal{B}_a(s). \end{cases}$$

For $B_a(s) = 1$, $\nu_a(\tilde{a}^*(s) \mid s) = 1$ and $\nu_a(\cdot \mid s) = 0$ otherwise. The decision maker executes the policy $\sigma_a(\tilde{a} \mid s) = u_a(s)\nu_a(\tilde{a} \mid s) + (1 - u_a(s))\mu(\tilde{a} \mid s)$, inducing the transition matrix $P_a(s' \mid s) = \sum_{\tilde{a} \in \mathcal{A}} \sigma_a(\tilde{a} \mid s)P(s' \mid s, \tilde{a})$. We denote the long-run action support with $C_a := \sum_{s \in \mathcal{S}} |\text{supp } \sigma_a(\cdot \mid s)|$.

In E_0 and E_2, the decision maker faces a task without access to advice. This may be a similar MDP, or another type of task. Importantly, the decision maker's performance in E_2 depends on knowledge formed in E_1 (and in E_0, although this is unaffected by advice, so we can omit it).

Information and learning. As in the one-shot model, we assume that decision makers learn about different actions by putting effort into *deliberating* actions suggested by the advice. Unlike in the one-shot model, decision makers also *experience* multiple state-action-outcome realizations because of the multiplicity of decision points. This can generate experiential learning, although its marginal contribution typically diminishes as experience accumulates (KC and Staats 2012). We assume that diminishing returns equally apply to deliberation. Hence, we define several constructs: the count of state visits under regime a, $N_s^a(T) := \sum_{t=1}^T \mathbf{1}\{s_t = s\}$, the count of state-action visits, $N_{s,\tilde{a}}^a := \sum_{t=1}^T \mathbf{1}\{s_t = s, \tilde{a}_t = \tilde{a}\}$, and the function controlling diminishing returns, $h : [0, \infty) \to [0, 1]$ with $h(0) = 0$, $h'(n) > 0$, $h''(n) < 0$, and $\lim_{n \to \infty} h(n) = 1$.

Using the one-shot model, we have $m_a(s) := u_a(s)M_{B_a(s)}(q_a(s))$, so we define possible learning from deliberation as $L_T^{del}(a) := \mathbb{E}_a\left[\sum_{s \in \mathcal{S}} m_a(s)h\left(N_s^a(T)\right)\right]$. Thus, a new state-specific comparison can add

knowledge, but resolving the same comparison has diminishing value. For experiential learning, we assume that every state-action pair contains the same total amount of transferable information $\bar{I} > 0$. The decision maker internalizes a fraction $h(n)$ of that after experiencing the pair n times. We define $L_T^{exp}(a) := \bar{I}\,\mathbb{E}_a\left[\sum_{s\in\mathcal{S}}\sum_{\tilde{a}\in\mathcal{A}} h\left(N_{s,\tilde{a}}^a(T)\right)\right]$. Then, total learning from E_1 is $L_T(a) := \eta_{del} L_T^{del}(a) + \eta_{exp} L_T^{exp}(a)$ for some $\eta_{del} \geq 0$ and $\eta_{exp} \geq 0$.

Learning under long horizons. Broad advice benefits learning through both deliberation and experience. Holding fixed the number of visits to a state, broad advice produces greater expected coverage of different actions and, thus, more experiential learning (Lemmas EC.1–EC.2 in Appendix A.3).

This does not necessarily translate into more learning, because different advice regimes may lead to different sequences of states visited. However, it turns out that there is more learning under broad advice, both from deliberation and experience, if the decision sequence is sufficiently long:

THEOREM 1. *Assume that P_a is irreducible on the same finite state space $\mathcal{S}$ for $a \in \{p, b\}$, $\rho_1 > 0$, and there exists a state $s^\circ \in \mathcal{S}$ such that $B_b(s^\circ) > 1$ and $u_b(s^\circ) = 1$. Let $a \in \{p, b\}$. Then,*

$$L_\infty(a) := \lim_{T\to\infty} L_T(a) = \eta_{del} \sum_{s\in\mathcal{S}} m_a(s) + \eta_{exp}\bar{I}C_a.$$

Moreover, let $\ell_{del} := \sum_{s\in\mathcal{S}} m_b(s)$ and $\ell_{exp} := \bar{I}(C_b - C_p)$. Then, $\ell_{del} > 0$, $\ell_{exp} > 0$, and $L_\infty(b) - L_\infty(p) = \eta_{del}\ell_{del} + \eta_{exp}\ell_{exp}$. Consequently, if $\eta_{del} + \eta_{exp} > 0$, there exists $T_0 < \infty$, such that $L_T(b) > L_T(p)\ \forall T \geq T_0$.

Theorem 1 (shown in Appendix A.3) indicates that broad advice has a learning advantage, even when considering the actual states visited, if the decision horizon is long enough. This stems from more deliberation about suggested actions, a broader set of experiences, or both. Note that this does not necessarily translate into an advantage in E_2, because of the trade-offs between learning and imitation, and the moderation of both effects through environmental novelty. Modeling a sequential decision making task in E_2 adds substantial complexity, but does not affect these trade-offs, so we refer the reader to our one-shot model above.

3. Experimental setting

We study advice precision in a repeated electric vehicle (EV) charging game. In each round, a participant travels from an origin to a destination and decides at each highway exit whether to recharge and by how much. The objective is to minimize elapsed game time. Decisions are sequential: adding more charge can avoid a later stop, but charging becomes slower at higher battery levels, and realized traffic determines how much charge the next segment consumes. Good performance therefore requires a state-contingent charging policy rather than a collection of independent choices.

The setting provides the variation needed to examine the model. We can obtain the optimal action for each state through dynamic programming and generate all advice formats from the same underlying policy. We

observe both outcomes and decision processes, including decision time, information gathering efforts, and the state-action combinations participants experience. We can also remove advice (E_2) and vary the novelty of the environment encountered.

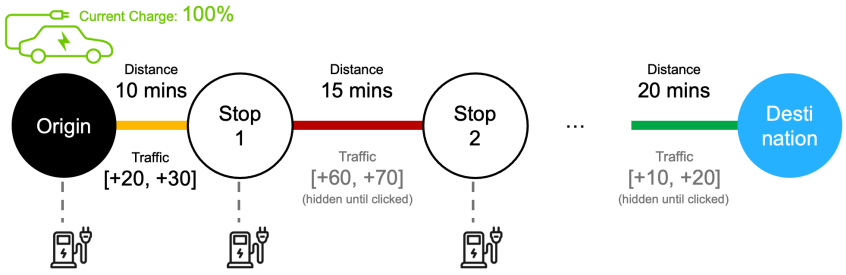
The main rounds follow the model's three environments. Participants first act without advice in E_0, receive advice in E_1, and again act without advice in E_2. The model treats the fallback policy at the start of E_1 as a sufficient statistic for experience accumulated beforehand. Empirically, the E_0 rounds measure the baseline policy and provide covariates or matched comparisons for the later E_2 rounds. Because the experiments randomize advice format rather than the availability of advice during E_1, all post-advice treatment effects are relative comparisons across advice formats. They do not identify what would have happened under continued unaided practice.

3.1. The EV charging task

A map consists of an origin, a destination, and a sequence of highway segments connected by exits, which we call stops. Distances are expressed in in-game minutes. Battery charge is measured in percentage points, and one percentage point of charge permits one in-game minute of driving. The state at a decision point is defined by the current stop and remaining charge. The participant chooses whether to continue or exit and, conditional on exiting, how much charge to add.

Traffic creates uncertainty about the charge required. Participants observe the next segment's traffic range automatically and may reveal the ranges for later segments before acting. After a segment is traversed, its traffic is realized and the participant observes the resulting charge and elapsed game time. The task therefore supplies the repeated state-action-outcome feedback that can support experiential learning. Figure 2 illustrates the interface at a decision point.

Figure 2 Illustration of the experimental task.



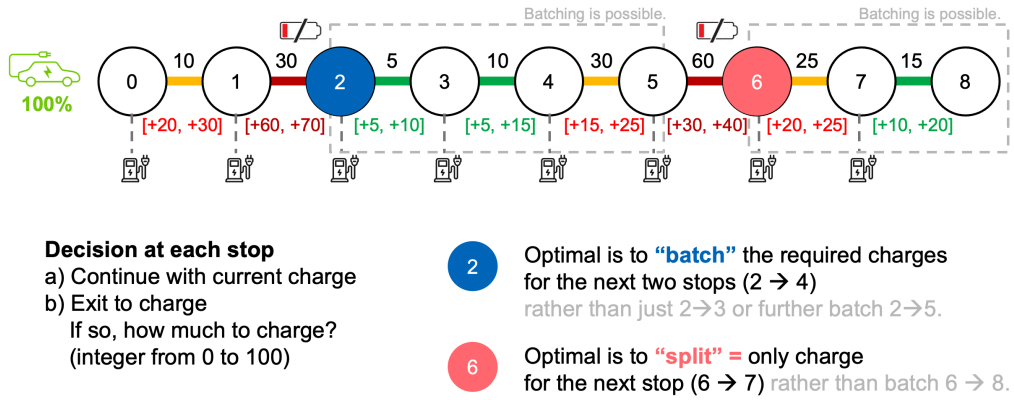
Note. Each circle represents a highway exit, or stop, and each line represents a highway segment. Each segment is labeled with its length in in-game minutes and, when revealed, its traffic range. The car symbol indicates the current location, and the current charge level is displayed next to it.

Exiting the highway to charge adds 30 in-game minutes. Charging time is increasing and convex in the amount charged (Appendix B.1 reports the exact function). Stopping frequently keeps individual charging

amounts small but incurs the fixed exit cost repeatedly. Charging for several future segments at once saves exits but requires a larger charge at a higher marginal time cost. If realized traffic leaves the vehicle without enough charge to complete the next segment, the participant incurs a 300-minute emergency penalty and arrives at the next stop with zero charge.

For each map, we compute the policy that minimizes expected elapsed game time. The dynamic program accounts for current stop and charge, displayed traffic ranges, fixed exit cost, nonlinear charging time, and emergency penalty. The resulting policy generally chooses between charging for the next segment only, which we call *splitting*, and charging for two or more consecutive segments, which we call *batching*. Figure 3 illustrates this trade-off. Advice is generated from this policy before the experiment and never uses traffic realizations that are hidden from participants.

Figure 3 Illustration of batching and splitting.



Note. Splitting uses more stops and smaller charging amounts. Batching uses fewer stops but requires enough charge to cover multiple future segments.

3.2. Advice treatments

All treatment conditions draw on the same optimal policy. The treatments change what the interface leaves for the participant to do. Precise advice states whether to exit and, when charging is recommended, gives the numerical target charge. Broad advice states the charging rule, such as charging for the current and next segment, but leaves the numerical target to the participant. Precise advice is therefore the experimental counterpart of a singleton consideration set in the model, whereas broad advice leaves several numerical implementations available.

Study 2 adds specific-broad advice. This treatment retains the qualitative rule but supplies the relevant operating condition, such as using the maximum traffic within the displayed range. It provides more detail than broad advice while still requiring the participant to calculate the charge. Study 2 also varies whether

the recommendation is accompanied by an explanation, or *rationale*. This details the logic of the recommendation, such as the fixed cost of leaving the highway or the nonlinear charging-time function, without changing the recommendation itself.

3.3. Procedure

Participants first read the instructions and complete a guided training round, teaching the interface, traffic display, charging controls, and emergency penalty. This is separate from the main task and is excluded from all analyses. Participants then complete the main rounds after passing the study's comprehension and attention checks. Study-specific recruitment platforms, exclusion rules, treatment cell counts, and payment schedules are reported in Appendix B.

The main task consists of a pre-advice phase (E_0), an advice phase (E_1), and a post-advice phase (E_2). The pre-advice rounds establish baseline performance and behavior. Advice format is assigned between participants before the advice phase. During the post-advice rounds, the interface no longer displays a recommendation. The two studies differ in the maps used in each phase and in the treatment factors crossed with advice format. Sections 4 and 5 describe the specific designs.

Performance is incentivized. Participants receive a fixed payment and a bonus that increases as elapsed game time decreases. Elapsed game time includes driving, exit, charging, and emergency-penalty time. After the game, participants report the strategy they used, their response to the advice, prior driving and EV experience, navigation-app use, experience with similar games, and demographic information. Appendix B provides the full instructions, screenshots, comprehension checks, participant flow, compensation, duration, and survey measures.

3.4. Measures and estimation

Our performance outcome is normalized elapsed game time, computed at the participant-round level. Let T_{ir}^{actual} denote participant i's realized elapsed game time in round r, T_{ir}^{opt} the elapsed time obtained by the dynamic-programming benchmark for the same realized traffic path, and T_r^{worst} the worst benchmark for that map and traffic condition. We define

$$Y_{ir} = \frac{T_r^{\text{worst}} - T_{ir}^{\text{actual}}}{T_r^{\text{worst}} - T_{ir}^{\text{opt}}}. \tag{1}$$

The measure equals one under the optimal policy and zero at the worst benchmark. Normalization permits comparisons across maps and traffic realizations, but it does not make two maps the same decision problem. Appendix B.4 reports the benchmark values used for every round.

The advice-phase analysis examines performance under advice, agreement with advice (a proxy for compliance), processing effort, and experienced state-action support. Performance in E_1 is the empirical counterpart of the contemporaneous reward. To define agreement, let x_{id} denote the action taken (the

additional charge selected) by participant i at decision d, $\tilde{a}^*_{id}$ the precise advice's numerical target, and $\{\ell_{id}, \ldots, u_{id}\}$ the set of feasible charge values implied by the broad advice, the relevant route distances, and the displayed best-case and worst-case traffic values. We then define action agreement as $G^a_{id} := \mathbf{1}\{x_{id} \in X^a_{id}\}$, where a indicates the advice type.

In particular, we assume a participant agrees with precise advice if they perform the stated action, that is, $X^p_{id} = \{\tilde{a}^*_{id}\}$. A participant agrees with broad advice if they perform any of the actions consistent with the advice, so $X^b_{id} = \{\ell_{id}, \ldots, u_{id}\}$. Specific-broad advice specifies the operating condition to consider "worst case traffic." Hence, it effectively proposes only one action. However, different from precise advice, participants need to map the advice to the actions, so we allow for a broader set, the worse-traffic half of the implied interval: $X^{sb}_{id} = \{h_{id}, \ldots, u_{id}\}$, where $h_{id} = \lceil (\ell_{id}+u_{id})/2 \rceil$. Our sensitivity analyses allow for 5 and 10 percentage point deviations, as well as for a singular set in the case of specific-broad. Note that, when no charging is recommended, $X^a_{id} = \{0\}$ for all advice types, and we do not consider deviations here.

For participant i, the action-agreement rate is the mean of the applicable decision-level indicators over E_1 decisions. We analyze this rate as an outcome of the randomized advice-format assignment. It records whether observed actions are consistent with the displayed recommendation, but it only proxies the model's latent engagement parameter. Appendix B.5 gives the interval construction, worked examples, and measurement sensitivities.

Decision time and interface activity provide proxies for information-processing effort. Our active-effort index gives equal weight to log decision time, slider search, and acquisition of future-traffic information. Finally, experienced state-action support measures exposure to different experiences. To obtain relevant measures, we reduce the state and action space through binning. For bin width w, define $b_w(z) := \lfloor z/w \rfloor$. Then, we measure $S_i(w) := |\{(s_{id}, b_w(c_{id}), b_w(x_{id})) : d \in \mathrm{E}_1\}|$, where s_{id} is the current stop (if a map is repeated, the same stop leads to the same indicator here), c_{id} the current charge, and x_{id} the action taken at decision d within the advice-phase E_1. The main specification uses $w = 10$, so the charge bins are anchored at zero, such as $[0, 10)$, $[10, 20)$, and so forth. Sensitivity analyses for $w = 5$ and $w = 20$ are omitted for brevity, but leave results qualitatively unchanged. Note that these are unrelated to the recommendation interval of the action-agreement definition. Appendix E.2 reports the component measures.

We characterize the post-advice phase using observable task features rather than assigning an assumed value of the model's novelty parameter. The dimensions are route topology and distances, displayed traffic support, realized traffic, overlap in feasible states and actions, overlap in the optimal state-action mapping, and whether the qualitative rule remains directly applicable. Study 1 supplies two matched comparisons: a return post-advice to a map previously seen without and with advice, and a return to a map previously seen only without advice. In Study 2's "familiar sequence" of post-advice maps, Round 5 repeats the advice-phase

map, distances, and displayed traffic ranges; Round 6 retains the topology and distances but changes the displayed ranges; and Round 7 retains Round 6's ex ante map and ranges but has four of five realized traffic values outside the advice-phase displayed support. We therefore estimate Rounds 5, 6, and 7 separately and report the mean of Rounds 5–6 as an additional summary. The "unseen sequence" of post-advice maps changes the map, state space, and charging opportunities completely and has no matched pre-advice round.

When a post-advice round has a corresponding pre-advice environment, we report both matched improvement and a baseline-adjusted post-advice level. For matched environment m, let $\Delta_{im} := Y_{im}^{\text{post}} - Y_{im}^{\text{pre}}$. The baseline-adjusted specification, commonly termed analysis of covariance (ANCOVA), is $Y_{im}^{\text{post}} := \alpha + \boldsymbol{\tau}^T \mathbf{T}_i + \gamma Y_{im}^{\text{pre}} + \boldsymbol{\eta}^T \mathbf{Z}_i + \varepsilon_{im}$, where $\mathbf{T}_i$ contains the randomized treatment indicators and $\mathbf{Z}_i$ contains the applicable randomized design factors. ANCOVA compares post-advice levels conditional on baseline performance and can improve precision relative to relying only on change scores when baseline and follow-up outcomes are correlated (Vickers and Altman 2001, Lin 2013). For the unseen sequence, we use pre-advice performance as a precision covariate rather than as a matched outcome. The advice-format-by-sequence interaction tests whether the relative effect of advice format differs between the familiar and unseen sequences.

We also report the change from the final E_1 round to the first E_2 round. This immediate-removal contrast measures the loss or persistence of assisted performance and is distinct from pre-to-post improvement. A smaller decline after qualitative advice is consistent with lower reliance on an executable recommendation, but it does not separately identify learning or skill loss.

Round- and decision-level regressions use participant-clustered standard errors. Participant-level mean and change contrasts use Welch confidence intervals, and ANCOVA specifications use HC1 heteroskedasticity-robust standard errors. Study 1 models include the randomized traffic condition where applicable. Study 2 factorial models use all randomized advice-format, rationale, and sequence cells and report every pairwise advice-format comparison. We also report advice-format simple effects when rationales are hidden, which isolates the format manipulation from the separate rationale treatment. Treatment means and 95% confidence intervals accompany the regression estimates. The main text reports raw p-values; Appendix C reports Holm and Benjamini–Hochberg adjustments within the corresponding analysis families.

4. Study 1: Precise versus broad advice

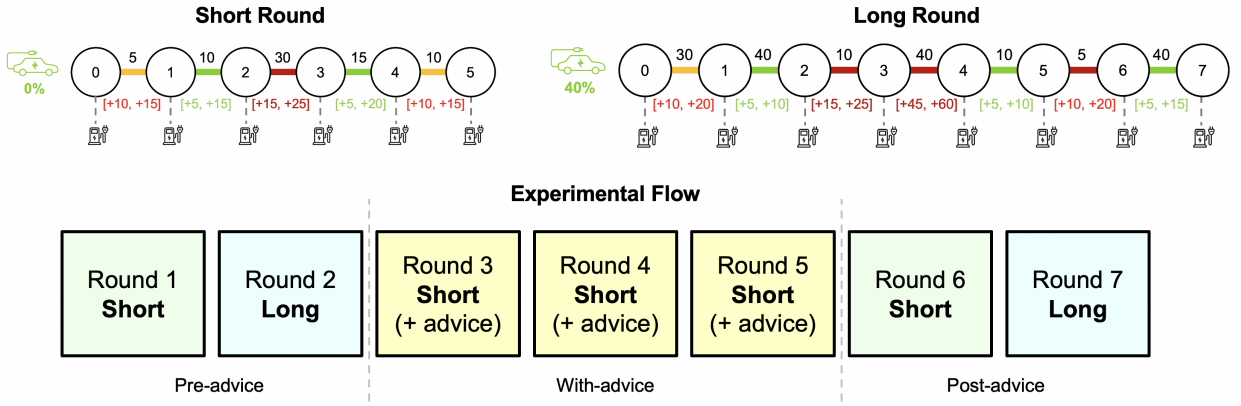
Study 1 provides the first test of the model. Participants receive precise or broad advice on a *short* map and later return, in E_2, to the *short* and *long* maps they encountered in E_0. The advice phase tests the predicted differences in action agreement, performance, effort, and experienced state-action support. The two rounds in E_2 permit map-specific pre-to-post comparisons.

4.1. Design

Study 1 uses a 2×2 between-subject design. The focal factor is advice format. Precise advice states whether to exit and, when charging is recommended, gives the target charge. Broad advice states whether to split or batch but leaves the target charge to the participant. The second factor changes the traffic realization pattern in Rounds 4–6. Realized delays tend to fall near the endpoints of the displayed ranges in the simple condition and in the interior of those ranges in the complex condition, making endpoint-based heuristics less reliable. The traffic factor tests the robustness of the advice-format results; it does not change the advice policy or its breadth.

Figure 4 shows the seven-round sequence. Participants first complete one short-map and one long-map round pre-advice. They then complete three short-map rounds with advice. Advice is removed for the final two rounds: Round 6 returns to the short map, with new traffic realizations. Round 7 returns to the long map from Round 2. Thus, each post-advice round has a participant-specific baseline on the same map, although participants never receive advice on the long map.

Figure 4 Study 1 maps and round sequence.



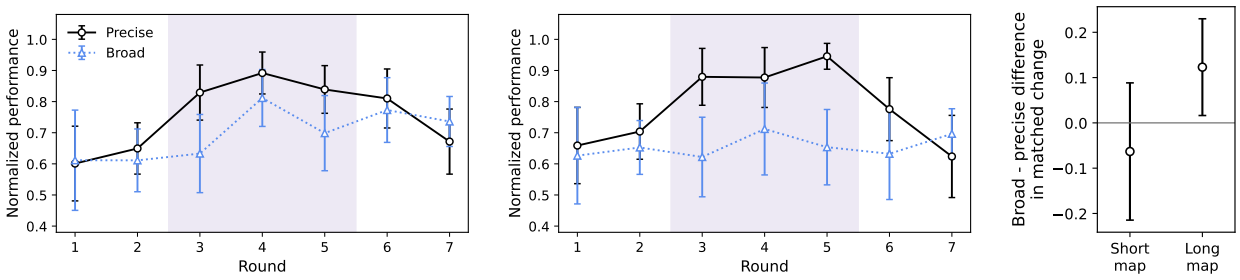
Note. Participants complete two E_0 rounds pre-advice, three advised rounds (E_1) on the short map, and two E_2 post-advice rounds. Round 6 returns to the short map and Round 7 returns to the long map first encountered in Round 2.

We preregistered Study 1 and recruited participants on Amazon Mechanical Turk. The final analytic sample contains 90 participants, with 44 assigned to broad advice and 46 to precise advice. The four advice-by-traffic cells contain 18 to 26 participants.

4.2. Results

Figure 5 plots normalized performance by round and traffic condition. Table 1 reports the focal randomized contrasts. Unless noted otherwise, estimates are broad minus precise. Appendix B.5 details the construction and sensitivity of the action-agreement measure.

Figure 5 Study 1 performance during and after advice.



(a) Simple traffic (b) Complex traffic (c) Matched changes

Note. Panels (a) and (b) plot normalized performance by round; the shaded region marks Rounds 3–5 (E_1). Points are treatment means and whiskers are 95% confidence intervals. Panel (c) plots the broad-minus-precise difference in matched pre-to-post changes for the short map (Round 6 minus Round 1) and long map (Round 7 minus Round 2). Table 1 also reports the corresponding baseline-adjusted post-advice levels.

Performance with advice. Treatment groups entered the advice phase with similar performance (Round 1: $p = 0.852$; Round 2: $p = 0.292$). The mean participant action-agreement rate was 0.643 under precise advice and 0.373 under broad advice. The broad-minus-precise difference was −0.271 (95% CI $[-0.407, -0.135]$, $p < 0.001$), supporting H_{1a}.

Across Rounds 3–5, mean normalized performance was 0.878 with precise advice and 0.693 with broad advice. The difference was −0.185 (95% CI $[-0.273, -0.097]$, $p < 0.001$). The final advice-round difference was −0.214 (95% CI $[-0.313, -0.115]$, $p < 0.001$). The randomized contrast establishes an assignment-level reward gap in the direction of H_{1c}.

Decision behavior with advice. Actions under broad advice were less concentrated around the precise policy target. The mean absolute distance was 7.19 charge points larger under broad advice (95% CI $[3.36, 11.02]$, $p < 0.001$). Participants assigned to broad advice also encountered 1.18 more distinct state-action tuples under the 10-point binning rule (95% CI $[0.13, 2.23]$, $p = 0.028$). The estimate remains positive with 5- and 20-point bins. This assignment-level evidence is consistent with the wider action dispersion in H_{1b}, although experienced support measures exposure rather than whether the additional observations improved participants' strategies.

Study 1 does not provide corresponding evidence for the effort difference in H_{1d}. The broad-minus-precise difference was 0.34 seconds for winsorized decision time (95% CI $[-4.57, 5.26]$, $p = 0.891$) and 0.058 standard deviations for the active-effort index (95% CI $[-0.115, 0.231]$, $p = 0.511$). Slider activity and future-traffic inspection were also statistically indistinguishable, and eventual treatment assignment did not predict these measures before advice appeared.

Performance after advice removal. Round 6 returns to the short map used during both E_0 and E_1. Mean performance was 0.792 after precise advice and 0.715 after broad advice, an unadjusted difference of −0.077

($p = 0.180$). Relative to Round 1, performance increased by 0.161 after precise advice and by 0.098 after broad advice. The difference in matched changes was −0.063 (95% CI $[-0.215, 0.088]$, $p = 0.409$). The baseline-adjusted level difference was similarly imprecise at −0.084 (95% CI $[-0.189, 0.022]$, $p = 0.119$). Pooling advice formats, short-map performance increased by 0.130 from Round 1 to Round 6 (95% CI $[0.055, 0.206]$, $p = 0.001$). Participants improved on the repeated short-map structure, consistent with H_2, but the improvement did not favor broad advice. Given prior exposure to the same decision task with advice, Round 6 likely represents a task with insufficient novelty for broad advice's learning benefits to kick in (see H_{3a}).

Round 7 returns to the longer map encountered before advice but never used during the advice phase. Mean performance was 0.720 after broad advice and 0.647 after precise advice, an unadjusted difference of 0.073 ($p = 0.163$). Relative to Round 2, performance changed by 0.092 after broad advice and by −0.031 after precise advice. The difference in matched changes was 0.123 (95% CI $[0.016, 0.230]$, raw $p = 0.024$). The baseline-adjusted level difference was smaller: 0.090 (95% CI $[-0.0003, 0.180]$, $p = 0.051$). The matched-change result does not survive correction across the transfer family (Holm-adjusted $p = 0.171$; Benjamini–Hochberg-adjusted $p = 0.054$). The long-map comparison therefore provides suggestive, specification-sensitive evidence in the direction of H_{3a}: Having not encountered the long map with advice, Round 7 does not allow for imitation of the actions suggested by precise advice during E_1. Pooling advice formats, long-map performance did not increase reliably from Round 2 to Round 7 (0.029, $p = 0.297$).

The first post-advice round also measures the immediate consequence of removing advice within the same setting. From Round 5 to 6, performance increased by 0.035 after broad advice and fell by 0.102 after precise advice. The difference was 0.138 (95% CI $[0.043, 0.233]$, $p = 0.005$). This contrast is consistent with greater reliance on the executable recommendation under precise advice. It is not a pre-to-post measure, so does not show that precise advice caused skill loss.

None of the advice-format contrasts varied detectably with the traffic pattern (Appendix D.3). Study 1 establishes the contemporaneous advantage of precise advice and shows that broad advice produces wider experienced action support without a detectable increase in measured effort. After advice removal, the repeated short-map comparison does not favor broad advice, while the long-map comparison is positive but sensitive to specification and multiplicity. The immediate-removal contrast is consistent with greater reliance under precise advice. Study 2 provides the main test of action mapping and post-advice performance across randomized subsequent environments.

5. Study 2: Specific-broad advice and transfer

Study 2 randomizes the post-advice sequence. This permits testing whether the relative effect of advice format changes when participants remain on a familiar decision structure or move to a map they have not

Table 1 Study 1 focal treatment contrasts.

Outcome	Broad − precise	95% CI	p-value
Advice phase (E_1)			
Normalized performance	−0.185	[−0.273, −0.097]	< 0.001
Format-consistent action agreement	−0.271	[−0.407, −0.135]	< 0.001
Experienced state-action support	1.184	[0.134, 2.234]	0.028
Active-effort index	0.058	[−0.115, 0.231]	0.511
Post-advice phase (E_2)			
Short-map matched change, Round 6 − Round 1	−0.063	[−0.215, 0.088]	0.409
Short-map ANCOVA level, Round 6	−0.084	[−0.189, 0.022]	0.119
Long-map matched change, Round 7 − Round 2	0.123	[0.016, 0.230]	0.024
Long-map ANCOVA level, Round 7	0.090	[−0.0003, 0.180]	0.051
First post-removal change, Round 6 − Round 5	0.138	[0.043, 0.233]	0.005

Note. Estimates are broad minus precise. Positive values therefore favor broad advice. Advice-phase performance specifications use participant-clustered standard errors and the controls described in §3.4. Format-consistent agreement, experienced state-action support, and matched changes are participant-level comparisons with Welch confidence intervals. Agreement requires the exact target under precise advice and membership in the full traffic-implied interval under broad advice. The state-action measure counts distinct tuples of stop, 10-point current-charge bin, and 10-point charging-action bin during Rounds 3–5. ANCOVA estimates are baseline-adjusted post-advice levels with HC1 standard errors. Raw p-values are shown; Appendix C reports Holm and Benjamini–Hochberg adjustments.

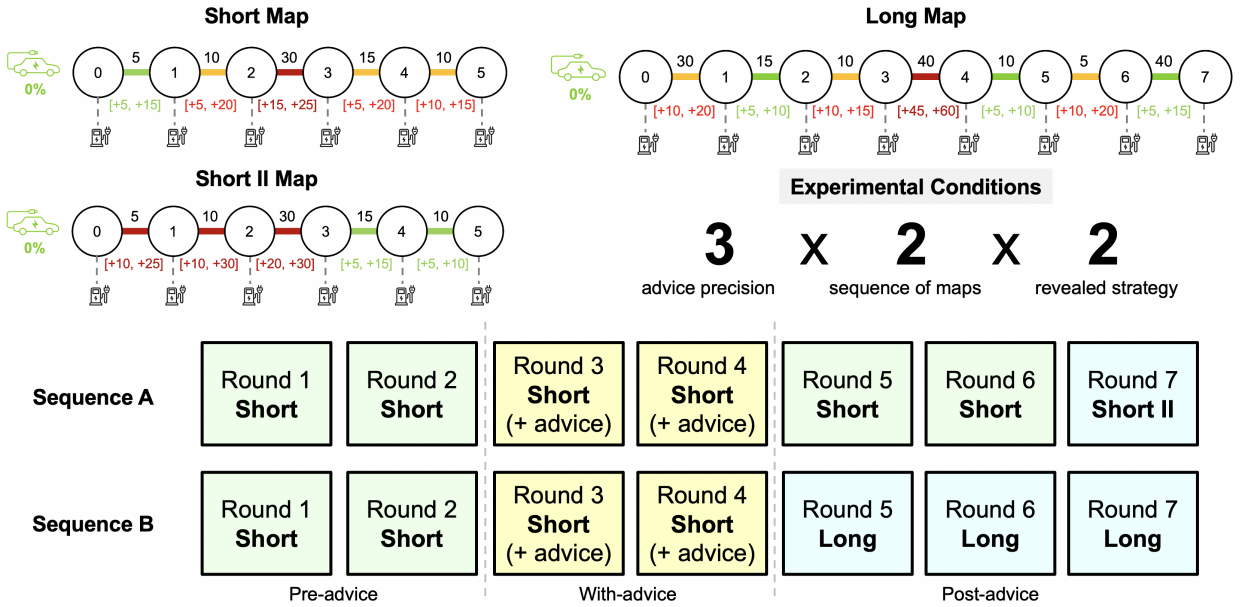
previously encountered. We also provide additional design options for the advice. First, we add a rationale manipulation that adds explanations to the advice. Second, we introduce specific-broad advice, which is more specific than broad advice but remains qualitative. This allows us to understand better how our model can guide designers in choosing the optimal advice. In particular, if participants observing specific-broad advice display behaviors that are distinctly in-between the behaviors of participants observing precise and those observing broad, this supports the continuity of advice breadth from $B = 1$ to $B > 1$ assumed in the model. If, however, behaviors under specific-broad and broad are highly similar, this indicates that there is a "fixed cost" of converting a qualitative description into a concrete action that is not currently reflected in the model.

5.1. Design

Study 2 uses a $3 \times 2 \times 2$ between-subject design. The factors are advice format (precise, broad, or specific-broad), rationale visibility (hidden or shown), and post-advice sequence (familiar or unseen). Participants complete seven main rounds: two short-map rounds without advice, two short-map rounds with advice, and three rounds without advice. Figure 6 shows the design.

Precise advice states whether to exit and gives the numerical target charge. Broad advice states whether to split or batch but leaves the target charge to the participant. Specific-broad advice adds an operating condition, specifically, to plan for the maximum traffic in the displayed range, while leaving the final calculation to the participant. The precise-versus-specific-broad comparison therefore holds strategic detail closer to constant while changing whether the interface supplies an executable action, or the participant needs to evaluate

Figure 6 Study 2 maps and round sequence.



Note. Participants complete two baseline rounds without advice and two advised rounds on the short map. Advice is then removed. The familiar sequence remains on the short-map structure, whereas the unseen sequence switches to a longer map. Advice format, rationale visibility, and post-advice sequence are assigned between participants.

different options. The broad-versus-specific-broad comparison asks whether a reduction in breadth within broad advice has a relevant effect, in particular compared with the contrast of broad and precise.

The rationale treatment is crossed with advice format. When shown, the rationale explains the logic behind the recommendation, such as the fixed time required to leave the highway or the nonlinear charging-time function. It does not change the recommended split-or-batch strategy or the numerical target, when one is provided.

The familiar sequence uses the short-map structure with new traffic realizations throughout E_2, with modified traffic ranges in the final round. Its pre-to-post comparison measures improvement within a familiar decision structure. The unseen sequence switches to the long map throughout E_2. Participants have no pre-advice observation on that map, so we analyze post-advice levels and control for short-map baseline performance. Because sequence assignment is randomized, the advice-format-by-sequence interaction directly tests whether the relative effect of advice format differs across the two post-advice environments.

We preregistered Study 2 and recruited participants on Prolific. The final analytic sample contains 400 participants, with 201 assigned to the familiar sequence and 199 to the unseen sequence. The 12 treatment cells contain 30 to 38 participants.

5.2. Results

Figure 7 plots normalized performance by round, advice format, rationale visibility, and post-advice sequence. Table 2 reports the focal advice-format contrasts. Advice-phase estimates use all randomized cells. Because rationale provision changes the familiar-sequence effect of advice format, the post-advice analysis reports both marginal format effects averaged across rationale assignment and when rationales are hidden. Both comparisons preserve random assignment.

Pre-advice performance is balanced across advice formats. Averaging Rounds 1–2, normalized performance is 0.816 under precise advice, 0.835 under broad advice, and 0.829 under specific-broad advice (all pairwise p-values exceed 0.29).

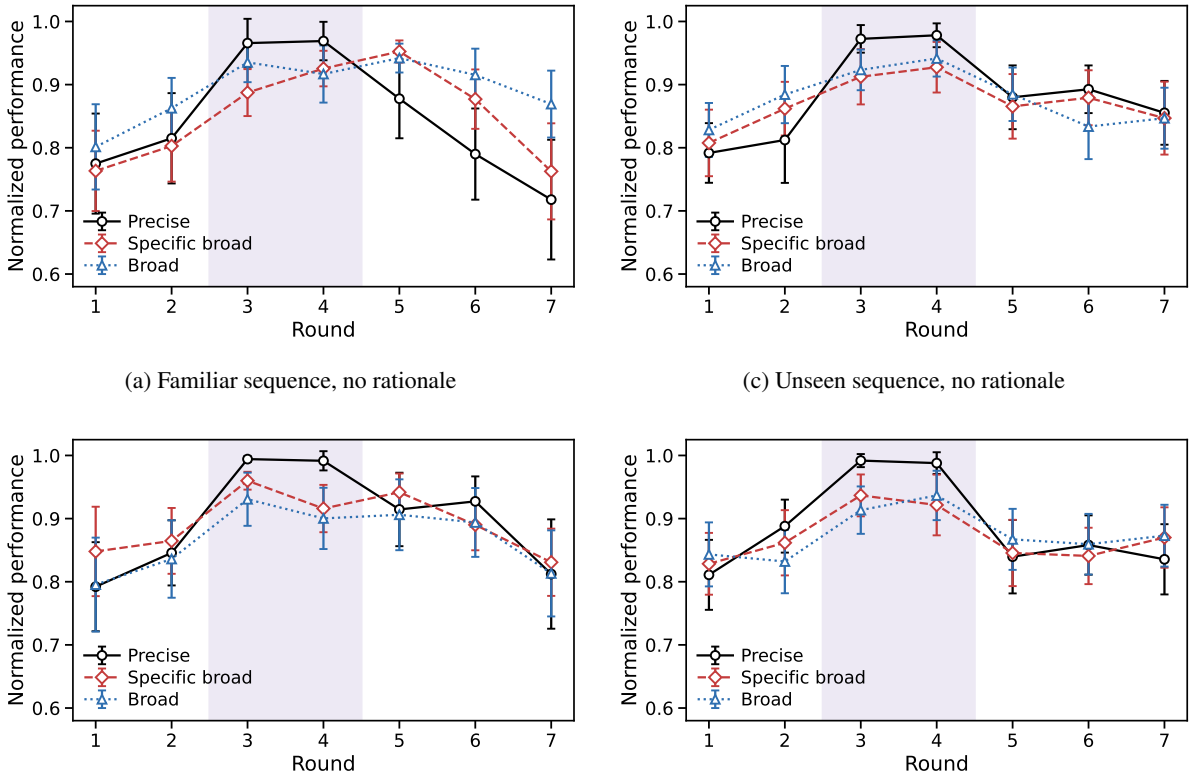
Performance with advice. Across Rounds 3–4, normalized performance is 0.981 under precise advice, 0.924 under broad advice, and 0.923 under specific-broad advice. Relative to precise advice, performance is 0.057 lower under broad advice (95% CI $[-0.076, -0.038]$, $p < 0.001$) and 0.058 lower under specific-broad advice (95% CI $[-0.076, -0.041]$, $p < 0.001$). Broad and specific-broad differ by -0.001 (95% CI $[-0.023, 0.020]$, $p = 0.902$). The comparisons with precise advice match the reward ordering in H_{1c}. However, the reduction in the number of actions implied by specific-broad, which still leaves participants to convert a textual recommendation into a numerical action, does not seem to be sufficient to translate into lower efforts to follow the advice, consistent with a missing "fixed cost" of converting the recommendation.

This is also evidenced by action agreement. Relative to precise, agreement is 19.3 percentage points lower under broad (95% CI $[-24.7, -13.9]$, $p < 0.001$) and 23.1 points lower under specific-broad advice (95% CI $[-28.4, -17.9]$, $p < 0.001$), in line with the advice-format pattern in H_{1a}. The two qualitative formats differ only by 3.8 points (95% CI $[-8.8, 1.3]$, $p = 0.142$), however, further indicating that specific-broad fails to reduce the translation effort from advice to action.

Decision behavior with advice. The interaction measures indicate that participants exert more work under the two qualitative formats. Broad advice increases winsorized decision time by 11.2 seconds relative to precise advice (95% CI $[8.4, 14.0]$, $p < 0.001$), and specific-broad increases it by 10.9 seconds (95% CI $[8.3, 13.5]$, $p < 0.001$), in line with H_{1d}. The two qualitative formats do not differ significantly (-0.3 seconds, $p = 0.834$). The active-effort index is 0.317 standard deviations higher under broad advice (95% CI $[0.235, 0.400]$, $p < 0.001$) and 0.315 standard deviations higher under specific-broad advice (95% CI $[0.239, 0.390]$, $p < 0.001$); the latter two's difference is effectively zero. Slider activity and future-traffic inspection produce the same pattern, and eventual treatment assignment does not predict these measures before advice appears.

Advice format also changes the set of state-action combinations participants experience. Participants assigned to broad advice encounter 1.18 more distinct stop-state-action tuples than those assigned to precise advice (95% CI $[0.85, 1.52]$, $p < 0.001$), while the specific-broad difference is 1.17 tuples (95% CI

Figure 7 Study 2 normalized performance across rounds.



(a) Familiar sequence, no rationale

(c) Unseen sequence, no rationale

(b) Familiar sequence, rationale shown

(d) Unseen sequence, rationale shown

Note. The shaded region marks Rounds 3–4, when advice was available. Points are treatment means and whiskers are 95% confidence intervals. Higher values indicate better performance.

$[0.82, 1.53]$, $p < 0.001$), in line with H_{1b} and the sequential model. Again, specific-broad does not differ from broad (-0.01, $p = 0.941$).

Performance after advice removal. The familiar sequence supports a pre-to-post comparison because participants remain on the short-map structure. Pooling advice and rationale conditions, normalized performance increases by 0.052 from the pre-advice to post-advice phases (95% CI $[0.033, 0.071]$, $p < 0.001$). The increase is positive within each advice format, consistent with H_2.

The three familiar rounds change different task features. Round 5 is the closest Study 2 analogue to Study 1's Round 6: both remove advice while repeating the previous short map and displayed traffic ranges. We therefore do not rank one as more novel than the other. Round 6 in Study 2 retains the route and distances but changes the displayed traffic ranges, without changing the optimal batch/split strategy per exit, thus adding a change that is absent from Study 1's Round 6. Round 7 retains the ex ante map, but changes traffic ranges such that the optimal batch/split strategy changes, increasing novelty. We report the three rounds separately.

Averaging across rationale assignment, the Round 5 broad-minus-precise estimate is 0.019 (95% CI $[-0.022, 0.059]$, $p = 0.361$), and the specific-broad-minus-precise estimate is 0.046 (95% CI $[0.007, 0.085]$, $p = 0.022$). In Round 6, the corresponding estimates are 0.035 (95% CI $[-0.010, 0.079]$, $p = 0.126$) and 0.020 (95% CI $[-0.026, 0.066]$, $p = 0.396$). In Round 7, the estimates are 0.065 (95% CI $[-0.007, 0.137]$, $p = 0.079$) and 0.025 (95% CI $[-0.051, 0.101]$, $p = 0.517$). However, the marginal effects are attenuated because rationales improve familiar post-advice performance primarily under precise advice, as we discuss below.

When rationales are hidden, the Round 5 broad-minus-precise estimate is 0.049 (95% CI $[-0.0005, 0.098]$, $p = 0.052$), and the specific-broad-minus-precise estimate is 0.080 (95% CI $[0.029, 0.130]$, $p = 0.002$), which survives Holm adjustment. In Round 6, the estimates are 0.104 (95% CI $[0.039, 0.170]$, $p = 0.002$) and 0.094 (95% CI $[0.025, 0.163]$, $p = 0.008$), both of which survive adjustment. Round 7 yields different estimates for the two qualitative formats: Broad advice exceeds precise advice by 0.137 (95% CI $[0.035, 0.238]$, raw $p = 0.008$), and the estimate again survives adjustment. The specific-broad-minus-precise estimate is 0.050 (95% CI $[-0.065, 0.164]$, $p = 0.395$). Broad advice is also 0.087 above specific-broad advice, although that contrast is imprecise (95% CI $[-0.005, 0.179]$, $p = 0.064$).

Looking at broad advice specifically, we find exactly the prediction from H_{3a}: in a near-identical environment, the benefit over precise is not significant, as in Study 1's Round 6. However, adding some novelty to the environment, as in Rounds 6–7 of Study 2, broad advice provides a clear advantage. The picture is more complex when contrasting specific-broad advice with precise advice, even though the direction of the effect sizes is in line with novelty, first increasing and then decreasing, as the model implies.

Round 5 also yields the immediate-removal contrast. From Round 4 to Round 5, performance changes by 0.117 more after broad than after precise advice (95% CI $[0.043, 0.191]$, $p = 0.002$) and by 0.118 more after specific-broad than after precise advice (95% CI $[0.054, 0.183]$, $p < 0.001$). Both contrasts survive Holm adjustment. The numerical advantage of precise advice falls sharply when the recommendation disappears, consistent with greater reliance on executable guidance.

The unseen sequence changes the route, state sequence, and charging opportunities and has no matched long-map baseline. Averaging across rationale assignment, post-advice performance is close across formats: broad minus precise is -0.012 (95% CI $[-0.047, 0.022]$, $p = 0.488$), and specific broad minus precise is -0.011 (95% CI $[-0.047, 0.025]$, $p = 0.559$). With rationales hidden, broad advice is 0.049 below precise advice (95% CI $[-0.094, -0.004]$, raw $p = 0.032$), but the contrast does not survive Holm adjustment. The specific-broad difference is -0.029 (95% CI $[-0.079, 0.021]$, $p = 0.254$). Neither qualitative format has a post-advice advantage. The relative advantage of the qualitative formats observed in the familiar sequence,

therefore, does not carry over to the unseen sequence, consistent with the boundary in H_{3b} (that a very large degree of novelty means neither learning nor imitation are useful).

The randomized sequence interactions in the no-rationale cells reinforce this boundary: On average across Rounds 5–6, the broad-minus-precise effect is 0.131 higher in the familiar sequence than in the unseen sequence (95% CI $[0.065, 0.196]$, Holm-adjusted $p < 0.001$); the corresponding specific-broad interaction is 0.117 (95% CI $[0.049, 0.185]$, Holm-adjusted $p = 0.004$). For Round 7, the broad interaction is 0.168 (95% CI $[0.049, 0.286]$, Holm-adjusted $p = 0.023$), whereas the specific-broad interaction is 0.073 (95% CI $[-0.059, 0.206]$, $p = 0.277$).

Table 2 Study 2 focal qualitative-versus-precise contrasts.

Outcome	Broad − precise	Specific broad − precise
Advice phase, all randomized cells		
Normalized performance	−0.057 [−0.076, −0.038]	−0.058 [−0.076, −0.041]
Format-consistent agreement	−0.193 [−0.247, −0.139]	−0.231 [−0.284, −0.179]
Decision time (seconds)	11.224 [8.421, 14.026]	10.913 [8.279, 13.547]
Active-effort index	0.317 [0.235, 0.400]	0.315 [0.239, 0.390]
State-action support	1.184 [0.851, 1.517]	1.171 [0.817, 1.525]
Post-advice ANCOVA, all rationale assignments		
Familiar Round 5	0.019 [−0.022, 0.059]	0.046 [0.007, 0.085]
Familiar Round 6	0.035 [−0.010, 0.079]	0.020 [−0.026, 0.066]
Familiar Rounds 5–6 mean	0.027 [−0.006, 0.060]	0.033 [−0.001, 0.066]
Familiar Round 7	0.065 [−0.007, 0.137]	0.025 [−0.051, 0.101]
Unseen Rounds 5–7	−0.012 [−0.047, 0.022]	−0.011 [−0.047, 0.025]
Post-advice ANCOVA, rationale hidden		
Familiar Round 5	0.049 [−0.0005, 0.098]	0.080 [0.029, 0.130]
Familiar Round 6	0.104 [0.039, 0.170]	0.094 [0.025, 0.163]
Familiar Rounds 5–6 mean	0.077 [0.031, 0.122]	0.087 [0.038, 0.135]
Familiar Round 7	0.137 [0.035, 0.238]	0.050 [−0.065, 0.164]
Unseen Rounds 5–7	−0.049 [−0.094, −0.004]	−0.029 [−0.079, 0.021]
First post-removal change	0.117 [0.043, 0.191]	0.118 [0.054, 0.183]
Familiar-minus-unseen, Rounds 5–6	0.131 [0.065, 0.196]	0.117 [0.049, 0.185]
Familiar-minus-unseen, Round 7	0.168 [0.049, 0.286]	0.073 [−0.059, 0.206]

Note. Entries are estimates with 95% confidence intervals. Positive values favor the qualitative format. Advice-phase performance, decision-time, and active-effort estimates use participant-clustered standard errors. Format-consistent agreement and state-action support are participant-level comparisons with Welch confidence intervals. Specific-broad agreement uses the worse-traffic half of the implied interval. Post-advice rows are baseline-adjusted levels with HC1 standard errors, except first post-removal change, which uses Welch inference. The all-rationale rows average over randomized rationale assignment. The rationale-hidden rows are simple effects within randomized no-rationale cells. The Rounds 5–6 mean is a precision summary rather than a distinct novelty level. Raw p-values appear in the text; Appendix C reports multiplicity adjustments. Broad-versus-specific-broad contrasts are reported in the text and Appendix E.3.

Rationales. Rationales do not have a common effect across advice formats or environments. In the familiar sequence, rationales raise post-advice performance under precise advice by 0.074 (95% CI $[0.017, 0.131]$, $p = 0.010$). The corresponding effects are −0.029 under broad ($p = 0.157$) and 0.011 under specific-broad advice ($p = 0.618$). The precise-advice effect exceeds the broad one by 0.108 (95% CI $[0.038, 0.179]$, $p = 0.003$) and the specific-broad one by 0.089 (95% CI $[0.014, 0.164]$, $p = 0.020$). The precise-advice effect remains positive after conditioning on advice-phase performance and agreement (0.063, 95% CI $[0.007, 0.118]$, $p = 0.026$). This pattern is closest to the model's effective-breadth channel (Oc). However,

the same rationale does not help precise advice on the unseen map. Its adjusted effect there is −0.056 (95% CI $[-0.104, -0.008]$, $p = 0.022$), with the broad and specific-broad effects statistically indistinguishable from zero. Moreover, none of the six format-by-sequence rationale effects survives Holm adjustment. This indicates that the benefits from explanations are more environment-specific than implied by our model.

Rationales also raise agreement with precise advice by 0.089 (95% CI $[0.011, 0.167]$, $p = 0.026$), in line with the perceived quality channel (Oa), but not reliably under the qualitative formats.

6. Recovering strategies using Inverse Reinforcement Learning

The analyses in §4–5 contrast performance to establish when each advice format performs better while available and after. In a sequential task, however, similar performance can arise from different policies. A participant may complete the route quickly by batching charges in a forward-looking way, by maintaining a large safety buffer, or by following a prescription. We use inverse reinforcement learning (IRL) to examine which decision criteria are most consistent with participants' actions.

The IRL analysis separates two objects that affect performance: compliance with advice and a participant's fallback heuristic. Precise advice should produce high compliance. Broad advice should produce lower contemporaneous compliance because the user must map a rule into an action, but it may be associated with an updated heuristic that places greater weight on the strategic charging logic after advice is removed.

Two features of the data make IRL useful. First, only few states are ever observed for a given participant because the state space combines the current stop and the remaining charge. Second, observed behavior under advice mixes the participant's own heuristic with the displayed recommendation. Our approach allows separating these forces by estimating both participants' fallback heuristics and compliance probabilities.

6.1. Estimating the fallback heuristic

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 at a stop, and transitions depend on the charging decision and traffic realization. We denote by $\mu_i : \mathcal{S} \to \Delta(\mathcal{A})$ the policy participant i uses absent advice, or *fallback heuristic* (we drop i when we discuss a single participant).

IRL takes the system dynamics as known and estimates the decision criteria that make the observed policy plausible (Hanawal et al. 2018). This approach is well suited for our setting because the dynamics are known by design, but participants may not minimize elapsed game time exactly. They may also value avoiding a breakdown, following simple rules, etc. In line with the literature, 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 $\tilde{r}_t = r_{s_t}(\tilde{a}_t)$, so 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. 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. 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.

The reward components $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ are chosen to be behaviorally interpretable. We derive them from responses in a pilot study (not reported here for brevity), in which participants described their strategy and the advice they would give future players. Using topic modeling (see Appendix F.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 a state-action pair. Components may be weighted positive or negative, so it can be more appropriate to think of some as costs rather than rewards.

Appendix F.2 details our theoretical framework. Let sc denote a scenario, defined by phase and advice condition. Estimating a participant's fallback heuristic corresponds to estimating $\boldsymbol{\theta}_{sc,i} = (\theta^1_{sc,i}, \ldots, \theta^m_{sc,i}) = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_{sc} + \boldsymbol{\Delta}_i$ for each participant i and scenario $sc \in \mathcal{SC} := \{pre\} \cup \bigcup_{tip \in \{p,pr,s,sr,b,br\}} \{with\text{-}tip, post\text{-}tip\}$. Here, p denotes precise advice, s denotes specific-broad advice, b denotes broad advice, and r denotes rationale visibility. The vector $\boldsymbol{\theta}_0$ captures average reward weights, $\boldsymbol{\Delta}_{sc}$ captures the scenario-specific shift, and $\boldsymbol{\Delta}_i$ captures participant-specific heterogeneity. We pool the pre-advice phase across eventual treatment assignments because participants have not yet observed differing treatments.

6.2. Estimating compliance with advice

The fallback heuristic describes what participants do absent advice. During the advice phase, observed behavior can also reflect compliance with the recommendation. Consistent with the model in §2 and the literature (e.g., Grand-Clément and Pauphilet 2026), we model advised behavior as a mixture of the optimal policy and the participant's heuristic. Let ν^* denote the deterministic policy that minimizes expected elapsed

game time. The algorithmic advice in all treatments is generated from this same policy; treatments differ in how the policy is communicated. In scenario sc, participant i chooses

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

We parameterize compliance as $\pi_{sc,i} = 1/[1+e^{-(\eta_{sc}+\eta_i)}]$ for scenarios $sc \in \bigcup_{tip\in\{p,pr,s,sr,b,br\}}\{with\text{-}tip\}$ and set compliance to zero in no-advice scenarios. The scenario component η_{sc} captures differences in implementability across advice formats, while η_i captures participant-level heterogeneity. An advice regime can therefore improve performance by increasing compliance, changing the fallback heuristic, or both. Appendix F.3 outlines our approach to identifying parameter distributions, which we validate in Appendix F.4.

6.3. How advice affects strategies

The quantities of interest are the scenario-level reward weights $\boldsymbol{\theta}_{sc} = \boldsymbol{\theta}_0 + \boldsymbol{\Delta}_{sc}$ and the compliance probabilities $\pi_{sc,i}$. Appendix F.5 reports the posterior distributions for $\pi_{sc,i}$. The estimates mirror the performance results: specific-broad advice increases compliance modestly compared to broad advice, while precise advice generates substantially higher compliance, consistent with H_{1a}. In contrast to the performance results, rationales have a large effect on compliance, especially for broad advice, which suggests that explanations can make a recommendation more credible, in line with the *perceived quality channel* (Oa).

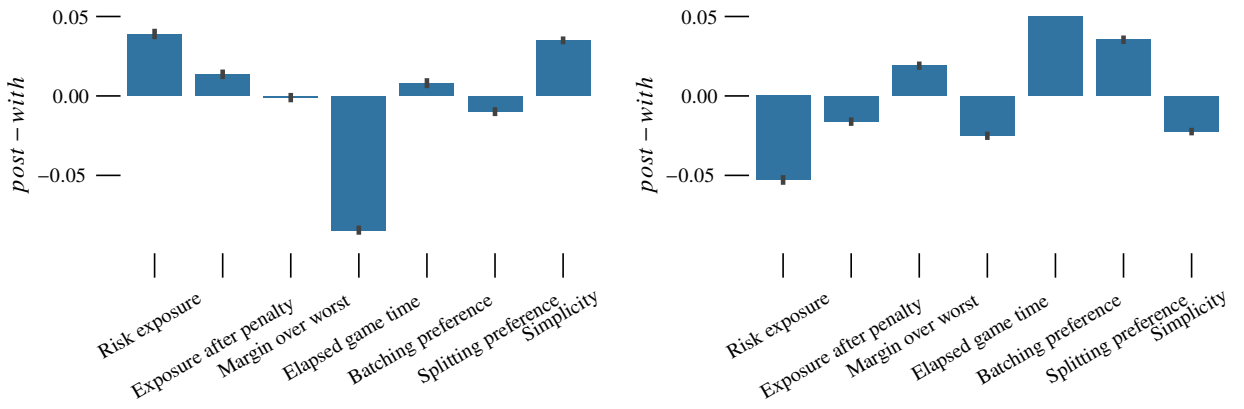
To compare $\boldsymbol{\theta}_{sc}$ across scenarios, we compute each component's relative feature importance,

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

We estimate the model using the pilot and Studies 1–2, yielding 3,666 trajectories across 13 scenarios and 549 participants. Each trajectory contains an average of 5.6 state-action pairs. We sample 300 times from each scenario-specific posterior to obtain the reported distributions.

Figure 8 compares how inferred feature importance changes from the advice phase (E_1) to the post-advice phase (E_2). After precise advice is removed, participants' inferred objectives shift toward simpler and more risk-focused criteria. The relative importance of risk exposure and simplicity increases, consistent with a reversion toward safety-oriented heuristics once the numerical prescription is absent. After broad advice is removed, the relative weights remain more closely tied to strategic charging components, especially batching and splitting, and less tied to avoiding any risk of running out of charge.

The IRL estimates complement the performance evidence by opening the black box of sequential decision making: First, they separate compliance from fallback changes: Precise advice appears to work primarily through high compliance. Broad advice produces lower compliance, but is associated with an updated heuristic that places more weight on the structure of the task. Second, the estimates show why outcome



(a) Precise advice (b) Broad advice

Figure 8 Change in relative feature importance FI^j_{sc} from the advice phase to the post-advice phase.

comparisons alone can understate mechanism differences: Similar post-advice completion times can reflect different heuristics, such that broad advice is associated with more desirable feature weights after its removal, even when it does not outperform precise advice.

7. Conclusion

Algorithmic advice can improve decision making but has the potential to limit engagement with the task. We provide a model that formalizes the resulting reward-learning trade-off as a function of the advice design. Precise advice converts a policy into an executable action and raises expected compliance and performance when guidance is available. Broad advice, on the other hand, leaves the final state-to-action mapping to the user. That burden can reduce immediate performance, but the effort it requires can generate learning that can be important, for instance, to retain supervisory capacity. Taking into account both learning and imitation, our model predicts that the benefit of broad advice is nonmonotone in environmental change: it can be limited when tasks remain unchanged, greatest when the underlying rule remains useful but the prior action does not, and lost when the environments differ too much.

Two online experiments reveal a consistent short-run trade-off and a more conditional post-advice pattern. Precise advice produces the strongest performance while recommendations are visible. Specific-broad advice helps locate the source of this advantage. Even when the qualitative rule specifies the relevant operating condition, performance remains close to broad advice and below precise advice if the participant must still determine the numerical action. The qualitative formats also involve longer decisions and more diverse experiences. Once advice is removed, their performance relative to precise advice depends on the subsequent decision-making environment. Broad advice, in particular, seems to fare better relative to precise advice when the subsequent environment is somewhat, but not completely familiar, but this advantage does not extend to the unseen sequence, where the environment is completely novel. The process and IRL analyses

are consistent with a distinction between following an executable recommendation and developing a decision rule that can guide later behavior, even when the latter does not fully translate into performance gains.

The experiments compare advice formats with one another, but not with continued unaided practice. They therefore identify how the formats perform relative to each other, but not whether either format improves unsupported capability in absolute terms. The process measures help interpret the observed differences but cannot separate learning from advice from learning through practice.

Within these limits, the results identify a practical design problem: Precise advice is attractive when immediate execution is the primary objective, the system is expected to remain available, and future decisions resemble those for which it can continue to prescribe an action. Leaving more of the mapping to the user may be worth the immediate performance cost when people must later act without executable guidance in related situations. That case weakens when the later environment changes enough that the earlier rules no longer apply. The relevant choice is therefore not between precise and broad advice in general, but between different allocations of decision work under different expectations about system availability and future task similarity. Our analytical framework also provides guidance on how to formalize and fine-tune advice design.

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.

Bader R, Woerndl W, Karitnig A, Leitner G (2011) Designing an explanation interface for proactive recommendations in automotive scenarios. *International Conference on User Modeling, Adaptation, and Personalization*, 92–104 (Springer).

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.

Bauer K, von Zahn M, Hinz O (2023) Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research* 34(4):1582–1602.

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.

Boyacı T, Canyakmaz C, De Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.

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.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.

Chromik M, Eiband M, Buchner F, Krüger A, Butz A (2021) I think I get your point, AI! the illusion of explanatory depth in explainable AI. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 307–317.

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.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

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 (2026) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* 72(1):667–692.

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.

Hoong R, Dreyfuss B (2025) Calibrated coarsening: Designing information for ai-assisted decisions. *SSRN* URL https://dx.doi.org/10.2139/ssrn.5286198.

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

Kagan E, Leider S, Sahin O (2026) Sequential decision making: From decision elicitation to strategy identification. *Management Science* 72(6):4889–4908.

KC DS, Staats BR (2012) Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management* 14(4):618–633.

Kluge A, Frank B (2014) Counteracting skill decay: four refresher interventions and their effect on skill and knowledge retention in a simulated process control task. *Ergonomics* 57(2):175–190.

Lapré MA, Tsikriktsis N (2006) Organizational learning curves for customer dissatisfaction: Heterogeneity across airlines. *Management Science* 52(3):352–366.

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

Lehmann M, Cornelius PB, Sting FJ (2024) AI meets the classroom: When do large language models harm learning? *arXiv* URL https://arxiv.org/abs/2409.09047.

Lin W (2013) Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7(1):295–318.

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.

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–298.

Nadler J, Thompson L, Boven LV (2003) Learning negotiation skills: Four models of knowledge creation and transfer. *Management Science* 49(4):529–540.

Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. *Management Science* 57(6):1111–1137.

Passi S, Vorvoreanu M (2022) Overreliance on AI: Literature review. *Microsoft Research Technical Report* URL https://microsoft.com/research/publication/overreliance-on-ai/.

Poulidis S, Bastani H, Bastani O (2025a) Self-regulated AI use hinders long-term learning. *SSRN* URL https://dx.doi.org/10.2139/ssrn.5604932.

Poulidis S, Ge H, Bastani H, Bastani O (2025b) Action vs. attention signals for human-AI collaboration: Evidence from chess. *The Wharton School Research Paper* URL https://dx.doi.org/10.2139/ssrn.5128584.

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.

Schilling MA, Vidal P, Ployhart RE, Marangoni A (2003) Learning by doing something else: Variation, relatedness, and the learning curve. *Management Science* 49(1):39–56.

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

Siderius J, Shumsky RA, Taylor A (2026) Use it or slowly lose it: Expertise atrophy with organizational AI usage. *SSRN* URL https://dx.doi.org/10.2139/ssrn.6398398.

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.

Todd P, Benbasat I (1991) An experimental investigation of the impact of computer based decision aids on decision making strategies. *Information Systems Research* 2(2):87–115.

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

Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R (2023) Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1):1–38.

Vereschak O, Bailly G, Caramiaux B (2021) How to evaluate trust in ai-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–39.

Vickers AJ, Altman DG (2001) Analysing controlled trials with baseline and follow up measurements. *BMJ* 323:1123–1124.

Wang W, Xu J, Wang M (2018) Effects of recommendation neutrality and sponsorship disclosure on trust vs. distrust in online recommendation agents: Moderating role of explanations for organic recommendations. *Management Science* 64(11):5198–5219.

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 "*Precise or Broad? The Reward-Learning Frontier in Algorithmic Advice Design*"

Appendix A: Proofs and additional results for §2

A.1. Derivation of the required information

To derive the information-processing cost, recall that the decision maker's prior is uniform: $P(A_\tau = A^*_\tau \mid A^*_\tau \in \mathcal{B}_\tau) = 1/B_\tau$, where A_τ denotes the (random) action chosen by the decision maker and A^*_τ the optimal action. Processing a signal Z about A^*_τ is costly according to the standard rational inattention cost $\kappa I(A^*_\tau; Z)$, where $\kappa \geq 0$ and $I(\cdot;\cdot)$ denotes mutual information. Because only the decision maker's choice A_τ is payoff relevant, any signal that induces A_τ has weakly higher information cost than A_τ itself. We can therefore write the minimum information cost for a given accuracy in terms of $I(A^*_\tau; A_\tau) = H(A_\tau) - H(A_\tau \mid A^*_\tau)$, where $H(\cdot)$ denotes entropy.

Condition on the event $A^*_\tau \in \mathcal{B}_\tau$. We impose a symmetric accuracy structure. With probability q, the decision maker selects the best option: $P(A_\tau = a^*_\tau \mid A^*_\tau = a^*_\tau) = q$. With probability $1-q$, the error is spread uniformly over the remaining $B-1$ options: $P(A_\tau = a \mid A^*_\tau = a^*_\tau) = \frac{1-q}{B-1}$, $\forall a \in \mathcal{B}_\tau \setminus \{a^*_\tau\}$. Under the uniform prior, the chosen action A_τ is also uniformly distributed, so $H(a_\tau) = \log B$. Moreover, $H(a_\tau \mid a^*_\tau) = -q \log q - (1-q)\log\left(\frac{1-q}{B-1}\right)$. Therefore,

$$M_{B_\tau}(q) := I(A^*_\tau; A_\tau) = \log B + q \log q + (1-q)\log\left(\frac{1-q}{B-1}\right)$$

and the information-processing cost is $\kappa M_{B_\tau}(q)$.

A.2. Proofs of mathematical statements for §2.1

Proof of Lemma 1. For (i), note that $q^*(B_\tau, \rho_\tau) = \frac{K_\tau}{K_\tau + B_\tau - 1}$ for any $B_\tau \in \mathbb{N}_{>0}$. We have $q^*(B_\tau + 1, \rho_\tau) = \frac{K_\tau}{K_\tau + B_\tau + 1 - 1} = \frac{K_\tau}{K_\tau + B_\tau} < \frac{K_\tau}{K_\tau + B_\tau - 1} = q^*(B_\tau + 1, \rho_\tau)$.

For (ii), $u = u(p_\tau, B_\tau, \rho_\tau) = 1 \Leftrightarrow \kappa \log\left(\frac{K_\tau + B_\tau - 1}{B_\tau}\right) \geq p_\tau r \Leftrightarrow K_\tau + B_\tau - 1 \geq B_\tau \exp(p_\tau r/\kappa) \Leftrightarrow B_\tau \leq \frac{K_\tau - 1}{\exp(p_\tau r/\kappa) - 1} = \frac{\exp(\rho_\tau r/\kappa) - 1}{\exp(p_\tau r/\kappa) - 1} = \bar{B}_\tau$. Clearly, this is greater than one if and only if $\rho_\tau \geq p_\tau$.

For (iii), first assume $B_\tau > \bar{B}_\tau$. Then, $u = 0$, so $A(p_\tau, B_\tau, \rho_\tau) = p_\tau$, which is independent of B_τ. Next, assume $B_\tau \leq \bar{B}_\tau$, so $u = 1$ and $A(p_\tau, B_\tau, \rho_\tau) = q^*(B_\tau, \rho_\tau)$, which is strictly decreasing, following (i).

For (iv), $u(p_{i,\tau}, B_\tau, \rho_\tau) = 1 \Rightarrow \kappa \log\left(\frac{K_\tau + B_\tau - 1}{B_\tau}\right) \geq p_{i,\tau} r \geq p_{i,\tau} r$. Thus, $u(p_{j,\tau}, B_\tau, \rho_\tau) = 1$.

For (v), first note that $L(B_\tau) = L(p_\tau, B_\tau, \rho_\tau) \geq 0$ follows from the assumptions on Ψ. Moreover, $M_1(q) = 0$ by assumption, so $L(1) = \Psi(0) = 0$. Next, for any $B_\tau > \bar{B}_\tau$, the decision maker does not engage with action set $\mathcal{B}_\tau$, so, there is no effort and, again, $L(B_\tau) = \Psi(0) = 0$. Next, consider $B_\tau \in \{2, \ldots, \lfloor \bar{B}_\tau \rfloor\}$ and note that, by the fact that Ψ is strictly increasing, $L(B_\tau)$ moves in the same way as $M^*_{B_\tau} = M_{B_\tau}(q^*(B_\tau, \rho_\tau))$. In particular, $\Delta := M^*_{B_\tau + 1} - M^*_{B_\tau} = [\log(B_\tau + 1) - \log(B_\tau)] - [\log(K_\tau + B_\tau) - \log(K_\tau + B_\tau - 1)] - K_\tau \log K_\tau / [(K_\tau + B_\tau - 1)(B_\tau + K_\tau)]$. Note that $K_\tau = \exp(r\rho_\tau/\kappa) > 1$ and the derivative to K_τ is

$$\frac{d\Delta}{dK_\tau} = \frac{(K_\tau^2 - B_\tau^2 + B_\tau)\log K_\tau}{(K_\tau + B_\tau - 1)^2 (K_\tau + B_\tau)^2}.$$

With $\lim_{K_\tau \downarrow 1} \Delta = 0$ and $\lim_{K_\tau \downarrow 1} d\Delta/dK_\tau = 0$. Say $B_\tau = 1$: Here, the derivative is positive for any $K_\tau > 1$, so Δ is positive, too. Now consider $B_\tau \geq 2$. Here, $d\Delta/dK_\tau > 0 \Leftrightarrow K_\tau > \sqrt{B_\tau(B_\tau - 1)}$. Hence, for small enough K_τ, the difference is negative. That K_τ-threshold increases in B_τ. It follows that $M^*_{B_\tau}$ is first strictly increasing, then strictly decreasing (with the second part only if $\bar{B}_\tau$ is large enough). Hence, there is either a single peak, or two values B_τ and $B_\tau + 1$

that both lead to the same maximum value. If we assume that B_τ is continuous on $[1, \bar{B}_\tau]$. Then, $dM^*_{B_\tau}/dB_\tau = 1/B_\tau - 1/(K_\tau+B_\tau-1) - K_\tau \log K_\tau/(K_\tau+B_\tau-1)^2$. Hence, $dM^*_{B_\tau}/dB_\tau > 0 \Leftrightarrow B_\tau < (K_\tau-1)^2/[K_\tau(\log K_\tau - 1)+1]$, where the right-hand side is strictly greater than 1. The result then follows from the boundedness. □

Proof of Proposition 1. Assume $u = u(p_\tau, B_\tau, \rho_\tau) = 0$. Then, $A = A(p_\tau, B_\tau, \rho_\tau) = p_\tau$. Otherwise, if $u = 1$, $r p_\tau \le V(B_\tau, \rho_\tau) \le r q^*$, so $q^* \ge p_\tau$ and $A \ge p_\tau$. Hence, the second term in $p_{\tau+1}$ is non-negative. The third term is also non-negative, by definition of L. Moreover, both $[A(p_\tau, B_\tau, \rho_t a u) - p_t]$ and $(1-p_\tau)L(p_\tau, B_\tau, \rho_\tau)$ are upper-bounded by $(1-p_\tau)$, while $D(\delta_{\tau+1}) + H(\delta_{\tau+1}) \le 1$, so the upper bound on $p_{\tau+1}$ follows immediately. □

Proof of Lemma 2. Recall from Lemma 1 that $A(p_1, B, \rho_1)$ is strictly decreasing on $B \in \{1, 2, \dots, \bar{B}_1\}$. If $B < \bar{B}_1$, this immediately implies $\alpha_B > 0$. If $B > \bar{B}_1$ but $1 < \bar{B}_1$, we have $A(p_1, 1, \rho_1) = 1$, while $A(p_1, B, \rho_1) = p_1$. Hence, $\alpha_B > 0 \Leftrightarrow p_1 < 1$. Because $\bar{B}_\tau > 1 \Leftrightarrow p_1 < \rho_1$ and $\rho_1 \le 1$, this is true. Finally, if $\bar{B}_1 \ge B$, then $A(p_1, B, \rho_1) = q^*(B, \rho_1) = \frac{K_1}{K_1+B-1}$. Hence, $\alpha_B = \frac{B-1}{K_1+B-1}$. □

Proof of Proposition 2. Recall that $p_2(B) = p_1 + D(\delta)\,[A(p_1, B, \rho_1) - p_1] + H(\delta)(1-p_1)L(p_1, B, \rho_1)$. Hence, $p_2(B) - p_2(1) = H(\delta)(1-p_1)\,[L(p_1, B, \rho_1) - L(p_1, 1, \rho_1)] - D(\delta)\,[A(p_1, 1, \rho_1) - A(p_1, B, \rho_1)]$. Note that $L(p_1, 1, \rho_1) = 0$ and $L(p_1, B, \rho_1) \ge 0$ (Lemma 1), so $\beta_B \ge 0$ follows. Also, $\alpha_B \ge 0$ follows from Lemma 2. Finally, as $\delta \to 1$, both coefficients approach zero, while β_B and α_B are upper-bounded: $\alpha_B \le A(p_1, 1, \rho_1) \le 1$ and $\beta_B \le (1-p_1)L(\rho_1, B_1^*, \rho_1) < \infty$ (where B_1^* can be found from Lemma 1 again). □

Proof of Proposition 3. Assume $B \le \bar{B}_1$. Then, $A_1 := A(p_1, B, \rho_1) = K_1/(K_1+B-1)$, $M_B^* = \log B + K_1 \log K_1/(K_1+B-1) - \log(K_1+B-1)$, and $p_2 = p_1 + D(\delta)\,[K_1/(K_1+B-1) - p_1] + H(\delta)(1-p_1)\Psi(M_B^*)$. Because $|\mathcal{A}_2| > \bar{B}_2$, the decision maker does not exert effort in E_2. Hence, $A_2 := A(p_2, |\mathcal{A}_2|, 1) = p_2$. The designer optimizes $J(B) = \mathbb{E}[R^-] + r[\gamma A_1 + (1-\gamma)A_2]$ over $B \in \{1, 2, \dots, |\mathcal{A}_1|$, which has the same maximum as $T(B) := \gamma A_1 + (1-\gamma)A_2$.

Assume, first, that B can be chosen from $[0, \infty)$. We have

$$T'(B) = \frac{(1-\gamma)(1-p_1)\,[(K_1-1)(K_1+B-1) - BK_1 \log K_1]\,H(\delta) - BK_1\,[\gamma + (1-\gamma)D(\delta)]}{B(K_1+B-1)^2}.$$

Denote the numerator with $N(B)$. The derivative is positive if and only if $N(B) > 0$ and $B > 0$. We have $N'(B) = -(1-\gamma)(1-p_1)\,[K_1 \log K_1 - K_1 + 1]\,H(\delta) - K_1\,[\gamma + (1-\gamma)D(\delta)] < 0$ and $N(B) \le N(0) = (1-\gamma)(1-p_1)(K_1-1)^2 H(\delta) > 0$. It follows that $T'(B) > 0 \Leftrightarrow B < B^*$ and $T'(B) < 0 \Leftrightarrow B > B^*$. Due to the unimodularity of $T(B)$, the best discrete value is either directly above or below B^*. Moreover, if $\lceil B^* \rceil \le \bar{B}_1$, the designer never prefers to induce the decision maker not to exert effort, since exerting effort leads to a higher accuracy.

Now let $B^*(\delta) := B^*$ and consider a change in δ:

$$(B^*)'(\delta) = \frac{(1-\gamma)(1-p_1)K_1(K_1-1)^2\,[H'(\delta)\,(\gamma + (1-\gamma)D(\delta)) - H(\delta)(1-\gamma)D'(\delta)]}{[(1-\gamma)(1-p_1)(K_1 \log K_1 - K_1 + 1)H(\delta) + K_1\,[\gamma + (1-\gamma)D(\delta)]]^2}.$$

Clearly all terms are positive, except the square brackets in the numerator. Recall that $\lim_{\delta\to 1} H(\delta) = \lim_{\delta \to 1} D(\delta) = 0$, so the limit of the square brackets, as $\delta \to 1$, is $H'(\delta)\gamma < 0$. The fact that $(B^*)'(\delta) < 0$ for all $\delta > \delta_0$ for some $\delta_0 \in [0, 1]$ follows from the continuity of B^* in δ. Moreover, because $D'(\delta) < 0$, $(B^*)'(\delta) > 0 \Leftrightarrow \frac{H'(\delta)}{(1-\gamma)D'(\delta)} < \frac{H(\delta)}{\gamma+(1-\gamma)D(\delta)}$. As $\frac{D(\delta)}{H(\delta)}$ is assumed decreasing for $H(\delta) > 0 \Leftrightarrow \delta < 1$, we have $\frac{H'(\delta)}{(1-\gamma)D'(\delta)} < \frac{H(\delta)}{(1-\gamma)D(\delta)}$. Hence, $(B^*)'(\delta) > 0$ whenever $\delta < 1$ and γ is sufficiently small. □

A.3. Proofs of mathematical statements and additional results for §2.2

We start with a key definition used throughout the proofs in this section:

DEFINITION EC.1. Consider two probability vectors, v and w, of equal dimensions. We write $v \succeq w$ when the partial sums of the components of v, sorted in decreasing order, are weakly larger than the corresponding partial sums of w, with equality for the full sum.

In particular, a vector that majorizes another is more concentrated. We will consider all relevant action distributions to be padded with zeros, so that they have dimension $|\mathcal{A}|$. We then have:

LEMMA EC.1. *Assume $\rho_1 > 0$ and let $1 \le B' < B$, where B' is the breadth of advice regime b' and B is the breadth of advice regime b for some state $s \in \mathcal{S}$. Then, the conditional choice distribution under breadth B' majorizes that under breadth B, that is, $\nu_{b'}(\cdot \mid s) \succeq \nu_b(\cdot \mid s)$. If $b' = p$, this translates into the executed policy: $\sigma_p(\cdot \mid s) \succeq \sigma_b(\cdot \mid s)\ \forall s \in \mathcal{S}$. Consequently, $|supp\ \sigma_p(\cdot \mid s)| \le |supp\ \sigma_b(\cdot \mid s)|$*

Proof. Recall $K_\tau = \exp(r\rho_1/\kappa) > 1$ and that $q_b^*(s) = \frac{K_\tau}{K_\tau + B - 1}$. Conditional on engagement with advice of breadth B, the probability vector, sorted in decreasing order, is

$$\nu_B = \left(\frac{K_\tau}{K_\tau + B - 1}, \frac{1}{K_\tau + B - 1} \ldots, \frac{1}{K_\tau + B - 1}, 0, \ldots, 0 \right),$$

where the first B entries are positive. For $k \le B'$, the sum of the first k largest components is $\frac{K_\tau + k - 1}{K_\tau + B - 1}$. Meanwhile, the sum of of the first k largest components $\nu_{B'}$ is $\frac{K_\tau + k - 1}{K_\tau + B' - 1}$. Because $B' < B$, the latter is larger. For $k > B'$, the sum of the first k largest components $\nu_{B'}$ is one, which is again (weakly) greater than $\frac{K_\tau + k - 1}{K_\tau + B - 1}$.

For the second part, note that $V(B, \rho_1) = \kappa \log\left(1 + \frac{K_\tau - 1}{B}\right)$ is decreasing in B. Since $B_p(s) = 1 \le B_b(s)$, broad advice cannot be used in a state where precise advice is not used. Hence, there are three cases: First, both regimes lead to engagement. Then, $\sigma_p(\cdot \mid s)$ is a point mass and majorizes $\sigma_b(\cdot \mid s)$. Second, only precise advice leads to engagement. Then, $\sigma_p(\cdot \mid s)$ is again a point mass, while decision makers use $\mu(\cdot \mid s)$ under broad advice. The former majorizes the latter for any possible distribution. Third, neither regime leads to engagement, so the decision maker uses $\mu(\cdot \mid s)$ in either case, and the distributions are equal. The support ordering follows from the same three cases. □

Next, we show how a distribution that majorizes another will lead to fewer state visits. Together with Lemma EC.1, this directly implies that learning from experience is higher under broad advice than under precise advice, conditional on the number of state visits being the same.

LEMMA EC.2. *Let h be the previously defined function controlling diminishing returns to learning. Let π be a probability distribution over a finite set of actions $\{1, \ldots, m\}$. After n independent draws from π, let $N_x(n)$ denote the number of times that action x is selected, and define $\zeta_h(n, \pi) := \mathbb{E}\left[\sum_{x=1}^m h\left(N_x(n)\right)\right]$.*

If $\pi \succeq \pi'$, then $\zeta_h(n, \pi) \le \zeta_h(n, \pi')$ for every $n \ge 1$. If $n \ge 2$, then the inequality is strict.

Proof. For $y \in [0, 1]$, define $F_n(y) := \mathbb{E}[h(Z)]$, where $Z \sim Binomial(n, y)$. By linearity of expectations, and the fact that the marginals of a multionomial distribution are binomial distributions, we have $\zeta_h(n, \pi) = \sum_{x=1}^m F_n(\pi_x)$, where π_x is the entry of π corresponding to action x. Assume first that $n = 1$. Then, $F_1(y) = yh(1)$, so $\zeta_h(1, \pi) = h(1) \sum_{x=1}^m \pi_x = h(1) = \zeta_h(1, \pi')$.

Assume now that $n \geq 2$. Consider first the binomial pdf: $p_{n,k}(y) = \binom{n}{k} y^k (1-y)^{n-k}$ $\forall k = 0, \ldots, n$, where we let $p_{n,k}(y) = 0$ for other values of k. Differentiating yields

$$\frac{d}{dy} p_{n,k}(y) = \binom{n}{k} \left[k y^{k-1} (1-y)^{n-k} - (n-k) y^k (1-y)^{n-k-1} \right] = n \left[p_{n-1,k-1}(y) - b_{n-1,k}(y) \right].$$

Consequently,

$$F_n'(y) = n \sum_{k=0}^{n} h(k) \left[b_{n-1,k-1}(y) - b_{n-1,k}(y) \right] = n \sum_{k=0}^{n-1} \left[h(k+1) - h(k) \right] b_{n-1,k}(y) = n \sum_{k=0}^{n-1} \Delta h(k) b_{n-1,k}(y),$$

where $\Delta h(k) := h(k+1) - h(k)$. Differentiating once more yields

$$F_n''(y) = n(n-1) \sum_{k=0}^{n-2} \left[\Delta h(k+1) - \Delta h(k) \right] b_{n-2,k}(y) = n(n-1) \sum_{k=0}^{n-2} \Delta_2 h(k) \binom{n-2}{k} y^k (1-y)^{n-2-k},$$

where $\Delta_2 h(k) := h(k+2) - 2h(k+1) + h(k)$.

Because h is strictly concave, its increments over adjacent unit intervals are decreasing, i.e., $h(k+2) - h(k+1) < h(k+1) - h(k)$, so $\Delta_2 h(k) < 0$ for all k. Every other term is nonnegative. Therefore, $F_n''(y) < 0$ $\forall y \in [0,1]$ and F_n is strictly concave. Now define $G_n(\pi) := \sum_{x=1}^{m} F_n(\pi_x)$, so $G_n(\pi) = \zeta_h(n, \pi)$. Because F_n is strictly concave, so is G_n. Indeed, for any two probability vectors $\pi, \tilde{\pi}$ and any $\lambda \in [0,1]$,

$$G_n\left(\lambda\pi + (1-\lambda)\tilde{\pi}\right) = \sum_{x=1}^{m} F_n\left(\lambda\pi_x + (1-\lambda)\tilde{\pi}_x\right) > \lambda \sum_{x=1}^{m} F_n(\pi_x) + (1-\lambda) \sum_{x=1}^{m} F_n(\tilde{\pi}_x) = \lambda G_n(\pi) + (1-\lambda) G_n(\tilde{\pi}).$$

Moreover, G_n is invariant to relabeling the actions: For every permutation matrix P, $G_n(P\pi) = G_n(\pi)$, because the permutation only changes the order of the terms in the sum.

Consider now two probability distributions with $\pi \succeq \pi'$. By the Hardy-Littlewood-Polya theorem (Le Breton 2006), there exists a doubly stochastic matrix D such that $\pi' = D\pi$. By the Birkhoff-von Neumann theorem, every doubly stochastic matrix can be represented as a convex combination of permutation matrices. Thus, for some permutation matrices $P_1, \ldots, P_J$ and weights $\lambda_j \geq 0$ satisfying $\sum_{j=1}^{J} \lambda_j = 1$, $D = \sum_{j=1}^{J} \lambda_j P_j$. Putting together the strict concavity and the permutation invariance, we have

$$\zeta_h(n, \pi') = G_n(\pi') = G_n\left(\sum_{j=1}^{J} \lambda_j P_j \pi \right) > \sum_{j=1}^{J} \lambda_j G_n(P_j \pi) = \sum_{j=1}^{J} \lambda_j G_n(\pi) = G_n(\pi) = \zeta_h(n, \pi). \quad \square$$

Proof of Theorem 1. Fix regime a. By irreducibility of the finite-state chain, every state $s \in \mathcal{S}$ is visited infinitely often almost surely, so $N_s^a(T) \overset{a.s.}{\to} \infty$ and $h\left(N_s^a(T)\right) \overset{a.s.}{\to} 1$.

Now, fix $(s, \tilde{a})$. If $\sigma_a(\tilde{a} \mid s) = 0$, then $N_{s,\tilde{a}}^a(T) = 0$ for every T. If $\sigma_a(\tilde{a} \mid s) > 0$, state s is visited infinitely often and, at each visit, action $\tilde{a}$ is selected with the same strictly positive conditional probability. Therefore, $\tilde{a}$ is selected infinitely often at state s almost surely, and $N_{s,\tilde{a}}^a(T) \overset{a.s.}{\to} \infty$, as well as $h\left(N_{s,\tilde{a}}^a(T)\right) \overset{a.s.}{\to} 1$.

Because $\mathcal{S}$ and $\mathcal{A}$ are finite, and $0 \leq h \leq 1$, we can apply the dominated convergence theorem to obtain $\lim_{T\to\infty} L_T^{del}(a) = \sum_{s\in\mathcal{S}} m_a(s)$ and $\lim_{T\to\infty} L_T^{exp}(a) = \bar{I} \sum_{s\in\mathcal{S}} |\text{supp}\, \sigma_a(\cdot \mid s)|$.

Under precise advice, $B_p(s) = 1$ and $M_1 = 0$, so $m_p(s) = 0$ for every $s \in \mathcal{S}$. Under broad advice, $m_b(s) \geq 0$. From Lemma EC.1, we have $C_b \geq C_p$. At state s°, we have $B_b(s^\circ) > 1$ and advice is used under regime b. Since $V(B_b(s^\circ), \rho_1) \leq V(1, \rho_1)$, precise advice is also used here. Therefore, $\sigma_p(\cdot \mid s^\circ)$ is a point mass, while $\sigma_b(\cdot \mid s^\circ)$ has support of size $B_b(s^\circ) > 1$. Hence, $C_b > C_p$. In addition, $\rho_1 > 0$ implies $q_b(s^\circ) > 1/B_b(s^\circ)$ and, therefore, $m_b(s^\circ) > 0$. Hence, $\ell_{exp} > 0$ and $\ell_{del} > 0$. If at least one of η_{del} and η_{exp} is positive, then $L_\infty(b) > L_\infty(a)$. Then, convergence of $L_T(b) - L_T(a)$ to a strictly positive limit implies that there exists $T_0 < \infty$, such that $L_T(b) > L_T(p)$ $\forall T \geq T_0$. $\square$

A.4. Model extension: Designer concern for decision-maker effort

The baseline model assumes that the designer values performance in environments E_1 and E_2 but does not internalize the decision maker's information-processing cost. We relax this assumption here. Let $\omega \geq 0$ denote the designer's weight on cognitive cost. For any fallback accuracy p_1, consideration-set breadth B, and perceived advice quality ρ_1, define realized information-processing effort as $e(p_1, B, \rho_1) := u(p_1, B, \rho_1) M_B\left(q^*(B, \rho_1)\right)$. Thus, $\kappa e(p_1, B, \rho_1)$ is the cognitive cost borne by the decision maker. We apply the same temporal weights to performance and effort. The designer's objective becomes $J_\omega(B) := \mathbb{E}[R^-] + r\left[\gamma A(p_1, B, \rho_1) + (1-\gamma) A\left(p_2(B), |\mathcal{A}_2|, 1\right)\right] - \omega\kappa\left[\gamma e(p_1, B, \rho_1) + (1-\gamma) e\left(p_2(B), |\mathcal{A}_2|, 1\right)\right]$. The case $\omega = 0$ recovers the baseline model, $\omega = 1$ means that the designer fully internalizes the decision maker's processing costs, while $\omega > 1$ permits cognitive burden to carry additional organizational consequences.

Comparison with precise advice. Maintain the no-engagement condition in E_2 used in Proposition 3, $|\mathcal{A}_2| > \bar{B}_2$, for every advice breadth under consideration. Then $A\left(p_2(B), |\mathcal{A}_2|, 1\right) = p_2(B)$ and $e\left(p_2(B), |\mathcal{A}_2|, 1\right) = 0$. For $B > 1$, define $\alpha_B := A(p_1, 1, \rho_1) - A(p_1, B, \rho_1)$, $\beta_B := (1-p_1) L(p_1, B, \rho_1)$, $m_B := e(p_1, B, \rho_1)$, and $\chi(\delta, \gamma) := \gamma + (1-\gamma) D(\delta)$. Because precise advice has breadth one, $m_1 = 0$. Combining $J_\omega(B)$ with Proposition 2 gives

$$J_\omega(B) - J_\omega(1) = r(1-\gamma) H(\delta) \beta_B - r\chi(\delta, \gamma)\alpha_B - \omega\gamma\kappa m_B. \tag{EC.1}$$

Hence, broad advice is preferred to precise advice if and only if $(1-\gamma) H(\delta)\beta_B > \chi(\delta, \gamma)\alpha_B + \omega\gamma\frac{\kappa}{r} m_B$. Define $\bar{\omega}_B(\delta, \gamma) := \frac{r[(1-\gamma)H(\delta)\beta_B - \chi(\delta,\gamma)\alpha_B]}{\gamma\kappa m_B}$.

PROPOSITION EC.1. *Maintain the assumptions of Proposition 3. For every fixed $B > 1$:*

1. *$J_\omega(B) - J_\omega(1)$ is weakly decreasing in ω and is strictly decreasing whenever $\gamma > 0$ and $m_B > 0$.*
2. *The set of advice breadths that the designer prefers to precise advice is nested and weakly shrinking in ω.*
3. *If $\gamma > 0$ and $m_B > 0$, the designer prefers advice of breadth B to precise advice exactly when $\bar{\omega}_B(\delta, \gamma) \geq 0$ and $0 \leq \omega \leq \bar{\omega}_B(\delta, \gamma)$.*

Proof. Under $|\mathcal{A}_2| > \bar{B}_2$, the decision maker does not process information in E_2, so second-period accuracy is $p_2(B)$ and second-period effort is zero. Therefore, $J_\omega(B) - J_\omega(1) = r\gamma\left[A(p_1, B, \rho_1) - A(p_1, 1, \rho_1)\right] + r(1-\gamma)\left[p_2(B) - p_2(1)\right] - \omega\gamma\kappa(m_B - m_1)$. Proposition 2 gives $p_2(B) - p_2(1) = H(\delta)\beta_B - D(\delta)\alpha_B$, while $m_1 = 0$. Substitution yields (EC.1). Its derivative with respect to ω is $-\gamma\kappa m_B \leq 0$, with strict inequality whenever $\gamma > 0$ and $m_B > 0$. The nesting result and the threshold $\bar{\omega}_B(\delta, \gamma)$ follow immediately. □

For fixed B, the effort term in (EC.1) does not depend on environmental novelty δ. Internalizing effort therefore shifts the broad-minus-precise objective difference downward without changing its shape as a function of δ.

Optimal breadth. Restrict attention to the engagement region $B \leq \bar{B}_1$. Up to terms that do not depend on B, the objective is $r\chi(\delta, \gamma) q^*(B, \rho_1) + r(1-\gamma)(1-p_1) H(\delta)\Psi\left(M_B(q^*(B, \rho_1))\right) - \omega\gamma\kappa M_B(q^*(B, \rho_1))$. An interior solution, assuming B is continuous, satisfies

$$-r\chi(\delta, \gamma)\frac{K_1}{(K_1 + B - 1)^2} + \left[r(1-\gamma)(1-p_1) H(\delta)\Psi'\left(M_B(q^*(B, \rho_1))\right] - \omega\gamma\kappa\right] \frac{(K_1 - 1)(K_1 + B - 1) - BK_1 \log K_1}{B(K_1 + B - 1)^2} = 0.$$

To obtain a closed form analogous to Proposition 3, we assume again a linear learning function $\Psi(m) = m$ and define $c_\omega := (1-\gamma)(1-p_1) H(\delta) - \omega\gamma\kappa/r$. If $c_\omega \leq 0$, precise advice is weakly optimal: broad advice lowers accuracy and creates no positive net value from processing effort. If $c_\omega > 0$, the continuous interior candidate is $B^*_\omega =$

$\frac{c_\omega(K_1-1)^2}{c_\omega(K_1\log K_1-K_1+1)+K_1\chi(\delta,\gamma)}$. When $B^*_\omega \le 1$, precise advice is optimal. When $B^*_\omega > 1$ and $\lceil B^*_\omega \rceil \le \bar{B}_1$, the best integer breadth is one of $\lfloor B^*_\omega \rfloor$ and $\lceil B^*_\omega \rceil$. Moreover, whenever the solution is interior,

$$\frac{\partial B^*_\omega}{\partial \omega} = -\gamma \frac{\kappa}{r} \frac{K_1\chi(\delta,\gamma)(K_1-1)^2}{[c_\omega(K_1\log K_1 - K_1 + 1) + K_1\chi(\delta,\gamma)]^2} \le 0.$$

Thus, placing greater weight on the decision maker's cognitive burden shifts the designer toward more precise advice. At the same time, whenever broad advice remains optimal, its relative advantage continues to be concentrated at intermediate environmental novelty because the added effort penalty is independent of δ.

Appendix B: Experimental protocol and additional definitions

This appendix reports the details of the implementation of Studies 1 and 2. Both main studies were preregistered before data collection (Study 1 at https://aspredicted.org/7LP_J9G, Study 2 at https://aspredicted.org/23C_WBQ).

B.1. Task mechanics and charging-time function

In each round, elapsed game time is the sum of driving time, exit time, charging time, and any emergency penalty. Exiting the highway to charge imposes a fixed cost of 30 in-game minutes. Let the current charge be s and the additional charge be ℓ. The charging-time component used in the experiment is

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

If the EV does not have enough charge to complete the next segment after traffic is realized, the participant incurs an emergency charging penalty of 300 in-game minutes and arrives at the next stop with 0% charge. The charging function, fixed exit cost, and emergency penalty are common across all treatment conditions.

B.2. Implementation, screening, and incentives

Participants first reviewed the task instructions and completed a guided training round before entering the main task. The training round introduced the map, traffic-range display, charging slider, fixed exit cost, nonlinear charging time, and emergency penalty. It is separate from the pre-advice phase (E_0) and excluded from every analysis. Participants then answered comprehension and attention questions concerning the interface, traffic uncertainty, emergency penalty, and performance objective.

Table EC.1 reports participants at each screening stage. Study 1 contains 219 recorded responses, of which 210 are non-preview responses, 180 consented, 112 finished, 106 passed the recorded attention check, and 90 enter the analytic sample. Study 2 contains 422 responses in the qualitative export used to define the study cohort, of which 415 pass the recorded attention check and 400 enter the analytic sample. Among those passing the recorded attention check, 16 Study 1 and 15 Study 2 observations were removed under additional study-specific quality and usability criteria.

The main task was incentivized by elapsed game time. Study 1 participants received a fixed payment of either \$2.50 or \$3.00, depending on recruitment wave, plus a performance bonus averaging \$2.21. Study 2 participants received a \$4.00 fixed payment plus a performance bonus averaging \$1.84. The bonus paid was increasing in decreasing elapsed game time. Table EC.2 reports total pay, bonus amounts, and completion times. Treatment assignment was between subjects within each study. Tables EC.3 and EC.4 report the final analytic cell counts.

Table EC.1 Participant flow.

Study	Stage	N
Study 1	Raw responses	219
Study 1	Non-preview/IP responses	210
Study 1	Consented	180
Study 1	Finished	112
Study 1	Passed attention check	106
Study 1	Final analytic sample	90
Study 2	Raw responses	422
Study 2	Passed attention check	415
Study 2	Final analytic sample	400

Table EC.2 Sample characteristics, compensation, and duration.

Measure	Study 1	Study 2
Final analytic sample	90	400
Fixed participation payment	$2.50 or $3.00	$4.00
Mean total pay	5.14	5.84
Median total pay	4.50	5.82
Mean bonus	2.21	1.84
Median bonus	1.50	1.83
Mean duration, minutes	50.67	43.00
Median duration, minutes	39.50	36.22
Female, %	40.00	49.50
Age 25–44, %	66.67	57.00
Driver's license, %	95.56	89.00
Car owner, %	91.11	81.25
EV owner, %	40.00	13.25
Frequent driver, %	67.78	76.25
Frequent navigation-app user, %	60.00	45.50
Similar-game experience, %	64.44	38.50

Note. Total pay equals fixed participation payment plus performance-based bonus. Duration is measured in minutes. Prior-experience variables are based on post-game survey responses.

Table EC.3 Study 1 treatment cell counts.

Advice	Simple traffic	Complex traffic	Total
Broad	26	18	44
Precise	22	24	46

Table EC.4 Study 2 treatment cell counts.

Advice	Rationale	Familiar	Unseen	Total
Broad	Rationale hidden	31	34	65
Broad	Rationale shown	38	32	70
Precise	Rationale hidden	33	36	69
Precise	Rationale shown	34	31	65
Specific-broad	Rationale hidden	35	32	67
Specific-broad	Rationale shown	30	34	64

B.3. Advice wording and rationale manipulation

All advice formats are generated from the same optimal policy, obtained through dynamic programming. The treatments vary how the recommendation is communicated. In the precise condition, participants receive an executable charging target. In the broad condition, participants receive a qualitative charging rule that is generated from the optimal policy but leaves the numerical mapping to the participant. In the specific-broad condition, participants receive the same type of qualitative rule with additional operating detail, such as the relevant traffic contingency, while still having to compute the charge amount. In the rationale-shown condition in Study 2, the advice is accompanied by a short justification explaining the logic behind the recommendation, such as the fixed overhead from exiting or the nonlinear charging-time function. Rationale visibility does not change whether the recommendation itself is precise, broad, or specific-broad.

Study 1: Advice Wording

Decision context	Exact wording shown to participants
No charge	You should not charge at this stop.
Precise advice	You should charge [optimal charge]% at this stop.
Broad advice: split	You should charge at this stop enough to cover just this road segment.
Broad advice: batch	You should charge **extra** at this stop to cover the travel through this segment and the next one.

Study 2: Advice Wording

Decision context	Exact wording shown to participants
No charge	You should not charge at this stop.
Precise advice	You should charge [optimal charge]% at this stop.
Broad advice: split	You should charge at this stop enough to cover just this road segment.
Specific-broad advice: split	You should charge just enough at this stop to cover the coming road segment, considering the worst-case traffic.
Broad advice: batch	You should charge **extra** at this stop to cover the travel through this segment and the next one.
Specific-broad advice: batch	You should charge **extra** at this stop to cover the travel through this segment and the next one, considering the worst-case traffic.

Study 2: Rationale Wording

Decision context	Exact wording shown to participants
Batch	Looking ahead, if you need to charge for a segment but the sum of charges required for multiple segments is less than or equal to 50%, charging is relatively fast so you should charge enough for these segments in one stop.
Split before the last stop	Looking ahead, if you need to charge for a segment and the sum of charges required for multiple segments is greater than 50%, charging is relatively slow so you should charge enough only for the next immediate segment.
Last stop	As there is only one segment left, you should charge only for it.

Representative task screens are shown in Figures 2, 3, 4, and 6 in the main text. The complete Qualtrics instruments, screenshots, and wording files are retained in the study materials and are available from the authors.

B.4. Performance benchmarks used for normalization

This appendix reports the optimal and worst-case elapsed game time (EGT) benchmarks used to construct normalized performance. Recall that we compute

$$Y_{ir} = \frac{T_r^{\text{worst}} - T_{ir}^{\text{actual}}}{T_r^{\text{worst}} - T_r^{\text{opt}}},$$

with round and condition-specific benchmark values. The optimal EGT is the optimal policy benchmark under the realized traffic path used for the corresponding round. The worst-case EGT is the elapsed game time of the worst performer for that round and condition. Tables EC.5–EC.6 report the relevant values.

Table EC.5 Study 1 optimal, oracle, and worst-case EGT benchmarks.

Round	Traffic condition	Optimal EGT	Worst EGT
1	All	492.07	1364
2	All	1049.83	2401
3	All	512.13	1646
4	Centered	498.26	1640
5	Centered	505.79	1643
6	Centered	503.38	1641
4	Skewed	470.45	1624
5	Skewed	522.24	1653
6	Skewed	491.58	1635
7	All	1051.69	2403

Note. Study 1 contains traffic-condition-specific benchmarks in Rounds 4–6 because the traffic realization treatment affects the relevant elapsed-time benchmark.

Table EC.6 Study 2 optimal and worst-case EGT benchmarks.

Round	Sequence	Optimal EGT	Worst EGT
1	All	531.40	1651
2	All	515.50	1644
3	All	492.50	1631
4	All	491.70	1631
5	Familiar	460.30	1619
6	Familiar	498.40	1632
7	Familiar	556.40	1661
5	Unseen	978.70	2357
6	Unseen	1000.70	2371
7	Unseen	1020.00	2381

Note. Study 2 uses common benchmarks in Rounds 1–4 and sequence-specific benchmarks in Rounds 5–7, when participants are assigned to the familiar or unseen post-advice sequence.

B.5. Action agreement

G^a_{id} defines format-consistent action agreement. This appendix makes more explicit the interval construction, gives worked examples, and reports sensitivity to alternative decision-level definitions.

Let c_{id} denote current charge at advised decision d. If the qualitative recommendation covers $K_{id} \in \{0, 1, 2\}$ upcoming segments, let d_{idk} be the distance of segment k and let $[\underline{t}_{idk}, \bar{t}_{idk}]$ be its displayed traffic range. The feasible additional-charge interval implied by the recommendation is

$$\ell_{id} := \min\left\{100 - c_{id}, \max\left[0, \sum_{k=1}^{K_{id}} \left(d_{idk} + \underline{t}_{idk}\right) - c_{id}\right]\right\},$$
$$u_{id} := \min\left\{100 - c_{id}, \max\left[0, \sum_{k=1}^{K_{id}} \left(d_{idk} + \bar{t}_{idk}\right) - c_{id}\right]\right\}.$$

For a no-charge recommendation, $K_{id} = 0$ and $\ell_{id} = u_{id} = 0$. Precise agreement requires $x_{id} = \tilde{a}^*_{id}$. Broad agreement requires $x_{id} \in \{\ell_{id}, \ldots, u_{id}\}$. Specific-broad advice identifies worst-case traffic but leaves the numerical mapping to the participant. We therefore define specific-broad agreement as $x_{id} \in \{h_{id}, \ldots, u_{id}\}$, where $h_{id} = \lceil (\ell_{id}+u_{id})/2 \rceil$. The tolerance analyses expand a nonzero point or interval by 5 or 10 percentage points, subject to the feasible action set. A no-charge recommendation always requires $x_{id} = 0$. Table EC.7 provides an example.

Table EC.7 Worked examples of format-consistent action agreement.

Advice format	State and recommendation	Consistent action(s)	Inconsistent action(s)
Precise	Numerical target is 45	45	Any integer other than 45
Broad, one segment	Current charge is 5; distance is 10; traffic is 5–20. Thus $[\ell, u] = [10, 25]$.	Any integer from 10 through 25	9 or below; 26 or above
Broad, two segments	Current charge is 0. The first segment has distance 5 and traffic 5–15; the second has distance 10 and traffic 5–20. Thus $[\ell, u] = [25, 50]$.	Any integer from 25 through 50	24 or below; 51 or above
Specific broad	For $[\ell, u] = [10, 25]$, the worse-traffic half is $[18, 25]$.	Any integer from 18 through 25	17 or below; 26 or above
No charge	The recommendation implies $[\ell, u] = [0, 0]$.	0	Any positive charge

Note. The two-segment example includes displayed uncertainty on both segments covered by the recommendation.

Table EC.8 reports participant-level action-agreement rates, including tolerance analyses. In Study 1, exact precise agreement averages 0.643, compared with 0.373 for broad interval agreement. The broad-minus-precise difference is −0.271 (95% CI $[-0.407, -0.135]$, $p < 0.001$). In Study 2, the corresponding rates are 0.808 for precise advice and 0.615 for broad advice. Specific-broad agreement averages 0.577. Relative to precise advice, the difference is −0.231 (95% CI $[-0.284, -0.179]$, $p < 0.001$). Relative to broad advice, the difference is −0.038 (95% CI $[-0.088, 0.013]$, $p = 0.142$).

Appendix C: Multiple-testing adjustments for focal tests

The main text reports raw p-values and confidence intervals. To assess sensitivity to multiplicity, Table EC.9 applies Holm and Benjamini–Hochberg adjustments within four analysis families: advice-phase reward, post-advice transfer, immediate change after removal, and total rationale effects. Note that the unseen-sequence outcome is an ANCOVA-adjusted post-advice level, not a short-map-to-long-map change.

Table EC.8 Participant-level action-agreement rates under alternative decision-level definitions.

Study	Advice format and definition	Primary	±5	±10
1	Precise, exact target	0.643	0.736	0.759
1	Broad, full implied interval	0.373	0.485	0.552
2	Precise, exact target	0.808	0.912	0.927
2	Broad, full implied interval	0.615	0.722	0.776
2	Specific broad, worse-traffic half interval	0.577	0.690	0.737

Note. Entries are means of participant-level action-agreement rates. Tolerance definitions leave no-charge recommendations at exactly zero.

Table EC.9 Focal tests with multiple-testing adjustments.

Family	Study	Test	Estimate	Raw p	Holm p	BH p
Reward gap	Study 1	Advice-phase performance: broad − precise	-0.185	< 0.001	< 0.001	< 0.001
Reward gap	Study 2	Advice-phase performance: broad - precise	-0.057	< 0.001	< 0.001	< 0.001
Reward gap	Study 2	Advice-phase performance: specific-broad - precise	-0.058	< 0.001	< 0.001	< 0.001
Reward gap	Study 2	Advice-phase performance: specific-broad - broad	-0.001	0.902	0.902	0.902
Transfer	Study 1	Matched short-map change: broad − precise	-0.063	0.409	1.000	0.483
Transfer	Study 1	Matched long-map change: broad − precise	0.123	0.024	0.171	0.054
Transfer	Study 2	Familiar pre-to-post change: broad - precise	0.077	0.016	0.144	0.050
Transfer	Study 2	Familiar pre-to-post change: specific-broad - precise	0.081	0.018	0.146	0.050
Transfer	Study 2	Familiar pre-to-post change: specific-broad - broad	0.004	0.903	1.000	0.903
Transfer	Study 2	Unseen post-advice level, ANCOVA: broad - precise	-0.049	0.032	0.192	0.059
Transfer	Study 2	Unseen post-advice level, ANCOVA: specific-broad - precise	-0.029	0.254	1.000	0.349
Transfer	Study 2	Unseen post-advice level, ANCOVA: specific-broad - broad	0.020	0.439	1.000	0.483
Transfer	Study 2	Familiar-minus-unseen interaction: broad - precise	0.143	< 0.001	< 0.001	< 0.001
Transfer	Study 2	Familiar-minus-unseen interaction: specific-broad - precise	0.103	0.005	0.051	0.028
Transfer	Study 2	Familiar-minus-unseen interaction: specific-broad - broad	-0.040	0.226	1.000	0.349
Removal	Study 1	First post-removal change: broad − precise	0.138	0.005	0.010	0.007
Removal	Study 2	Familiar first post-removal: broad - precise	0.117	0.002	0.007	0.005
Removal	Study 2	Familiar first post-removal: specific-broad - precise	0.118	< 0.001	0.002	0.002
Removal	Study 2	Familiar first post-removal: specific-broad - broad	0.002	0.951	0.951	0.951
Rationale	Study 2	Total rationale effect: precise, familiar	0.074	0.010	0.062	0.062
Rationale	Study 2	Total rationale effect: broad, familiar	-0.029	0.157	0.627	0.314
Rationale	Study 2	Total rationale effect: specific-broad, familiar	0.011	0.618	1.000	0.618
Rationale	Study 2	Total rationale effect: precise, unseen	-0.056	0.022	0.111	0.067
Rationale	Study 2	Total rationale effect: broad, unseen	0.022	0.382	1.000	0.573
Rationale	Study 2	Total rationale effect: specific-broad, unseen	-0.017	0.532	1.000	0.618

Note. Estimates are reported in normalized-performance units. Adjustments are computed across every row within the family shown in the first column. Holm controls the family-wise error rate; Benjamini–Hochberg controls the false discovery rate. Process measures are exploratory and are reported with confidence intervals rather than included in these performance families.

The assisted-performance results and the immediate post-removal contrasts are robust to both adjustments. The transfer evidence is less uniform. The Study 1 long-map estimate and the two familiar-sequence simple effects do not survive Holm adjustment across the full transfer family. The broad advice-format-by-sequence interaction remains significant under Holm adjustment, whereas the specific-broad interaction is marginal ($p = 0.051$). None of the six format-by-sequence total rationale effects survives Holm adjustment.

Appendix D: Additional analyses for Study 1

This appendix reports the estimates underlying §4, robustness to alternative specifications and measurement choices, and descriptive analyses of participants' charging strategies. The focal treatment effects are intention-to-treat comparisons by assigned advice format. Analyses that condition on observed agreement or use process measures are exploratory.

D.1. Focal estimates and matched changes

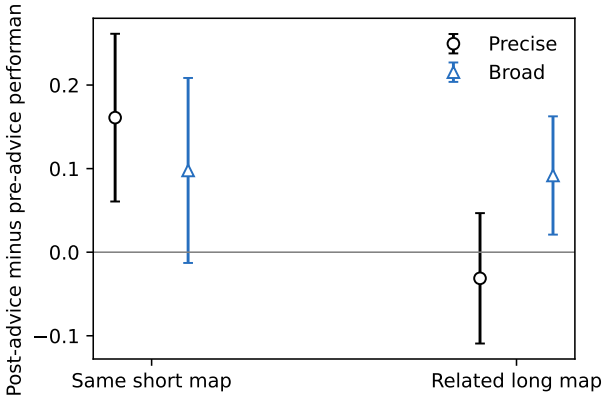
Table EC.10 reports performance contrasts used in the main text. The matched short-map change compares Round 6 with Round 1; the matched long-map change compares Round 7 with Round 2. Figure EC.1 plots treatment-specific changes.

Table EC.10 Study 1 performance estimates and robustness checks.

Outcome	Comparison	Estimate	95% CI	Raw p
Advice-phase performance, Rounds 3–5	Broad − precise	−0.185	[−0.273, −0.097]	< 0.001
Final advice round, Round 5	Broad − precise	−0.214	[−0.313, −0.115]	< 0.001
Short-map post-advice level, Round 6	Broad − precise	−0.077	[−0.190, 0.036]	0.180
Matched short-map change, Round 6 − Round 1	Broad − precise	−0.063	[−0.215, 0.088]	0.409
Long-map post-advice level, Round 7	Broad − precise	0.073	[−0.030, 0.177]	0.163
Matched long-map change, Round 7 − Round 2	Broad − precise	0.123	[0.016, 0.230]	0.024
First post-removal change, Round 6 − Round 5	Broad − precise	0.138	[0.043, 0.233]	0.005
Short-map post level, ANCOVA	Broad − precise	−0.084	[−0.189, 0.022]	0.119
Long-map post level, ANCOVA	Broad − precise	0.090	[−0.0003, 0.180]	0.051

Note. Performance is normalized as in Equation (1). The advice-phase estimate uses participant-clustered standard errors. Participant-level contrasts use Welch confidence intervals. The ANCOVA specifications regress the post-advice level on assigned advice format, the corresponding pre-advice map performance, and traffic assignment, with HC1 standard errors. Across the unified transfer family reported in Appendix C, the long-map change has Holm-adjusted $p = 0.171$ and Benjamini–Hochberg-adjusted $p = 0.054$; the short-map change has adjusted p-values of 1.000 and 0.483. The post-removal contrast has adjusted p-values of 0.010 and 0.007.

Figure EC.1 Study 1 changes in without-advice performance relative to the matched pre-advice round.



Note. Points are treatment means and whiskers are 95% confidence intervals. The short-map change is Round 6 minus Round 1; the long-map change is Round 7 minus Round 2.

The matched changes can also be read within treatment. Short-map performance increased by 0.161 after precise (95% CI [0.058, 0.264]) and by 0.098 after broad advice (95% CI [−0.016, 0.212]). Long-map performance changed by −0.031 after precise (95% CI [−0.111, 0.049]) and by 0.092 after broad advice (95% CI [0.019, 0.165]). These within-treatment changes are descriptive because the design has no group without advice during E_1.

D.2. Advice-phase behavior

Table EC.11 reports the advice-phase behavioral contrasts. Action agreement uses the format-specific primary definitions in Appendix B.5. Experienced state-action support counts distinct tuples of stop, current-charge bin, and added-charge bin during Rounds 3–5. Absolute distance from the precise policy target describes action dispersion around the numerical optimum; it is not the action-agreement measure for broad advice.

Table EC.11 Study 1 advice-phase process contrasts.

Outcome	Broad − precise	95% CI	p-value
Format-consistent action agreement	−0.271	[−0.407, −0.135]	< 0.001
Absolute distance from precise policy target	7.192	[3.361, 11.022]	< 0.001
Experienced state-action support, 5-point bins	1.383	[0.430, 2.337]	0.005
Experienced state-action support, 10-point bins	1.184	[0.134, 2.234]	0.028
Experienced state-action support, 20-point bins	1.733	[0.567, 2.900]	0.004
Decision time, winsorized seconds	0.343	[−4.571, 5.257]	0.891
Log decision time	−0.004	[−0.201, 0.193]	0.970
Active-effort index	0.058	[−0.115, 0.231]	0.511
Active slider movements	0.366	[−0.587, 1.320]	0.452
Looked ahead before acting	0.050	[−0.012, 0.113]	0.115

Note. Action agreement is a participant-level action-agreement-rate comparison with a Welch confidence interval. The effort and distance estimates come from decision-level regressions with participant-clustered standard errors and controls for round, exit, current charge, precise policy target, traffic-range width, and recommendation type. State-action support is measured at the participant level and compared using Welch confidence intervals. The active-effort index gives equal weight to log decision time, slider search, and future-traffic acquisition.

Broad advice yields lower format-consistent agreement and wider experienced state-action support than precise advice. The effort proxies are not statistically distinguishable across formats in Study 1. Eventual advice assignment also does not predict pre-advice process measures. The broad-minus-precise placebo estimates have $p = 0.660$ for log decision time, $p = 0.515$ for active slider movements, $p = 0.999$ for reveal clicks, $p = 0.537$ for looking ahead, and $p = 0.671$ for the effort index.

D.3. Traffic-pattern robustness

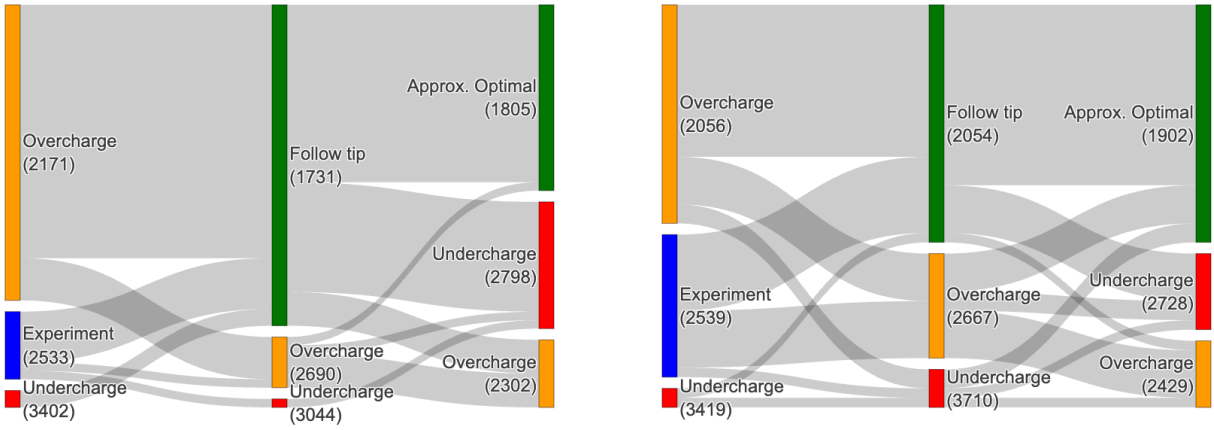
The traffic factor changes realized delays in Rounds 4–6 but leaves the maps, displayed traffic ranges, and advice policy unchanged. We estimate advice-format-by-traffic interactions for E_1 performance (matched changes and first post-removal change). None is statistically significant: the interaction estimates are 0.099 for advice-phase performance ($p = 0.261$), 0.063 for the short-map change ($p = 0.675$), −0.021 for the long-map change ($p = 0.837$), and −0.045 for the post-removal change ($p = 0.636$). The focal results therefore do not appear to be driven by the traffic realization.

D.4. Strategy-cluster transitions

We classify each 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). We then apply sequence clustering (Gabadinho et al. 2011) to participants' label sequences within each phase, to summarize strategy profiles. Next, we consider how participants transition between profile as they move from pre-advice to advice phase, and from advice to post-advice phase. Figure EC.2 reports the

resulting transitions. Precise advice moves more participants into the near-policy cluster while recommendations are visible, consistent with its higher action agreement and performance. Behavior remains more dispersed under broad advice. Among participants assigned to the "Follow tip" cluster during the advice phase, 76% of those receiving broad, respectively 55% of those receiving precise advice are assigned to a near-policy cluster after removal.

Figure EC.2 Study 1 cluster transitions by advice condition.



(a) Precise advice

(b) Broad advice

Note. The left, middle, and right bars indicate cluster assignments in the pre-advice, advice, and post-advice phases. Numbers in parentheses report average elapsed game time within clusters.

Appendix E: Additional analyses for Study 2

This appendix reports the estimates underlying §5, as well as several additional analyses.

E.1. Focal performance contrasts

Table EC.12 reports the focal advice-format contrasts. Advice-phase models pool Rounds 3–4. Familiar-sequence pre-to-post and removal estimates use the no-rationale cells. The unseen sequence does not have a pre-advice long-map observation; its post-advice outcome is therefore analyzed in levels with short-map baseline performance as a covariate. The final rows report the advice-format-by-sequence interactions from the ANCOVA model described in §3.4.

The advice-phase performance and action-agreement contrasts remain significant under both Holm and Benjamini–Hochberg adjustment. The familiar-sequence pre-to-post contrasts for broad and specific-broad advice relative to precise advice remain significant under Benjamini–Hochberg adjustment but not under Holm adjustment. The broad-versus-precise format-by-sequence interaction remains significant under both adjustments. The specific-broad-versus-precise interaction remains significant under Benjamini–Hochberg adjustment and is marginal under Holm adjustment ($p = 0.051$). The post-removal contrasts in the familiar sequence are also robust to both adjustments.

E.2. Decision process and experienced support

Decision time is the elapsed time on each decision screen. The distribution is right-skewed, and the longest observations may include browser inactivity. The primary specification winsorizes duration at the advice-phase 99th percentile and reports both seconds and the log. The conclusions are unchanged with untrimmed log duration and alternative cutoffs.

Table EC.12 Detailed Study 2 focal contrasts.

Outcome	Comparison or sample	Estimate [95% CI]	Raw p
Advice-phase normalized performance	Broad - precise	-0.057 [-0.076, -0.038]	< 0.001
Advice-phase normalized performance	Specific-broad - precise	-0.058 [-0.076, -0.041]	< 0.001
Advice-phase normalized performance	Specific-broad - broad	-0.001 [-0.023, 0.020]	0.902
Action agreement	Broad - precise	-0.176 [-0.217, -0.135]	< 0.001
Action agreement	Specific-broad - precise	-0.205 [-0.249, -0.161]	< 0.001
Action agreement	Specific-broad - broad	-0.029 [-0.073, 0.015]	0.192
Familiar-structure pre-to-post change	Broad - precise	0.077 [0.015, 0.138]	0.016
Familiar-structure pre-to-post change	Specific-broad - precise	0.081 [0.014, 0.147]	0.018
Familiar-structure pre-to-post change	Specific-broad - broad	0.004 [-0.063, 0.071]	0.903
Familiar-structure pre-to-post change	Post - pre	0.052 [0.033, 0.071]	< 0.001
Familiar-structure pre-to-post change	Post - pre	0.033 [0.000, 0.066]	0.047
Familiar-structure pre-to-post change	Post - pre	0.065 [0.036, 0.094]	< 0.001
Familiar-structure pre-to-post change	Post - pre	0.058 [0.020, 0.096]	0.004
First post-removal change	Broad - precise	0.117 [0.043, 0.191]	0.002
First post-removal change	Specific-broad - precise	0.118 [0.054, 0.183]	< 0.001
First post-removal change	Specific-broad - broad	0.002 [-0.054, 0.057]	0.951
Unseen-sequence post-advice performance, ANCOVA	Broad - precise	-0.049 [-0.094, -0.004]	0.032
Unseen-sequence post-advice performance, ANCOVA	Specific-broad - precise	-0.029 [-0.079, 0.021]	0.254
Unseen-sequence post-advice performance, ANCOVA	Specific-broad - broad	0.020 [-0.031, 0.071]	0.439
Familiar-minus-unseen advice-format effect, ANCOVA	Broad - precise	0.143 [0.078, 0.208]	< 0.001
Familiar-minus-unseen advice-format effect, ANCOVA	Specific-broad - precise	0.103 [0.031, 0.175]	0.005
Familiar-minus-unseen advice-format effect, ANCOVA	Specific-broad - broad	-0.040 [-0.106, 0.025]	0.226
Familiar pre-to-post trajectory	All familiar participants	0.052 [0.033, 0.071]	< 0.001
Familiar pre-to-post trajectory	precise, familiar	0.033 [0.000, 0.066]	0.047
Familiar pre-to-post trajectory	broad, familiar	0.065 [0.036, 0.094]	< 0.001
Familiar pre-to-post trajectory	specific-broad, familiar	0.058 [0.020, 0.096]	0.004

Note. Familiar-sequence advice-format contrasts use the no-rationale cells. The unseen-sequence outcome is the post-advice level adjusted for pre-advice short-map performance. The interaction is positive when an advice-format effect is more favorable in the familiar than in the unseen sequence. Appendix C reports the adjusted p-values used in the main text.

Slider activity measures search over numerical actions. Additional slider measures are the number of distinct values inspected, the search range, and direction reversals. Information acquisition is measured by future-traffic reveal clicks and an indicator for looking ahead before acting.

The active-effort index gives equal weight to log decision time, slider search, and future-information acquisition. Within the slider domain, active movements, distinct values, search range, and direction reversals receive equal weight. Within the information domain, reveal clicks and the look-ahead indicator receive equal weight. Decision time and interface activity are proxies for processing effort; they do not distinguish productive deliberation from confusion.

Table EC.13 reports the advice-phase pairwise comparisons. Broad and specific-broad advice both increase decision time, active interaction, distance from the precise target, and state-action support relative to precise advice. The qualitative formats are similar on decision time, the composite effort index, and state-action support. Broad advice produces somewhat more slider movement than specific-broad advice.

Eventual advice assignment does not predict pre-advice timing, slider, or look-ahead measures in the corresponding placebo specifications. This reduces concern that the advice-phase process differences reflect baseline imbalance. It does not establish that the additional time or interaction produces learning.

E.3. Transfer specifications and environmental comparisons

We characterize post-advice environments through observable task features rather than assigning each condition an assumed value of the model's novelty parameter. Table EC.14 records relevant characteristics of each E_2 environment that contribute to novelty.

Table EC.13 Study 2 advice-phase process contrasts.

Measure	Comparison	Estimate [95% CI]	p-value
Decision time (seconds)	Broad − precise	11.224 [8.421, 14.026]	< 0.001
Decision time (seconds)	Specific broad − precise	10.913 [8.279, 13.547]	< 0.001
Decision time (seconds)	Specific broad − broad	-0.311 [-3.210, 2.588]	0.834
Active-effort index (SD)	Broad − precise	0.317 [0.235, 0.400]	< 0.001
Active-effort index (SD)	Specific broad − precise	0.315 [0.239, 0.390]	< 0.001
Active-effort index (SD)	Specific broad − broad	-0.003 [-0.080, 0.074]	0.943
Active slider movements	Broad − precise	1.478 [1.018, 1.937]	< 0.001
Active slider movements	Specific broad − precise	0.949 [0.538, 1.359]	< 0.001
Active slider movements	Specific broad − broad	-0.529 [-1.027, -0.032]	0.037
Looked ahead	Broad − precise	0.095 [0.048, 0.143]	< 0.001
Looked ahead	Specific broad − precise	0.117 [0.070, 0.163]	< 0.001
Looked ahead	Specific broad − broad	0.021 [-0.024, 0.066]	0.361
Distance from precise target	Broad − precise	6.019 [4.612, 7.426]	< 0.001
Distance from precise target	Specific broad − precise	7.498 [5.928, 9.068]	< 0.001
Distance from precise target	Specific broad − broad	1.480 [-0.105, 3.064]	0.067
State-action support	Broad − precise	1.184 [0.851, 1.517]	< 0.001
State-action support	Specific broad − precise	1.171 [0.817, 1.525]	< 0.001
State-action support	Specific broad − broad	-0.013 [-0.365, 0.338]	0.941

Note. Decision-level estimates include round, exit, current charge, precise policy target, traffic-range width, recommendation type, rationale, and sequence controls, with participant-clustered standard errors. State-action support counts distinct tuples of stop, binned current charge, and binned added charge using 10-percentage-point bins.

Table EC.14 Task features of the post-advice comparisons.

Comparison	Prior exposure and structure	Traffic change	Realizations outside advice support	Interpretation
S1 R6	Same short-map structure as baseline and advice rounds	Same range displayed	0 of 5	Matched return to the advised structure
S2 familiar R5	Same short map, distances, and displayed ranges as advice rounds	Same range displayed	0 of 5	Immediate removal under the same displayed structure
S2 familiar R6	Same topology and distances	Different range displayed, same optimal batch/split arrangement	1 of 5	Related environment with changed uncertainty
S2 familiar R7	Same ex ante map and ranges as Round 6	Different range displayed, different optimal batch/split arrangement	4 of 5	Distributional surprise within the same route structure
S1 R7	Long map seen before advice; advice was shown only on the short map	Not comparable to advice map	Not comparable	Return to a previously seen but unadvised structure
S2 unseen R5–7	New route, state sequence, and charging opportunities	Not comparable to advice map	Not comparable	Unseen structure; only general charging logic remains portable

Note. The categories are multidimensional task comparisons, not estimates of a continuous structural parameter. The Rounds 5–6 average combines two related but nonidentical environments and is reported as a precision summary rather than as a distinct novelty level.

For the familiar sequence, we estimate Round 5, Round 6, their average, and Round 7 separately. The direct equality tests in Table EC.15 compare participant-level differences across rounds.

Table EC.16 reports ANCOVA estimates both across rationale assignments and within the rationale-hidden cells.

E.4. Rationale effects

Rationale visibility is randomized. Within each advice-format-by-sequence cell, the total-effect specification regresses mean post-advice performance on rationale assignment and pre-advice performance. The conditional specification additionally controls for advice-phase performance and action agreement. The latter controls are affected by rationale

Table EC.15 Direct tests of advice-format effect differences across familiar rounds.

Sample	Effect comparison	Format contrast	Difference	p-value
All rationale assignments	Round 6 − Round 5	Broad − precise	0.016	0.565
All rationale assignments	Round 6 − Round 5	Specific broad − precise	−0.026	0.335
All rationale assignments	Round 7 − mean Rounds 5–6	Broad − precise	0.038	0.281
All rationale assignments	Round 7 − mean Rounds 5–6	Specific broad − precise	−0.008	0.843
Rationale hidden	Round 6 − Round 5	Broad − precise	0.056	0.123
Rationale hidden	Round 6 − Round 5	Specific broad − precise	0.014	0.702
Rationale hidden	Round 7 − mean Rounds 5–6	Broad − precise	0.060	0.280
Rationale hidden	Round 7 − mean Rounds 5–6	Specific broad − precise	−0.037	0.544

Note. Each row is the change in the indicated advice-format contrast across rounds, estimated from participant-level outcome differences with the same baseline adjustment used in the round-specific ANCOVA specifications. The joint tests of the broad and specific-broad changes have $p = 0.219$ for Round 6 versus Round 5 and $p = 0.304$ for Round 7 versus the Rounds 5–6 mean across rationale assignments. The corresponding rationale-hidden joint p-values are 0.202 and 0.116.

Table EC.16 Study 2 round-specific advice-format estimates.

Sequence and outcome	Comparison	All rationale assignments	Rationale hidden
Familiar Round 5	Broad − precise	0.019 [−0.022, 0.059], $p = 0.361$	0.049 [−0.000, 0.098], $p = 0.052$
Familiar Round 5	Specific broad − precise	0.046 [0.007, 0.085], $p = 0.022$	0.080 [0.029, 0.130], $p = 0.002$
Familiar Round 6	Broad − precise	0.035 [−0.010, 0.079], $p = 0.126$	0.104 [0.039, 0.170], $p = 0.002$
Familiar Round 6	Specific broad − precise	0.020 [−0.026, 0.066], $p = 0.396$	0.094 [0.025, 0.163], $p = 0.008$
Familiar mean, Rounds 5–6	Broad − precise	0.027 [−0.006, 0.060], $p = 0.112$	0.077 [0.031, 0.122], $p = 0.001$
Familiar mean, Rounds 5–6	Specific broad − precise	0.033 [−0.001, 0.066], $p = 0.055$	0.087 [0.038, 0.135], $p < 0.001$
Familiar Round 7	Broad − precise	0.065 [−0.007, 0.137], $p = 0.079$	0.137 [0.035, 0.238], $p = 0.008$
Familiar Round 7	Specific broad − precise	0.025 [−0.051, 0.101], $p = 0.517$	0.050 [−0.065, 0.164], $p = 0.395$
Unseen mean, Rounds 5–7	Broad − precise	-0.012 [−0.047, 0.022], $p = 0.488$	-0.049 [−0.094, −0.004], $p = 0.032$
Unseen mean, Rounds 5–7	Specific broad − precise	-0.011 [−0.047, 0.025], $p = 0.559$	-0.029 [−0.079, 0.021], $p = 0.254$

Note. Each estimate is a post-advice level adjusted for pre-advice short-map performance. All-rationale models include rationale assignment. Rationale-hidden estimates use randomized cells in which no rationale was shown. Raw p-values are displayed; Appendix C reports multiplicity adjustments.

assignment, so the conditional estimates are descriptive mechanism comparisons rather than causal mediation effects. Table EC.17 reports both specifications and direct comparisons of the rationale effects across advice formats.

Appendix F: Inverse Reinforcement Learning (IRL) approach in detail

F.1. Deriving the reward components

At the end of our pilot study, we asked 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?"

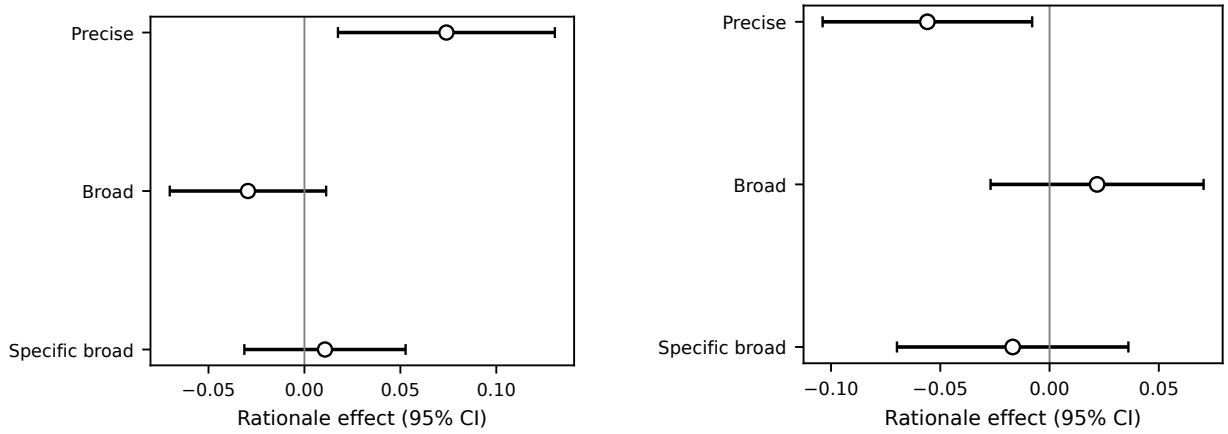
We make answering these questions obligatory, and most answers are comprehensive with 94 (resp. 64 – 155) words at the median (resp. IQR). Each answer forms a "document" ($n = 122$, we also include participants' answers if they were excluded from the main analysis). We then use BERTopic (Grootendorst 2022) 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.18.

Table EC.17 Study 2 rationale effects on post-advice performance.

Specification	Sequence	Advice format or contrast	Estimate [95% CI]	p-value
Total effect, ANCOVA	Familiar	Precise	0.074 [0.017, 0.131]	0.010
Total effect, ANCOVA	Familiar	Broad	-0.029 [-0.070, 0.011]	0.157
Total effect, ANCOVA	Familiar	Specific-broad	0.011 [-0.031, 0.053]	0.618
Total effect, ANCOVA	Unseen	Precise	-0.056 [-0.104, -0.008]	0.022
Total effect, ANCOVA	Unseen	Broad	0.022 [-0.027, 0.071]	0.382
Total effect, ANCOVA	Unseen	Specific-broad	-0.017 [-0.070, 0.036]	0.532
Conditional comparison	Familiar	Precise	0.063 [0.007, 0.118]	0.026
Conditional comparison	Familiar	Broad	-0.031 [-0.072, 0.011]	0.146
Conditional comparison	Familiar	Specific-broad	-0.000 [-0.035, 0.034]	0.981
Conditional comparison	Unseen	Precise	-0.052 [-0.101, -0.003]	0.039
Conditional comparison	Unseen	Broad	0.014 [-0.033, 0.062]	0.554
Conditional comparison	Unseen	Specific-broad	-0.018 [-0.059, 0.024]	0.405
Difference in effects	Familiar	Precise - Broad	0.108 [0.038, 0.179]	0.003
Difference in effects	Familiar	Precise - Specific-broad	0.089 [0.014, 0.164]	0.020
Difference in effects	Familiar	Specific-broad - Broad	0.019 [-0.044, 0.082]	0.550
Difference in effects	Unseen	Precise - Broad	-0.077 [-0.146, -0.009]	0.027
Difference in effects	Unseen	Precise - Specific-broad	-0.039 [-0.111, 0.033]	0.284
Difference in effects	Unseen	Specific-broad - Broad	-0.038 [-0.109, 0.033]	0.293

Note. Total-effect models regress post-advice performance on rationale assignment and pre-advice performance within each advice-format-by-sequence cell. Conditional models additionally control for advice-phase performance and action agreement. Because these are post-treatment variables, the conditional estimates are descriptive mechanism comparisons rather than causal mediation effects.

Figure EC.3 Total rationale effects on post-advice performance.



(a) Familiar sequence

(b) Unseen sequence

Note. Within each advice-format-by-sequence cell, estimates control for pre-advice performance. Whiskers are 95% heteroskedasticity-robust confidence intervals.

F.2. 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).

Table EC.18 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.

MaxCausalEnt can be formulated as a constrained optimization problem,

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

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 $\boldsymbol{f} = (f_1,\dots,f_m)$.

Introducing Lagrange multipliers $\boldsymbol{\theta} = (\theta_1,\dots,\theta_m)$ 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

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

F.2.1. 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:

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

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 $\boldsymbol{z}_i \in \mathbb{R}^d$ 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: $\boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$ 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 $\boldsymbol{z}_i$ 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.

F.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 $\mathcal{L}(\psi) \;=\; \mathbb{E}_{q_\psi}\big[\log p(Z, \mathcal{D})\big] - \mathbb{E}_{q_\psi}\big[\log q_\psi(Z)\big]$ is the *evidence lower bound* (ELBO). This can be re-expressed in terms of the Kullback–Leibler (KL) divergence. Using $p(Z \mid \mathcal{D}) = \frac{p(Z, \mathcal{D})}{p(\mathcal{D})}$,

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

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

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

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

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

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.

F.4. 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?

F.4.1. Synthetic data: Identification. The first validation exercise evaluates whether the approach recovers model parameters when misspecification is not a concern. We create a synthetic data set according to the assumed hierarchical model. To generate data comparable to 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 F.2.1. For each 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.

We then estimate the posterior distributions of model parameters. We then correlate each "true" parameter with the median of the corresponding posterior distribution. Table EC.19 provides an overview of the results. The model accurately recovers 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.

F.4.2. Real data: Log-likelihoods and posterior predictive checks. We then run the 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.

Table EC.19 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 advice scenarios, that is, scenarios in which algorithmic advice is available. Note that $\theta^j_{sc} = \theta^j_0 + \Delta^j_{sc}$.

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, the same statistic on the test set in the synthetic experiment (Appendix F.4.1) 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.

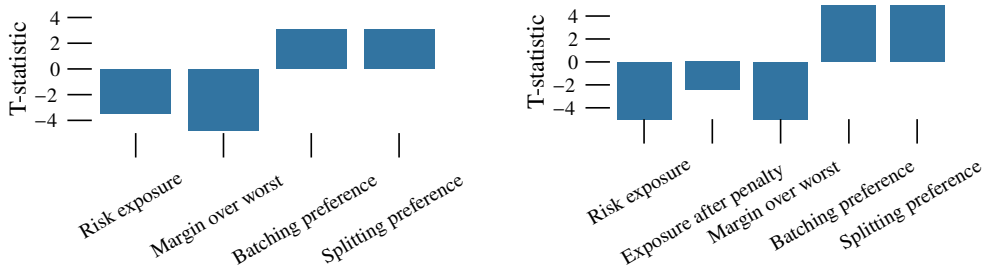
We also conduct a posterior predictive check (PPC), a Bayesian counterpart to a goodness-of-fit test (Gelman et al. 1996). For each observation, the model returns a vector of predicted choice probabilities over the set of feasible actions. Concatenating those, we obtain a "probability matrix," with observations 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 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 compute the Brier score for the empirical data (36.60). We then draw 300 replicated datasets by generating model parameters from the posterior distribution, then using them to simulate actions. 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 ($p \approx 0.93$).

F.4.3. Real data: Consistency with qualitative evidence. We also compare the IRL results with qualitative observations from sequence clustering (see Figure EC.2 for the Study 1 results). 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 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.4 shows the t-statistic when comparing $(\Delta^j_i)_{i \in N^*_{sc}}$ and $(\Delta^j_i)_{i \notin N^*}$ for a given component j, where N^* 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.4a) focus more on batching and splitting decisions, and are less concerned with risk exposure and simplicity. This is consistent 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.4b).

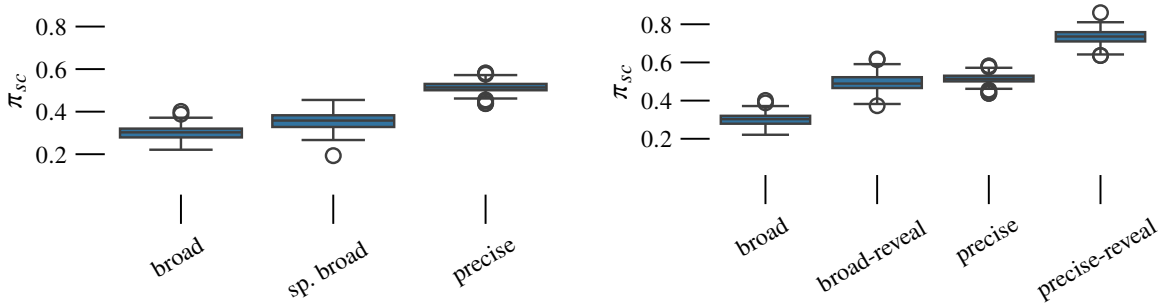
Figure EC.4 Comparison of Δ_i^j across clusters (t-test).



(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 ($p < 0.05$).

Figure EC.5 Estimated compliance probability by advice format and rationale visibility.



(a) Advice formats without rationale (b) Advice formats with and without rationale

We also run a t-test on the compliance probability in the 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, the model estimates are consistent with the qualitative evidence on participants' behavior, supporting the validity of the IRL approach.

F.5. Compliance estimates by advice format

Figure EC.5 reports the posterior distribution of compliance probabilities estimated from the model in §6.2. The estimates separate the probability of executing the displayed optimal action from the fallback heuristic.

Precise advice generates the highest compliance among the no-rationale formats. Specific-broad advice increases compliance only modestly relative to broad advice, whereas rationales produce a larger shift. This pattern is consistent with the interpretation that action-level implementability and perceived justification are distinct channels.

References for the Appendices

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* URL https://arxiv.org/abs/2203.11409.

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

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* URL https://arxiv.org/abs/2203.05794.

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

Le Breton M (2006) Notes on inequality measurement: Hardy, Littlewood and Polya, Schur convexity and majorization. URL https://dse.univr.it/it/documents/it2/le_breton_it2_notes.pdf.

Phan D, Pradhan N, Jankowiak M (2019) Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv* URL https://arxiv.org/abs/1912.11554.

Ramachandran D, Amir E (2007) Bayesian inverse reinforcement learning. *International Joint Conference on Artificial Intelligence*, 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.