SOCIAL INFERENCE IS NOT ENOUGH: A DECISION-VALUE FRAMEWORK FOR DIAGNOSIS AND INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In strategic interactions, more accurate inference about an opponent does not necessarily translate into better decisions or higher payoffs. We develop a value-guided social inference framework to diagnose where this connection holds or breaks down. Building on value of information, the framework distinguishes three payoff comparisons relative to a prior-informed baseline: potential value, the improvement available from perfect knowledge; acquired value, the expected improvement enabled by an evidence procedure; and realized value, the informed policy's advantage in deployment, where evidence and opponent responses may differ from the source model. We characterize when remaining uncertainty has no decision value and connect payoff-relevant representations to cost-sensitive evidence selection and stopping. Controlled 2×2 games isolate the absence of opportunities for better decisions, difficulty acquiring decision-relevant evidence, and loss of predicted gains under response mismatch. In two negotiation protocols with large language model agents, targeted interventions improve net return in the reported diagnostic comparisons by avoiding redundant identification, choosing cost-effective evidence, and verifying consequential but unreliable claims. Together, these results support a diagnostic account of how social inference contributes to interaction outcomes and illustrate how that account can guide what an agent learns, how it acts, and when it reassesses its model of the interaction.

1 INTRODUCTION

In game-theoretic interaction, an agent's outcomes depend on both its own choices and those of others. Negotiating an agreement, coordinating a joint plan, or competing for a resource involves uncertainty about other participants' preferences, intentions, and likely responses. We refer to the inference of these unobserved attributes from observed evidence, such as behavior and communication, as **social inference**. A decision-centered view asks which uncertainties affect the agent's choices, what evidence would be useful, and when the agent has enough information to act.

Existing research evaluates social inference—whether framed as Theory of Mind (ToM), opponent modeling, or intention recognition—through prediction accuracy and interactive performance (Albrecht & Stone, 2018; Kim et al., 2023; Chen et al., 2024; Riemer et al., 2025). Yet prediction accuracy or representational richness alone does not determine decision quality. The resolved uncertainty may be irrelevant to the available choices, evidence may leave the optimal policy unchanged, or predicted gains may fail in deployment. These cases call for different interventions: act now, seek more useful evidence, or revise the interaction model. Where, specifically, does the contribution of social inference become limited?

Value of information (VoI) assesses information through the decisions it enables (Blackwell, 1953; Raiffa & Schlaifer, 1961; Howard, 1966), including what to ask and when to act (Mohammad & Nakadai, 2019; Dong et al., 2026). Applying it to social inference requires a fixed policy class and baseline, representations that preserve decision-relevant payoff differences, and evaluation of each policy under the opponent responses it induces. We therefore develop a **value-guided social inference framework** organized around **potential**, **acquired**, and **realized value** (Figure 1).

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.

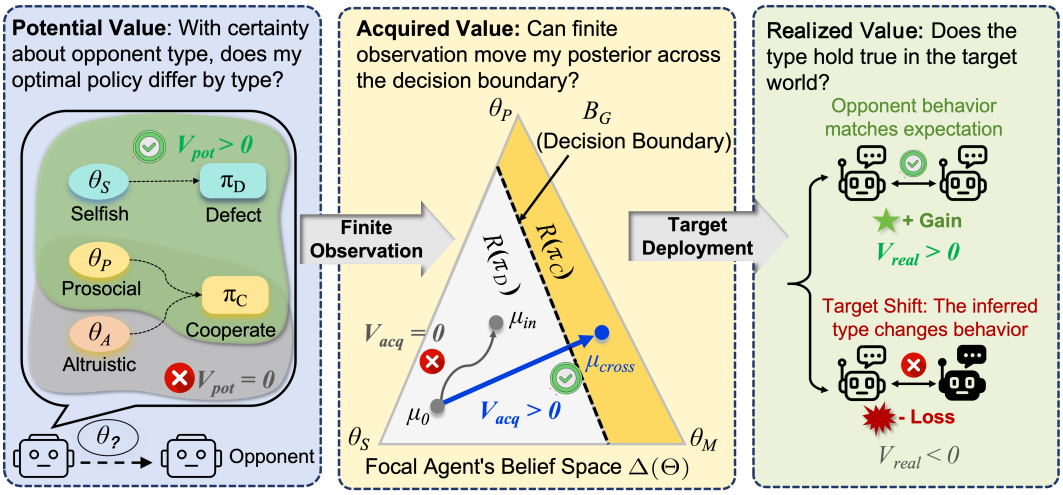


Figure 1: **Three comparisons of social-inference value. (1) Potential:** perfect type information helps only when supported types lack a common optimal policy, as for (θ_S, θ_P); (θ_P, θ_A) share π_C and have zero potential. **(2) Acquired:** evidence moves the belief from μ_0 to possible posteriors. Updates within $R(\pi_D)$ leave the decision unchanged; crossing B_G can enable improvement. Acquired value averages these improvements over evidence outcomes. **(3) Realized:** the predicted gain is retained when the target model matches the source. Changed type-to-response mappings can reduce or reverse it.

Potential value measures how much perfect knowledge could improve on the best decision available under the current belief. For example, in a one-shot game where the same focal action is optimal against every supported opponent type, further type information has zero potential value. **Acquired value** measures how much of the available opportunity a particular evidence channel (e.g., key points to communicate, questions to ask, opponent actions to observe) captures in expectation. For example, if cooperation becomes optimal only above a threshold belief in the opponent's cooperativeness, evidence that never moves the posterior across that boundary yields no acquired value. **Realized value** measures the informed policy's advantage in deployment, where the evidence distribution or the opponent's responses may differ from those anticipated by the model.

These comparisons guide what to learn, when to act, and when to reassess the interaction model. We develop them for partially observed interaction and examine controlled 2×2 games and language-mediated negotiations.

Our contributions are threefold. **(i) Diagnostic framework.** We define three consistent payoff comparisons for examining where information contributes to decisions, providing a common language for identifying failure modes of social inference. **(ii) Design criteria.** We characterize when resolving uncertainty improves decisions and how evidence acquisition and interaction responses affect that benefit, yielding criteria for acquisition and stopping. **(iii) Controlled evidence.** We conduct experiments to examine distinct bottlenecks in games and negotiations and test the corresponding interventions, establishing concrete settings in which adjusting the appropriate part of the social inference process improves decision performance.

2 Related Work

Social inference and interactive performance. Computational social inference includes inverse planning, belief–desire inference, and learned behavioral models (Baker et al., 2009; 2011; Rabinowitz et al., 2018). Mental-state benchmarks examine inference accuracy and robustness (Le et al., 2019; Kim et al., 2023; Chen et al., 2024; Ullman, 2023), while interactive evaluations examine the benefits of mental-state modeling in collaboration (Sclar et al., 2022; Li et al., 2023). Functional evaluation asks whether these capacities improve interaction (Riemer et al., 2025). Opponent modeling and interactive POMDPs connect latent models to control (Gmytrasiewicz & Doshi, 2005; Albrecht & Stone, 2018; He et al., 2016; Yu et al., 2022); recent LLM agents implement such models through revisable natural-language hypotheses or executable programs (Cross et al., 2025; Jha et al., 2026).

Building on this connection, we organize the payoff contribution of social inference into potential, acquired, and realized value.

VoI, acquisition, and metareasoning. Bayesian decision theory defines perfect and sample information by optimal expected utility (Raiffa & Schlaifer, 1961; DeGroot, 1970; Howard, 1966). Information gain and decision value measure different objectives (Lindley, 1956; Blackwell, 1953), and beliefs induce action regions in partially observable Markov decision processes (POMDPs) (Smallwood & Sondik, 1973; Kaelbling et al., 1998). Selecting computations and deciding what to learn connect acquisition to downstream objectives (Hay et al., 2012; Arumugam & Van Roy, 2021). In negotiation, Mohammad & Nakadai (2019) study eliciting a represented user's utility, and Dong et al. (2026) balance clarification benefits against user costs. Our acquisition criterion follows VoI theory and places it within a diagnostic framework that also examines the available opportunity and the informed policy's deployed return. We also use Naive VoI as a baseline in the experiments of Section 6.

Decision-focused prediction and interaction feedback. Decision-focused learning and predict-then-optimize losses train predictions for downstream decisions (Wilder et al., 2019; Elmachtoub & Grigas, 2022). We study evaluation quantities and interventions that connect inference to interaction outcomes. Opponent-learning awareness models adaptation (Foerster et al., 2018), performative prediction examines how deployment changes outcome distributions (Perdomo et al., 2020), and hypergame theory considers mismatched perceptions and nested beliefs (Trencsenyi et al., 2025). These perspectives motivate evaluating informed and baseline policies under their own induced responses. Our comparison accommodates fixed response rules and policy-dependent reactions. Appendix F discusses preference models, reasoning depth, and repeated interaction. Negotiation platforms such as NegMAS provide related alternating-offer mechanisms and negotiator abstractions (Mohammad et al., 2020); our controlled study uses the finite protocol specified in Appendix D.2.

3 A Value-guided Framework for Social Inference

3.1 Decision Setting and Baseline

Let $\theta \in \Theta$ denote unknown opponent characteristics, such as preferences, beliefs, or response rules; we use *type* as shorthand for their value. Types may be discrete, continuous, or mixed. Let Π be the focal policy class, with each $\pi \in \Pi$ a single action or complete contingent plan. The bounded expected focal return $u(\pi, \theta)$ includes modeled opponent responses and averages over remaining interaction randomness. This reference is the *source model* when compared with deployment. For the current belief μ over Θ, define

$$U(\pi, \mu) = \mathbb{E}_\mu[u(\pi, \theta)], \quad W(\mu) = \max_{\pi \in \Pi} U(\pi, \mu), \quad \pi^\star(\mu) \in \arg\max_{\pi \in \Pi} U(\pi, \mu). \tag{1}$$

Here U is belief-averaged return, W its optimum, and $\pi^\star$ a selected optimizer. We assume measurable returns and optimizers, attained maxima, and a fixed tie rule (Appendix A.1). The *prior-informed baseline* is $\pi_0 = \pi^\star(\mu_0)$ for initial prior μ_0; at a later belief μ, it is $\pi^\star(\mu)$. All comparisons hold the objective, policy class, and reference model fixed.

An evidence action q specifies an observation procedure with outcome $Y \sim L_q(\cdot \mid \theta)$ and posterior $\mu^{q,y}$ after observing y. Before observation, $\mu^{q,Y}$ is random, and $\mathbb{E}_Y$ averages over the source predictive law: draw $\theta \sim \mu$, then $Y \sim L_q(\cdot \mid \theta)$. Adaptive procedures are included by taking Y as the complete observation history. We first consider *pure information*, which changes beliefs but leaves the subsequent physical state, policy class, and payoff model fixed. These values are gross gains before acquisition costs; Section 4.2 adds costs and actions that change time, available contracts, or responses.

3.2 Potential Value: Opportunity for Better Decisions

Potential value at belief μ is the expected gain from learning the opponent characteristics θ exactly before selecting a policy:

$$V_{\mathrm{pot}}(\mu) = \mathbb{E}_\mu[\max_{\pi \in \Pi} u(\pi, \theta)] - W(\mu). \tag{2}$$

The first term permits a different optimal policy for each realized θ; $W(\mu)$ evaluates a policy chosen under the current uncertainty. Their difference bounds the gain available from resolving that uncertainty within the specified decision problem.

Proposition 1 (When Potential Value Is Zero)**.** *Under the assumptions above, $V_{\mathrm{pot}}(\mu) = 0$ if and only if there is a single policy $\bar{\pi} \in \Pi$ such that*

$$u(\bar{\pi}, \theta) = \max_{\pi \in \Pi} u(\pi, \theta) \quad \textit{for } \mu\textit{-almost every } \theta. \tag{3}$$

Equivalently, the current optimal policy $\pi^\star(\mu)$ has this property.

The probability-one condition covers discrete and continuous characteristics: a diffuse posterior can still support a common optimal policy. Appendix A.2 gives the proof and Appendix A.3 continuous examples; Section 4.1 examines the payoff distinctions a useful representation preserves.

3.3 Acquired Value: Expected Gains from Evidence

Acquired value evaluates an evidence procedure before its outcome is observed:

$$V_{\mathrm{acq}}(q; \mu) = \mathbb{E}_Y[W(\mu^{q,Y})] - W(\mu). \tag{4}$$

After observing y, the gain from revising the policy is evaluated under the updated belief, using the policy selected before that observation as the comparator: $W(\mu^{q,y}) - U(\pi^\star(\mu), \mu^{q,y})$. Averaging this conditional improvement over the source predictive distribution of Y yields $V_{\mathrm{acq}}(q; \mu)$. Potential and acquired value correspond to established perfect- and sample-information quantities (Raiffa & Schlaifer, 1961; DeGroot, 1970). They satisfy

$$V_{\mathrm{pot}}(\mu) = V_{\mathrm{acq}}(q; \mu) + \mathbb{E}_Y[V_{\mathrm{pot}}(\mu^{q,Y})]. \tag{5}$$

The procedure captures part of the initial opportunity and leaves the remainder as expected posterior potential. Hence $0 \le V_{\mathrm{acq}}(q; \mu) \le V_{\mathrm{pot}}(\mu)$. For a fixed nested observation process, let the nonnegative integer n be the observation budget, and write $V_{\mathrm{acq}}(n)$ for the acquired value of its first n observations at a fixed starting belief. This expected gross value is nondecreasing in n. Proofs are in Appendix A.2.

3.4 Realized Value: Gains in Deployment

For starting belief μ, fix the source-selected rule $\pi_q(y; \mu) = \pi^\star(\mu^{q,y})$. Let $u_{\mathrm{tar}}(\pi, \theta)$ be expected focal return when π faces its own induced opponent response in the target environment, averaging over remaining interaction randomness. Let $\mathbb{E}_{\mathrm{tar}}$ average over the target joint law of types θ and outcomes Y. The fixed source rule must specify an action for every target history (Appendix A.7). *Realized value* is

$$V_{\mathrm{real}}(q; \mu) = \mathbb{E}_{\mathrm{tar}}[u_{\mathrm{tar}}(\pi_q(Y; \mu), \theta) - u_{\mathrm{tar}}(\pi^\star(\mu), \theta)]. \tag{6}$$

Thus μ specifies the source belief used to select both policies, while the target law specifies the types and observations used to evaluate them. Realized value equals acquired value when that joint law and both policies' returns agree with the source model. Under mismatch, realized value can be lower, negative, or higher. Net deployed gain also subtracts the difference in acquisition costs, with other interaction costs accounted for consistently.

Evidence mismatch concerns how observations inform beliefs; response mismatch concerns how deployed policies produce returns. These suggest interventions such as independent verification, reliability updates, and response-model recalibration. Appendix A.7 expands the response-dependent comparison and relates it to a contrast holding the opponent response fixed.

4 Guiding Inference with Decision Value

4.1 Decision-Relevant Uncertainty

For finite Π, decision boundaries are ties among policies with maximal posterior expected return. Useful representations preserve policy payoff differences, including their magnitudes, so that beliefs over types support accurate policy comparisons. Appendix A.3 gives a sufficient representation

and an example showing how oracle-action labels can discard relevant payoff magnitudes. For two candidate policies $\pi_a, \pi_b \in \Pi$, the signed margin

$$\Delta(\mu) = \mathbb{E}_\mu[u(\pi_a, \theta) - u(\pi_b, \theta)]$$

measures the expected payoff difference between the policies and locates the belief relative to their decision boundary.

The value of a candidate channel depends on the policy improvements its possible observations enable. Using the current optimum $\pi^\star(\mu)$ as the comparator, Eq. (4) becomes

$$V_{\mathrm{acq}}(q; \mu) = \mathbb{E}_Y \left[\max_{\pi \in \Pi} \mathbb{E}_{\mu^{q,Y}} [u(\pi, \theta) - u(\pi^\star(\mu), \theta)] \right]. \tag{7}$$

This combines each observation's probability with the resulting payoff improvement. For discrete policies and a fixed tie rule, it also equals switch probability times mean improvement conditional on switching. Switching between tied policies gives zero improvement. With continuously varying policies, the same comparison evaluates gradual changes in the selected policy (Appendix A.5).

4.2 Evidence Selection and Stopping

Let $\mathcal{Q}$ be the set of feasible evidence actions, $c(q)$ the acquisition cost of action q expressed in the same units as return, and stop the option to act with the current information. With at most one pure-information acquisition opportunity, maximizing expected net return gives the one-step value-of-information rule (Hay et al., 2012, cf. Definition 6):

$$q^\star \in \operatorname*{arg\,max}_{q \in \mathcal{Q} \cup \{\mathrm{stop}\}} \{V_{\mathrm{acq}}(q; \mu) - c(q)\}, \qquad V_{\mathrm{acq}}(\mathrm{stop}; \mu) = c(\mathrm{stop}) = 0. \tag{8}$$

The calculation enumerates or samples each channel's outcomes, updates the belief, and averages the best downstream payoff. Its quality depends on both the channel likelihoods and the estimated policy returns. With several acquisition opportunities, continuation values account for complementary observations whose joint benefit exceeds their individual benefits (Hay et al., 2012). Appendix C.2 gives the recursion and a two-signal example.

Potential value supplies an upper bound on pure-information benefit. If every nonempty acquisition procedure has expected cost at least an upper bound on remaining potential, stopping is weakly optimal among those procedures. More generally, cost-sensitive stopping compares the expected net value of feasible continuations with acting now. *Decision completion* occurs when expected remaining potential is zero, or below a specified tolerance. The posterior can still retain uncertainty among types sharing an optimal policy. Cost-sensitive stopping can occur before decision completion while positive potential remains. The observation budget needed to approach completion depends on the evidence process (Appendix A.4).

Let s denote the current physical state and $J(s, \mu)$ the optimal expected net return from continuing interaction and acquisition at state s with belief μ. When an information action changes the physical state, denote the resulting state and updated belief by s' and μ'. Its continuation score is $-c(q) + \mathbb{E}[J(s', \mu')]$, with expectation over the next states and observations induced by q, compared with acting now. State s' captures elapsed time, available contracts, and changed responses. Verification consumes a decision in our negotiations; the standing offer remains available. Other protocols may also revoke offers.

4.3 Diagnosing Gaps in Decision Value

Fix the objective, policy class, initial state, and baseline, then ask: *Potential:* would perfect information improve the decision? Low potential limits the benefit of further identification. *Acquired:* which channel offers enough improvement to cover its cost? This guides what to ask or verify. *Realized:* does the informed policy retain its advantage when each policy faces its own target responses? A gap motivates checking the evidence likelihoods and policy-return model.

The comparisons can reveal several bottlenecks or failure modes in the same interaction. Implementation checks can additionally examine whether the belief updater and planner use the available evidence effectively, for example by comparing alternatives at a common public history (Appendix A.6).

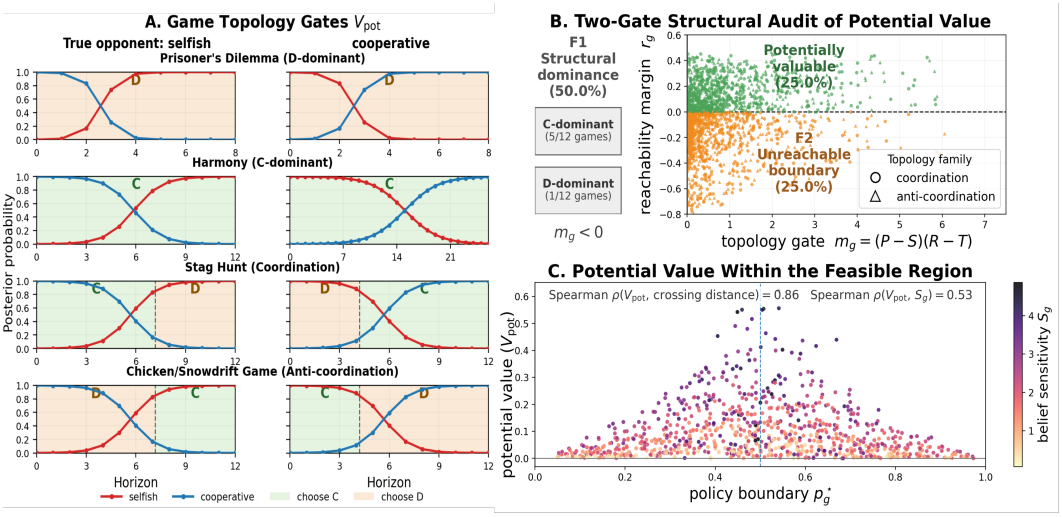


Figure 2: **Potential value depends on the decisions information can change. A:** Illustrative belief paths and focal best-response regions in four game structures; paths use mean log-evidence accumulation. **B:** Dominance (F_1) and unreachable boundaries (F_2) separate two reasons for zero potential. **C:** Potential value within the feasible region varies with boundary geometry and payoff sensitivity. Sampling, axis definitions, and numerical summaries are given in Appendix C.1.2.

Matched interventions strengthen the diagnosis. Comparing unobserved policy outcomes requires simulator access, randomized trials, or justified counterfactual estimation; a single observed trajectory usually leaves several explanations open (Appendix A.8).

5 Controlled Games: Diagnosing Limits to Social Inference

Strict ordinal 2×2 games make the three diagnostics directly auditable. The focal agent chooses C or D under uncertainty about type-dependent opponent cooperation. We examine decision opportunity, finite-evidence acquisition, and deployment mismatch; Appendix C.1 specifies the protocols, formulas, and numerical summaries.

Potential: When Further Identification Cannot Help. Figure 2A illustrates Proposition 1: beliefs can become more concentrated while the same action remains optimal against every supported opponent type. In a one-shot Prisoner's Dilemma evaluated by material payoff, for example, identifying a cooperative opponent does not remove the incentive to defect.

Across 12 canonical symmetric game structures, zero potential arises either from a dominant action or from supported types lying on one side of an existing boundary (Figure 2B). Both admit a common optimal action. When support spans the boundary, potential depends on the payoff improvements from type-specific decisions (panel C). Thus zero potential justifies acting now under nonnegative pure-information costs. The plotted proportions describe the specified sampling scheme, not the prevalence of these limitations in social interaction.

Acquired: When Evidence Has Yet to Improve the Decision. Positive potential leaves open whether the available observations can capture it. Figure 3A varies observation noise and belief-update temperature in a fixed acquisition task and reports the number of observations required for the first change in the selected policy. Panel B instead fixes exact Bayesian conditioning ($T_B = 1$) and reports the decision-value quantity $\eta_{\mathrm{acq}}(n) = \frac{V_{\mathrm{acq}}(n)}{V_{\mathrm{pot}}(\mu_0)}$, computed by exact finite integration. Increasing observation noise slows the rate at which the available potential value is captured.

First-switch time measures neither acquired value nor completion: a switch may offer little gain, and later evidence may still help. Further acquisition is justified by expected gains after costs (Section 4.2), not by positive potential alone. These fixed-channel plots illustrate acquisition difficulty without testing an optimal query or stopping policy.

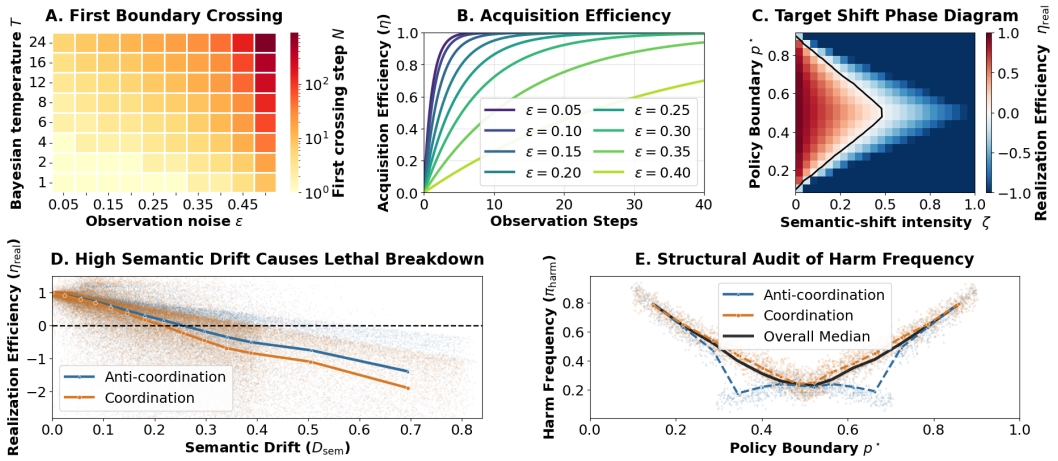


Figure 3: **Acquiring useful evidence and retaining its benefit are distinct requirements. A:** First-switch counts under observation noise and update temperature. **B:** Acquisition efficiency $\eta_{\text{acq}}(n) = V_{\text{acq}}(n)/V_{\text{pot}}(\mu_0)$, computed by exact finite integration. **C–E:** Deployment/source gain ratios and harm frequencies under prescribed changes in type-conditioned opponent behavior (*semantic shift*), with focal payoffs and policies fixed. Negative ratios indicate underperformance relative to the prior-informed baseline. Appendix C.1.3 gives generation and aggregation procedures for these descriptive ensemble summaries.

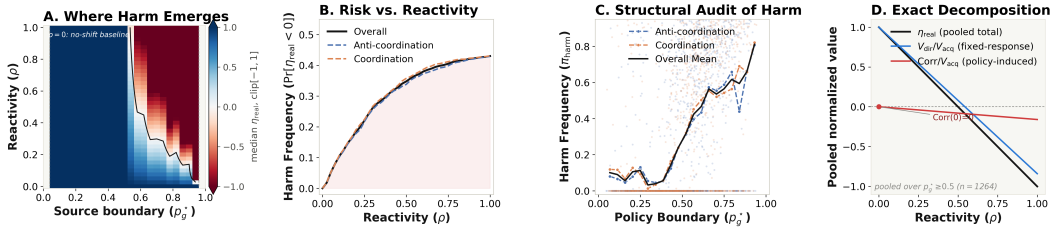


Figure 4: **Realized value under policy-dependent responses. A:** Median η_{real} by source boundary $p_g^\star$ and reactivity ρ; $\rho = 0$ is source-matched. **B:** Population harm frequency $\Pr_{\text{games}}[\eta_{\text{real}}(\rho) < 0]$. **C:** For each game, the fraction of $\rho \in [0, 1]$ giving negative realized value, by boundary and topology. **D:** Fixed-response term $V_{\text{dir}}/V_{\text{acq}}$ and response correction $\text{Corr}/V_{\text{acq}}$, where $\text{Corr} = V_{\text{real}} - V_{\text{dir}}$, pooled over $p_g^\star \geq 0.5$. Definitions and statistics appear in Appendices A.7 and C.1.4.

Realized: When the Predicted Advantage Fails in Deployment. Figure 3C–E changes the opponent behavior associated with each type while keeping the source observation process and selected policies fixed. The informed policy can then lose its source advantage and underperform the prior-informed baseline. This instantiates the response mismatch in Section 3.4: evidence can support a useful source decision while its predicted payoff consequences fail to hold in deployment.

Sensitivity to shift varies with the source decision boundary, and similar drift magnitudes can yield different payoff effects. The relevant diagnostic is therefore the change in the policies' relative returns, not behavioral drift alone (Appendix C.1.3).

Figure 4 adds policy-dependent responses: the opponent's cooperation probability is unchanged under D but multiplied by $1 - \rho$ under C, for reactivity $\rho \in [0, 1]$. Unlike exogenous shifts, this can produce a nonzero response correction in Eq. (34). With source policies fixed, the deployment/source gain ratio $\eta_{\text{real}} = V_{\text{real}}/V_{\text{acq}}$ is affine in ρ, and $\rho = 0$ recovers acquired value. Harm concentrates in, but is not confined to, high-boundary games. Numerical checks and subgroup statistics are in Appendix C.1.4.

These losses motivate validating the type-to-behavior mapping and reevaluating policy returns; further source observations alone do not validate that mapping. The games isolate mismatch without testing a learned repair. We next examine evidence and action interventions in negotiation.

Table 1: Net return by parameter-based diagnostic group and protocol. Entries report Full / the named baseline, with the paired mean difference in parentheses. Inefficient acquisition compares evidence-channel choice in final-offer negotiation and evidence acquisition versus immediate acceptance in standing-offer negotiation.

Protocol	Diagnostic group	Baseline	DeepSeek-V4-Pro	GPT-6-Sol
Final offer	Redundant identification	Prediction-first	$6.00/5.10$ $(+0.90)$	$6.00/5.10$ $(+0.90)$
	Inefficient acquisition	No language	$6.90/6.40$ $(+0.50)$	$6.90/6.40$ $(+0.50)$
	Misplaced trust	Naive VoI	$7.40/4.00$ $(+3.40)$	$7.40/4.00$ $(+3.40)$
Standing offer	Redundant identification	Prediction-first	$6.40/5.10$ $(+1.30)$	$6.40/5.10$ $(+1.30)$
	Inefficient acquisition	Aspiration	$6.16/3.20$ $(+2.96)$	$5.96/3.20$ $(+2.76)$
	Misplaced trust	Naive VoI	$7.23/3.40$ $(+3.83)$	$7.23/3.40$ $(+3.83)$

6 Diagnosing and Improving LLM Negotiation Decisions

We study bilateral negotiation over three indivisible items, where an LLM agent seeks an acceptable allocation under uncertainty about its opponent's private item values and can acquire evidence through costly questions or exact-value verification. We evaluate final-offer and standing-offer protocols; the final-offer implementation uses NegMAS (Mohammad et al., 2020) to execute proposals and utility-based acceptance. We examine how value diagnostics change the agent's evidence and contract choices. Negotiation instances are randomly sampled and then grouped by preference supports, contract payoffs, evidence channels and costs, and disclosure settings. See Appendix D and E for detailed environment settings.

6.1 From Failure Diagnosis to Agent Design

We examine three failure modes within the parameter-based groups. **Redundant identification** occurs when a baseline purchases further type information although the supported types share an optimal contract. **Inefficient acquisition** occurs when a baseline forgoes useful evidence or pays more to resolve the same contract-relevant distinction. **Misplaced trust** occurs when a baseline treats a misleading preference claim as sufficiently reliable to justify an unnecessary concession. Group assignment uses the instance parameters; trajectories are used to analyze the associated behavior, and improvement by our agent is evaluated separately. The labels identify the decision-value limitation examined by a comparison, not an intrinsic opponent type.

Value-guided agent. Our agent, *Full*, applies Section 4: stop when remaining potential cannot justify acquisition, select evidence by expected net continuation value, and reassess reliability when it affects contract choice. Both protocols divide three items under a public finite support of opponent preferences. The LLM extracts language evidence, estimates reliability and answerability, proposes questions, and selects and executes actions. A deterministic numerical module performs finite Bayesian conditioning and evaluates legal contracts, information actions, potential and acquired value, and continuation scores. The study tests whether these diagnostics improve an LLM controller's choices, rather than benchmarking unaided LLM planning.

Protocols and comparisons. In *final-offer negotiation*, evidence acquisition precedes one terminal proposal. In *standing-offer negotiation*, an incumbent agreement can be accepted immediately or revised after acquisition through a terminal counteroffer. Prediction-first supplies the identification-focused comparison; No language removes dialogue and questions but retains verification; Aspiration accepts a standing offer that meets its threshold; and Naive VoI uses the same VoI evaluation procedure and evidence actions as Full while holding linguistic reliability at $r_t = 0.97$. Table 1 states the comparator for every diagnostic group. Both protocols are evaluated with DeepSeek and GPT focal agents. Appendices D and E specify the sampling, grouping, evaluation units, and execution checks in detail.

6.2 Decision Improvements and Their Critical Steps

Full achieves higher mean net return than the matched baseline in each reported diagnostic comparison (Table 1). The trajectories explain how these gains arise: the intervention changes a stopping decision,

an evidence channel, or a concession, rather than uniformly increasing inference effort. We write $X \mid Y$ for an allocation giving items X to the focal agent and Y to the opponent.

Redundant identification: act with sufficient information. A common optimal contract gives zero potential despite type uncertainty (Proposition 1). Full commits, whereas Prediction-first pursues its type-confidence objective. In a representative final-offer trajectory, Full immediately proposes $A \mid BC$; Prediction-first verifies two values before proposing the same allocation. Likewise, Full accepts a decision-sufficient standing offer without identifying the exact type. Gains come entirely from lower acquisition cost, showing why sufficiency for acting should be assessed at the contract level.

Inefficient acquisition: resolve the contract distinction at acceptable cost. Supported preferences favor different contracts, giving positive potential. In final-offer negotiation, Full asks whether A is worth at least five, resolving the contract-relevant distinction more cheaply than No language's exact verification. It stops when further acquisition no longer pays despite residual type uncertainty. In standing-offer negotiation, Aspiration accepts $B \mid AC$ at its threshold; Full instead verifies A, observes seven, and selects $BC \mid A$, recovering C while meeting the opponent's reservation commitment. The net-value criterion thus improves channel choice in one protocol and prevents premature acceptance in the other. One GPT trial verifies but retains the incumbent; it remains in the results, separating evidence acquisition from implementation of the available improvement (Appendix A.6).

Misplaced trust: verify a consequential claim. A misleading claim that A is important sustains the concession $B \mid AC$ under Naive VoI. Full retains enough source uncertainty for independent verification to have positive net value. Observing A's true value of two supports $AB \mid C$, recovering an unnecessarily conceded item while preserving acceptance. Naive VoI has the same verification action, but its high reliability assumption makes verification appear unprofitable. The contrast therefore concerns the evidence model, not audit access or a changed acceptance rule (Section 3.4). Response-threshold controls show that Full does not verify the unchanged claim when both types share an optimal contract (Appendix D.5). Verification pays only when source uncertainty changes the payoff comparison; Appendix D.6 derives the threshold for the displayed context.

The gains arise from changing a stopping decision, an evidence channel, or a concession. Their scope depends on channel availability, source assessment, and execution of the selected intervention; Appendices D.6 and E provide the detailed trajectories and controls.

7 Conclusion

Social inference improves interaction through the decisions it enables, the costs it incurs, and the conditions under which its benefits hold. Our potential, acquired, and realized value comparisons locate limitations in the decision problem, evidence process, and deployment model. Controlled games isolate these limitations; two LLM negotiation protocols show that corresponding interventions can improve net return in matched diagnostic comparisons by avoiding redundant identification, selecting cost-effective evidence, and verifying consequential claims. Broader application requires reliable value estimates and credible evaluation of alternative policies in deployment. More broadly, our work provides a decision-centered perspective for evaluating social intelligence and designing interactive agents, shifting the focus from how much an agent knows about others to how effectively it determines what to learn, when to act, and when to revise its understanding of the interaction.

Limitations. The framework is diagnostic and model-relative: it introduces neither a new opponent-modeling architecture nor guarantees of improved interaction. Conclusions depend on the prior, policy class, evidence model, payoffs, and acquisition costs. Controlled games and finite negotiations with known response rules and verification support mechanism-specific claims, not failure-prevalence estimates or general agent rankings. Generalization to open-ended, long-horizon, multi-party, or human interaction remains untested. Learning decision-sufficient abstractions and estimating their value at scale remain open, since our experiments make payoff-relevant distinctions explicit. Finally, value follows the focal objective; higher return does not imply greater fairness, collective welfare, social understanding, or social intelligence.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, improve the manuscript's readability and organization, ensure consistent technical terminology, simplify and condense sentences, identify relevant literature, and assist with reference formatting and LaTeX preparation. Large language models were also used as components of the negotiation agents evaluated in this study. Their roles, inputs, model configurations, and interaction protocols are described in Section 6 and Appendices D and E. We have reviewed all AI-assisted work. In particular, we checked and corrected AI-assisted descriptions of the experimental procedures to ensure that they accurately reflected the actual study protocol. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices A and B state the assumptions and provide proofs and worked instantiations. Appendix C documents the matrix-game protocols and details enabling reproduction. Appendices D and E specify the negotiation environments, evidence channels, method objectives, model settings, evaluation units, parameter-based grouping rules, and recorded execution checks.

REFERENCES

Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018. doi: 10.1016/j.artint.2018.01.002.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 373–382. PMLR, 2021. URL https://proceedings.mlr.press/v139/arumugam21a.html.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009. doi: 10.1016/j.cognition.2009.07.005.

Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 2469–2474. Cognitive Science Society, 2011. URL https://escholarship.org/uc/item/5rk7z59q.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. doi: 10.1198/016214505000000907.

Pierpaolo Battigalli and Martin Dufwenberg. Guilt in games. *American Economic Review*, 97(2):170–176, 2007. doi: 10.1257/aer.97.2.170.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Gary E. Bolton and Axel Ockenfels. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193, 2000. doi: 10.1257/aer.90.1.166.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019. doi: 10.1126/science.aay2400.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004. doi: 10.1162/0033553041502225.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869, 2002. doi: 10.1162/003355302760193904.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–1451, 1982. doi: 10.2307/1913390.

Vincent P. Crawford, Miguel A. Costa-Gomes, and Nagore Iriberri. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *Journal of Economic Literature*, 51(1): 5–62, 2013. doi: 10.1257/jel.51.1.5.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel L. K. Yamins, and Nick Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/12f483f624b378f9f3058d8ecd3c7ff5-Abstract-Conference.html.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human–agent communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42879–42896. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1987. URL https://aclanthology.org/2026.acl-long.1987/.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3922.

Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. doi: 10.1162/003355399556151.

Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018. URL https://www.ifaamas.org/Proceedings/aamas2018/pdfs/p122.pdf.

Peter I. Frazier and Warren B. Powell. Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403, 2010. doi: 10.1287/deca.1100.0190. URL https://pubsonline.informs.org/doi/10.1287/deca.1100.0190.

Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005. doi: 10.1613/jair.1579.

Uri Gneezy. Deception: The role of consequences. *American Economic Review*, 95(1):384–394, 2005. doi: 10.1257/0002828053828662.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. SuspicionAgent: Playing imperfect information games with theory of mind aware GPT-4. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F2yGbwXJAi.

Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 346–355, 2012. URL https://www.auai.org/uai2012/papers/123.pdf.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé, III. Opponent modeling in deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1804–1813. PMLR, 2016. URL https://proceedings.mlr.press/v48/he16.html.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Kunal Jha, Aydan Yuenan Huang, Eric Ye, Natasha Jaques, and Max Kleiman-Weiner. Modeling others' minds as code. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6edd46d69ef91f4555d67f7b321d6902-Abstract-Conference.html.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Max Kleiman-Weiner, Alejandro Vientós, David G. Rand, and Joshua B. Tenenbaum. Evolving general cooperation with a Bayesian theory of mind. *Proceedings of the National Academy of Sciences*, 122(25):e2400993122, 2025. doi: 10.1073/pnas.2400993122.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598/.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017. URL https://www.ifaamas.org/Proceedings/aamas2017/pdfs/p464.pdf.

Huao Li, Yu Chong, Simon Stepputtis, Joseph P Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 180–192, 2023.

Terence X. Lim, Sidney Tio, and Desmond C. Ong. Improving multi-agent cooperation using theory of mind. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020. URL https://escholarship.org/uc/item/5tj7v8zq.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Yasser Mohammad and Shinji Nakadai. Optimal value of information based elicitation during negotiation. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pp. 242–250. International Foundation for Autonomous Agents and Multiagent Systems, 2019. URL https://www.ifaamas.org/Proceedings/aamas2019/pdfs/p242.pdf.

Yasser Mohammad, Shinji Nakadai, and Amy Greenwald. Negmas: A platform for automated negotiations. In *International Conference on Principles and Practice of Multi-Agent Systems*, pp. 343–351. Springer, 2020.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *American Economic Review*, 85(5):1313–1326, 1995.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302, 1993.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82091–82130. PMLR, 2025. URL https://proceedings.mlr.press/v267/riemer25a.html.

Melanie Sclar, Graham Neubig, and Yonatan Bisk. Symmetric machine theory of mind. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19450–19466. PMLR, 2022. URL https://proceedings.mlr.press/v162/sclar22a.html.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973. doi: 10.1287/opre.21.5.1071.

Dale O. Stahl, II and Paul W. Wilson. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327, 1994. doi: 10.1016/0167-2681(94)90103-1.

Vince Trencsenyi, Agnieszka Mensfelt, and Kostas Stathis. A survey on hypergame theory: Modelling misaligned perceptions and nested beliefs for multi-agent systems, 2025. URL https://arxiv.org/abs/2507.19593v3. Version 3, revised July 15, 2026.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023. URL https://arxiv.org/abs/2302.08399v5.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658. URL https://ojs.aaai.org/index.php/AAAI/article/view/3982.

Sarah A. Wu, Rose E. Wang, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021. doi: 10.1111/tops.12525.

Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. Model-based opponent modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28208–28221, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b528459c99e929718a7d7e1697253d7f-Abstract.html.

A Additional Theory and Proofs

A.1 Source model, continuous types, and decision regions

The latent variable θ describes whatever opponent characteristics are needed to specify the source return model. The type space Θ may be finite, a subset of Euclidean space, or a product of discrete and continuous spaces. Formally, take it to be a standard Borel space so that regular conditional distributions exist. For a fixed decision context x and source environment G, a type specifies an opponent strategy $\pi_{-i}^{G,\theta}$. If J_G denotes expected focal return under a pair of strategies, then

$$u(\pi, \theta) = J_G(\pi, \pi_{-i}^{G,\theta} \mid x). \tag{9}$$

Each opponent strategy maps interaction histories to actions, so different focal policies can induce different response sequences. The payoff objective, context, and focal policy class are held fixed in the information comparison. When contexts are sampled, the prior, observation model, and policy comparisons are conditioned on each context and then averaged over the declared population.

Assume that u is bounded and measurable, that the optimal values and selected optimizers are measurable, and that maxima are attained. A finite policy class with measurable returns satisfies these requirements. A compact metric policy class with returns continuous in policy and measurable in type is another sufficient setting, with a measurable optimizer selection. We fix this selection rule throughout. The formulation accommodates continuous type spaces directly. Problems with unattained optima require a separate formulation using suprema and approximate optimizers.

For the fixed-channel sequential special case used in this appendix, let $H_n = O_{1:n}$ denote the observed history, let μ_n be the conditional distribution of θ given H_n, and set $\pi_n = \pi^\star(\mu_n)$, including $\pi_0 = \pi^\star(\mu_0)$. Write $V_{\mathrm{acq}}(n)$ for the channel value in Eq. (4) when the acquired outcome is H_n. For an observation model with likelihood $\ell_n(h \mid \theta)$ relative to a common dominating measure, Bayes' rule is

$$\mu_n(d\theta \mid h) = \frac{\ell_n(h \mid \theta)\mu_0(d\theta)}{\int_\Theta \ell_n(h \mid \vartheta)\mu_0(d\vartheta)} \tag{10}$$

where the denominator is positive and ϑ is an integration variable. More generally, a regular conditional distribution defines the update; null histories contribute zero to source expectations. Adaptive queries can be included in H_n when their selection rule and observation laws are specified and acquisition leaves the subsequent return model fixed. Increasing n reveals more of this same statistical experiment.

For finitely many policies, the weak optimality region of a policy π is

$$\mathcal{W}(\pi) = \bigcap_{\pi' \in \Pi} \{\mu : \mathbb{E}_\mu[u(\pi, \theta) - u(\pi', \theta)] \geq 0\}. \tag{11}$$

Weak regions overlap on ties, and the fixed selection rule assigns each belief one policy, including ties on sets of positive probability. For finite types, the regions are polyhedral subsets of the belief simplex. Decision boundaries are ties among maximizing policies. We use "crossing" to denote an update that changes the selected optimal policy, including discrete jumps between belief regions.

At a later interaction history h, write u_h for continuation return and Π_h for the policies still available. Conditional potential value is

$$V_{\mathrm{pot}}(h) = \mathbb{E}_{\mu_h}\left[\max_{\pi \in \Pi_h} u_h(\pi, \theta)\right] - \max_{\pi \in \Pi_h} \mathbb{E}_{\mu_h}[u_h(\pi, \theta)]. \tag{12}$$

This permits reassessment at each decision point. Across physical states, continuation values also reflect changes in time, available contracts, and future responses. The pure-information identities below compare beliefs within a fixed return model and policy class.

A.2 Remaining potential, acquisition, and zero-value information

For brevity in the proofs, let $M(\theta) = \max_{\pi \in \Pi} u(\pi, \theta)$ be the oracle return and $W(\mu) = \max_{\pi \in \Pi} U(\pi, \mu)$ the best return under belief μ. Then

$$V_{\mathrm{pot}}(\mu) = \mathbb{E}_\mu[M(\theta)] - W(\mu). \tag{13}$$

Proof of Proposition 1. Since $W(\mu) = \mathbb{E}_\mu[u(\pi^\star(\mu), \theta)]$, potential value can be written as

$$V_{\mathrm{pot}}(\mu) = \int_\Theta \big[M(\theta) - u(\pi^\star(\mu), \theta)\big]\, \mu(d\theta).$$

The integrand is nonnegative. If $V_{\mathrm{pot}}(\mu) = 0$, it vanishes μ-almost surely, so $\bar{\pi} = \pi^\star(\mu)$ satisfies Eq. (3). Conversely, suppose a policy $\bar{\pi}$ attains $M(\theta)$ for μ-almost every θ. Then

$$\mathbb{E}_\mu[M(\theta)] = \mathbb{E}_\mu[u(\bar{\pi}, \theta)] \le W(\mu) \le \mathbb{E}_\mu[M(\theta)],$$

giving $V_{\mathrm{pot}}(\mu) = 0$. The same nonnegative-gap argument applies to every policy maximizing $U(\pi, \mu)$. □

The integral is with respect to the belief distribution μ and accommodates discrete, continuous, and mixed characteristics. For a finite type space, the condition applies to every positive-probability type. For a continuous distribution, it applies on a set of total probability one, even when each individual parameter value has probability zero. If Θ is a subset of Euclidean space, Π is finite, and every $u(\pi, \theta)$ is continuous in θ, the nonnegative gap is continuous and the condition also holds throughout the topological support of μ. Appendix A.3 illustrates the general probability-one condition.

Conditional on H_n, the expected gain over the prior baseline is

$$\Delta_n = U(\pi_n, \mu_n) - U(\pi_0, \mu_n) \ge 0. \tag{14}$$

The law of total expectation gives $V_{\mathrm{acq}}(n) = \mathbb{E}[\Delta_n]$. Evaluating both policies under μ_n isolates the benefit of revising the decision. The alternative difference $W(\mu_n) - W(\mu_0)$ also reflects learning whether the situation itself has favorable returns. Since π_0 is fixed,

$$\mathbb{E}[V_{\mathrm{pot}}(\mu_n)] = \mathbb{E}[M(\theta)] - \mathbb{E}[W(\mu_n)], \tag{15}$$

$$V_{\mathrm{acq}}(n) = \mathbb{E}[W(\mu_n)] - W(\mu_0). \tag{16}$$

Adding the equations proves Eq. (5), and nonnegativity of its components proves the source bounds.

The same argument is conditional on any current history. For $m \ge n$, let π_m use the additional observations from the same experiment. Then

$$\mathbb{E}[u(\pi_m, \theta) - u(\pi_n, \theta) \mid H_n] = V_{\mathrm{pot}}(\mu_n) - \mathbb{E}[V_{\mathrm{pot}}(\mu_m) \mid H_n] \ge 0. \tag{17}$$

The inequality follows because a later optimizer can retain π_n. Remaining potential decreases in conditional expectation, while individual histories can increase it. For instance, with two types, actions that correctly identify the type paying one and errors paying zero, potential is the smaller type probability. Starting from probabilities $(0.9, 0.1)$ gives potential 0.1; an observation with likelihoods $(1/9, 1)$ yields posterior $(0.5, 0.5)$ and potential 0.5. Other outcomes compensate in expectation. Once potential is exactly zero, however, later pure information cannot make it positive except on a null event in this fixed model.

If the currently implemented policy is $\widehat{\pi}$ rather than $\pi^\star(\mu)$, its oracle shortfall separates as

$$\mathbb{E}_\mu[M(\theta) - u(\widehat{\pi}, \theta)] = V_{\mathrm{pot}}(\mu) + \big[W(\mu) - U(\widehat{\pi}, \mu)\big]. \tag{18}$$

The bracketed term measures the improvement available from better planning at the current belief, before acquiring further evidence.

Costs and a valid stopping certificate. Consider any further pure information experiment starting at belief μ, possibly with adaptive queries and a bounded stopping time. Conditional on its final history, even an optimal policy cannot exceed the oracle return. Consequently its expected gross improvement is at most $V_{\mathrm{pot}}(\mu)$. If C is its nonnegative total acquisition cost, its expected net improvement is at most

$$V_{\mathrm{pot}}(\mu) - \mathbb{E}[C]. \tag{19}$$

If every nontrivial continuation has expected cost at least $c_{\min}$ and $V_{\mathrm{pot}}(\mu) \le c_{\min}$, stopping is weakly optimal among these pure-information continuations. When cheaper queries are available, the stopping decision compares their expected gains with their costs. Physical benefits of continued interaction, such as later concessions, enter the full continuation value separately.

A.3 PAYOFF-RELEVANT UNCERTAINTY AND SUFFICIENT REPRESENTATIONS

For a finite policy class, choose a reference policy π_{ref} and define the vector of type-dependent payoff differences

$$d(\theta) = \big(u(\pi, \theta) - u(\pi_{\mathrm{ref}}, \theta)\big)_{\pi \in \Pi}. \tag{20}$$

Policy selection depends on $\mathbb{E}_\mu[d(\theta)]$; any action-independent component of the type's returns cancels in this comparison.

Proposition 2 (A sufficient payoff representation)**.** *Let $Z = f(\theta)$ be a measurable representation such that $d(\theta) = \widetilde{d}(f(\theta))$ for a measurable function $\widetilde{d}$. The conditional distribution of Z determines every expected policy payoff difference, the selected policy, and its conditional improvement over any specified reference policy.*

Proof. For each policy π, the expected difference from π_{ref} is $\mathbb{E}[\widetilde{d}(Z)_\pi]$. Subtracting two coordinates gives every pairwise policy comparison. Applying the fixed tie rule yields the selected policy. The same comparisons yield its gain over the specified reference. □

Equivalently, $u(\pi, \theta) = r(\pi, f(\theta)) + c(\theta)$, where c is independent of policy. This representation preserves policy comparisons for the fixed return model. Preserving every expected payoff difference for every belief requires types with equal representation values to have equal difference vectors, as can be checked using point-mass beliefs. Preserving only the selected policy can require less information. An observation that leaves the distribution of a sufficient representation unchanged also leaves the selected policy and its relative value unchanged. This construction applies directly to continuous preference vectors. For sequential plans, the differences concern complete continuation returns, including future transitions and observations.

Example 1 (Oracle-action labels do not preserve mixture decisions)**.** *Take policies a, b, three types, $u(b, \theta) = 0$, and $u(a, \theta) = (2, 0.1, -1)$. The prior is $(0.2, 0.2, 0.6)$, so the prior action is b. Let a binary observation have likelihood $P(L \mid \theta) = (1, 0, 1/2)$. The two equally likely posteriors are $(0.4, 0, 0.6)$ and $(0, 0.4, 0.6)$. Both assign probability 0.4 to types whose oracle action is a, but their expected $a - b$ payoff differences are 0.2 and -0.56. Only the first posterior selects a, giving acquired value 0.1.*

Both posterior supports contain types with conflicting optimal actions. Oracle-action probabilities discard payoff magnitudes that remain relevant to the mixture decision.

Continuous preferences and common optimal policies. Consider a continuous preference parameter $w \in [0, 1]$, for example the first coordinate of normalized weights $(w, 1 - w)$. In a stylized decision problem, two focal policies a, b have returns $u(a, w) = w$ and $u(b, w) = 1 - w$. If the belief μ is uniform on $[0.7, 0.9]$, policy a is optimal throughout that interval. Both the perfect-information return and $W(\mu)$ equal 0.8, so $V_{\mathrm{pot}}(\mu) = 0$ despite a nondegenerate continuous belief. This conclusion holds for any belief concentrated on that interval. If instead μ is uniform on $[0.2, 0.8]$, the optimal policy changes at $w = 0.5$; the perfect-information return is 0.65, $W(\mu) = 0.5$, and $V_{\mathrm{pot}}(\mu) = 0.15$.

The policy class also matters. With a continuously adjustable action $a \in [0, 1]$ and return $u(a, w) = -(a - w)^2$, perfect information selects $a = w$, while the current optimal action is $a = \mathbb{E}_\mu[w]$. Here $V_{\mathrm{pot}}(\mu) = \mathrm{Var}_\mu(w)$, where $\mathrm{Var}_\mu(w) = \mathbb{E}_\mu[(w - \mathbb{E}_\mu[w])^2]$ is the belief variance. Every positive-variance belief has positive potential in this problem. Proposition 1 covers both examples through the existence of a common optimal policy.

Approximate decision sufficiency. Suppose returns lie in an interval of length L, and a policy $\bar{\pi}$ is at most ϵ below the oracle throughout a set K with posterior probability at least $1 - \delta$. Here ϵ is a payoff tolerance and δ bounds the posterior mass outside K. Then

$$V_{\mathrm{pot}}(\mu) \le \mathbb{E}_\mu[M(\theta) - u(\bar{\pi}, \theta)] \le \epsilon + \delta L. \tag{21}$$

The first inequality follows because the Bayes optimizer is at least as good as $\bar{\pi}$ in expectation; splitting the oracle shortfall over K and its complement gives the second. Approximate decision sufficiency can therefore hold with substantial posterior uncertainty when a policy is close to optimal over most posterior mass.

Entropy reduction and decision value. Let $\theta = (Z, W_1, \ldots, W_k)$ consist of independent fair bits, with payoff $\mathbf{1}\{a = Z\}$ for action $a \in \{0, 1\}$. Observing all nuisance bits W_j leaves acquired value zero and potential $1/2$, although the fraction of type entropy removed is $k/(k+1)$. Observing Z alone removes all potential while leaving k bits of uncertainty. Entropy and decision value thus measure different objectives. For continuous variables, differential entropy additionally depends on coordinate scale: under $Y = cX$ it changes by $\log|c|$ when defined, while information value is invariant to relabeling the same decision problem. Payoff tolerances are assessed through remaining potential and task-specific return bounds.

A.4 Acquisition over time and finite decision completion

The source expectation averages over types and complete observation histories from the declared protocol. It is distinct from the conditional value on one observed history.

Policy switches and the incremental value of evidence. Let $B_n = \mathbf{1}\{\pi_n \neq \pi_0\}$, let $\beta(n) = \Pr(B_n = 1)$, and let $\Gamma(n) = \mathbb{E}[\Delta_n \mid B_n = 1]$ when this event has positive probability, with $\Gamma(n) = 0$ otherwise. Since $\Delta_n = 0$ whenever the policies coincide,

$$V_{\mathrm{acq}}(n) = \beta(n)\Gamma(n). \tag{22}$$

Here β measures disagreement with the initial policy at the evaluation endpoint, and Γ measures conditional improvement. At a tie, a policy change can have zero gain; a switch to a unique optimizer has strictly positive conditional gain on that history. Their product gives an algebraic decomposition of acquired value.

Define the marginal acquired value of observation step $n \geq 1$ by $\delta V_n = V_{\mathrm{acq}}(n) - V_{\mathrm{acq}}(n-1)$. Then

$$\delta V_n = \mathbb{E}[U(\pi_n, \mu_n) - U(\pi_{n-1}, \mu_n)] \geq 0, \qquad V_{\mathrm{acq}}(n) = \sum_{t=1}^{n} \delta V_t. \tag{23}$$

Subtract the acquired-value definitions and condition on H_n. The policy π_{n-1} is determined by the observed prefix and remains available. Evaluating both policies under μ_n gives the inequality by optimality. Averaging over histories and telescoping yields the identity, with $V_{\mathrm{acq}}(0) = 0$. Prior observations can enable the value of later evidence. Comparisons across channels require an ordering of the statistical experiments, such as Blackwell garbling, where one channel is obtained by type-independent post-processing of the other (Blackwell, 1953). Any proposed noise parameterization can be checked for this ordering.

Expected acquisition budgets and history-dependent stopping. For a payoff tolerance $\epsilon \geq 0$, define the expected acquisition budget

$$n_\epsilon = \inf\{n \geq 0 : V_{\mathrm{acq}}(n) \geq V_{\mathrm{pot}}(\mu_0) - \epsilon\}, \qquad \inf \varnothing = \infty. \tag{24}$$

The subscript $\epsilon = 0$ denotes exact completion. Equation (5) makes this the first budget at which expected remaining potential is at most ϵ. In contrast, the history-dependent stopping time is

$$\tau_\epsilon = \inf\{n \geq 0 : V_{\mathrm{pot}}(\mu_n) \leq \epsilon\}. \tag{25}$$

The budget n_ϵ controls the expected residual at a fixed horizon, while τ_ϵ records when a particular history first reaches the tolerance. Its distribution can include arbitrarily late or infinite completion times. A procedure stopped at $\min(\tau_\epsilon, N)$ for finite N can retain residual potential above ϵ on censored histories; these residuals are included in completion reports.

Proposition 3 (Exact acquisition and remaining uncertainty)**.** *For a fixed observation budget n, $V_{\mathrm{acq}}(n) = V_{\mathrm{pot}}(\mu_0)$ if and only if $V_{\mathrm{pot}}(\mu_n) = 0$ almost surely. Equivalently, for almost every history, its posterior-selected policy is oracle-optimal for almost every type under that posterior.*

Proof. By Eq. (5), equality is equivalent to zero expected remaining potential. Since remaining potential is nonnegative, it must vanish almost surely. Proposition 1 gives the equivalent policy statement. □

An observation can complete the decision by identifying a set of types that share an optimal policy. The posterior within that set may remain diffuse.

Conditions for persistent residual potential. Suppose two types have positive prior probability and no common optimal policy. If every finite history with positive marginal probability has positive likelihood under both types, both remain possible after each such history. Proposition 1 gives strictly positive remaining potential at every finite step, hence $n_0 = \infty$. For conditionally independent binary observations with type-conditioned probabilities strictly between zero and one, both types assign positive probability to every finite sequence. When their optimal actions differ, exact completion is therefore impossible at any finite observation budget. If the two observation probabilities coincide, the evidence is uninformative and residual potential stays at its initial value.

More generally, indistinguishable types with conflicting optimal policies obstruct completion even with indefinitely many observations. For finite Π, bounded returns and a nested infinite observation sequence imply convergence of each conditional policy return by the martingale convergence theorem. Taking the maximum of finitely many convergent returns then shows

$$\lim_{n\to\infty} V_{\mathrm{acq}}(n) = V_{\mathrm{pot}}(\mu_0) - \mathbb{E}[V_{\mathrm{pot}}(\mu_\infty)], \tag{26}$$

where μ_∞ conditions on the entire observation sequence. Persistent ambiguity is harmless only when it leaves a common optimal policy. This limit statement uses a finite policy class; extension to an arbitrary infinite class needs conditions allowing the limiting operation to commute with optimization.

Evaluating accessible information value. Candidate queries are evaluated through their possible posterior decisions, payoff gains, and costs. With a finite remaining budget, histories determine subsequent queries and their continuation values. This calculation uses the specified observation model to assess how much of the remaining opportunity the available procedures can acquire.

Information can be complementary. Let the type contain independent fair bits (Z_1, Z_2, W), and let action $a \in \{0, 1\}$ pay one exactly when a equals the parity of Z_1, Z_2. Observe Z_1 first and Z_2 second. The first observation has zero acquired value; the second completes the decision, giving acquired value $1/2$ while W remains unknown. A multi-query continuation can therefore be valuable even when its first observation has zero immediate benefit. Nonconcavity of information value in sample size is also studied by Frazier & Powell (2010).

A.5 Mixed policies and continuously varying decisions

Consider a finite set of pure plans $a \in \mathcal{A}$ and a mixed policy σ, where $\sigma(a)$ is the probability of selecting plan a. If the opponent follows the fixed source strategy and does not condition on the announced mixture itself, the expected return is linear:

$$u(\sigma, \theta) = \sum_{a\in\mathcal{A}} \sigma(a) u(a, \theta), \qquad \max_{\sigma} U(\sigma, \mu) = \max_{a\in\mathcal{A}} U(a, \mu). \tag{27}$$

Thus a pure optimizer exists; mixing among tied optimal plans does not strictly increase the optimized source return. The same argument applies to a private lottery over complete contingent plans, including the opponent's within-plan responses. Pure-policy decision regions suffice for this case.

If the declared problem includes constraints on randomization, an objective favoring diversity, or opponent responses to an announced distribution, it can instead have an interior mixed optimizer that varies continuously with beliefs. Such a model must specify the resulting return function consistently. The definitions of potential, acquired, and realized value and the remaining-potential identity still apply under the assumptions in Appendix A.1; finite polyhedral decision regions need not exist. Changing action probabilities can change value even when the modal action stays fixed. Conversely, changing the mixture does not guarantee a strict gain if the policies are payoff-equivalent.

An entropy-regularized controller can be evaluated against the common material-payoff objective through its implementation regret. When regularization is part of the evaluation objective, the same regularized objective is used for the informed controller and its baseline. In sequential interaction, policy evaluation includes continuation decisions beyond the first-action distribution.

A.6 Imperfect inference, control, and decision regret

Let $\widehat{\pi}_n(H_n)$ be an implemented policy and let its conditional decision regret be

$$r_n(H_n) = U(\pi_n, \mu_n) - U(\widehat{\pi}_n(H_n), \mu_n) \geq 0. \tag{28}$$

Its achieved source gain, denoted V_{src}, is

$$V_{\mathrm{src}}(\widehat{\pi}_n) = \mathbb{E}[u(\widehat{\pi}_n, \theta) - u(\pi_0, \theta)] = V_{\mathrm{acq}}(n) - \mathbb{E}[r_n]. \tag{29}$$

Conditioning on the history proves the identity. It separates an observation process's ideal value from the controller's ability to use it, and it allows negative achieved source gain. Randomized implementations add an expectation over their internal randomness. A controller's improvement evaluated under its own misspecified beliefs need not equal this gain. In particular, a likelihood temperature different from one is evaluated through achieved source gain unless that update is exact Bayesian conditioning for a correctly specified alternative observation model. A switch/gain factorization remains algebraically valid for the implemented policy, but its conditional gain can be negative.

For a prediction guarantee, let p be the true conditional type distribution and q its prediction, and let π_p, π_q maximize expected return under these distributions. Suppose returns lie in an interval of length L. Define total variation by $\mathrm{TV}(p, q) = \sup_A |p(A) - q(A)|$, where the supremum is over measurable sets, and write $D_{\mathrm{KL}}(p\|q)$ for the Kullback–Leibler divergence, with value infinity when the required absolute continuity fails. Then

$$\mathrm{Reg}(p, q) := U(\pi_p, p) - U(\pi_q, p) \le 2L\,\mathrm{TV}(p, q) \le L\sqrt{2D_{\mathrm{KL}}(p\|q)}. \tag{30}$$

To prove the first inequality, add and subtract $U(\pi_p, q)$ and $U(\pi_q, q)$. Their intervening difference is nonpositive by q-optimality, while each of the two distribution-change terms is bounded by $L\,\mathrm{TV}(p, q)$. Pinsker's inequality gives the second. Averaging and Jensen's inequality yield $\mathbb{E}[\mathrm{Reg}] \le L\sqrt{2\mathbb{E}[D_{\mathrm{KL}}]}$. With discrete types, or densities relative to a common measure and well-defined log losses, KL is conditional excess log loss relative to the true conditional model. For a finite class of at least two policies with a unique optimizer under p, define its payoff margin by

$$\gamma(p) = \min_{\pi \ne \pi_p} [U(\pi_p, p) - U(\pi, p)] > 0. \tag{31}$$

Every pairwise payoff difference changes by at most $2L\,\mathrm{TV}(p, q)$, so $2L\,\mathrm{TV}(p, q) < \gamma(p)$ guarantees $\pi_q = \pi_p$.

Comparing two predictors' losses involves a separate quantity. If the true state is deterministically one, actions receive payoff for matching it, and predictions assign probability $1/2 \pm \varepsilon$ to it, the induced actions have regrets zero and one. Their log-loss difference approaches zero, while both losses approach $\log 2$ and remain above the Bayes loss of zero. Excess-risk consistency concerns convergence to the Bayes-optimal loss (Bartlett et al., 2006). At a fixed payoff scale, Eq. (30) bounds expected decision regret as distribution error vanishes.

A.7 Deployment and each policy's induced opponent response

Let G_{tar} and x_{tar} specify the target environment and context. Write $\Psi_{-i}(\theta, \pi)$ for the opponent response to deploying π, including initialization and response selection where necessary. The response rule can describe a behavioral model, a learned reaction, or an equilibrium response. The compact target return in the main text expands as

$$u_{\mathrm{tar}}(\pi, \theta) = J_{G_{\mathrm{tar}}}(\pi, \Psi_{-i}(\theta, \pi) \mid x_{\mathrm{tar}}). \tag{32}$$

The target joint law of types and observations is part of the evaluation specification, and the deployed rule is defined throughout its support. In the response-shift comparisons of Section 5, the type–observation law and selected source policies are fixed while subsequent opponent behavior changes. The baseline remains the source prior-informed policy.

If the informed-policy response is held fixed in both branches, the comparison instead becomes

$$\begin{aligned} V_{\mathrm{dir}} = \mathbb{E}_{\mathrm{tar}}\big[&J_{G_{\mathrm{tar}}}(\pi_n, \Psi_{-i}(\theta, \pi_n) \mid x_{\mathrm{tar}}) \\ &- J_{G_{\mathrm{tar}}}(\pi_0, \Psi_{-i}(\theta, \pi_n) \mid x_{\mathrm{tar}})\big]. \end{aligned} \tag{33}$$

Adding and subtracting this second return in the actual deployment comparison yields

$$\begin{aligned} V_{\mathrm{real}}(n) = V_{\mathrm{dir}} + \mathbb{E}_{\mathrm{tar}}\big[&J_{G_{\mathrm{tar}}}(\pi_0, \Psi_{-i}(\theta, \pi_n) \mid x_{\mathrm{tar}}) \\ &- J_{G_{\mathrm{tar}}}(\pi_0, \Psi_{-i}(\theta, \pi_0) \mid x_{\mathrm{tar}})\big]. \end{aligned} \tag{34}$$

The correction decomposes the deployment comparison into a fixed-response contrast and the effect of changing responses. We write Corr for this second, response-change term, so that $V_{\mathrm{real}}(n) = V_{\mathrm{dir}} + \mathrm{Corr}$. Estimating its components requires the corresponding counterfactual returns. For example, policies a, b induce responses r_a, r_b, with $J(a, r_a) = 1$, $J(b, r_a) = 0$, and $J(b, r_b) = 2$. The common-response contrast is $+1$, while the policy deployment contrast is -1. A policy-independent opponent makes the correction zero. Equality with acquired value additionally requires agreement of source and target policy returns and the type–observation law. Correctly modeled stochastic responses are included in these expectations.

For an implemented controller, replace π_n by $\widehat{\pi}_n$ in the target comparison and keep its source rule fixed. Comparing with $V_{\mathrm{src}}(\widehat{\pi}_n)$ separates deployment effects from source implementation regret. A comparison beginning before acquisition includes its costs in complete-policy returns. A comparison after a common pre-collected trace measures the deployment contrast conditional on that acquisition protocol.

A conditional-loss summary. Let $D_{\mathrm{tar}} = u_{\mathrm{tar}}(\pi_n, \theta) - u_{\mathrm{tar}}(\pi_0, \theta)$ be the target return difference, let $S = \{D_{\mathrm{tar}} < 0\}$ be the event that the informed policy is worse, and write $p_S = \Pr_{\mathrm{tar}}(S)$. If $0 < p_S < 1$ and positive constants m, M satisfy $\mathbb{E}[D_{\mathrm{tar}} \mid S] \le -m$ and $\mathbb{E}[D_{\mathrm{tar}} \mid S^c] \le M$, then

$$V_{\mathrm{real}}(n) \le -p_S m + (1 - p_S) M. \tag{35}$$

Conditioning on S gives the bound. In particular, $p_S > M/(m + M)$ is sufficient for a negative mean. The endpoint cases use the corresponding unconditional bounds. These conditions concern mean losses and gains under the declared target distribution; using exact conditional means gives an identity. Generalization from empirical estimates to a new target population requires assumptions connecting the two populations.

A.8 Paired evaluation and interpretation

An evaluation specifies the objective, policy class, type and context population, observation protocol, source return model, and prior-informed baseline. Oracle potential uses type-conditioned optimization within the same policy class. A heuristic oracle provides a lower estimate of the optimized type-informed return; an information-value ceiling requires a valid optimization bound. Achieved gains can also be reported directly when the oracle value is unavailable.

For K independent held-out source draws, write $(\theta_k, H_{n,k})$ for the type and history and $\pi_{n,k}$ for the resulting posterior policy. Let $D_k = u(\pi_{n,k}, \theta_k) - u(\pi_0, \theta_k)$ and $B_k = \mathbf{1}\{\pi_{n,k} \neq \pi_0\}$. Then

$$\widehat{V}_{\mathrm{acq}} = \frac{1}{K} \sum_k D_k, \qquad \widehat{\beta} = \frac{1}{K} \sum_k B_k, \qquad \widehat{\Gamma} = \frac{\sum_k B_k D_k}{\sum_k B_k}. \tag{36}$$

Estimating Γ requires observed switches. When none occur, report a switch count of zero and leave the conditional-gain estimate undefined; the estimated product can be set to zero. Individual type-conditioned differences D_k can be negative even for exact Bayesian policies. For an approximate controller, the paired mean estimates achieved source gain.

Initial potential is estimated by replacing $\pi_{n,k}$ with the type-conditioned oracle. Current potential uses draws from the current posterior and the current Bayes policy as baseline, with approximation and model error reported. Realized value evaluates both policies under their own target responses, resetting opponents to the prescribed initial condition in each branch. Identical policies can share a rollout or be assigned zero difference. For distinct policies, shared random seeds can reduce variance while preserving each branch's marginal law.

Exact finite integration gives the values under the specified model directly. Empirical estimates report independent sample counts, uncertainty in paired differences, observed switch counts, and support for conditional estimates. Multiple traces from a common case are clustered at the case level. When initial potential is positive, acquisition efficiency is

$$\eta_{\mathrm{acq}}(n) = \frac{V_{\mathrm{acq}}(n)}{V_{\mathrm{pot}}(\mu_0)} \in [0, 1]. \tag{37}$$

Acquisition efficiency uses initial potential as its denominator. The deployment ratio $V_{\mathrm{real}}(n)/V_{\mathrm{acq}}(n)$ is defined for positive acquired value and may be negative or exceed one. Reporting ratios involves

excluding zero denominators and identifying sensitivity to very small denominators. Completion reports distinguish expected-value budgets, history-specific stopping times, and censored runs.

The final-offer negotiation diagnostics use the LLM-estimated source channel and the known acceptance rule. Numerical values characterize that working model; deployed net returns evaluate the LLM's actual actions. Standing-offer negotiation adds an incumbent contract, case-specific acquisition costs, and a terminal counteroffer (Appendix E). Paired method improvements are distinct from the theoretical comparison against the prior-informed baseline.

B Closed-Form Instantiations

These instantiations connect the general definitions to the controlled games. In experimental summaries, an unqualified V_{pot} denotes the initial value $V_{\mathrm{pot}}(\mu_0)$. The same decision problem can also be evaluated at a later belief without changing the definitions.

B.1 Binary-action games

Let C, D be the focal and opponent actions, and let $(R, S; T, P)$ be the focal payoff matrix, with focal actions indexing rows. Write $q_{g,\theta}$ for the opponent's cooperation probability in game g under type θ. A belief μ gives the mean cooperation probability $p_\mu = \mathbb{E}_\mu[q_{g,\theta}]$. For any cooperation probability p, define the two action returns by

$$U_g(C, p) = pR + (1-p)S, \qquad U_g(D, p) = pT + (1-p)P. \tag{38}$$

Thus $U(a, \mu) = U_g(a, p_\mu)$ for either action a. The signed action margin and, when its denominator is nonzero, the boundary are

$$g(p) = p(R-T) + (1-p)(S-P), \qquad p_g^\star = \frac{P-S}{R-T-S+P}. \tag{39}$$

Choose C when $g(p) > 0$ and D when $g(p) < 0$, using the fixed rule at ties. With $a_g = R-T-S+P$, a nonzero a_g gives $g(p) = a_g(p - p_g^\star)$, so the threshold direction depends on the sign of a_g. If $a_g = 0$, the gap is constant and information cannot improve action selection.

With strict payoffs, $m_g = (P-S)(R-T) < 0$ identifies a dominant action, whereas $m_g > 0$ gives a threshold inside $(0, 1)$. In the latter case, the current potential value is positive precisely when

$$\Pr_\mu(q_{g,\theta} < p_g^\star) > 0 \quad \text{and} \quad \Pr_\mu(q_{g,\theta} > p_g^\star) > 0. \tag{40}$$

Indeed, if all probability lies on one side or at the threshold, one action is optimal almost surely. If both strict sides have positive probability, no action is optimal almost surely; Proposition 1 gives positive potential. This criterion applies to discrete types and continuous preferences. For finite type support, let $q_{\min}, q_{\max}$ be the extrema over positive-probability types. It reduces to $r_g = \min\{p_g^\star - q_{\min}, q_{\max} - p_g^\star\} > 0$, recovering the main-text initial-support test at $\mu = \mu_0$.

Let δ_θ be the point mass at type θ, let $a_\theta = \pi^\star(\delta_\theta)$ be its oracle action, and let $a_\mu = \pi^\star(\mu)$ be the action selected at the current belief. When $a_g \neq 0$, define the expected crossing distance

$$d_\times(\mu) = \mathbb{E}_\mu\left[|q_{g,\theta} - p_g^\star| \mathbf{1}\{a_\theta \neq a_\mu\}\right]. \tag{41}$$

When oracle and current actions differ, their payoff difference is $|g(q_{g,\theta})|$; otherwise it is zero. Hence

$$V_{\mathrm{pot}}(\mu) = |a_g| d_\times(\mu). \tag{42}$$

The plotted initial quantities use $a_0 = \pi_0$ and $d_\times = d_\times(\mu_0)$. Away from a prior tie, action disagreement is equivalent to lying strictly on the opposite side of the boundary. The action-based definition also covers a prior exactly on the boundary. Potential, payoff slope, and crossing distance are linked by Eq. (42).

B.2 Normal-form games and sequential plans

In a finite normal-form game, let a, b be focal and opponent actions and $u_i(a, b)$ the focal payoff. The posterior-averaged opponent action distribution is $\mathbf{p}_\mu(b) = \mathbb{E}_\mu[\pi_{-i}^{G,\theta}(b)]$, giving

$U(a, \mu) = \sum_b \mathbf{p}_\mu(b) u_i(a, b)$. Decision boundaries are ties among maximizing actions. A representation preserving every action's expected payoff difference is sufficient for selection.

In a finite-horizon partially observed game with finite action and observation alphabets and fixed opponent strategy models, choose Π as the complete contingent policy trees available after the initial information experiment. Each tree has a type-conditioned expected return, so the finite-policy results apply to the entire continuation plan.

A belief-state formulation is also possible when the observable state and posterior together determine the distribution of future outcomes. Let s be the current state, s' the next state, t the decision time, $r_i(s, a, b)$ the immediate focal reward, and $\gamma \in [0, 1]$ a discount factor. Let y collect the newly observed information, including any observed actions and next state, and let $\mathcal{F}$ be the Bayesian belief update. The continuation value can then satisfy

$$V_t(s, \mu) = \max_a \mathbb{E}[r_i(s, a, b) + \gamma V_{t+1}(s', \mathcal{F}(\mu, y \mid s, a))], \tag{43}$$

where the expectation integrates $\theta \sim \mu$, opponent actions, state transitions, and observations under the specified model. A zero terminal continuation value gives the standard finite-horizon recursion. A sequential abstraction preserves the information needed to evaluate future transitions, observations, and continuation returns.

C Experimental Protocols, Parameters, and Reproducibility

This section documents the matrix-game studies used in Section 5: the structural opportunity audit, finite-evidence summaries, and exogenous response shifts. The general evidence-action calculations following these protocols support Section 4.2; The final-offer and standing-offer negotiation studies retain their separate specifications in Appendices D and E.

Table 2: Experimental scope and evaluation units.

Study	Scope
Potential-value audit	Four illustrative games; 12 canonical symmetric topologies with 300 cardinal assignments each. Figure 2.
Finite-evidence summaries	Fixed-channel first-switch counts over observation noise and update temperature; exact acquired-value calculations. Figure 3A–B.
Exogenous response shifts	Source observation process and policies held fixed; behavior-reversal grid and a separate mixture of target shifts. Figure 3C–E.
Final-offer negotiation	96 evaluation episodes per focal model: 12 type/disclosure conditions, two phrasings, and four methods; 16 additional response-boundary control episodes per model.
Standing-offer negotiation	15 randomly sampled cases, grouped retrospectively by decision-relevant parameters; three repetitions each and a matched Full–baseline pair per case.

Table 3: Parameters and reproduction inputs for the matrix-game plots.

Panel	Parameters
Figure 2A	Four payoff matrices and type-response pairs listed in Appendix C.1.2; initial type mass 10^{-3} or $1 - 10^{-3}$, conditioned on the designated true type. Stop after three posterior changes ≤ 0.02 with true-type mass ≥ 0.99, or at 500 steps. Deterministic expected-log-likelihood accumulation; no random draws.
Figure 2B–C	12 canonical symmetric orders, 300 cardinal draws each, total 3,600; payoffs uniform on $[-1.5, 1.5]$; seed 20260506. Four-profile mixture in Table 5.

Panel	Parameters
Figure 3A	$T_{\mathrm{B}} \in \{1, 2, 4, 6, 8, 12, 16, 24\}$; $\varepsilon \in \{.05, .10, .15, .20, .25, .30, .35, .40, .45, .48\}$. Maximum displayed first-switch count 894.
Figure 3B	Exact Bayesian conditioning ($T_B = 1$); acquisition efficiency $V_{\mathrm{acq}}(n)/V_{\mathrm{pot}}(\mu_0)$ from finite integration over observation counts using Eqs. (46)–(47).
Figure 3C	24 payoff orders, 200 cardinal draws/order; seed 20260506; $\varepsilon = .10$, $n = 8$ exact observations; 101 reversal levels on $[0, 1]$; 25×25 bins; median ratio color-clipped to $[-1, 1]$; support contour count ≥ 8.
Figure 3D–E	24 orders, 500 cardinal draws/order; retain feasible crossing and $V_{\mathrm{acq}} \geq 10^{-3}$; $\varepsilon = .10$, $n = 8$, seed 20260506. 100 shifts/source: flip, neutral, permutation, random, local with probabilities $(.25, .15, .25, .25, .10)$. First four mix with $\lambda \sim U[0, 1]$; local noise has $\sigma \sim U[0, .45]$. 4,369 retained games, 436,900 shifts; plot at most 100,000 (seed 42); jitter SDs .015 and .005.

C.1 Controlled-Game Protocols and Plotting Details

Section 5 uses these studies to distinguish opportunity, acquisition, and deployment limitations. The payoff objective, focal action set, and prior-informed baseline are common to each within-instance comparison. Potential is evaluated before finite evidence, acquisition concerns policy improvements under the source model, and deployment compares the source-selected informed and baseline policies under the target behavior. The observation-budget summaries contain no cost-optimized stopping intervention; the response-shift summaries contain no learned response-repair intervention. Their decision implications follow from the criteria in Section 4.

C.1.1 Topology Enumeration and Coverage

The topology evaluation in Figure 2 uses all 12 strict ordinal symmetric 2×2 classes after quotienting by simultaneous action relabeling $C \leftrightarrow D$, which maps (R, S, T, P) to (P, T, S, R). Each orbit is represented by the orientation $R > P$, placing the larger mutual-action payoff at (C, C). The target-shift panels use the 24 ordered focal-payoff permutations.

Table 4 lists the 12 symmetric representatives used in the topology evaluation. Larger rank numbers correspond to larger sampled cardinal payoffs.

Table 4: Canonical strict ordinal symmetric 2×2 topology representatives used in the topology evaluation. Each action-relabeling orbit is oriented so that $R > P$, i.e., the (C, C) diagonal payoff is larger than the (D, D) diagonal payoff. Larger rank numbers correspond to larger sampled cardinal payoffs.

Topology ID	Payoff order	Family label	Common name / mnemonic
T01	$T > R > P > S$	D-dominant	Prisoner's Dilemma
T02	$R > S > T > P$	C-dominant	Harmony
T03	$S > R > T > P$	C-dominant	Harmony variant
T04	$R > T > S > P$	C-dominant	Harmony variant
T05	$R > S > P > T$	C-dominant	Harmony / no-conflict variant
T06	$S > R > P > T$	C-dominant	Harmony / no-conflict variant
T07	$S > T > R > P$	Anti-coordination	Chicken / Snowdrift variant
T08	$T > R > S > P$	Anti-coordination	Chicken / Snowdrift
T09	$T > S > R > P$	Anti-coordination	Chicken / Hawk–Dove variant
T10	$R > T > P > S$	Coordination	Stag Hunt / Assurance
T11	$R > P > S > T$	Coordination	Assurance variant
T12	$R > P > T > S$	Coordination	Coordination / Assurance variant

C.1.2 Potential: Structural Gates, Belief Paths, and Payoff Magnitudes

Diagnostic quantities. The topology gate is $m_g = (P - S)(R - T)$. The displayed failure mode F_1 is dominance, $m_g < 0$, for which a common action is optimal against all opponent responses. Conditional on an interior boundary, $m_g > 0$, define

$$r_g = \min\{p_g^\star - q_{\min}, q_{\max} - p_g^\star\},$$

with extrema over positive-prior types. The displayed F_2 condition is $r_g \leq 0$: the supported responses do not lie strictly on both sides. Either condition gives zero potential by Proposition 1; positive potential requires $m_g > 0$ and $r_g > 0$. These labels distinguish two structural causes of the same diagnostic outcome and do not describe inaccurate type inference.

Belief-path illustrations.

Figure 2A uses four illustrative payoff matrices, with rows denoting focal C, D and columns denoting opponent C, D:

$$M_{\text{PD}} = \begin{pmatrix} 0.5 & -1.5 \\ 1.5 & -0.5 \end{pmatrix}, \quad M_{\text{Harmony}} = \begin{pmatrix} 1.5 & 0.5 \\ -0.5 & -1.5 \end{pmatrix},$$

$$M_{\text{Coord}} = \begin{pmatrix} 1.5 & -0.5 \\ -1.5 & 0.5 \end{pmatrix}, \quad M_{\text{Anti}} = \begin{pmatrix} -1.5 & 0.5 \\ 1.5 & -0.5 \end{pmatrix}.$$

These illustrative matrices are used only for the posterior-trajectory examples in Figure 2A; the canonical $R > P$ orientation used for the topology enumeration is specified in Table 4.

The two latent opponent types are labeled selfish and cooperative, with cooperation probabilities

Game	$q_{\text{selfish}}(C)$	$q_{\text{coop}}(C)$	Family
PD	0.05	0.95	D-dominant
Harmony	0.60	0.99	C-dominant
Coordination	0.15	0.85	Coordination
Anti-coordination	0.15	0.85	Anti-coordination

Each trajectory is conditioned on a designated data-generating type and a nearly dogmatic initialization against it: $P_0(\text{selfish}) = 10^{-3}$ for selfish truth and $P_0(\text{selfish}) = 1 - 10^{-3}$ for cooperative truth. The plotted horizon stops after three consecutive posterior changes of at most 0.02 with true-type probability at least 0.99, or at a maximum of 500 steps.

The curves use deterministic mean log-evidence accumulation. With $s = q_{\text{selfish}}(C)$, $c = q_{\text{coop}}(C)$, and $q_\star$ the designated true type's cooperation probability, the plotted probability is

$$\widetilde{p}_h = \text{logistic}\left(\text{logit}(p_0) + h\left[q_\star \log\frac{s}{c} + (1 - q_\star)\log\frac{1-s}{1-c}\right]\right). \tag{44}$$

The curve transforms expected log posterior odds, yielding a deterministic illustration of belief evolution. The expected posterior probability is a separate quantity because the logistic map is nonlinear. The eight cutoffs, ordered by game and then selfish/cooperative truth, are $(8, 8, 12, 26, 12, 12, 12, 12)$. Predictive cooperation is $s\widetilde{p}_h + c(1 - \widetilde{p}_h)$; exact payoff ties select C. In the two crossing games the cooperation boundary is $.25$, the selfish-type probability boundary is $6/7$, and the continuous crossing horizons are 7.16386565 and 4.21257449. Shading uses continuous crossings and markers use integer horizons.

Panels B–C of Figure 2 enumerate the 12 canonical topology representatives and draw 300 ordinal-preserving cardinal samples per topology, for 3600 total samples. For each topology, four cardinal payoffs are sampled uniformly from $[-1.5, 1.5]$, sorted, and assigned according to the strict ordinal rank tuple. The random seed is 20260506.

Each cardinal sample draws four opponent cooperation probabilities using the mixture in Table 5. Add offsets $(-1.2, -0.35, 0.35, 1.2)$ times the sampled spread to the sampled center, add Gaussian noise at the stated standard deviation, clip to $[0.02, 0.98]$, and sort the four values. The prior over the four sampled opponent types is uniform:

$$\mu_0(\theta_j) = \frac{1}{4}, \qquad j = 1, \ldots, 4.$$

Table 5: Type-profile for the potential-value cardinal sweep. Center and spread are uniform on the listed intervals.

Component	Probability	Center	Spread	Noise SD
Broad central	.55	[.35, .65]	[.18, .42]	.030
Narrow central	.17	[.45, .55]	[.03, .10]	.015
Low cooperation	.14	[.08, .28]	[.03, .12]	.015
High cooperation	.14	[.72, .92]	[.03, .12]	.015

The prior-informed action, potential value, and expected oracle crossing distance are all computed under this same prior.

For each sample, the sweep records m_g, r_g, $p_g^\star$, V_{pot}, $S_g = |R - T - S + P|$, and expected oracle crossing distance. Figure 2C displays at most 1,800 feasible samples.

Numerical summaries and their interpretation. Under equal weighting of the 12 canonical topologies, six admit a dominant action, accounting for the displayed 50.0% share. Under the $R > P$ orientation, these are five C-dominant classes and one D-dominant class. The displayed 25.0% unreachable-boundary and 25.0% potentially valuable shares additionally depend on the sampled type profiles; they are not prevalence estimates for social interactions. Within the feasible subset, the Spearman correlations of potential with oracle crossing distance and payoff slope are 0.86 and 0.53. Equation (42) links these quantities algebraically, so the correlations describe this ensemble rather than establish independent predictors of information value.

Holding the evaluation objective fixed. For material payoffs in a one-shot Prisoner's Dilemma, $T > R > P > S$ makes defection dominant, so perfect type information has zero potential. A prosocial objective changes this comparison. For example, with $u_i^\lambda = u_i + \lambda u_{-i}$ in a symmetric game, the cooperate-minus-defect gaps against C and D become

$$(1+\lambda)R - T - \lambda S, \qquad S + \lambda T - (1+\lambda)P,$$

respectively. Changing λ can change the effective decision regions. Any resulting information benefit is evaluated under that modified objective. Repeated interaction likewise changes the available policies and continuation returns. These distinctions concern the scope of the fixed-decision comparison; they are not additional experiments.

C.1.3 Acquired and Realized Value: Finite Evidence and Response Shift

Figure 3 combines fixed-channel acquisition summaries (A–B) and exogenous response-shift comparisons (C–E). Panel B evaluates $\eta_{\mathrm{acq}}(n) = V_{\mathrm{acq}}(n)/V_{\mathrm{pot}}(\mu_0)$ by exact finite integration with $T_B = 1$. In panels C–E, *semantic shift* denotes a change in type-conditioned opponent behavior. Drift magnitude alone does not determine the sign of realized value.

Acquisition observations and exact value calculations

Figure 3A uses a single fixed acquisition instance. For likelihood temperature T_B, we use the tempered posterior

$$\mu_n^{(T_B)}(\theta) \propto \mu_0(\theta)\, p(H_n \mid \theta)^{1/T_B},$$

so that $T_B = 1$ recovers exact Bayesian conditioning. All policy comparisons under tempered updates are evaluated using the source payoff model. The grid sweeps likelihood temperature $T_{\mathrm{B}} \in \{1, 2, 4, 6, 8, 12, 16, 24\}$ and noise $\varepsilon \in \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.48\}$. It records $N = \inf\{n \geq 1 : \pi^\star(\mu_n) \neq \pi_0\}$, taking $N = \infty$ if no change occurs; the largest displayed value is 894. Settings with $T_{\mathrm{B}} \neq 1$ use tempered beliefs, whose selected policies are evaluated through source payoff differences.

Figure 3B uses exact Bayesian conditioning ($T_B = 1$) and finite integration over the source channel's observation outcomes at each integer budget n. For each noise level, Eqs. (46)–(47) give $V_{\mathrm{acq}}(n)$; division by $V_{\mathrm{pot}}(\mu_0) > 0$ gives the displayed acquisition efficiency. This value averages payoff improvements over possible observations and is distinct from the first-switch counts in panel A.

First switching, acquisition, and completion. The first-switch count records when a selected action first differs from the prior action. It differs from the endpoint switch probability in Eq. (22), from the payoff gain conditional on switching, and from the completion criteria in Appendix A.4.

A first switch need not remove remaining potential and may be followed by later switches. With exact conditioning, Eq. (7) evaluates the probability-weighted payoff improvements over all possible observations. Tempered updates instead define an implemented rule whose source gain is evaluated as in Appendix A.6. Neither a large first-switch count nor positive initial potential alone establishes that continued acquisition has positive net value.

Exogenous target responses and aggregation Figure 3C uses a grid over source boundary $p^\star$ and semantic-shift intensity ζ. Source games are generated by enumerating the 24 strict ordinal payoff orders and drawing 200 ordinal-preserving cardinal resamples per order from $[-1.5, 1.5]$ with seed 20260506. The opponent response kernel sets $q^{\mathrm{obs}}(C) = 1 - \varepsilon$ if C has higher subjective utility for that type and $q^{\mathrm{obs}}(C) = \varepsilon$ otherwise, with $\varepsilon = 0.10$. Acquisition integrates exactly over $n = 8$ observations. Target behavior is

$$q^{\mathrm{tar}}(C; \zeta) = (1 - \zeta) q^{\mathrm{obs}}(C) + \zeta(1 - q^{\mathrm{obs}}(C)),$$

with ζ on a 101-point grid over $[0, 1]$. Cells plot median η_{real}, clipped to $[-1, 1]$ for color normalization, over 25 bins along $p^\star$ and 25 bins along ζ. Cells with insufficient support are masked; The valid-support contour uses cells with count at least 8, and the zero-efficiency contour is drawn at $\eta_{\mathrm{real}} = 0$.

Panels D–E of Figure 3 use source-side coordination and anti-coordination games paired with target shifts. The source set is generated from 500 ordinal-preserving cardinal resamples per strict ordinal order, $\varepsilon = 0.10$, $n = 8$, and the same seed 20260506; only feasible source games with $V_{\mathrm{acq}} \geq 10^{-3}$ and positive boundary-crossing probability are retained. For each retained source game, 100 target shifts are drawn. The target-shift family is sampled from $\{\mathrm{flip}, \mathrm{neutral}, \mathrm{permute}, \mathrm{random}, \mathrm{local}\}$ with probabilities $(0.25, 0.15, 0.25, 0.25, 0.10)$. For flip, neutral, permute, and random shifts, $\lambda \sim \mathrm{Unif}[0, 1]$ mixes q^{obs} with $1 - q^{\mathrm{obs}}$, the neutral vector 0.5, a random permutation of q^{obs}, or an iid $\mathrm{Unif}[0, 1]$ vector, respectively. For local shifts, $\sigma \sim \mathrm{Unif}[0, 0.45]$ and q^{tar} is q^{obs} plus iid Gaussian noise with standard deviation σ, clipped to $[0, 1]$. This yields 4369 retained source games and 436900 target-shift rows. The realized-efficiency cloud is subsampled to at most 100000 points for rendering with random seed 42; the semantic-drift axis uses the behavioral drift between source and target type profiles. The harm frequency of a source game is

$$\pi_{\mathrm{harm}} = \Pr_{\mathrm{shift}}[\eta_{\mathrm{real}} < 0],$$

estimated over its target-shift samples. For visualization, Gaussian jitter with standard deviation 0.015 is added to the drift cloud and jitter with standard deviation 0.005 is added to the $p^\star$ coordinate in the harm-frequency scatter.

Value definitions for finite observation counts Once the type prior and cooperation probabilities are given, finite-sample value can be computed without fitting a growth curve. Suppose the types in this reference calculation are finite and observations are conditionally independent Bernoulli draws with probabilities $q^{\mathrm{obs}}_{g,\theta}$. For k cooperative observations out of n, let $w_n(k)$ be the marginal count probability and $\mu_{n,k}$ its posterior:

$$w_n(k) = \sum_\theta \mu_0(\theta) \binom{n}{k} (q^{\mathrm{obs}}_{g,\theta})^k (1 - q^{\mathrm{obs}}_{g,\theta})^{n-k}, \tag{45}$$

$$\mu_{n,k}(\theta) = \frac{\mu_0(\theta)\binom{n}{k}(q^{\mathrm{obs}}_{g,\theta})^k (1 - q^{\mathrm{obs}}_{g,\theta})^{n-k}}{w_n(k)}. \tag{46}$$

The posterior is used only when $w_n(k) > 0$; zero-mass counts do not contribute. Let $a_{n,k} = \pi^\star(\mu_{n,k})$ and $a_0 = \pi_0$. Then

$$V_{\mathrm{acq}}(n) = \sum_{k : w_n(k) > 0} w_n(k) \left[U(a_{n,k}, \mu_{n,k}) - U(a_0, \mu_{n,k}) \right]. \tag{47}$$

These equations instantiate the general acquired-value definition for a finite observation model; they specify an evaluation method rather than an additional experimental instance. For an implemented rule $\widehat{a}_{n,k}$, source or target gain is evaluated directly as

$$\sum_{\theta,k} \mu_0(\theta) P(k \mid \theta) \left[U_g(\widehat{a}_{n,k}, q^{\mathrm{eval}}_{g,\theta}) - U_g(a_0, q^{\mathrm{eval}}_{g,\theta}) \right], \tag{48}$$

where U_g is defined in Eq. (38), $P(k \mid \theta)$ is the source binomial count law, and $q^{\mathrm{eval}}_{g,\theta}$ is the evaluation response probability. The expression also evaluates policies selected using tempered beliefs. The negotiation study separately specifies its endogenous language process, working-model values, and true-type trajectory evaluation.

Interpretation of target-shift frequencies. The plotted target-shift summaries describe the experimenter's specified ensemble. Each source game contributes multiple shifts, giving clustered game–shift rows. Panels D–E condition on feasible crossing and $V_{\mathrm{acq}} \geq 10^{-3}$, which stabilizes the ratio and defines the population summarized.

Decision interpretation. On the retained positive-acquired-value population, $\eta_{\mathrm{real}} = V_{\mathrm{real}}/V_{\mathrm{acq}}$ has the same sign as realized value. A negative value means that the frozen informed policy underperforms the source prior-informed baseline under the specified target evaluation. It does not imply that every individual target outcome is harmful or that an implementable agent knows the target-optimal policy. The phase and harm-frequency profiles vary with the source boundary and the prescribed shifts. Similar drift magnitudes can yield different payoff effects because their direction and location relative to the decision boundary matter.

This diagnosis motivates validating the type-to-behavior mapping and recomputing policy comparisons using a supported target model. The plotted comparisons neither estimate such a model nor measure the benefit or cost of repairing it. Holding the source observation law fixed makes additional source evidence insufficient, by itself, to identify the changed deployment mapping.

C.1.4 Policy-Dependent Target Response

Figure 4 extends the response-shift comparisons in this appendix to a policy-dependent opponent, instantiating $\Psi_{-i}(\theta, \pi)$ in Eq. (32) rather than the exogenous shifts used in Figure 3C–E.

Mechanism and protocol. The target response is $q^{\mathrm{tar}}(\theta, \pi; \rho) = q^{\mathrm{obs}}(\theta)$ if $\pi = D$ and $(1-\rho)q^{\mathrm{obs}}(\theta)$ if $\pi = C$, with reactivity $\rho \in [0, 1]$: opponent behavior is when the focal agent commits to cooperation. Sampling follows Figure 3D–E except for the opponent type distribution: seed 20260506, the 24 strict ordinal payoff orders, 500 cardinal resamples per order from $\mathrm{Unif}[-1.5, 1.5]$, $n = 8$ exact observations, and the $V_{\mathrm{acq}} \geq 10^{-3}$ feasibility filter. The opponent type distribution uses the four-component mixture in Table 5; 2527 games are retained.

Panels and statistics. Panel A confirms that ρ=0 recovers $\eta_{\mathrm{real}} \equiv 1$ and $\mathrm{Corr} \equiv 0$ to machine precision, the exact policy-independent limit used elsewhere in this appendix. Panel B shows the population harm frequency $\Pr_{\mathrm{games}}[\eta_{\mathrm{real}}(\rho) < 0]$ rising monotonically to 43.0% at ρ=1. Because $\eta_{\mathrm{real}}(\rho)$ is exactly affine in ρ for each game (verified to 2.8×10^{-14}), Panel C reports the closed-form harm fraction of $\rho \in [0, 1]$ per game against $p^\star_g$: harm concentrates in, but is not exclusive to, high-boundary games, with 68.1% of games at $p^\star_g \geq 0.5$ harmed at ρ=1 versus 17.9% below that threshold (minimum harmed $p^\star_g = 0.054$), so $p^\star_g \geq 0.5$ marks a high-incidence region rather than a strict boundary. Panel D pools V_{dir}, Corr, and V_{real} over the $p^\star_g \geq 0.5$ subset ($n = 1264$); at ρ=1 the aggregate $\eta_{\mathrm{real}} = -1.00$ splits exactly into $V_{\mathrm{dir}}/V_{\mathrm{acq}} = -0.84$ and $\mathrm{Corr}/V_{\mathrm{acq}} = -0.16$, confirming Eq. (34) by direct construction rather than by fit.

C.2 Evidence-Action Values and Sequential Decisions

This subsection supplies the finite-channel calculation and continuation recursion used in Section 4.2. The worked complementarity example is analytical.

C.2.1 Posterior branches, costs, and sequential complementarity

For a finite type space and channel likelihood $L_q(y \mid \theta)$, compute

$$p_q(y) = \sum_\theta \mu(\theta) L_q(y \mid \theta), \qquad \mu^{q,y}(\theta) = \frac{\mu(\theta) L_q(y \mid \theta)}{p_q(y)}.$$

Terms with $p_q(y) = 0$ are omitted. Substitution in Eq. (7) gives an exact finite calculation under the specified model. Sampling posterior branches gives a Monte Carlo alternative, with model and maximization bias considered in its evaluation.

For an arbitrary channel, set $M(\theta) = \max_\pi u(\pi, \theta)$. Conditional expectation gives $\mathbb{E}_Y \mathbb{E}_{\mu^{q,Y}} M = \mathbb{E}_\mu M$. Subtracting and adding $\mathbb{E}_Y W(\mu^{q,Y})$ yields Eq. (5). The same argument covers adaptive pure-information procedures through their complete observation histories.

With at most k pure-information actions remaining, terminal decision value $J_0(\mu) = W(\mu)$, and nonnegative costs,

$$J_k(\mu) = \max\left\{W(\mu), \max_q \left[-c(q) + \sum_y p_q(y) J_{k-1}(\mu^{q,y})\right]\right\}.$$

Here J_k is the net continuation value for the specified horizon and channel library, including acquisition costs. Its maximizing branch determines whether to act now or acquire further evidence.

For an explicit complementarity example, let a binary type have prior $1/2$. A safe action pays zero; relying pays $+1$ for type one and -2 for type zero. A conditionally independent binary signal is correct with probability 0.7. After one positive signal the posterior is 0.7, so relying has payoff 0.1; after a negative signal the safe action remains optimal. One-signal gross value is 0.05. Two signals produce posterior $0.49/0.58$ after two positives, with outcome probability 0.29, yielding gross value 0.155; the other outcomes select safe. At cost 0.06 per signal, one signal has negative net value -0.01, whereas two forced signals have positive net value 0.035. The two-signal procedure therefore has positive net value even though a one-signal procedure loses value.

D Final-Offer Negotiation: Protocol and Diagnostics

This study evaluates whether value-guided decisions address three decision-value limitations in LLM negotiation. Negotiation instances are randomly sampled and then grouped retrospectively as redundant identification, inefficient acquisition, or misplaced trust according to their preference supports, contract payoffs, evidence channels and costs, and disclosure settings. Baseline behavior is summarized within these groups before evaluating the corresponding interventions. Within each evaluated context, every declared type and disclosure condition is assessed. Enumerating those conditions is an evaluation procedure within sampled contexts, not a mechanism for generating the diagnostic groups.

D.1 Diagnostic rules and matched baselines

The post-sampling grouping follows the decision-value distinctions in Section 4.3. It uses the instance's preference support, payoff and acceptance structure, evidence channels and costs, and disclosure setting. Group membership is determined from these sampled instance parameters before trajectory outcomes are analyzed; the associated baseline trajectories are then used to examine how the corresponding limitation manifests in behavior.

Redundant identification. The parameter condition is a common optimal contract over the supported opponent types under the specified payoff and acceptance model. This makes potential value zero even when type uncertainty remains. The paired comparison then examines whether Prediction-first purchases further information and whether value-guided stopping avoids that cost; these behaviors are evaluation outcomes, not prerequisites for assigning the context to this group.

Inefficient acquisition. The preference support contains contract-relevant distinctions, and the available channels and costs permit a beneficial information-dependent decision or a cheaper way to resolve the same distinction. In final-offer negotiation, the comparison uses No language, which lacks the cheaper question channel but retains independent verification. This contrast measures channel availability rather than a planning error under identical access. In standing-offer negotiation, the comparison uses Aspiration to test whether immediate acceptance forgoes a beneficial evidence-dependent revision. The case parameters identify the opportunity; the agents' acquisition and revision decisions are evaluated afterward.

Misplaced trust. The preference and disclosure settings permit a misleading claim to support an extra concession that benefits the opponent. The grouping concerns this contract-relevant source-reliability problem; whether an agent relies on the claim, verifies it, or improves its return is assessed from the subsequent trajectories. The comparator is Naive VoI, whose fixed source-reliability assumption

can make verification appear unprofitable even when a reliability-sensitive model assigns that action positive net value. Both methods retain the same verification action, so the contrast concerns source assessment rather than audit access.

The labels index the decision-value limitation examined by each paired comparison and need not be mutually exclusive descriptions of an interaction. Honest and withholding disclosure conditions remain controls within the same preference context, rather than additional instances of a misleading-claim failure. Standing-offer negotiation uses the same parameter-based grouping logic, with its sampling and cross-model evaluation described in Appendix E.

D.2 Environment and evidence channels

For items A, B, and C, let $f, \theta \in \{0, \dots, 10\}^3$ denote focal and opponent values, each summing to ten. A contract $d \in \{0,1\}^3$ assigns item j to the focal agent when $d_j = 1$. Utilities are $u_F(d) = f^\top d$ and $u_O(d, \theta) = \theta^\top (1 - d)$. The public prior is uniform on the evaluated context's support in Table 6. The opponent accepts exactly when $u_O(d, \theta) \geq \tau$, with $\tau = 5.5$ in the primary evaluation. A rejected contract or voluntary exit gives focal gross payoff zero. The opponent's disagreement payoff is τ.

Table 6: Preference supports of the evaluated contexts, grouped after sampling by decision-relevant parameter conditions. The focal agent observes the support but not the current opponent type or disclosure condition. Tuples give values of (A,B,C).

Diagnostic comparison	Focal values f	Opponent support Θ
Redundant identification	$(6, 2, 2)$	$(0, 5, 5), (1, 4, 5), (1, 5, 4), (2, 4, 4)$
Inefficient acquisition	$(5, 4, 1)$	$(6, 3, 1), (7, 2, 1), (1, 3, 6), (1, 2, 7)$
Misplaced trust	$(4, 4, 2)$	Core: $(2, 2, 6)$; dual-demand: $(5, 0, 5)$

Before one final proposal, the focal agent has at most three information actions. A language question asks whether an item value meets a specified integer threshold and costs 0.1. The available questions cover the distinct nontrivial threshold partitions of the public support. An independent verification costs 0.6 and returns the exact value of one selected item. Fees are paid when requested, including unanswered questions. Information actions leave the acceptance rule unchanged. The agent can instead submit any of the eight allocations or exit. NegMAS (version 0.15.7) executes the final offer and the true-utility response; the information stage is implemented separately. There are no counteroffers in this task. Total return is

$$R_F = u_F(d)\mathbf{1}\{u_O(d, \theta) \geq \tau\} - 0.1N_{\text{ask}} - 0.6N_{\text{verify}}, \tag{49}$$

with zero gross payoff on exit. These costs are in game utility units; they do not measure API expenditure or latency.

Disclosure conditions. Opponents in the redundant-identification and inefficient-acquisition contexts invite threshold questions and answer according to their true values. In the misplaced-trust context, the core type either reports honestly, withholds preference statements and declines questions, or consistently reports the dual-demand profile $(5, 0, 5)$. The dual-demand type also has an honest control. The strategic opening states that A is worth at least five and requests A and C, leaving B to the focal agent. Subsequent answers follow that same reported profile, while acceptance always uses true values. This deception has a payoff incentive: the true core opponent receives utility eight from A+C, compared with six from C alone; the focal agent's gross payoff falls from eight to four when it concedes A. Each condition has two fixed phrasings of the opening message.

Language-enabled agents also observe independent historical source audits: 48 matching and zero mismatching reported profiles in the redundant-identification and inefficient-acquisition contexts, and four matching and two mismatching profiles in the misplaced-trust context. These are past-transaction records, not labels of the current opponent's honesty. They provide observable reliability evidence before verification. No language removes both these language-source records and all opponent text. The strategic report is internally consistent before audit; this task tests whether source uncertainty justifies verification and how an audited contradiction changes the contract decision.

D.3 LLM DECISION FLOW AND WORKING-MODEL CALCULATIONS

Inputs and responsibilities. At each decision, the LLM sees its own values, the public type support and prior, all eight explicitly labeled allocations, the visible history, costs, acceptance threshold, and remaining information budget. It receives no true-type or disclosure-condition label. In a first call, it extracts quoted value predicates from opponent messages and estimates source reliability r_t and question-answer probability a_t. Verification results are supplied as independent observations. A numerical tool conditions the working model and evaluates every legal contract and information action. In a second call, the same LLM reads those values and submits an executable action with a short explanation. The selected action is executed as submitted. Format and legality failures can be corrected through the same interface; legal departures from a method's objective remain in the results.

Correlated language evidence. Let $\phi \in \Theta$ be the opponent's reported preference profile, held constant within a conversation. For $K = |\Theta|$, the working channel is

$$P_t(\phi \mid \theta) = r_t \mathbf{1}\{\phi = \theta\} + \frac{1 - r_t}{K - 1} \mathbf{1}\{\phi \neq \theta\}. \tag{50}$$

Let C_t be the extracted language predicates and Z_t the exact verification observations. The joint working posterior is

$$b_t(\theta, \phi) \;\propto\; \mu_0(\theta) P_t(\phi \mid \theta) \mathbf{1}\{\theta \text{ agrees with } Z_t\} \mathbf{1}\{\phi \text{ satisfies all } C_t\}, \qquad \mu_t(\theta) = \sum_{\phi} b_t(\theta, \phi). \tag{51}$$

Consequently, repeated claims constrain one profile rather than acting as independent confirmations. An answered question reveals its predicate on ϕ; a verification reveals an item value of θ. A refusal, assigned probability $1 - a_t$, leaves the belief unchanged. The reliability estimate lies in $[0.05, 0.99]$ and can be revised by the LLM after new observations. Conditioning is exact within this finite working model; reliability and willingness-to-answer estimates remain LLM judgments.

From the paper's values to action scores. The contract return table is

$$w(d, \theta) = u_F(d) \mathbf{1}\{u_O(d, \theta) \geq \tau\}, \qquad W(b) = \max_d \sum_{\theta, \phi} b(\theta, \phi) w(d, \theta).$$

It directly instantiates Eqs. (1) and (7). The tool computes potential value and, for each information action q, the expected improvement after conditioning, its policy-switch probability β_q, and conditional improvement Γ_q when switching has positive probability. With no switch, acquired value is zero. With ℓ information actions left, it also evaluates the finite continuation recursion

$$J_0(b) = W(b), \qquad J_\ell(b) = \max \left\{ W(b), \max_q \left[-c(q) + \sum_y P(y \mid b, q) J_{\ell - 1}(b^{q,y}) \right] \right\}. \tag{52}$$

These hypothetical branches hold the current reliability and answer-rate estimates fixed. The LLM can reassess those estimates after an actual observation. Contract values, V_{pot}, $V_{\text{acq}} = \beta\Gamma$, and the continuation scores are provided together, giving the agent explicit stopping and acquisition criteria. Computing J_ℓ does not force the LLM to implement its maximizing action.

Method differences. Full is instructed to use decision sufficiency and expected net continuation value. No language uses the same LLM analysis and action interface, with opponent text, source-language audits, and questions removed; it still receives exact verifications and value calculations. Naive VoI (fixed reliability) receives the same VoI evaluation procedure, acquisition actions, and payoff model as Full but holds $r_t = 0.97$ for language evidence throughout instead of reassessing that reliability from the available source evidence. It remains a VoI decision method, not an information-gain or no-verification baseline. Prediction-first uses the same evidence analysis but, while maximum type probability is below 0.90 and informative actions remain, receives type-information scores and is instructed to maximize information gain. Cost breaks ties only within 0.01 bit. It switches to payoff-based choice once its confidence target is met or identification cannot continue. The common allocation catalog remains visible in both phases.

D.4 EVALUATION INSTANCES AND MODEL SETTINGS

The final-offer evaluation includes all four types in each of the redundant-identification and inefficient-acquisition contexts. For the misplaced-trust context, it includes honest, strategic, and withholding

disclosure by the core type, plus honest disclosure by the dual-demand type. Twelve conditions, two phrasings, and four methods give 96 episodes per focal model. An additional 16 response-boundary control episodes apply two acceptance thresholds, two phrasings, and four methods to the strategically reporting core type. The environments are paired across methods; fresh LLM responses supply decision variability. Phrasing variants are not independent draws of negotiation environments, and the exhaustive within-context evaluation does not provide a prevalence estimate for the diagnostic categories.

DeepSeek V4 Pro and GPT-6 Sol are the two focal-model configurations used in the evaluation and response-boundary controls. DeepSeek requests enable thinking with low reasoning effort and a 16,384-token cap. GPT uses the gateway's default thinking behavior, temperature 0.2, and a 4,096-token cap. Gateway model labels do not independently attest upstream weights, and the configurations do not provide an equal-compute comparison.

D.5 Results, controls, and objective adherence

Table 7 gives all final-offer evaluation conditions. Means average the two phrasings within each true-type condition, then weight the four types equally within the redundant-identification and inefficient-acquisition contexts. Disclosure conditions in the misplaced-trust context remain separate so that misleading-claim gains and honest-case verification costs are both visible. Net return evaluates each complete implemented procedure, including its information costs. These are paired method comparisons; their differences need not equal the formal V_{real} comparison with a prior-informed baseline.

Table 7: All final-offer evaluation conditions. Entries are mean net return. Each redundant-identification or inefficient-acquisition cell averages eight episodes; each disclosure-control cell averages two.

Model	Condition	Full	No language	Naive VoI	Prediction-first
DeepSeek	Redundant identification	6.000	6.000	6.000	5.100
	Inefficient acquisition	6.900	6.400	6.838	6.025
	Misplaced trust: strategic core	7.400	7.400	4.000	7.400
	Honest core control	7.400	7.400	8.000	7.400
	Withholding core control	7.400	7.400	7.400	7.400
	Honest dual-demand control	3.400	3.400	4.000	3.400
GPT	Redundant identification	6.000	6.000	6.000	5.100
	Inefficient acquisition	6.900	6.400	6.900	6.163
	Misplaced trust: strategic core	7.400	7.400	4.000	7.400
	Honest core control	7.400	7.400	8.000	7.400
	Withholding core control	7.400	7.400	7.400	7.400
	Honest dual-demand control	3.400	3.400	4.000	3.400

Response-dependent value. With $\tau = 4.9$, giving C to the opponent is sufficient for both types in the misplaced-trust context, so retaining A+B is optimal in either case. With $\tau = 6.1$, retaining B and giving A+C is optimal for both types, although the core type also admits another tied optimum. In both settings Full stops without information despite the same strategic opening. Table 8 shows that the value of resolving a reported preference depends on its consequences under the response rule. These controls change a known threshold across tasks; they do not estimate an unknown response law online.

Table 8: Response-boundary controls. DeepSeek and GPT have identical means in every cell; each model contributes two episodes per method and threshold.

Threshold	Full	No language	Naive VoI	Prediction-first
$\tau = 4.9$	8.000	8.000	8.000	7.400
$\tau = 6.1$	4.000	4.000	4.000	3.400

Implementation and adherence checks. Every DeepSeek/GPT trajectory in the final-offer evaluation and response-boundary controls passes checks of quoted predicate extraction, visible-input

restrictions, action execution, cost accounting, final acceptance, and exact replay. DeepSeek has no recorded objective departures. GPT has one Prediction-first departure in its 96-episode evaluation: after narrowing to two types in the inefficient-acquisition context, it chooses a 0.1 question with information gain 0.943456 bit over a 0.6 verification with gain one bit, exceeding the prescribed tie tolerance. Its resulting net return, 8.3, is included in the reported results. Thus one of the 24 GPT Prediction-first evaluation episodes departs from its identification objective; the other methods have no recorded objective departures.

A separate source-assessment error remains visible in DeepSeek: one Naive VoI trajectory in the inefficient-acquisition context estimates answer probability zero despite the opponent's invitation to ask and therefore purchases verification. This yields 4.4 instead of 4.9 and lowers that method's mean in this context to 6.8375. The action is consistent with its stated working model, illustrating the distinction between evidence-model assessment and action adherence.

D.6 Observed decision points and value criteria

Let $X \mid Y$ denote a contract assigning items X to the focal agent and Y to the opponent. Figure 5 summarizes the decisive steps reported in the DeepSeek final-offer evaluation. The baseline in each panel matches Table 1. These trajectory summaries illustrate the aggregate comparisons rather than adding independent evidence.

Redundant identification

True opponent: $(1, 4, 5)$.
Full
$V_{\text{pot}} = 0$ despite type uncertainty.
↓
Propose $A \mid BC$; accepted.
Net: 6.0

Prediction-first
Verify A → 1, then B → 4.
Total acquisition cost: 1.2.
↓
Propose $A \mid BC$; accepted.
Net: 4.8

Inefficient acquisition

True opponent: $(1, 3, 6)$.
Full
Ask "Is A worth at least 5?"
↓ answer: no; cost 0.1.
Residual $V_{\text{pot}} = 0.0667$.
Propose $AB \mid C$; accepted.
Net: 8.9

No language
Resolve the contract group by exact verification; cost 0.6.
↓
Propose $AB \mid C$; accepted.
Net: 8.4

Misplaced trust

Claim: "A is worth at least 5."
True opponent: $(2, 2, 6)$.
Full
Core-type probability 0.3333.
Verify A → 2; cost 0.6.
↓
Core-type probability 1.
Propose $AB \mid C$; accepted.
Net: 7.4

Naive VoI
Core-type probability 0.03.
Propose $B \mid AC$; accepted.
Net: 4.0

Figure 5: Key decision points in final-offer negotiation. Full stops when identification cannot improve the contract, uses a cheaper contract-relevant question, or verifies a consequential claim. Utterances and paths are abbreviated to the evidence channels, contract decisions, and acquisition costs relevant to each comparison.

Redundant identification: a common optimal contract. For every type in this context, $A \mid BC$ gives focal utility six and is accepted. No higher-valued focal allocation is accepted. Potential value is therefore zero under every posterior on this support. In the shown case, Prediction-first verifies two items before proposing the same contract, so its entire 1.2 loss is information cost. The diagnostic criterion concerns the equality of optimal contracts across types; Full's cost saving is the separately evaluated intervention outcome.

Inefficient acquisition: a contract-relevant partition. The first two types in this context require A, favoring $BC \mid A$ with focal utility five; the other two favor $AB \mid C$ with utility nine. At the uniform prior, the best expected contract return is 4.5, while perfect information gives seven, so $V_{\text{pot}} = 2.5$. The question about A separates the two contract groups. In the displayed low-A trajectory, Full's posterior is approximately $(0.0067, 0.0067, 0.4933, 0.4933)$: full type identification is incomplete, but remaining potential 0.0667 is below the cheapest further query cost 0.1. This instantiates the cost ceiling in Eq. (19). High-A answers instead shift the preferred contract from $AB \mid C$ to $BC \mid A$. A realized low-A answer need not change the prior-selected contract for the question to have positive acquired value before observation. No language independently verifies an item and reaches the same type-appropriate contract, paying 0.5 more. Prediction-first is an additional identification-focused comparison: in the reported low-A trajectory it verifies A and B at total cost 1.2, then proposes $AB \mid C$ for net return 7.8.

Misplaced trust: reliability changes whether verification pays. Let p be current probability of the core type after the misleading opening. Giving A+C away guarantees focal utility four; retaining A+B gives expected utility $8p$. When $p \leq 1/2$, acting now therefore gives four. Exact verification of A distinguishes the types and has net continuation value $4 + 4p - 0.6$, exceeding four exactly when $p > 0.15$. More generally, immediate return is $\max\{4, 8p\}$, so verification strictly beats acting now exactly for $0.15 < p < 0.85$ (with indifference at the endpoints). In the observed Full trace, $p = 0.3333$ makes verification worthwhile. Naive VoI instead yields $p = 0.03$ and value 3.52, so the same available verification is declined. The verified value A=2 subsequently identifies the core type and supports $AB \mid C$. The two methods differ in their assessed value of verification, not in access to it. Honest-source controls show the complementary limit: when the report is true, verification can lower net return by its fee rather than improving the contract.

E Standing-Offer Negotiation and Cross-Model Evaluation

This study examines the same diagnostic categories when an agent can accept an incumbent agreement or seek evidence before revising it. Instances are randomly sampled and then grouped retrospectively according to their decision-relevant parameters. We summarize baseline behavior within each group and evaluate the corresponding interventions.

E.1 Protocol, opponent, and focal methods

Three indivisible items have integer values summing to ten. The focal agent knows its own values, a uniform prior over a public finite opponent domain, all eight allocations, a binding standing offer, and the opponent's reservation commitment. The true opponent vector remains private. An episode has four focal decisions. The focal agent may accept the standing offer, ask a threshold question, purchase an exact-value verification, terminate, or submit one terminal counteroffer. Asking or verifying consumes a decision and incurs its declared case cost without revoking the standing offer. A terminal counteroffer is accepted exactly when it gives the opponent at least its committed true utility; either response ends the episode. Net return is focal contract utility minus communication and verification costs.

The opponent is a bounded LLM negotiator without the focal value tool, posterior calculation, or access to focal values. In the redundant-identification and inefficient-acquisition cases it gives a guarded opening and answers questions truthfully. In the misplaced-trust cases it follows a strategic brief: it falsely states that secondary item A has high value, reinforcing the standing offer that gives it both A and its true core item C. It may maintain or weaken that claim when questioned, but its actual utility and reservation commitment are unchanged. Thus a successful deception transfers a genuinely valuable item from the focal agent to the opponent; it does not merely hide the opponent's preferred allocation.

Full, Prediction-first, Naive VoI, and No language retain the definitions in Appendix D.3. Aspiration adds one stopping rule: it immediately accepts a binding offer whose focal utility reaches a declared aspiration level. The matched baselines are Prediction-first for redundant identification, Aspiration for inefficient acquisition, and Naive VoI for misplaced trust. Each primary Full–baseline pair shares the allocation catalog, public type domain, and acquisition actions. Naive VoI retains verification, so its comparison changes source assessment rather than audit access. The supplementary No-language runs remove dialogue and questions as specified in the shared method definition.

These comparators are fixed by the diagnostic mechanism under study before trajectory outcomes are examined: Prediction-first isolates continued identification, Aspiration isolates premature stopping, and Naive VoI isolates source-reliability assessment. The matched comparisons are therefore mechanism-specific intervention tests rather than an overall ranking of the methods.

E.2 Random Sampling, Parameter-Based Grouping, and Paired Evaluation

After random sampling, we group instances according to the parameter conditions in Appendix D.1: a common optimal contract across the type support; contract-relevant uncertainty that available evidence can resolve at acceptable cost; or a disclosure setting in which a misleading priority claim can sustain an unnecessary concession. Grouping is retrospective and parameter-based: group

membership is determined from the sampled instance parameters before trajectory outcomes are analyzed. We then summarize baseline behavior and evaluate the matched interventions within these groups.

The evaluated sample contains 15 numerically distinct cases, with five in each of the three parameter-based groups. Each case has three repetitions, identified by 21001–21003, and a matched Full–baseline comparison. Repetitions measure behavioral variation on the same sampled case rather than independent draws of new negotiation environments. All completed case–repetition pairs enter the reported means and win counts.

The sampled cases vary in focal values and item orderings. Redundant-identification cases use eight-type domains, standing allocation $A \mid BC$, reservation commitments of six or seven, communication-cost parameters 0.42–0.55, and verification-fee parameters 0.15–0.20. Inefficient-acquisition cases use two-type domains, standing allocation $B \mid AC$, commitments of six to eight, communication-cost parameters 0.06–0.10, and verification-fee parameters 0.13–0.20. Misplaced-trust cases use two-type domains, the same standing allocation, commitments of six or seven, communication-cost parameters 0.10–0.12, and verification-fee parameters 0.30–0.55. All net-return comparisons include the total communication and verification charges incurred along the trajectory.

Case identifiers are D01–D05 for redundant identification, I01–I05 for inefficient acquisition, and S01–S04 plus S06 for misplaced trust. These identifiers are used only for traceability and do not define the diagnostic categories.

The sampled cases, parameter-based groups, prompts, costs, method objectives, and opponent briefs are held fixed across the two focal-model evaluations. DeepSeek contributes 45 paired comparisons, or 90 trajectories, and GPT evaluates the same 45 case–repetition pairs; its opponent remains DeepSeek V4 Pro. Repetition identifiers randomize domain order and request nonces but do not control service-side sampling.

Table 9: Standing-offer comparisons on the randomly sampled cases, grouped retrospectively by decision-relevant parameters. Each group names its matched baseline. Pairs count completed case–repetition comparisons; wins count pairs in which Full achieves strictly higher net return. All completed pairs enter the reported means and counts.

Diagnostic group	Comparator	Pairs	Full	Baseline	Difference	Wins
DeepSeek V4 Pro: sampled-case evaluation						
Redundant identification	Prediction-first	15	6.400	5.096	1.304	15
Inefficient acquisition	Aspiration	15	6.160	3.200	2.960	15
Misplaced trust	Naive VoI	15	7.230	3.400	3.830	15
GPT-6 Sol: same-case evaluation						
Redundant identification	Prediction-first	15	6.400	5.096	1.304	15
Inefficient acquisition	Aspiration	15	5.960	3.200	2.760	14
Misplaced trust	Naive VoI	15	7.230	3.400	3.830	15

E.3 Observed trajectories and value criteria

Figure 6 summarizes the common action sequence in the redundant-identification comparisons and one logged example for each of the other diagnoses. These are explanatory trajectory summaries, not additional evaluation episodes. The accompanying text retains the decisive observations, contract changes, and reported returns.

Redundant identification: potential value stops identification. Across these cases, all eight supported types accept the standing allocation $A \mid BC$, and no accepted alternative improves focal utility. Full therefore assigns no contract value to identifying the exact type. Prediction-first verifies two item values before accepting the same standing allocation. Its pairwise loss of 1.16–1.40 is entirely total acquisition cost. The relevant decision point precedes any purchase: uncertainty remains, but the optimal contract does not depend on resolving it.

Inefficient acquisition: evidence supports a contract revision. The two supported types require different opponent bundles to meet the reservation commitment. Aspiration stops because the incumbent gives the focal agent slightly more than its threshold. Full compares that return with

Diagnostic group	Baseline trajectory	Full's decisive steps
Redundant identification	Prediction-first verifies two values, then accepts the incumbent $A \mid BC$.	Recognizes a common optimal contract; accepts $A \mid BC$ without acquisition. The final allocation is unchanged.
Inefficient acquisition	Aspiration accepts $B \mid AC$ because its return meets the aspiration level.	Verifies A $\rightarrow$ 7; replaces $B \mid AC$ with $BC \mid A$. Net return is 5.78 rather than 3.00 in `I01`.
Misplaced trust	Naive VoI relies on the false priority for A and accepts $B \mid AC$.	Verifies A $\rightarrow$ 2; selects $AB \mid C$. Net return is 7.45 rather than 4.00 in `S01`, while the opponent still accepts.

Figure 6: Critical decisions in standing-offer negotiation. The first row describes the shared sequence across the sampled redundant-identification cases; the other rows summarize recorded cases. The gains arise from avoiding unnecessary identification, revising a contract after useful evidence, and removing an excess concession sustained by a misleading claim.

the outcome-conditioned continuation value of verification. In `I01`, verifying A=7 changes the preferred allocation from the incumbent $B \mid AC$ to $BC \mid A$; Full returns 5.78 after cost versus Aspiration's three. The acquisition is useful because the subsequent contract recovers a valuable item, not because a more complete preference description is intrinsically rewarded. The GPT exception in `I03/21002` localizes the remaining failure after acquisition: the correct observation is obtained, but the implemented action does not realize the available contract improvement.

Misplaced trust: source uncertainty makes verification valuable. The opponent's false priority for A supports the incumbent $B \mid AC$, which yields it utility eight in `S01`. Naive VoI treats the claim as reliable and accepts. Full assigns enough probability to the competing explanation that verifying A has positive net continuation value. The observed A=2 rules out the inflated claim and supports $AB \mid C$, which still meets the opponent's reservation utility six. Full returns 7.45 after cost instead of four, while the opponent loses the two-unit gain created by its deception. In `S02/21002`, Full verifies C rather than A; C=7 uniquely determines the type in the public domain and resolves the same contract choice. This observation illustrates selection by decision relevance rather than by lexical overlap with the disputed statement.

E.4 Execution checks

All 90 DeepSeek evaluation trajectories and all 90 GPT trajectories replay from recorded model outputs with identical histories, actions, costs, acceptance decisions, and returns. The checks compare claim extraction with visible utterances, restrict focal inputs to public state, match every selected action to the executed event, and verify $V_{\text{acq}} = \beta\Gamma$ and the potential-value bound. Format-correction requests address malformed or contract-inconsistent outputs; they do not select or override the focal action. GPT's inefficient-acquisition exception is an implemented decision outcome, not an interface failure.

F Context: Preferences, Strategic Types, and Repeated Interaction

Preference models and strategic reasoning. Social-preference models describe fairness, inequity aversion, reciprocity, and belief-dependent motives (Rabin, 1993; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Battigalli & Dufwenberg, 2007). Bounded strategic reasoning concerns how agents model others' choices, as in level-based accounts and cognitive hierarchies (Stahl & Wilson, 1994; Nagel, 1995; Camerer et al., 2004; Crawford et al., 2013). These approaches offer ways to specify opponent types and their source strategies. The focal evaluation objective is specified separately. Belief-dependent utilities additionally require a consistent model of those beliefs. Holding the focal objective fixed permits comparison of policies with different information.

Strategic communication and source reliability. Strategic information transmission studies communication when a sender's preferences differ from a receiver's (Crawford & Sobel, 1982); experimental work also examines how the consequences of deception affect disclosure (Gneezy, 2005). ToM-aware LLM agents have been evaluated in imperfect-information games (Guo et al., 2024). These connections motivate distinguishing an opponent's statements from its payoff-relevant characteristics. Our negotiation protocols use a declared working language channel and binding acceptance rule; they neither estimate an equilibrium of unrestricted communication nor establish how people would disclose information.

Sequential interaction and policy evaluation. Bayesian intention models have also been used to coordinate cooperative task allocation (Lim et al., 2020; Wu et al., 2021). Sequential social dilemmas and Bayesian reciprocity examine cooperation over repeated encounters (Leibo et al., 2017; Kleiman-Weiner et al., 2025). Repeated interaction expands the policy class to include contingent responses such as retaliation and reciprocity, so its information value is evaluated over complete interaction plans. Strong imperfect-information game play, exemplified by DeepStack and multiplayer poker systems (Moravčík et al., 2017; Brown & Sandholm, 2019), similarly involves complete decision systems. Isolating the contribution of a latent-type model within such systems requires matched comparisons between informed and prior-informed policies under a common objective and policy class.