SAFE PREFERENCE IDENTIFICATION THROUGH STRATEGIC INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a partner's preferences through strategic interaction can fail for two distinct reasons: safety constraints may remove the interactions that distinguish preferences, while a misspecified behavioral model may interpret informative responses in the wrong direction. We formalize these bottlenecks through an available-information rate Γ_i and a learned-evidence rate $\widetilde{\Gamma}_i$. Their separation exposes a central failure regime, $\Gamma_i > 0$ but $\widetilde{\Gamma}_i \leq 0$, in which the admissible environment contains identifying information but the learner cannot convert it into correctly directed evidence.

We characterize constraint-induced non-identifiability, derive information-theoretic sample-complexity lower bounds governed by Γ_i, and show that signed learned evidence determines asymptotic posterior drift under likelihood misspecification. Experiments in an analytic instance, an exactly enumerable strategic game, and continuous HighwayEnv merging show that future commitments can change current responses, trajectory-level evidence can improve identification over one-step evidence, and natural likelihood mismatch can reverse learned evidence. Safe strategic preference identification therefore requires both informative admissible interactions and a behavioral model that interprets their responses faithfully.

1 INTRODUCTION

Consider an autonomous vehicle trying to infer whether a merging driver is cautious or assertive. Yielding may be safe but uninformative: both driver types may accept the gap. A more assertive commitment may reveal the driver's preference, but at the cost of collision risk or short-term task loss. Even when such an interaction is admissible, a misspecified behavioral model can interpret the resulting response in the wrong direction. Strategic preference identification therefore faces two distinct bottlenecks: the interaction may fail to reveal the preference, or the learner may misinterpret the information that it reveals.

The first bottleneck is *available information*. Safety and task constraints restrict the strategic probes the learner may attempt. If every admissible probe induces the same response law under two preferences, then no admissible identification procedure can distinguish them (Figure 1). More generally, we quantify the discrimination remaining under the admissible interaction family by an *available-information rate* Γ_i.

The second bottleneck is *learned evidence*. Even when admissible probes generate preference-dependent behavior, inference may rely on a misspecified behavioral likelihood. An informative response can then favor the wrong preference under the learner's model. We quantify the direction of this learned evidence by a distinct rate $\widetilde{\Gamma}_i$. The central separation is

$$\Gamma_i > 0 \qquad \text{while} \qquad \widetilde{\Gamma}_i \leq 0.$$

In this regime, some admissible randomized design separates the true preference from every alternative in the true response laws, yet no stationary admissible design yields strictly positive expected learned evidence against every alternative. Thus information can remain available without becoming usable evidence: admissibility determines what the environment can reveal, while likelihood fidelity determines what the learner actually learns.

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

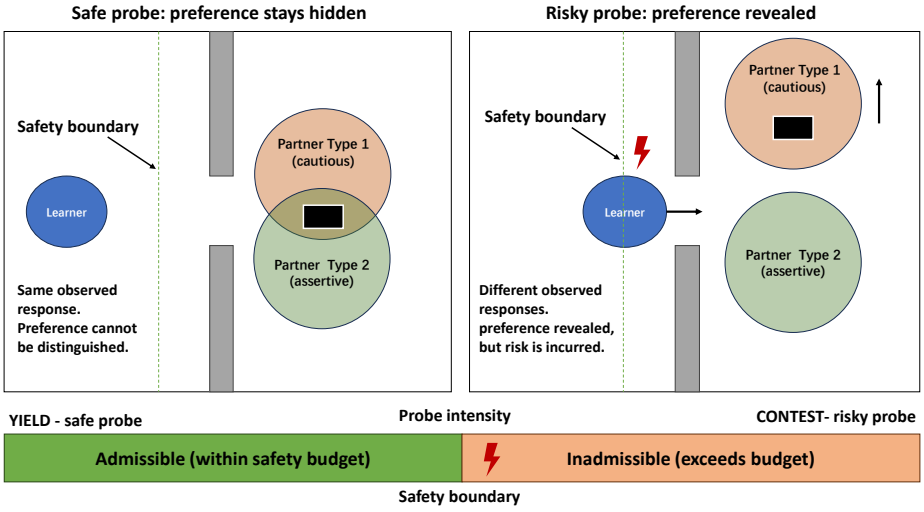


Figure 1: **Safety can remove the informative experiment.** A safe yielding probe leaves two preferences observationally equivalent, while a more confrontational probe separates them but may violate the safety budget.

Our setting lies at the intersection of inverse and preference learning, active information acquisition, and strategic interaction Ziebart et al. (2008); Christiano et al. (2017); Sadigh et al. (2017); Brown et al. (2018); Hadfield-Menell et al. (2016); Büning et al. (2022); Goktas et al. (2024). Existing preference-learning methods typically treat information acquisition as the choice of an informative query or exploration policy. Here, the probe is itself a strategic intervention: it changes the partner's decision problem and must satisfy safety and task constraints before the partner's preference is known. At the same time, the learner may reason with a misspecified model of the partner's response. Our setting therefore couples strategic intervention, robust admissibility, and inference under behavioral-model misspecification.

Our main contributions are:

- **Safe identifiability under constrained interaction.** We formalize robustly admissible strategic probing, characterize constraint-dependent preference indistinguishability, and derive an information-theoretic sample-complexity lower bound governed by the admissible discrimination rate Γ_i.
- **A separation between available information and learned evidence.** We define signed learned evidence under likelihood misspecification, characterize its asymptotic effect on posterior odds, and identify the regime $\Gamma_i > 0$ but $\widetilde{\Gamma}_i \leq 0$, where identifying information remains available but no stationary admissible design accumulates correctly directed evidence against every alternative.
- **Validation across strategic models.** We validate the theory in a one-dimensional analytic instance, an exactly enumerable repeated game, and continuous HighwayEnv merging. The experiments show that future commitments can alter current responses, full trajectories can retain more identifying information than one-step observations, and natural behavioral-model mismatch can reverse learned evidence.

2 Related Work

Preference learning and active acquisition. IRL infers latent objectives from behavior Ziebart et al. (2008), while preference-based RL and RLHF learn from comparisons or related feedback Christiano et al. (2017); Lee et al. (2021); Zhu et al. (2023); Rafailov et al. (2023). Active preference and inverse-learning methods select informative queries or exploration policies Sadigh et al. (2017); Biyik & Sadigh (2018); Brown et al. (2018); Lindner et al. (2022). In these settings, in-

formation acquisition is typically treated as a query or exploration choice. Here, a probe is itself a strategic intervention that changes the partner's decision problem and must satisfy robust risk and task constraints before the preference is known.

Strategic interaction and model misspecification. Interactive IRL, CIRL, and multi-agent inverse learning make behavior and information acquisition endogenous Hadfield-Menell et al. (2016); Büning et al. (2022); Yu et al. (2019); Goktas et al. (2024). Separately, inverse and preference learning can fail under reward misidentification or behavioral-model misspecification Tien et al. (2022); Skalse & Abate (2023). Our focus is the interaction between these two issues: even when a strategic experiment remains admissible and informative in the true environment, a misspecified response model can turn that information into evidence for the wrong preference.

Sequential experimental design. Our lower bounds build on change-of-measure arguments from sequential design and controlled sensing Chernoff (1959); Nitinawarat et al. (2013), while posterior behavior under model misspecification connects to classical KL-projection results Berk (1966). The new structural distinction is between the information made available by the admissible experiment family, summarized by Γ_i, and the direction of evidence produced by the learner's likelihood, summarized by $\widetilde{\Gamma}_i$.

3 Theory of Safe Preference Identification

We consider a finite set of candidate preferences $\Theta = \{\theta_1, \dots, \theta_M\}$, where M is the number of candidates and $\theta^\star \in \Theta$ denotes the unknown true preference. At each round t, the learner selects a *strategic probe* $u_t \in \mathcal{U}$, where $\mathcal{U}$ is the candidate probe set. A probe is a short commitment that changes the partner's decision problem, rather than merely requesting an observation.

If $\theta^\star = \theta_i$, probe u_t induces a response $Y_t \sim P_i^{u_t}$, where P_i^u denotes the response distribution under preference θ_i and probe u. After observing Y_t, the learner may update its belief and choose subsequent probes adaptively. Each probe also affects task performance and interaction safety: $J_i(u)$ denotes the learner's expected task value under preference θ_i, while L_i^u denotes the corresponding random interaction loss. Thus, a probe is simultaneously an experiment for preference identification and a task action with safety consequences.

Two bottlenecks organize the theory. First, safety and task constraints restrict which probes are admissible and therefore how much preference information remains available. Second, even an informative admissible probe may yield misleading learned evidence when the behavioral likelihood used for inference is misspecified. We formalize these two effects below through an available-information rate and a learned-evidence rate. In the highway setting, a probe is a four-step speed commitment and the response is the merging vehicle's resulting action sequence.

3.1 Admissible Strategic Probes

Because the true preference is unknown when a probe is selected, admissibility cannot depend on knowing which candidate preference is correct. Let $[M] = \{1, \dots, M\}$. For each preference θ_i, define the reference task value $J_i^{\mathrm{ref}} := \max_{u' \in \mathcal{U}} J_i(u')$, namely, the best learner task value achievable within the same candidate probe set when only task performance is optimized. The task degradation caused by using probe u is therefore $[J_i^{\mathrm{ref}} - J_i(u)]_+$, where $[x]_+ := \max\{x, 0\}$.

For an interaction-risk budget $c \geq 0$ and a task-degradation budget $\epsilon \geq 0$, we define the robustly admissible probe set as

$$\mathcal{U}(c, \epsilon) = \left\{ u \in \mathcal{U} : \max_{i \in [M]} \mathrm{CVaR}_\alpha(L_i^u) \leq c,\ \max_{i \in [M]} [J_i^{\mathrm{ref}} - J_i(u)]_+ \leq \epsilon \right\}. \tag{1}$$

Here CVaR_α is the conditional value at risk of a loss at confidence level $\alpha \in (0, 1)$, measuring its upper-tail severity. For any integrable loss random variable L, $\mathrm{CVaR}_\alpha(L) := \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\alpha} \mathbb{E}[(L - \eta)_+] \right\}$, where η is an auxiliary threshold and the expectation is taken under the distribution of L.

The maxima in equation 1 enforce *robust* admissibility: before the preference is identified, a probe must satisfy both budgets for every candidate preference, rather than only on average under the learner's current belief. The learner may then adaptively select probes from $\mathcal{U}(c, \epsilon)$.

3.2 Safe Identifiability and Information Limits

Once the constraints determine the admissible probe family, we ask whether the remaining interactions can distinguish candidate preferences.

Safe preference equivalence. For fixed budgets (c, ϵ), define

$$\theta_i \sim_{c,\epsilon} \theta_j \quad \Longleftrightarrow \quad P_i^u = P_j^u \text{ for all } u \in \mathcal{U}(c, \epsilon). \tag{2}$$

Thus θ_i and θ_j are safely indistinguishable if every admissible probe induces the same response distribution under the two preferences. This ambiguity is constraint-induced when the pair can be distinguished by some probe in $\mathcal{U}$ but not by any probe in $\mathcal{U}(c, \epsilon)$. Tightening either budget can only shrink the admissible set and hence merge safe equivalence classes; the proof is given in Appendix A.1.

To quantify how strongly an admissible probe separates a pair, let

$$D_{ij}^{\star}(c, \epsilon) := \sup_{u \in \mathcal{U}(c,\epsilon)} D_{\mathrm{KL}}(P_i^u \| P_j^u),$$

where $D_{\mathrm{KL}}(P\|Q) := \mathbb{E}_{Y\sim P}[\log(p(Y)/q(Y))]$ denotes Kullback–Leibler divergence whenever P is absolutely continuous with respect to Q. Then $\theta_i \sim_{c,\epsilon} \theta_j$ if and only if $D_{ij}^{\star}(c, \epsilon) = 0$.

For multiple preferences, however, the probe that best separates one alternative need not separate another. We therefore define the *available-information rate*

$$\Gamma_i(c, \epsilon) := \sup_{q \in \Delta(\mathcal{U}(c,\epsilon))} \min_{j \in [M]\setminus\{i\}} \mathbb{E}_{u\sim q}\left[D_{\mathrm{KL}}(P_i^u \| P_j^u)\right], \tag{3}$$

where $\Delta(\mathcal{U}(c, \epsilon))$ is the set of probability distributions over admissible probes. Hence Γ_i is the largest worst-alternative KL rate achievable by a stationary randomized admissible probe design.

Theorem 1 (Information lower bound)**.** *Assume the admissible probe family is finite. Any adaptive identification procedure that uses only probes in $\mathcal{U}(c, \epsilon)$ and returns the correct preference with probability at least $1 - \delta$ under every $\theta_i \in \Theta$, for $\delta < 1/2$, satisfies*

$$\mathbb{E}_i[N] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\Gamma_i(c, \epsilon)} = \Omega\left(\frac{\log(1/\delta)}{\Gamma_i(c, \epsilon)}\right), \tag{4}$$

where N is the stopping time, $\mathbb{E}_i$ denotes expectation when $\theta^{\star} = \theta_i$, and $\mathrm{kl}(a, b) := a \log(a/b) + (1-a)\log((1-a)/(1-b))$ is binary KL divergence.

Thus $\Gamma_i(c, \epsilon) = 0$ means that at least one alternative cannot be reliably separated using admissible probes, whereas small positive $\Gamma_i(c, \epsilon)$ implies a large information-theoretic sample requirement. The pairwise specialization replaces Γ_i by $D_{ij}^{\star}$; proofs of both statements are given in Appendices A.2 and A.3.

The corresponding identifiability region is $\mathcal{B}_i := \{(c, \epsilon) : \Gamma_i(c, \epsilon) > 0\}$. Equivalent critical-budget characterizations are deferred to Appendix A.4.

3.3 Likelihood Fidelity under Misspecification

The available-information rate Γ_i is defined entirely by the *true* response laws. Identification can nevertheless fail when the learner performs inference with a misspecified behavioral model. We write $\widehat{P}_i^u$ for the response distribution assumed by the learner under preference θ_i and probe u. Let p_i^u and $\widehat{p}_i^u$ denote the corresponding probability mass functions or densities with respect to a common dominating measure.

For a true preference θ_i and an alternative θ_j, define the expected learned log-evidence from one probe as

$$\Delta_{ij}(u; \widehat{P}) := \mathbb{E}_{Y\sim P_i^u}\left[\log \frac{\widehat{p}_i^u(Y)}{\widehat{p}_j^u(Y)}\right] = D_{\mathrm{KL}}(P_i^u \| \widehat{P}_j^u) - D_{\mathrm{KL}}(P_i^u \| \widehat{P}_i^u). \tag{5}$$

We assume the displayed likelihood ratios are well defined and integrable. Unlike true KL divergence, Δ_{ij} is signed: positive values favor the true preference θ_i, while negative values favor the alternative θ_j. Thus true KL measures whether preferences generate distinguishable behavior, whereas Δ_{ij} measures whether the learner interprets that behavior in the correct direction.

Mirroring Γ_i, define the *learned-evidence rate*

$$\widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) := \sup_{q \in \Delta(\mathcal{U}(c,\epsilon))} \min_{j \in [M]\setminus\{i\}} \mathbb{E}_{u\sim q}\Big[\Delta_{ij}(u; \widehat{P})\Big]. \tag{6}$$

Hence Γ_i measures the discrimination available in the true environment, whereas $\widetilde{\Gamma}_i$ measures the worst-alternative direction of evidence produced by the learner's likelihood.

Theorem 2 (Posterior drift under misspecification)**.** *Suppose $\theta^\star = \theta_i$, the learner starts from a full-support prior $\widehat{b}_0$ with $\widehat{b}_0(k) > 0$ for every $k \in [M]$, and updates its posterior using $\widehat{P}$. Under a stationary randomized design $u_t \overset{\text{i.i.d.}}{\sim} q$, conditionally independent responses $Y_t \sim P_i^{u_t}$, and integrable log-likelihood ratios,*

$$\frac{1}{T}\log\frac{\widehat{b}_T(i)}{\widehat{b}_T(j)} \xrightarrow{\text{a.s.}} \mathbb{E}_{u\sim q}\Big[\Delta_{ij}(u; \widehat{P})\Big], \tag{7}$$

where $\widehat{b}_T(k)$ is the learner's posterior probability of θ_k after T rounds.

Thus positive learned evidence against every alternative implies posterior concentration on θ_i under that stationary design, whereas a negative rate against some θ_j asymptotically favors that alternative. More generally, under adaptive probing the conditional expected increment in log-posterior odds at round t remains $\Delta_{ij}(u_t; \widehat{P})$. If the likelihood is exact, $\widehat{P}_k^u = P_k^u$ for all k and u, then

$$\Delta_{ij}(u; \widehat{P}) = D_{\mathrm{KL}}(P_i^u \| P_j^u), \qquad \widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) = \Gamma_i(c, \epsilon).$$

Corollary 1 (Information without usable evidence)**.** *If*

$$\Gamma_i(c, \epsilon) > 0 \qquad \textit{but} \qquad \widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) \le 0,$$

then the admissible probe family contains enough true information to separate θ_i from every alternative, but no stationary randomized admissible design yields strictly positive expected learned log-evidence against every alternative.

This regime cannot occur under an exact likelihood and isolates the second failure mode studied in this paper: information may remain available even when the learner cannot turn it into correctly directed evidence. Quantitative likelihood-fidelity bounds, safety-model robustness bounds, and the continuous-preference analogue are deferred to Appendices A.5 and A.6.

Operational policy. For experiments, we instantiate the theory with Safe-ID, a simple confidence-targeted probing heuristic. Before the learner's posterior reaches confidence $1 - \delta$, Safe-ID selects the admissible probe with highest Bayesian information gain under its current behavioral model; afterward, it switches to the admissible probe with highest posterior-expected task value. Under likelihood mismatch, both posterior updating and information-gain selection use $\widehat{P}$, while the controlled experiments keep the true environment and admissible set fixed. Full specification and baselines are given in Appendix A.7.

4 Analytic Instance and Theory Validation

We make the first bottleneck concrete in a one-dimensional model that isolates three consequences of constrained interaction: admissibility can destroy identifiability, preserve identifiability while making it statistically expensive, and make the most informative probe less aggressive than the largest admissible one.

4.1 Model and Constraint Effects

The learner chooses an interaction intensity $x \in [0, 1]$, and the partner produces a binary response $Y \in \{0, 1\}$. Under scalar preference θ, let $p_\theta(x) := \Pr(Y = 1 \mid \theta, x) = \sigma\big(\beta\theta g(x)\big)$, where

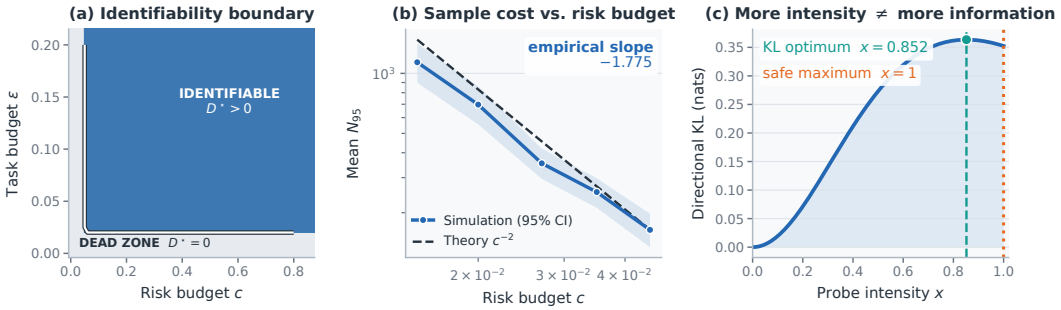


Figure 2: **Constraint geometry controls identifiability and cost.** (a) Tight budgets can confine admissible probes to an information dead zone. (b) Identification cost follows the predicted inverse-power scaling. (c) Pairwise KL can peak before the largest admissible intensity.

$\sigma(z) = (1+e^{-z})^{-1}$ is the logistic function, $\beta > 0$ controls response rationality, and $g(x)$ determines how strongly the probe affects the partner's response. We assume $g(0) = 0$, so $p_\theta(0) = 1/2$ for every preference; whenever $g(x) \neq 0$, distinct preferences induce distinct response probabilities.

We specialize the two constraints from Section 3.1 to this model. If $Y = 1$ triggers a binary adverse-event loss $L \in \{0, 1\}$ with conditional probability $h(x)$, then $\Pr(L = 1 \mid \theta_i, x) = h(x)p_{\theta_i}(x)$. Define the robust risk and task degradation as

$$R(x) := \max_{i \in [M]} \mathrm{CVaR}_\alpha(L \mid \theta_i, x), \qquad C_T(x) := \max_{i \in [M]} [J_i^{\mathrm{ref}} - J_i(x)]_+.$$

The admissible intensity set is therefore

$$\mathcal{X}(c, \epsilon) := \{x \in [0, 1] : R(x) \leq c,\ C_T(x) \leq \epsilon\}.$$

For the clean analytic instance, we choose $J_i^{\mathrm{ref}} - J_i(x) = t_0 x^n$ for every i, with $t_0, n > 0$, so that $C_T(x) = t_0 x^n$. For preferences θ_i and θ_j, let $d_{ij}(x) := D_{\mathrm{KL}}(\mathrm{Bern}(p_{\theta_i}(x)) \| \mathrm{Bern}(p_{\theta_j}(x)))$. Then $D_{ij}^\star(c, \epsilon) = \sup_{x \in \mathcal{X}(c,\epsilon)} d_{ij}(x)$, the pairwise safe separation from Section 3.2. Full model and risk derivations are given in Appendix B.

Constraints can destroy identifiability. Suppose $g(x) = 0$ for $x \leq x_{\mathrm{id}}$, where $x_{\mathrm{id}} \in [0, 1]$ is an information dead-zone threshold. If $\mathcal{X}(c, \epsilon) \subseteq [0, x_{\mathrm{id}}]$, all admissible intensities induce the same response distribution across preferences, so $D_{ij}^\star(c, \epsilon) = 0$ for every $i \neq j$. Once an admissible intensity satisfies $g(x) \neq 0$, distinct scalar preferences become distinguishable. The formal statement is deferred to Appendix B.

Constraints can make identification statistically expensive. Identifiability may persist even as admissible probes become weakly informative. More generally, suppose as $x \to 0$,

$$g(x) = g_0 x^r + o(x^r), \qquad h(x) = h_0 x^m + o(x^m), \qquad C_T(x) = t_0 x^n + o(x^n),$$

with $g_0, h_0, t_0 > 0$ and $r, m, n > 0$. Then $d_{ij}(x) = \Theta(x^{2r})$, while the largest admissible intensity satisfies $\bar{x}(c, \epsilon) = \Theta(\min\{c^{1/m}, \epsilon^{1/n}\})$. Applying the information lower bound from Section 3.2 gives

$$\mathbb{E}_i[N] = \Omega\Big(\log(1/\delta) \max\{c^{-2r/m}, \epsilon^{-2r/n}\}\Big). \tag{8}$$

Thus identification can remain possible while its sample requirement diverges as either budget tightens. Constants and asymptotic expansions are given in Appendix B.5.

More aggressive need not mean more informative. Information need not increase monotonically with probe intensity. For same-sign preferences, logistic responses can saturate at large x, causing pairwise KL to peak strictly before the largest admissible intensity.

Figure 2 verifies all three effects. For $r = m = n = 1$, equation 8 predicts c^{-2} scaling in the risk-limited regime, consistent with the fitted behavior over the tested budget range. The same instance exhibits an interior KL maximum despite the availability of more aggressive admissible probes. Experimental parameters, confidence intervals, and the separate confidence-targeting study are reported in Appendix B.6.

5 SEQUENTIAL STRATEGIC INTERACTION

The analytic instance captures how constraints restrict available information, but not a defining feature of strategic interaction: a partner may respond to what the learner commits to doing in the future. We test this mechanism first in an exactly enumerable repeated game and then in continuous highway merging.

5.1 EXACT REPEATED BOTTLENECK GAME

We consider a two-agent bottleneck game in which each agent progresses through four states (start, approach, bottleneck, completion) and chooses either WAIT or ADVANCE. Simultaneous attempts to enter the bottleneck can create conflict. The learner has a fixed task objective, while the partner has latent preference $\rho \in \{0.5, 1, 2, 4\}$ controlling the relative cost of conflict and delay.

A probe is an announced $H = 4$ learner commitment $u = (a_{A,0}, \ldots, a_{A,H-1})$. The partner observes the entire commitment before responding and follows a finite-horizon soft-rational policy computed by exact dynamic programming. Consequently, its current action distribution can depend on future components of u, even when the learner's current action is unchanged. For every preference and commitment, we exactly enumerate the resulting distribution P_ρ^u over complete interaction trajectories. The transition, reward, likelihood, and robust-admissibility specifications are given in Appendix C.

Future commitments change current behavior. Among commitment pairs with the same first learner action but different futures, the largest first-step total-variation distance between partner responses is 0.629. Thus the learner's current action alone does not determine the current response. Removing this strategic lookahead nearly doubles the mean number of encounters required to reach 95% confidence, from 85.7 to 173.3 in the symmetric state and from 98.0 to 170.0 in the asymmetric state. The matched ablation is reported in Appendix C.8.

Trajectories retain additional identifying information. Using the full interaction trajectory instead of only the first response reduces the mean encounters to 95% confidence from 102.7 to 85.7 in the symmetric state and from 109.1 to 98.0 in the asymmetric state. A safe max–min-KL (Chernoff-style) baseline performs similarly ($84.1/99.5$), indicating that the trajectory advantage is not specific to Safe-ID. Paired intervals and additional baselines are given in Appendix E.5.

Safety reshapes the available information. For aggregate reporting across possible true preferences, define $\Gamma_{\min}(c, \epsilon) := \min_{i \in [M]} \Gamma_i(c, \epsilon)$. The exact game directly realizes the theoretical chain

$$c \longrightarrow \mathcal{U}(c, \epsilon) \longrightarrow \Gamma_{\min}(c, \epsilon) \longrightarrow \text{identification time.}$$

In the asymmetric state, increasing c from 0.25 to 0.60 to 0.80 expands the admissible probe set from 4 to 6 to 9, raises the minimum available-information rate over true preferences, $\Gamma_{\min}$, from 0.0008 to 0.0163 to 0.0349, and reduces mean identification time from 410.9 to 176.0 to 98.0 encounters. Unconstrained and posterior-averaged-risk variants can be faster but violate the robust admissibility criterion in equation 1; full comparisons are in Appendix C.9.

5.2 CONTINUOUS HIGHWAY-MERGE SCALE-UP

We next test whether the same strategic effects persist beyond the exactly enumerable game. In HighwayEnv `merge-v1`, a main-road learner selects among 12 four-step speed commitments, while a merging partner with one of four latent reward vectors responds through a finite-horizon soft planner. Confirmatory evaluation uses six disjoint geometry classes with 50 seeded perturbations each; simulation and enumeration details are given in Appendix D.

The future-commitment effect persists across the confirmatory geometries: among commitments sharing the same current action but differing in their future, mean first-step total-variation distance is 0.142 (95% CI $[0.129, 0.156]$), with every clustered cell exceeding the predeclared 0.01 threshold (Figure 3(a–b)). Relaxing the risk budget increases the median admissible-set size across geometry–perturbation cells from 7 to 10, while the median of the $1{,}200$ cell-specific available-information rates $\Gamma_{g,s,i}$ increases from 0.0845 to 0.0932.

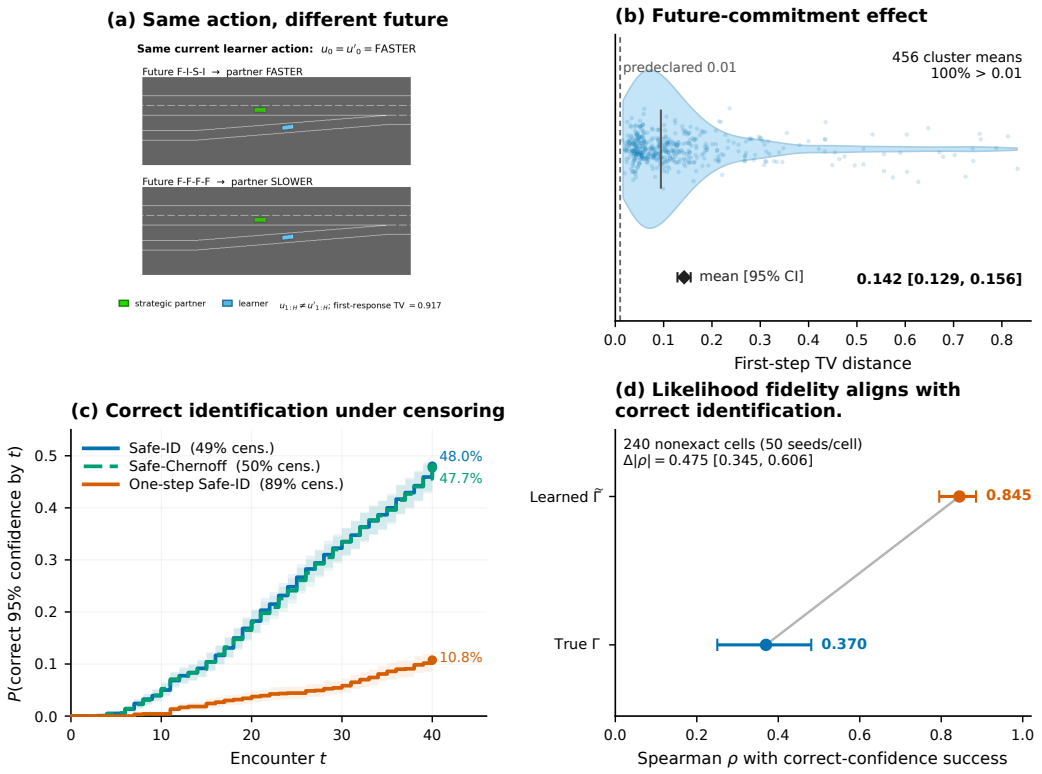


Figure 3: **Strategic identification scales to continuous highway merging.** (a–b) Commitments with the same current action but different futures induce different current responses. (c) Full-trajectory evidence substantially improves correct identification over one-step evidence and performs comparably to Safe-Chernoff. Panel (d), concerning likelihood mismatch, is analyzed in Section 6.

Trajectory-level evidence also remains important in the continuous setting. By 40 encounters, Safe-ID reaches correct 95% confidence in 48.0% of runs, compared with 10.7% using one-step evidence; Safe-Chernoff is practically tied at 47.7% (Figure 3(c)). Because many runs do not reach the threshold within 40 encounters, full censoring-aware RMST results are deferred to Appendix D.

6 When True Information Becomes Misleading Evidence

Section 5 showed that strategic interaction can create identifying information. We now isolate the second bottleneck: whether the learner's likelihood interprets that information in the correct direction. Throughout this section, we hold the true environment and admissible probe family fixed and perturb only the learner's behavioral model. Thus the true response laws P_i^u—and hence the available-information rate Γ_i—remain unchanged, while the learned model $\widehat{P}_i^u$ changes Δ_{ij} and $\widetilde{\Gamma}_i$.

6.1 Information Can Exist without Usable Evidence

Under the exact likelihood, the implementation recovers $\Delta_{ij} = D_{\mathrm{KL}}(P_i^u \| P_j^u)$ and $\widetilde{\Gamma}_i = \Gamma_i$ to numerical precision; full checks are given in Appendix A.5.

We then perturb only the learner's behavioral model while keeping the true response laws and admissible set fixed. Across the controlled sweep, 115 configurations satisfy

$$\Gamma_i(c, \epsilon) > 0 \qquad \text{but} \qquad \widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) \leq 0,$$

directly realizing the failure regime of Corollary 1. At $c = 0.8$, for example, the minimum true discrimination rate remains 0.03488, while the minimum learned-evidence rate becomes negative

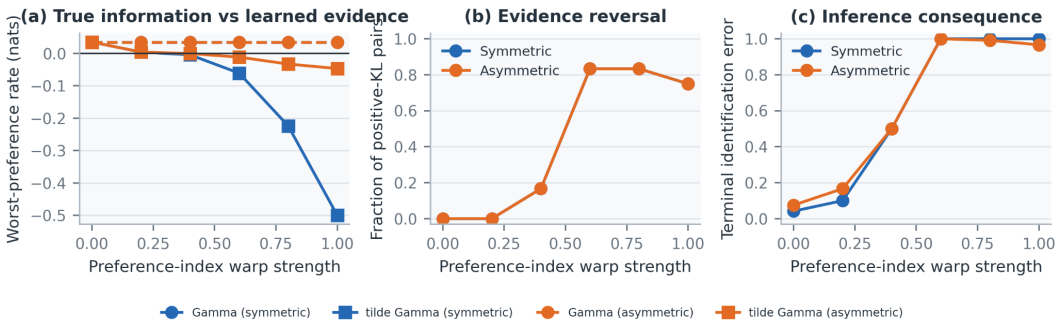


Figure 4: **True information can remain while learned evidence reverses.** Increasing likelihood mismatch leaves the true discrimination rate positive while the learned-evidence rate crosses zero, producing reversed evidence and high identification error.

as likelihood mismatch increases (Figure 4). Rationality and contamination mismatch also produce negative learned evidence on positive-KL pairs, showing that reversal is not specific to a single misspecification mechanism. Additional sweeps are reported in Appendix A.5.

6.2 Signed Evidence Predicts Posterior Failure

The sign of learned evidence has the posterior consequence predicted by Theorem 2. Under stationary probe designs, predicted and empirical mean log-odds slopes are respectively $0.575/0.580$ for exact positive evidence, $0.093/0.092$ for degraded positive evidence, and $-0.575/-0.566$ for reversed evidence. Thus the one-step quantity Δ_{ij} accurately predicts whether posterior odds grow toward or away from the true preference. Under severe mismatch, the learner can become highly confident in the wrong preference, with terminal error reaching 96.7%–100%.

The same distinction appears under natural HighwayEnv mismatch. Perturbing only the learner's assumed rationality, reaction delay, or acceleration dynamics leaves the true environment, admissible set, and Γ_i unchanged, but can drive $\widetilde{\Gamma}_i$ as low as -0.553. Across 240 non-exact mismatch cells, learned evidence is substantially more aligned with correct-confidence success than true available information ($\rho = 0.845$ versus 0.370; Figure 3(d)). We interpret this correlation as a mechanistic diagnostic rather than a finite-sample prediction theorem. Full mismatch, censoring, and likelihood-repair analyses are given in Appendices D and E.4.

7 Discussion and Limitations

Our results separate two requirements for safe strategic preference identification: admissible interactions must contain enough information to distinguish preferences, and the learner's behavioral model must turn that information into correctly directed evidence. The distinction is structural rather than algorithm-specific. In particular, Γ_i characterizes information available in the true environment, whereas $\widetilde{\Gamma}_i$ characterizes the evidence induced by the learner's likelihood; the regime $\Gamma_i > 0$ but $\widetilde{\Gamma}_i \leq 0$ shows that these need not coincide.

Several limitations remain. The posterior-drift result characterizes stationary randomized designs asymptotically and does not establish finite-sample optimality of Safe-ID. The exactly enumerable game is deliberately small for auditability, while the HighwayEnv evaluation remains a controlled two-vehicle, short-horizon setting. Extending the framework to high-dimensional learned response models, richer multi-agent interactions, and adaptive guarantees under model uncertainty is an important direction for future work.

Overall, safe preference identification requires both *informative admissible interaction* and *faithful behavioral inference*: preserving one without the other is insufficient for reliable identification.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, were used to assist with manuscript editing and organization, literature search, code review, and consistency checks between the theoretical definitions, experimental implementation, and reported results. They were also used to provide feedback on mathematical exposition, experimental methodology, and interpretation of results. All theoretical claims and proofs, experimental designs, code, and numerical results were reviewed and verified by the authors; reported quantities were checked against the underlying implementations and experimental outputs. The authors take full responsibility for the final content of this work, including all text, claims, code, and results produced with the aid of generative AI.

REFERENCES

Robert H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966. doi: 10.1214/aoms/1177699597.

Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pp. 519–528. PMLR, 2018.

Daniel S Brown, Yuchen Cui, and Scott Niekum. Risk-aware active inverse reinforcement learning. In *Conference on robot learning*, pp. 362–372. PMLR, 2018.

Thomas Kleine Büning, Anne-Marie George, and Christos Dimitrakakis. Interactive inverse reinforcement learning for cooperative games. In *International Conference on Machine Learning*, pp. 2393–2413. PMLR, 2022.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

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

Denizalp Goktas, Amy Greenwald, Sadie Zhao, Alec Koppel, and Sumitra Ganesh. Efficient inverse multiagent learning. In *International Conference on Learning Representations*, 2024.

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

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, 2022.

Sirin Nitinawarat, George K Atia, and Venugopal V Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on automatic control*, 58(10):2451–2464, 2013.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active preference-based learning of reward functions. 2017.

Joar Skalse and Alessandro Abate. Misspecification in inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15136–15143, 2023.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. *arXiv preprint arXiv:2204.06601*, 2022.

Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In *International conference on machine learning*, pp. 7194–7201. PMLR, 2019.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Proofs and Additional Identifiability Results

A.1 Monotonicity of Safe Equivalence

Suppose $c_1 \le c_2$ and $\epsilon_1 \le \epsilon_2$. If $u \in \mathcal{U}(c_1, \epsilon_1)$, then by definition

$$\max_{i\in[M]} \mathrm{CVaR}_\alpha(L_i^u) \le c_1 \le c_2, \qquad \max_{i\in[M]} [J_i^{\mathrm{ref}} - J_i(u)]_+ \le \epsilon_1 \le \epsilon_2.$$

Hence $u \in \mathcal{U}(c_2, \epsilon_2)$, and therefore

$$\mathcal{U}(c_1, \epsilon_1) \subseteq \mathcal{U}(c_2, \epsilon_2).$$

Now let $\theta' \in [\theta]_{c_2,\epsilon_2}$. By definition,

$$P_{\theta'}^u = P_\theta^u \qquad \text{for every } u \in \mathcal{U}(c_2, \epsilon_2).$$

Since $\mathcal{U}(c_1, \epsilon_1) \subseteq \mathcal{U}(c_2, \epsilon_2)$, the same equality holds for every $u \in \mathcal{U}(c_1, \epsilon_1)$. Thus $\theta' \in [\theta]_{c_1,\epsilon_1}$, so

$$[\theta]_{c_2,\epsilon_2} \subseteq [\theta]_{c_1,\epsilon_1}.$$

□

A.2 Pairwise Information Lower Bound

Let $\mathbb{P}_i$ and $\mathbb{P}_j$ denote the probability laws of the complete stopped interaction history under hypotheses θ_i and θ_j, respectively, and let $\widehat{\theta}$ denote the terminal preference estimate. Because the probing and stopping rules depend only on the observed history and are identical under the two hypotheses, the likelihood ratio between the stopped processes accumulates through the response likelihoods. Under the required absolute-continuity and integrability conditions,

$$D_{\mathrm{KL}}(\mathbb{P}_i \| \mathbb{P}_j) = \mathbb{E}_i \left[\sum_{t=1}^{N} \log \frac{p_i^{u_t}(Y_t)}{p_j^{u_t}(Y_t)} \right].$$

Conditioning on the history available before round t and the selected probe gives

$$\mathbb{E}_i \left[\log \frac{p_i^{u_t}(Y_t)}{p_j^{u_t}(Y_t)} \middle| h_t, u_t \right] = D_{\mathrm{KL}} \left(P_i^{u_t} \| P_j^{u_t} \right) \le D_{ij}^\star(c, \epsilon),$$

where the inequality follows because $u_t \in \mathcal{U}(c, \epsilon)$. Consequently,

$$D_{\mathrm{KL}}(\mathbb{P}_i \| \mathbb{P}_j) \le D_{ij}^\star(c, \epsilon)\, \mathbb{E}_i[N].$$

Let

$$A := \{\widehat{\theta} = \theta_i\}.$$

If the procedure has error probability at most δ under both hypotheses, then

$$\mathbb{P}_i(A) \ge 1 - \delta, \qquad \mathbb{P}_j(A) \le \delta.$$

Applying data processing to the binary statistic $\mathbf{1}\{A\}$ yields

$$D_{\mathrm{KL}}(\mathbb{P}_i \| \mathbb{P}_j) \ge \mathrm{kl}(1 - \delta, \delta).$$

Combining the two bounds gives

$$\mathbb{E}_i[N] \ge \frac{\mathrm{kl}(1 - \delta, \delta)}{D_{ij}^\star(c, \epsilon)}.$$

Since $\mathrm{kl}(1 - \delta, \delta) = \Theta(\log(1/\delta))$ as $\delta \to 0$, the corresponding asymptotic lower bound follows. □

A.3 Proof of Theorem 1

Fix the true preference θ_i. For every alternative θ_j, $j \neq i$, the pairwise change-of-measure argument above gives

$$\mathbb{E}_i\left[\sum_{t=1}^{N} D_{\mathrm{KL}}\left(P_i^{u_t} \| P_j^{u_t}\right)\right] \geq \mathrm{kl}(1-\delta, \delta).$$

If $\mathbb{E}_i[N] = \infty$, the claimed lower bound is immediate, so assume $\mathbb{E}_i[N] < \infty$. Because the admissible probe family is finite, define for each $u \in \mathcal{U}(c, \epsilon)$

$$N_u := \sum_{t=1}^{N} \mathbf{1}\{u_t = u\}, \qquad q_i(u) := \frac{\mathbb{E}_i[N_u]}{\mathbb{E}_i[N]}.$$

Since $\sum_u N_u = N$, we have $q_i \in \Delta(\mathcal{U}(c, \epsilon))$. Moreover, for every $j \neq i$,

$$\mathbb{E}_i[N] \sum_{u \in \mathcal{U}(c,\epsilon)} q_i(u) D_{\mathrm{KL}}\left(P_i^u \| P_j^u\right) \geq \mathrm{kl}(1-\delta, \delta).$$

Taking the minimum over alternatives gives

$$\mathbb{E}_i[N] \min_{j \in [M]\setminus\{i\}} \sum_{u \in \mathcal{U}(c,\epsilon)} q_i(u) D_{\mathrm{KL}}\left(P_i^u \| P_j^u\right) \geq \mathrm{kl}(1-\delta, \delta).$$

By the definition of $\Gamma_i(c, \epsilon)$,

$$\min_{j \in [M]\setminus\{i\}} \sum_{u \in \mathcal{U}(c,\epsilon)} q_i(u) D_{\mathrm{KL}}\left(P_i^u \| P_j^u\right) \leq \Gamma_i(c, \epsilon).$$

Therefore

$$\mathbb{E}_i[N]\, \Gamma_i(c, \epsilon) \geq \mathrm{kl}(1-\delta, \delta),$$

and hence

$$\mathbb{E}_i[N] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\Gamma_i(c, \epsilon)}.$$

□

Extensions to measurable probe spaces require the corresponding measurability and occupation-measure assumptions; all experimental probe families considered in this paper are finite.

A.4 Critical Identification Budgets

The identifiability region introduced in Section 3.2 admits equivalent critical-budget representations. For fixed task budget ϵ, define

$$c_i^\star(\epsilon) := \inf\{c : \Gamma_i(c, \epsilon) > 0\}, \qquad \epsilon_i^\star(c) := \inf\{\epsilon : \Gamma_i(c, \epsilon) > 0\}.$$

The corresponding pairwise thresholds are

$$c_{ij}^\star(\epsilon) := \inf\{c : D_{ij}^\star(c, \epsilon) > 0\}, \qquad \epsilon_{ij}^\star(c) := \inf\{\epsilon : D_{ij}^\star(c, \epsilon) > 0\}.$$

These quantities give the minimum interaction-risk or task-degradation budget required to cross the global or pairwise safe-identifiability boundary.

A.5 Additional Results for Learned Evidence

A.5.1 Signed-Evidence Decomposition

Assume that, for every relevant probe u, P_i^u is absolutely continuous with respect to both $\widehat{P}_i^u$ and $\widehat{P}_j^u$, and that the corresponding log-likelihood ratios are integrable. Then

$$\begin{aligned}
\Delta_{ij}(u; \widehat{P}) &= \mathbb{E}_{Y \sim P_i^u}\left[\log \frac{\widehat{p}_i^u(Y)}{\widehat{p}_j^u(Y)}\right] \\
&= \mathbb{E}_{Y \sim P_i^u}\left[\log \frac{p_i^u(Y)}{\widehat{p}_j^u(Y)} - \log \frac{p_i^u(Y)}{\widehat{p}_i^u(Y)}\right] \\
&= D_{\mathrm{KL}}\left(P_i^u \| \widehat{P}_j^u\right) - D_{\mathrm{KL}}\left(P_i^u \| \widehat{P}_i^u\right).
\end{aligned}$$

Thus learned evidence depends on relative model fit: observations generated under θ_i favor the learner's model for θ_i only when that model fits the true response law better than the learner's model for the competitor.

A.5.2 Proof of Theorem 2

Bayesian updating under the learner's likelihood $\widehat{P}$ gives the exact posterior-odds recursion

$$\log \frac{\widehat{b}_T(i)}{\widehat{b}_T(j)} = \log \frac{\widehat{b}_0(i)}{\widehat{b}_0(j)} + \sum_{t=1}^{T} \log \frac{\widehat{p}_i^{u_t}(Y_t)}{\widehat{p}_j^{u_t}(Y_t)}.$$

Conditioning on the history available before round t and on u_t, while the true preference is θ_i, gives

$$\mathbb{E}_i \left[\log \frac{\widehat{p}_i^{u_t}(Y_t)}{\widehat{p}_j^{u_t}(Y_t)} \middle| h_t, u_t \right] = \Delta_{ij}(u_t; \widehat{P}).$$

Consequently, for an arbitrary adaptive probing rule,

$$\mathbb{E}_i \left[\log \frac{\widehat{b}_T(i)/\widehat{b}_T(j)}{\widehat{b}_0(i)/\widehat{b}_0(j)} \right] = \mathbb{E}_i \left[\sum_{t=1}^{T} \Delta_{ij}(u_t; \widehat{P}) \right].$$

Now suppose $u_t \overset{\text{i.i.d.}}{\sim} q$ and, conditional on the probes, the responses are independent with $Y_t \sim P_i^{u_t}$. Under the stated log-likelihood integrability condition, the strong law applied to the probe–response pairs gives

$$\frac{1}{T} \sum_{t=1}^{T} \log \frac{\widehat{p}_i^{u_t}(Y_t)}{\widehat{p}_j^{u_t}(Y_t)} \longrightarrow \mathbb{E}_{u \sim q} \left[\Delta_{ij}(u; \widehat{P}) \right] \qquad \text{a.s.}$$

Since the prior log-odds term divided by T vanishes,

$$\frac{1}{T} \log \frac{\widehat{b}_T(i)}{\widehat{b}_T(j)} \longrightarrow \mathbb{E}_{u \sim q} \left[\Delta_{ij}(u; \widehat{P}) \right] \qquad \text{a.s.}$$

If this limiting rate is positive for every $j \neq i$, then $\widehat{b}_T(i)/\widehat{b}_T(j) \to \infty$ for every alternative and hence $\widehat{b}_T(i) \to 1$. If the rate is negative for some j, then $\widehat{b}_T(i)/\widehat{b}_T(j) \to 0$, so the posterior cannot concentrate on θ_i.

Finally, if $\widehat{P}_k^u = P_k^u$ for every candidate k and probe u, then

$$\Delta_{ij}(u; \widehat{P}) = D_{\mathrm{KL}}(P_i^u \| P_j^u),$$

and therefore

$$\widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) = \Gamma_i(c, \epsilon).$$

□

A.5.3 Likelihood-Distortion Stability

Assume that the true and learned likelihoods have common support and satisfy

$$|\log \widehat{p}_k^u(y) - \log p_k^u(y)| \leq \eta$$

for every admissible probe $u \in \mathcal{U}(c, \epsilon)$, every $k \in [M]$, and every response y in the common support. Then, for any $j \neq i$,

$$\Delta_{ij}(u; \widehat{P}) - D_{\mathrm{KL}}(P_i^u \| P_j^u) = \mathbb{E}_{Y \sim P_i^u} \left[\log \frac{\widehat{p}_i^u(Y)}{p_i^u(Y)} - \log \frac{\widehat{p}_j^u(Y)}{p_j^u(Y)} \right].$$

Hence

$$\left| \Delta_{ij}(u; \widehat{P}) - D_{\mathrm{KL}}(P_i^u \| P_j^u) \right| \leq 2\eta.$$

Averaging over any admissible design q, taking the minimum over $j \neq i$, and finally taking the supremum over q preserves the same uniform bound. Therefore

$$\left| \widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) - \Gamma_i(c, \epsilon) \right| \leq 2\eta.$$

In particular, $\Gamma_i(c, \epsilon) > 2\eta$ is sufficient to ensure $\widetilde{\Gamma}_i(c, \epsilon; \widehat{P}) > 0$.

A.5.4 CVaR Stability under Response-Model Error

Let

$$D_{\mathrm{TV}}(P,Q) := \sup_A |P(A) - Q(A)|$$

denote total-variation distance. Consider two models P and Q that induce possibly different distributions of the same loss random variable L, and assume

$$0 \le L \le L_{\max} \qquad \text{almost surely.}$$

Using the variational representation of CVaR, the minimization may be restricted to $\eta \in [0, L_{\max}]$. For every such η, $(L-\eta)_+ \in [0, L_{\max}]$, so

$$|\mathbb{E}_P[(L-\eta)_+] - \mathbb{E}_Q[(L-\eta)_+]| \le L_{\max} D_{\mathrm{TV}}(P,Q).$$

Since $|\inf_\eta a_\eta - \inf_\eta b_\eta| \le \sup_\eta |a_\eta - b_\eta|$, the two CVaR objectives satisfy

$$\left|\mathrm{CVaR}_\alpha^P(L) - \mathrm{CVaR}_\alpha^Q(L)\right| \le \frac{L_{\max}}{1-\alpha} D_{\mathrm{TV}}(P,Q).$$

In particular, if the true and learned response models induce the corresponding loss laws and

$$D_{\mathrm{TV}}\left(P_i^u, \widehat{P}_i^u\right) \le \zeta_i^u,$$

then

$$\left|\mathrm{CVaR}_\alpha^{P_i^u}(L) - \mathrm{CVaR}_\alpha^{\widehat{P}_i^u}(L)\right| \le \frac{L_{\max}}{1-\alpha}\zeta_i^u.$$

A sufficient model-based safety condition is therefore

$$\mathrm{CVaR}_\alpha^{\widehat{P}_i^u}(L) \le c - \frac{L_{\max}}{1-\alpha}\zeta_i^u.$$

This calibration margin addresses response-model uncertainty and is distinct from the max-over-preferences robustness in the primary admissible set.

A.6 Continuous-Preference Identifiability

For completeness, we give a local continuous-preference analogue of the finite-hypothesis results. This extension is not required for the main theory or experiments.

For this subsection, let the preference parameter vary over $\Theta_{\mathrm{cont}} \subseteq \mathbb{R}^d$. For probe u, let $p_\theta^u(y)$ denote the response density or probability mass function and define the score

$$s_\theta^u(y) := \nabla_\theta \log p_\theta^u(y).$$

Whenever the score is well defined and square integrable, the Fisher information contributed by probe u is

$$F_u(\theta) := \mathbb{E}_{Y\sim P_\theta^u}\left[s_\theta^u(Y) s_\theta^u(Y)^\top\right].$$

For a randomized admissible design $q \in \Delta(\mathcal{U}(c,\epsilon))$, define

$$F_q(\theta) := \mathbb{E}_{u\sim q}[F_u(\theta)], \qquad \Lambda(c,\epsilon;\theta) := \sup_{q\in\Delta(\mathcal{U}(c,\epsilon))} \lambda_{\min}(F_q(\theta)),$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.

For a finite admissible probe family,

$$\Lambda(c,\epsilon;\theta) > 0 \quad \Longleftrightarrow \quad \bigcap_{u\in\mathcal{U}(c,\epsilon)} \ker F_u(\theta) = \{0\}.$$

Proof. If $\Lambda(c, \epsilon; \theta) > 0$, then there exists an admissible design q such that $F_q(\theta) \succ 0$. Hence, for every nonzero $v \in \mathbb{R}^d$,

$$v^\top F_q(\theta) v = \mathbb{E}_{u \sim q} \left[v^\top F_u(\theta) v \right] > 0,$$

so no nonzero direction can belong to the kernel of every probe used by the design. In particular,

$$\bigcap_{u \in \mathcal{U}(c,\epsilon)} \ker F_u(\theta) = \{0\}.$$

Conversely, suppose

$$\bigcap_{u \in \mathcal{U}(c,\epsilon)} \ker F_u(\theta) = \{0\}.$$

Choose any design assigning strictly positive probability to every admissible probe, for example the uniform distribution. If a nonzero vector v satisfied $v^\top F_q(\theta) v = 0$, then every nonnegative term $v^\top F_u(\theta) v$ receiving positive weight would have to equal zero. Since each $F_u(\theta)$ is positive semidefinite, this would imply $v \in \ker F_u(\theta)$ for every admissible u, contradicting the assumed trivial intersection. Thus $F_q(\theta) \succ 0$, and consequently $\Lambda(c, \epsilon; \theta) > 0$. □

A.6.1 Soft-Rational Partner Responses

Let $\mathcal{A}$ be a finite partner action set and let $\beta > 0$ denote the inverse-temperature parameter. Suppose the partner follows

$$P_\theta(a \mid u) = \frac{\exp(\beta Q_\theta(u, a))}{\sum_{a' \in \mathcal{A}} \exp(\beta Q_\theta(u, a'))}, \qquad a \in \mathcal{A},$$

with preference-dependent action value

$$Q_\theta(u, a) = \theta^\top \phi(u, a),$$

where $\phi(u, a) \in \mathbb{R}^d$ is the feature vector associated with probe u and partner action a.

The score is

$$\nabla_\theta \log P_\theta(a \mid u) = \beta \left(\phi(u, a) - \mathbb{E}_{A \sim P_\theta(\cdot \mid u)} [\phi(u, A)] \right),$$

and therefore

$$F_u(\theta) = \beta^2 \operatorname{Cov}_{A \sim P_\theta(\cdot \mid u)} \left[\phi(u, A) \right].$$

Thus a preference direction is locally observable only if some admissible probe induces variation in partner behavior along features that depend on that direction. Directions lying in the kernel of every admissible probe-wise Fisher matrix are locally unidentifiable.

A.6.2 Estimation Precision

Under the standard Cramér–Rao regularity conditions, consider a design q such that $F_q(\theta) \succ 0$. Any unbiased estimator $\widehat{\theta}$ based on N independent observations generated under q satisfies

$$\operatorname{Cov}(\widehat{\theta}) \succeq \frac{1}{N} F_q(\theta)^{-1}.$$

Consequently,

$$\sup_{\|v\|_2 = 1} \operatorname{Var}\left(v^\top \widehat{\theta} \right) \geq \frac{1}{N \lambda_{\min}(F_q(\theta))} \geq \frac{1}{N \Lambda(c, \epsilon; \theta)}.$$

Thus the admissible interaction family determines not only local identifiability but also the best attainable estimation precision along the least observable preference direction.

A.7 Safe-ID: Operational Probing Policy

The theory characterizes when identifying information is available and when the learner's likelihood converts that information into correctly directed evidence, but it does not prescribe an online probing strategy. We therefore use Safe-ID as a simple operational policy in the experiments.

Let h_t denote the interaction history available before round t, and let

$$\widehat{b}_t(\theta_i) := \Pr_{\widehat{P}}(\theta^\star = \theta_i \mid h_t), \qquad i \in [M],$$

denote the learner's posterior computed using its behavioral likelihood $\widehat{P}$. For an admissible probe $u \in \mathcal{U}(c, \epsilon)$, define the learner's Bayesian information gain by

$$\widehat{\mathcal{I}}_t(u) := I_{\widehat{b}_t, \widehat{P}}(\theta; Y \mid u) = \sum_{i \in [M]} \widehat{b}_t(\theta_i) D_{\mathrm{KL}} \left(\widehat{P}_i^u \middle\| \sum_{j \in [M]} \widehat{b}_t(\theta_j) \widehat{P}_j^u \right).$$

Thus probe selection is based on the learner's current behavioral model, which may differ from the true response laws P_i^u. When the likelihood is exact, $\widehat{P}_i^u = P_i^u$ for every i and u, and $\widehat{\mathcal{I}}_t$ reduces to the usual Bayesian information gain under the true response model.

For confidence level $1 - \delta$, define

$$\widehat{\tau}_\delta := \inf \left\{ t : \max_{i \in [M]} \widehat{b}_t(\theta_i) \geq 1 - \delta \right\}.$$

Safe-ID selects

$$u_t \in \begin{cases} \arg\max_{u \in \mathcal{U}(c,\epsilon)} \widehat{\mathcal{I}}_t(u), & t < \widehat{\tau}_\delta, \\ \arg\max_{u \in \mathcal{U}(c,\epsilon)} \sum_{i \in [M]} \widehat{b}_t(\theta_i) J_i(u), & t \geq \widehat{\tau}_\delta. \end{cases} \tag{9}$$

Thus Safe-ID uses the learner's assumed response model both to update its posterior and to select informative probes. The admissible set $\mathcal{U}(c, \epsilon)$ is held fixed by the true environment in the controlled likelihood-mismatch experiments; only inference and probe selection use $\widehat{P}$. This isolates likelihood misspecification from changes in the underlying safety constraints.

The threshold $\widehat{\tau}_\delta$ is a Bayesian confidence criterion and does not, by itself, imply frequentist δ-correct identification. We therefore report empirical correctness separately from time to threshold.

B Analytic Strategic-Response Model

This appendix gives the derivations and experimental specification for the analytic instance in Section 4.

B.1 Response Distribution and Endogenous Risk

The learner selects an interaction intensity $x \in [0, 1]$, and the partner produces a binary response $Y \in \{0, 1\}$. Let $r_H^\theta(x, y)$ denote the partner's utility from response y under preference θ. We choose

$$r_H^\theta(x, 1) - r_H^\theta(x, 0) = \theta g(x),$$

which under a logistic response model gives

$$p_\theta(x) := \Pr(Y = 1 \mid \theta, x) = \sigma\big(\beta \theta g(x)\big), \qquad \sigma(z) := (1 + e^{-z})^{-1}.$$

Here $\beta > 0$ controls response rationality. Since $g(0) = 0$, $p_\theta(0) = 1/2$ for every θ; if $\theta_i \neq \theta_j$ and $g(x) \neq 0$, strict monotonicity of σ implies $p_{\theta_i}(x) \neq p_{\theta_j}(x)$.

Suppose $Y = 1$ triggers a binary adverse-event loss $L \in \{0, 1\}$ with conditional probability $h(x)$. Then

$$q_i(x) := \Pr(L = 1 \mid \theta_i, x) = h(x) p_{\theta_i}(x).$$

For $L \sim \mathrm{Bern}(q)$, the CVaR definition from Section 3.1 reduces to

$$\mathrm{CVaR}_\alpha(L) = \min \left\{ 1, \frac{q}{1 - \alpha} \right\},$$

obtained by evaluating the variational objective at its breakpoints $\eta \in \{0, 1\}$. Hence the robust interaction risk is

$$R(x) = \max_{i \in [M]} \min \left\{ 1, \frac{h(x) p_{\theta_i}(x)}{1 - \alpha} \right\}. \tag{10}$$

B.2 Task Cost and Admissible Intensities

Let

$$J_i(x) := \mathbb{E}[r_A(x, Y) \mid \theta_i, x]$$

denote the learner's expected task value, where $r_A(x, Y)$ is the learner's task reward. Using the reference value J_i^{ref} defined in Section 3.1, the robust task degradation is

$$C_T(x) := \max_{i \in [M]} [J_i^{\mathrm{ref}} - J_i(x)]_+. \tag{11}$$

Thus the admissible intensity set is

$$\mathcal{X}(c, \epsilon) = \{x \in [0, 1] : R(x) \le c,\ C_T(x) \le \epsilon\}.$$

In the clean analytic instance, we impose

$$J_i^{\mathrm{ref}} - J_i(x) = t_0 x^n \qquad \text{for every } i,$$

with $t_0, n > 0$, so $C_T(x) = t_0 x^n$. The implementation uses the equivalent normalization $J_i^{\mathrm{ref}} = 0$ and $J_i(x) = -t_0 x^n$.

When R and C_T are nondecreasing, the admissible set is an interval $\mathcal{X}(c, \epsilon) = [0, \bar{x}(c, \epsilon)]$. Defining

$$R^{\leftarrow}(c) := \sup\{x \in [0, 1] : R(x) \le c\}, \qquad C_T^{\leftarrow}(\epsilon) := \sup\{x \in [0, 1] : C_T(x) \le \epsilon\},$$

its upper endpoint is

$$\bar{x}(c, \epsilon) = \min\{R^{\leftarrow}(c), C_T^{\leftarrow}(\epsilon)\}. \tag{12}$$

B.3 Pairwise Information and Saturation

For preferences θ_i and θ_j, define

$$\begin{aligned} d_{ij}(x) &:= D_{\mathrm{KL}}\big(\mathrm{Bern}(p_{\theta_i}(x)) \| \mathrm{Bern}(p_{\theta_j}(x))\big) \\ &= p_{\theta_i}(x) \log \frac{p_{\theta_i}(x)}{p_{\theta_j}(x)} + (1 - p_{\theta_i}(x)) \log \frac{1 - p_{\theta_i}(x)}{1 - p_{\theta_j}(x)}. \end{aligned} \tag{13}$$

The pairwise safe separation is therefore

$$D_{ij}^{\star}(c, \epsilon) = \sup_{x \in \mathcal{X}(c, \epsilon)} d_{ij}(x).$$

For same-sign preferences, increasing x can drive both logistic response probabilities toward the same saturation boundary. Pairwise KL therefore need not increase monotonically with intensity and may attain its maximum strictly inside the admissible interval, as demonstrated in Section 4.

B.4 Dead-Zone Identifiability

Proposition 1. Suppose $g(x) = 0$ for all $x \le x_{\mathrm{id}}$. If $\mathcal{X}(c, \epsilon) \subseteq [0, x_{\mathrm{id}}]$, then $D_{ij}^{\star}(c, \epsilon) = 0$ for every $i \ne j$. Conversely, if there exists $x \in \mathcal{X}(c, \epsilon)$ with $g(x) \ne 0$, then every pair of distinct scalar preferences is distinguishable.

Proof. If $\mathcal{X}(c, \epsilon) \subseteq [0, x_{\mathrm{id}}]$, then $g(x) = 0$ for every admissible x, so $p_{\theta_i}(x) = 1/2$ for all i. Hence $d_{ij}(x) = 0$ throughout the admissible set and therefore $D_{ij}^{\star}(c, \epsilon) = 0$.

Conversely, let $x \in \mathcal{X}(c, \epsilon)$ satisfy $g(x) \ne 0$. For $\theta_i \ne \theta_j$ and $\beta > 0$, $\beta\theta_i g(x) \ne \beta\theta_j g(x)$. Strict monotonicity of the logistic function gives $p_{\theta_i}(x) \ne p_{\theta_j}(x)$, hence $d_{ij}(x) > 0$. □

B.5 Local Safety–Information Scaling

More generally, suppose that as $x \to 0$,

$$g(x) = g_0 x^r + o(x^r), \qquad h(x) = h_0 x^m + o(x^m), \qquad C_T(x) = t_0 x^n + o(x^n),$$

where $g_0, h_0, t_0 > 0$ and $r, m, n > 0$. Since $\sigma(z) = 1/2 + z/4 + o(z)$,

$$p_\theta(x) = \frac{1}{2} + \frac{\beta\theta g_0}{4} x^r + o(x^r). \tag{14}$$

Writing $a_i := \beta\theta_i g_0/4$, a second-order expansion of Bernoulli KL around $1/2$ gives

$$D_{\mathrm{KL}}\big(\mathrm{Bern}\big(\tfrac{1}{2} + a_i x^r + o(x^r)\big) \,\|\, \mathrm{Bern}\big(\tfrac{1}{2} + a_j x^r + o(x^r)\big)\big) = 2(a_i - a_j)^2 x^{2r} + o(x^{2r}),$$

and therefore

$$d_{ij}(x) = \frac{\beta^2 g_0^2 (\theta_i - \theta_j)^2}{8} x^{2r} + o(x^{2r}). \tag{15}$$

Moreover, $p_\theta(x) = 1/2 + o(1)$ implies $\max_i q_i(x) = \frac{h_0}{2} x^m + o(x^m)$. For sufficiently small x this lies below $1 - \alpha$, so

$$R(x) = \frac{h_0}{2(1-\alpha)} x^m + o(x^m). \tag{16}$$

Solving the local risk and task constraints yields

$$\bar{x}(c, \epsilon) = (1 + o(1)) \min\left\{ \left(\frac{2(1-\alpha)c}{h_0}\right)^{1/m}, \left(\frac{\epsilon}{t_0}\right)^{1/n} \right\}. \tag{17}$$

Because the leading KL term is increasing for sufficiently small x, its maximum over the local admissible interval occurs at $\bar{x}(c, \epsilon)$:

$$D_{ij}^\star(c, \epsilon) = \frac{\beta^2 g_0^2 (\theta_i - \theta_j)^2}{8} \bar{x}(c, \epsilon)^{2r} (1 + o(1)). \tag{18}$$

Combining this expression with the pairwise information lower bound from Section 3.2 gives

$$\mathbb{E}_i[N] \geq \frac{8\ \mathrm{kl}(1-\delta, \delta)}{\beta^2 g_0^2 (\theta_i - \theta_j)^2} \bar{x}(c, \epsilon)^{-2r} (1 + o(1)),$$

and hence

$$\mathbb{E}_i[N] = \Omega\Big(\log(1/\delta) \max\{c^{-2r/m}, \epsilon^{-2r/n}\}\Big). \tag{19}$$

B.6 Frozen Analytic Experiments

The principal configuration uses $\alpha = 0.90$, $h(x) = 0.08x$, $C_T(x) = 0.20x$, and $g(x) = x$, with $\Theta_A = \{-1, +1\}$, $\Theta_B = \{0.5, 1.5\}$, and $\Theta_M = \{-1.5, -0.5, 0.5, 1.5\}$. We sweep $\beta \in \{0.5, 1, 2, 4\}$. Theory quantities are evaluated on a 4001-point intensity grid and policies on a 501-point grid. Sequential experiments use 30 seeds, posterior thresholds 0.90, 0.95, and 0.99, and a maximum horizon of $10{,}000$ rounds; all information quantities use natural logarithms.

For the $r = m = n = 1$ local-scaling experiment, regressions are restricted to the preregistered regime $\bar{x} \leq 0.2$, where the leading-order approximation error is at most 10%. The saturation experiment uses $\Theta_B = \{0.5, 1.5\}$ and $\beta = 4$.

Prediction	Empirical result	Theory
Dead-zone non-identifiability	$216/216$: $D^\star = 0$	$D^\star = 0$
Information vs. ID time	Spearman $= -0.987$	negative
Risk-limited slope	-1.775 $[-2.018, -1.534]$	-2
Task-limited slope	-2.042 $[-2.178, -1.897]$	-2
Interior KL optimum	$x^\star = 0.852$	possible

B.7 Confidence-Targeted and Utility-Weighted Identification

We additionally compare confidence-targeted information probing with a fixed utility–information tradeoff while holding the environment, hypothesis sets, constraints, seeds, and response model fixed. The utility-weighted controller selects

$$x_t \in \arg\max_{x \in \mathcal{X}(c,\epsilon)} \{J_{b_t}(x) + 10 I_{b_t}(\theta; Y \mid x)\}.$$

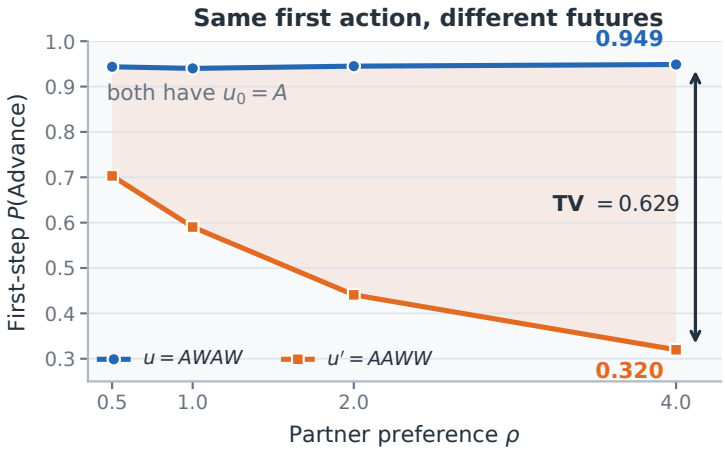


Figure 5: **Future commitments change the current response.** In the symmetric initial state, commitments $AWAW$ and $AAWW$ share the same first learner action but induce first-step advance probabilities 0.949 and 0.320 at $\rho = 4$, yielding $D_{\mathrm{TV}} = 0.629$.

In the $\{-1, +1\}$ and four-hypothesis conditions, this controller is censored in all runs at the 0.95 posterior threshold. In a representative two-hypothesis run, its posterior plateaus at 0.938197 and it selects $x_{\mathrm{UW}} = 0$, although the information-maximizing admissible action is $x_{\mathrm{MI}} = 0.170$. At this point, the task advantage $\Delta J = 0.034000$ exceeds the weighted forgone information $10\Delta I = 0.033158$, so the fixed scalarized objective prefers to stop probing before the confidence target is reached.

C Exact Sequential Shared-Bottleneck Game

This appendix gives the complete specification of the exactly enumerable shared-bottleneck environment used in Section 5, together with the evaluation protocols and ablations underlying the reported strategic-interaction results.

C.1 Environment and Transition Dynamics

Each agent $k \in \{A, H\}$ has progress state $z_k \in \{0, 1, 2, 3\}$, corresponding respectively to start, approach, bottleneck, and completion. The nonterminal physical state is $s = (z_A, z_H)$; the environment also contains safe and adverse terminal states. Each agent chooses $a_k \in \{\textsc{Wait}, \textsc{Advance}\}$, where Advance attempts to increase progress by one.

A conflict attempt occurs whenever the tentative transition would place both agents in bottleneck state 2. Conditional on such an attempt, the adverse terminal state is reached with probability $\kappa_{\mathrm{col}} = 0.25$; otherwise the conflict is resolved safely. When both agents simultaneously attempt to enter an unoccupied bottleneck, the successful entrant is chosen uniformly. If one agent already occupies the bottleneck, that agent remains and the entering agent is blocked. An encounter terminates when both agents complete, the adverse terminal state is reached, or the horizon expires.

C.2 General-Sum Reward Model

Let $\Delta_{k,t} \in \{0, 1\}$ indicate successful progress by agent k at round t, $D_{k,t} \in \{0, 1\}$ indicate that agent k remains incomplete, $C_t \in \{0, 1\}$ indicate a conflict attempt, and $B_t \in \{0, 1\}$ indicate transition to the adverse terminal state.

The partner's latent preference

$$\rho \in \{0.5, 1, 2, 4\}$$

controls the relative cost of conflict and delay. Its reward is

$$r_H^\rho(t) = \Delta_{H,t} - \rho(B_t + 0.25C_t) - \frac{0.20}{\rho} D_{H,t}. \tag{20}$$

Larger ρ therefore increases aversion to conflict and adverse outcomes while reducing the relative penalty on delay.

The learner objective is fixed across partner preferences:

$$r_A(t) = \Delta_{A,t} - 3B_t - 0.25C_t - 0.20D_{A,t}. \tag{21}$$

Because $r_A \neq r_H^\rho$, the interaction is general-sum.

C.3 Strategic Partner Response Model

A learner probe is an announced commitment

$$u = (u_0, \ldots, u_{H-1}), \qquad H = 4,$$

which the partner observes in full before responding. Let $P(s' \mid s, u_t, a_H)$ denote the transition kernel and let $\gamma = 0.95$. With terminal value $V_H^\rho(s; u) = 0$, backward dynamic programming gives

$$Q_t^\rho(s, a_H; u) = \sum_{s'} P(s' \mid s, u_t, a_H) \Big[r_H^\rho(s, u_t, a_H, s') + \gamma V_{t+1}^\rho(s'; u) \Big], \tag{22}$$

$$V_t^\rho(s; u) = \frac{1}{\beta} \log \sum_{a_H} \exp(\beta Q_t^\rho(s, a_H; u)), \tag{23}$$

with inverse temperature $\beta = 2$. The resulting soft-rational partner policy is

$$\pi_H^\rho(a_H \mid s, t, u) = \frac{\exp(\beta Q_t^\rho(s, a_H; u))}{\sum_{\widetilde{a}_H} \exp(\beta Q_t^\rho(s, \widetilde{a}_H; u))}. \tag{24}$$

Because the continuation value depends on the full commitment u, two probes with identical current learner action can induce different current partner responses. All soft values are evaluated using numerically stable log-sum-exp.

C.4 Exact Trajectory Distribution and Information

For each preference ρ and probe u, we enumerate every trajectory with nonzero probability. A trajectory is represented as

$$Y = (a_{H,0}, s_1, \ldots, a_{H,T-1}, s_T), \qquad T \leq H,$$

with probability

$$P_\rho^u(Y) = \prod_{t=0}^{T-1} \pi_H^\rho(a_{H,t} \mid s_t, t, u) P(s_{t+1} \mid s_t, u_t, a_{H,t}). \tag{25}$$

For every (ρ, u), enumeration satisfies $\sum_Y P_\rho^u(Y) = 1$ to numerical tolerance.

Given a posterior b over partner preferences, the exact trajectory-level mutual information is

$$I_b(\rho; Y \mid u) = \sum_i b(\rho_i) \sum_Y P_{\rho_i}^u(Y) \log \frac{P_{\rho_i}^u(Y)}{\sum_j b(\rho_j) P_{\rho_j}^u(Y)}. \tag{26}$$

C.5 Risk, Task Value, and Robust Admissibility

Trajectory loss is

$$L(Y) = \begin{cases} 1, & \text{if the adverse terminal state is reached,} \\ 0.25, & \text{if a conflict attempt occurs without an adverse outcome,} \\ 0, & \text{otherwise.} \end{cases}$$

For each (ρ, u), the complete finite-support distribution of L is obtained from trajectory enumeration. We then compute $\mathrm{CVaR}_\alpha(L^u_\rho)$ exactly by evaluating the standard finite-support CVaR objective at its support points.

The learner's expected discounted task value under preference ρ is

$$J_\rho(u) = \mathbb{E}^u_\rho \left[\sum_{t=0}^{T-1} \gamma^t r_A(t) \right],$$

where the expectation is evaluated exactly under P^u_ρ. Define

$$J^{\mathrm{ref}}_\rho := \max_{u' \in \mathcal{U}} J_\rho(u'), \qquad C_T(u) := \max_\rho [J^{\mathrm{ref}}_\rho - J_\rho(u)]_+.$$

The robustly admissible probe set is therefore

$$\mathcal{U}(c, \epsilon) = \left\{ u \in \mathcal{U} : \max_\rho \mathrm{CVaR}_\alpha(L^u_\rho) \le c,\ C_T(u) \le \epsilon \right\}.$$

Unless otherwise stated, experiments use $\alpha = 0.90$, $c = 0.80$, and $\epsilon = 1.00$. Admissibility is always evaluated over the full preference class rather than the current posterior support.

C.6 Identification Methods and Baselines

All non-oracle identification methods use the same Bayesian posterior update and confidence-threshold stopping rule; they differ only in probe selection or the evidence retained from each encounter.

Safe-ID variants. Trajectory Safe-ID maximizes trajectory-level mutual information over $\mathcal{U}(c, \epsilon)$. One-step Safe-ID uses only the first partner response and transition for both information calculation and the corresponding one-step evidence comparison. Static-prior repeats the information-maximizing probe computed under the initial prior.

Task, random, and boundary baselines. Safe-Task maximizes posterior-expected task value; utility-weighted probing maximizes $J_b(u) + 10 I_b(\rho; Y \mid u)$; Safe-Random samples uniformly from the admissible set; Risk-Boundary selects the admissible probe with largest robust CVaR; and Active-Unconstrained maximizes trajectory information without the risk or task constraints.

Information-design baselines. Safe-Chernoff greedily maximizes, over admissible probes, the minimum trajectory KL from the current MAP preference to its competitors. Γ-stationary uses a single truth-independent randomized design maximizing the minimum KL over all ordered preference pairs. The Interactive-IRL-style baseline maximizes mutual information in the partner's immediate policy action under the same admissibility filter while retaining the full-trajectory posterior update. This baseline is an adaptation for comparison and is not intended as an implementation of a specific published algorithm.

C.7 Primary Sequential Evaluation

The primary evaluation uses 30 seeds for each of four true preferences and two initial states. Unless otherwise noted, identification is assessed at the 95% posterior-confidence threshold.

Because some methods fail to reach the confidence threshold within the evaluation horizon, uncensored stopping-time means are not treated as complete summaries when censoring exceeds 5%. In those cases we additionally report Kaplan–Meier and restricted-mean-time diagnostics.

Among baselines omitted from the main-text summary, Safe-Task and the utility-weighted method attain symmetric/asymmetric mean N_{95} values of $132.0/391.6$, with 53.3% censoring for the utility-weighted asymmetric condition, while Risk-Boundary attains $135.1/148.1$. Complete per-method results are reported with the released experiment tables.

C.8 Matched Mechanism Ablations

To isolate strategic lookahead, the *immediate-reward* ablation removes the partner's continuation value and chooses partner responses using only the current-step reward. State dynamics, learner commitments, response temperature, admissibility constraints, Bayesian update, and stopping rule are otherwise unchanged.

To isolate robust preference protection, the *posterior-average-safety* ablation changes only the admissibility test, replacing the worst-preference CVaR constraint with its posterior-weighted counterpart. These matched ablations therefore modify only the mechanism being tested.

C.9 Risk-Budget Sweep

The asymmetric-state risk frontier evaluates

$$c \in \{0.25, 0.40, 0.60, 0.80\}$$

at fixed $\alpha = 0.90$ and $\epsilon = 1.00$. For each budget we recompute the robustly admissible probe set, the corresponding discrimination rate, and the identification statistics reported in Section 5. A stricter sensitivity setting with $\alpha = 0.95$ and $c = 0.80$ leaves the asymmetric state feasible with six admissible probes but makes the symmetric state infeasible.

D Continuous Highway-Merge Scale-Up

D.1 Environment, Preferences, and Frozen Probe Family

The scale-up uses HighwayEnv `merge-v1`. Physical micro-dynamics run at 16 Hz while the learner and strategic partner act at 2 Hz. The learner announces one of 12 frozen four-step commitments over $\{\text{SLOWER}, \text{IDLE}, \text{FASTER}\}$. The merging partner uses the simulator's low-level longitudinal/lateral controller but samples its macro-response from a finite-horizon soft planner whose value depends on the entire remaining learner commitment. Four ordered partner reward vectors vary conflict aversion, near-conflict aversion, progress, delay, and courtesy.

Two pilot geometries are used before freezing the environment, preference vectors, probe family, loss/task definitions, and budgets. Confirmatory evaluation uses six disjoint geometry classes crossing which vehicle is slightly ahead with small/medium/large gaps, with 50 saved speed/position perturbations per class. Loss is 1 for collision, 0.25 for a severe near-conflict, and 0 otherwise; primary safety is worst-preference $\text{CVaR}_{0.90}$. The task budget is $\epsilon = 0.22$ and the risk sweep is $c \in \{0, 0.10, 0.20, 0.25\}$.

For each geometry, preference, and learner commitment, all $3^4 = 81$ partner macro-response sequences are enumerated through the continuous simulator. The trajectory likelihood is the product of the soft response probabilities along each branch. Exact-model checks verify normalization and recover $\Delta = \text{KL}$ and $\widetilde{\Gamma} = \Gamma$ to numerical tolerance.

D.2 Future Commitment and Safety Filtration

For every matched commitment pair with the same first learner action and different futures, we compare the partner's first-step response distributions. Averaging the 50 seeded perturbations before clustered bootstrap, the confirmatory mean/median/90th-percentile/maximum TV distances are $0.1425/0.0943/0.3088/0.8315$. The 95% bootstrap interval for the mean is $[0.129, 0.156]$, and every clustered geometry–preference–commitment-pair cell exceeds the predeclared 0.01 threshold.

The safety filtration is heterogeneous across geometry. The confirmatory summary is:

c	min $\lvert\mathcal{U}\rvert$	median $\lvert\mathcal{U}\rvert$	min Γ	median Γ
0.00	1	7.0	0.03026	0.08448
0.10	3	7.5	0.03262	0.08448
0.20	3	8.0	0.03262	0.08448
0.25	10	10.0	0.03262	0.09322

Partner-ahead geometries lose informative probes under tighter budgets, whereas several learner-ahead geometries lie on a safety-saturated plateau; the intermediate plateaus are retained rather than smoothed away.

D.3 Censoring-Aware Identification

The primary $c = 0.20$ study contains 6,000 method runs forming 1,200 matched geometry–preference–seed units. Because the episode terminates at the first posterior crossing of 0.95 and many runs do not cross by encounter 40, we treat no crossing by 40 as administrative right censoring. Any-confidence curves use Kaplan–Meier estimation. Correct and wrong first crossings are treated as competing events; with common complete follow-up to encounter 40, their cumulative incidences equal the empirical cumulative cause proportions. The restricted mean time $\mathrm{RMST}(40) = \sum_{t=0}^{39} S(t)$ equals the mean capped time-to-threshold under this administrative censoring and is not an ordinary complete-case N_{95}.

Method	Censor.	Any 95%	Correct 95%	Terminal err.	RMST(40)
Safe-ID	0.494	0.506	0.480	0.149	31.96
Safe-Chernoff	0.498	0.502	0.477	0.142	32.11
One-step Safe-ID	0.887	0.113	0.107	0.361	38.43
Safe-Task	0.812	0.188	0.180	0.237	37.50
Safe-Random	0.735	0.265	0.253	0.217	36.19

Paired bootstrap over matched experiment units gives a Safe-ID minus one-step RMST difference of -6.467 encounters $[-7.019, -5.948]$ and a correct-confidence success difference at encounter 40 of $+0.372$ $[0.344, 0.401]$. Safe-ID versus Safe-Chernoff is unresolved: the RMST difference is -0.145 $[-0.308, 0.024]$ and the correct-confidence difference is 0.003 $[-0.003, 0.010]$.

Censoring is broad but heterogeneous. For Safe-ID it ranges from 15.0% to 74.5% across the six geometries and from 23.0% to 71.0% across the four preferences. At the run level, true Γ correlates with threshold success at Spearman $\rho = 0.509$ $[0.462, 0.554]$ and with capped time at $\rho = -0.528$ $[-0.570, -0.486]$; these are descriptive associations, not a claim that Γ determines finite-horizon adaptive performance.

D.4 Natural Mismatch and Signed Evidence

The natural-mismatch study perturbs only the learner's assumed response rationality, reaction delay, or acceleration capability while leaving the true physical world, admissible set, and Γ unchanged. No preference index reversal is used. Across non-exact levels, the minimum $\widetilde{\Gamma}$ is -0.5527; at the strongest affected level, 70.8% of geometry–preference design cells have nonpositive signed rate, and up to 47.3% of positive-true-KL pair–probe cells exhibit evidence reversal.

The time-to-confidence analysis preserves wrong high-confidence crossings as competing events. Under a predeclared moderate rationality mismatch ($0.70\times$ the true response rationality), any-confidence capped time is 31.68 encounters versus 32.49 under the exact likelihood, yet correct-confidence success falls from 46.6% to 27.2%. Thus confidence can arrive at least as quickly while correct confidence becomes substantially rarer (Figure 6).

Across 240 non-exact geometry–preference–mismatch cells, $\widetilde{\Gamma}$ is much more strongly aligned with correct-confidence success than the unchanged true rate:

$$\rho(\widetilde{\Gamma}, \text{correct-confidence}) = 0.845\ [0.795, 0.885], \qquad \rho(\Gamma, \text{correct-confidence}) = 0.370\ [0.250, 0.481].$$

The paired bootstrap difference in absolute correlations is 0.475 $[0.345, 0.606]$. This is an exploratory mechanistic diagnostic, not a finite-sample optimality theorem.

D.5 Scale-Up Limitations

The highway environment provides continuous geometry and micro-dynamics but remains a two-focal-vehicle interaction. The strategic partner chooses from three speed-level macro-responses over a four-step planner, with low-level lateral merge control supplied by the simulator. The mismatch

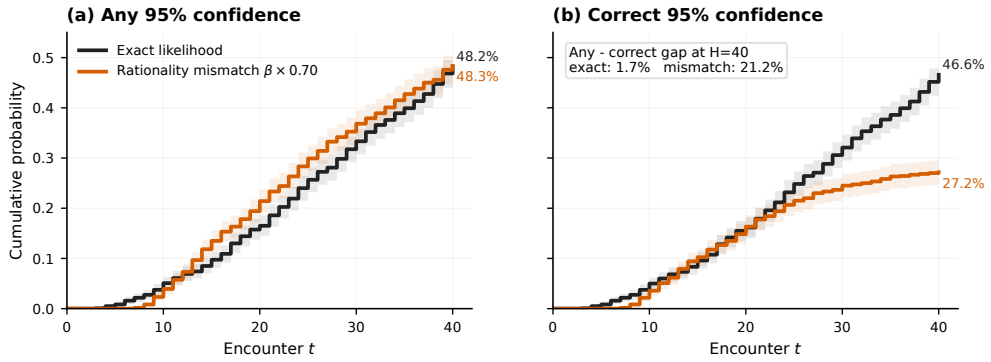


Figure 6: **Natural likelihood mismatch can preserve confidence speed while degrading its correctness.** Exact inference and the predeclared rationality mismatch $\beta_{\text{model}} = 0.70\beta_{\text{true}}$ reach any 95% confidence by encounter 40 at nearly identical rates (48.2% and 48.3%), while correct-confidence success falls from 46.6% to 27.2%. The any-minus-correct gap—wrong high-confidence cumulative incidence—therefore grows from 1.7% to 21.2%. Shaded bands are 95% experiment-unit bootstrap intervals.

study is controlled and parametric rather than a trained neural response model. These choices preserve auditable P_i^u, Γ_i, and $\widetilde{\Gamma}_i$ while leaving dense traffic, longer adaptive commitments, and learned response-model generalization for future work.

E Likelihood-Fidelity Theory Experiments

E.1 Exact Recovery and Controlled Mismatch

The exact likelihood-fidelity extension preserves the frozen horizon $H = 4$, preferences $\{0.5, 1, 2, 4\}$, $\alpha = 0.90$, $\epsilon = 1.0$, two initial states, 16 commitments, and risk budgets $c \in \{0.25, 0.40, 0.60, 0.80\}$. For each true preference, competitor, and probe, Δ_{ij} is evaluated exactly under the enumerated true trajectory law, and both Γ_i and $\widetilde{\Gamma}_i$ are solved as max–min linear programs over randomized probe distributions.

With $\widehat{P} = P$, the maximum absolute discrepancy $|\Delta_{ij} - D_{\mathrm{KL}}(P_i^u \| P_j^u)|$ is 6.661×10^{-16}, the maximum relative error is 5.270×10^{-15}, and the maximum $|\widetilde{\Gamma}_i - \Gamma_i|$ is 2.082×10^{-17}; every nonempty LP succeeds with zero recorded constraint violation. Controlled likelihood-only misspecification then varies a continuous preference-index warp, response rationality, contamination, and model collision dynamics while retaining every tested level, including levels with no evidence reversal. Across the full sweep, 115 state/budget/preference/model cells satisfy $\Gamma_i > 0$ but $\widetilde{\Gamma}_i \leq 0$.

At $c = 0.8$, the preference-warp sweep keeps the minimum true rate fixed at 0.03488 in both initial states. The minimum learned-evidence rate is 0.00406 at warp 0.2 and becomes -0.00455 (symmetric) and -0.00060 (asymmetric) at warp 0.4. At warp 0.6 and 0.8, the fraction of positive-KL pairs with negative signed evidence reaches 0.8333 in both states. Rationality mismatch also yields strongly negative individual evidence; for example, in the asymmetric state with model rationality 6, one $AAWW$ pair has true KL 0.04685 but learned evidence -0.49183.

E.2 Stationary Posterior-Drift Validation

The stationary validation samples probes i.i.d. from one frozen Γ-optimal design, generates responses from the true exact model, and updates posterior odds with the learned model. No adaptive Safe-ID trajectory is used for this theorem check. With 40 seeds per case:

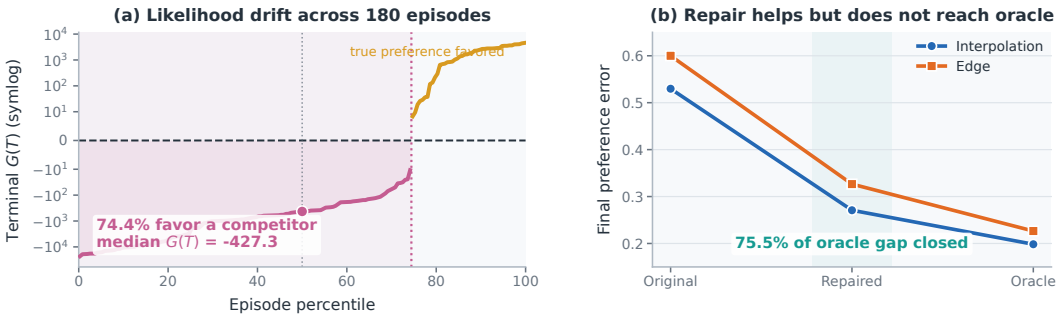


Figure 7: Continuous learned-likelihood diagnostics retained as supporting evidence. (a) Terminal learned log-likelihood ratios are negative in 74.4% of the original frozen diagnostic episodes despite collapsed posterior uncertainty. (b) Validation-only repair reduces final preference error and closes 75.5% of the original-to-oracle gap, but does not reach oracle performance.

Case	Theory rate	Empirical mean	Empirical std.
Exact positive	0.57535	0.57952	0.02554
Degraded positive	0.09253	0.09227	0.00416
Reversed	−0.57535	−0.56583	0.03085

The reversed case therefore validates not merely attenuation but a change in the sign of asymptotic learned posterior drift.

E.3 Likelihood-Fidelity Margin and CVaR Calibration

In the finite soft-policy game every relevant true-support trajectory has positive learned probability, so the uniform log-density error η is computable exactly. The sufficient condition $\Gamma_i > 2\eta$ is used only as a sufficient condition: failure of the margin is never interpreted as evidence reversal. In practice the supremum is conservative because rare trajectories can dominate η; 99th-percentile log error is therefore retained only as a diagnostic and is not substituted into the theorem.

For the model-safety bound, all exact hypothesis/probe/model rows are audited. The inequality

$$\left|\mathrm{CVaR}_\alpha^{P_i^u}(L) - \mathrm{CVaR}_\alpha^{\widehat{P}_i^u}(L)\right| \le \frac{L_{\max}}{1-\alpha} D_{\mathrm{TV}}(P_i^u, \widehat{P}_i^u)$$

has zero violations over 2,304 rows (largest numerical excess 0).

E.4 Continuous Repair as an Observed-Design Diagnostic

The continuous learned-response environment does not provide a fixed finite probe family or an enumerable true P_i^u, so we do not report a rigorous $\widetilde{\Gamma}$. Instead, Original, Repaired, and Oracle likelihoods are evaluated on the same 360 frozen Safe-Random trajectories per held-out regime against each true preference's nearest held-out competitor:

Model	Regime	Mean signed evidence	Reversal fraction
Oracle	Edge	0.751	0.006
Original	Edge	−0.572	0.489
Repaired	Edge	0.099	0.114
Oracle	Interpolation	0.220	0.011
Original	Interpolation	0.940	0.375
Repaired	Interpolation	0.029	0.211

The reduction in reversal is consistent across the two regimes, whereas mean signed evidence is non-monotone; this diagnostic is therefore intentionally kept distinct from the finite-design rate theorem.

E.5 Stronger Exact-Game Baselines

The new information-design baselines are evaluated on the same frozen seeds, admissible sets, Bayesian updates, and stopping rule as Safe-ID. Point estimates are:

Method	Symmetric N_{95}	Asymmetric N_{95}
Safe-Chernoff	84.1	99.5
Safe-ID	85.7	98.0
Interactive-IRL-style	102.7	109.1
One-step Safe-ID	102.7	109.1
Γ-stationary	116.5	121.5
Static-prior	135.1	148.1
Safe-Random	108.6	205.5
Active-Unconstrained	25.6	26.5

Paired Safe-ID-minus-Safe-Chernoff differences are 1.542 $[-2.310, 5.675]$ in the symmetric state and -1.475 $[-8.492, 5.967]$ in the asymmetric state, so both intervals include zero. The Γ-stationary design is significantly slower in both states, while the Interactive-IRL-style rule reproduces the one-step Safe-ID aggregate results. These comparisons are interpreted as criterion tradeoffs, not a ranking or a global-optimality claim.