Hypothesis Memo:
Information Advantage as a Bargaining Curse

Pre-design

February 2026

Status: Pre-design — ready for `/design-experiment`

1 Research Question

Does revealing the opponent's reservation price to one bargainer *reduce* the overall deal rate compared to symmetric ignorance, even when a zone of possible agreement exists?

Bilateral bargaining under incomplete information is a canonical setting in economics: Myerson and Satterthwaite (1983) proved that no mechanism achieves fully efficient trade when both valuations are private. The standard intuition is that *more* information should make deals easier—an informed bargainer can always find mutually acceptable terms. We hypothesize the opposite: that giving one player complete knowledge of the surplus space triggers exploitative anchoring behavior that paradoxically *reduces* deal rates. If confirmed, this result challenges a core assumption in mechanism design and connects the Myerson–Satterthwaite impossibility to a behavioral rather than strategic failure mode.

2 Interestingness Argument

This hypothesis is worth testing for three reasons:

1. **Counterintuitive direction.** The naïve prediction is "more information = more deals." An informed bargainer knows the zone of agreement exists and can compute acceptable terms. Our hypothesis predicts the opposite: information advantage triggers behavioral shifts (exploitative anchoring) that reduce efficiency. A confirmed result would be a genuine surprise.

2. **Observable mechanism.** The mediating variable—offer anchoring—is directly measurable in game transcripts. We can compute where the human's first offer falls relative to (a) the opponent's reservation price and (b) the midpoint of the zone of agreement. This lets us test *why* the effect occurs, not just *whether* it occurs. Formal mediation analysis is feasible.

3. **Bridges theory and behavioral economics.** Myerson–Satterthwaite says incomplete information prevents efficient trade for *strategic* reasons (incentive compatibility). Our hypothesis suggests that *providing* information can also prevent efficient trade, for *behavioral* reasons (exploitative anchoring, overconfidence). This is a distinct failure mode that the classical theory does not address.

3 Causal Model

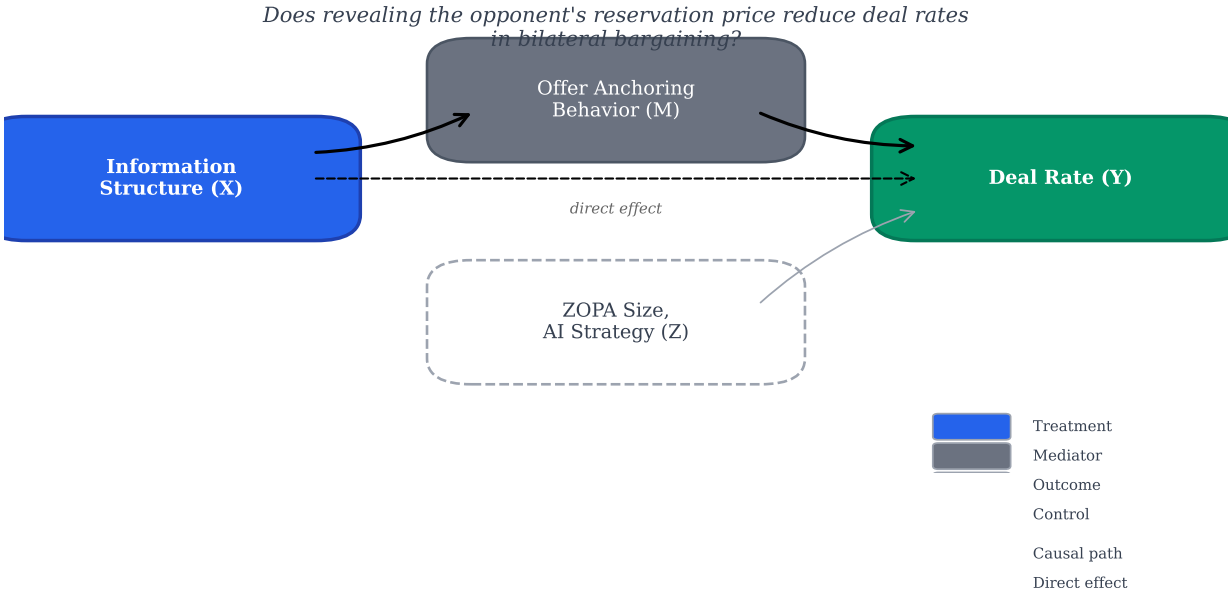

Figure 1: Causal DAG for the information-structure bargaining hypothesis. The treatment (blue) is the human player's information about the opponent's reservation price. The primary causal pathway runs through offer anchoring behavior (gray): informed players anchor offers near the opponent's limit, while uninformed players anchor toward the midpoint. The outcome (green) is whether the pair reaches a deal. A dashed arrow represents a possible direct effect of information on deal rates through channels other than anchoring (e.g., overconfidence, slower concession). ZOPA size and AI strategy (dashed border) are held constant across conditions.

3.1 Variable Definitions

Variable	Type	Operationalization	Game Measurement
X	Treatment	Human sees own valuation only (control) vs. both valuations (treatment)	Game rule: display own vs. both valuations
M	Mediator	Anchoring index: $\frac{\text{first offer} - V_s}{V_b - V_s}$	First offer from transcript; known V_s, V_b
Y	Outcome	Deal reached (binary); rounds to deal; surplus captured	Transcript: deal/no-deal, final price, round count
Z	Control	ZOPA size ($V_b - V_s$), round count, AI strategy	Identical parameters in both game configs

3.2 Testable Implications

1. **Deal rate is lower in the informed condition.** The human who knows V_s fails to reach agreement more often than the human who does not.

2. **First offers are closer to V_s in the informed condition.** The anchoring index is lower (closer to 0) when the human knows the opponent's valuation, indicating exploitative anchoring.

3. **Anchoring mediates the deal-rate effect.** Controlling for the anchoring index should attenuate the treatment effect on deal rate.

4. **Conditional on a deal, the informed human captures more surplus.** The informed player gets a better price *when* they succeed—but succeeds less often. The net effect on expected payoff is the key secondary comparison.

3.3 Identification Strategy

- **Randomized:** Information structure (treatment assignment) is randomized via game variant—each session is either control or treatment.
- **Held constant:** ZOPA size (same valuation draws), round count, AI strategy (identical `player.md`), payoff structure, turn structure. The AI does not know whether the human is informed.
- **Ruled out:** Confounds from AI behavioral differences (AI is identical across conditions), payoff structure differences (identical), ZOPA existence (ZOPA always exists in both conditions).
- **Limitations:** (1) The AI opponent is an LLM, not a human—results speak to human–AI bargaining, which may differ from human–human bargaining. (2) The human knows they face an AI, which may alter their willingness to exploit. (3) We cannot observe beliefs directly—only offers and accept/reject decisions.

4 Next Steps

This hypothesis is ready for `/design-experiment` to map to a concrete game design. The design phase will: select a bargaining game structure, define exact valuation distributions and payoff formulas, specify the AI's fixed strategy, and produce 12-item game specs for each condition (informed vs. uninformed). The key design decisions that remain:

- Valuation distribution and ZOPA size (fixed or drawn from a distribution each session)
- Number of bargaining rounds (tradeoff: more rounds = more data per session but introduces learning effects)
- AI strategy (must be fixed, reasonable, and identical across conditions—e.g., a concession-based strategy with aspiration levels)
- Whether to include a "no ZOPA" condition as a manipulation check