Hypothesis Memo:
Same Information, Different Source — Why Cheap Talk Outperforms Verified Disclosure in Bargaining

Pre-design

February 2026

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

1 Research Question

Does the *source* of partial information about the bargaining zone — game-verified disclosure vs. the opponent's unverified claim (cheap talk) — have opposite effects on deal rates, even when the information content is identical?

In two-player bargaining, a zone of possible agreement (ZOPA) may exist but remain hidden from one or both players. A natural intervention is to provide partial information about the opponent's valuation to help players find mutually acceptable terms. But information can arrive through different channels: it can be *verified* (disclosed by the game rules, known to be true) or *unverified* (claimed by the opponent, who may be lying). Standard economic theory treats these symmetrically — rational agents should update on verified information and discount cheap talk by the sender's incentive to deceive. Behaviorally, however, the source may matter enormously.

We hypothesize a *source-channel reversal*: game-verified disclosure of the opponent's approximate valuation *decreases* deal rates relative to no information, because it activates exploitative anchoring (the informed player pushes toward the opponent's limit). Conversely, the same information conveyed as an unverified opponent claim *increases* deal rates, because the act of voluntary disclosure triggers a reciprocal social process — the opponent "chose to share," creating perceived obligation and cooperative framing. If confirmed, this implies that the behavioral channel (reciprocity vs. exploitation) dominates the informational channel (content accuracy), and that *less credible* information can produce *better* outcomes.

2 Interestingness Argument

Triviality Scorecard

No sharpening required. The hypothesis already features a source-channel reversal (same content, opposite effects by source), two competing mechanisms cleanly separated by design, and a secondary moderator (ZOPA size). The core surprise — that less credible information produces better outcomes — is a strong, testable claim.

Dimension	Score	Reasoning
Prediction surprise	4/5	The dominant intuition is that verified information is more useful than cheap talk. The prediction that cheap talk *outperforms* verified disclosure — same content, opposite effects — is surprising. Experts would split: game theorists expect verified info to dominate; behavioral economists might see the reciprocity channel but would not confidently predict a full reversal.
Literature gap	4/5	Cheap talk in bargaining is well-studied (Farrell & Gibbons 1989; Valley et al. 2002), and information disclosure in auctions/bargaining is a major literature. But the *direct comparison* of same-content information from different sources (verified game rule vs. unverified opponent claim) in bilateral bargaining lacks experimental evidence. The source-credibility literature in persuasion does not extend cleanly to bargaining.
Mechanism specificity	4/5	Two competing mechanisms are identified and the design distinguishes them. Game disclosure activates *exploitative anchoring* (negative channel): the player with verified information knows where to push. Opponent claims activate *reciprocal trust* (positive channel): voluntary sharing creates cooperative obligations. Holding information content constant isolates the source channel.
Boundary conditions	4/5	The hypothesis is a moderated effect, not a main effect: information helps or hurts depending on *source*. Additionally, we predict this reversal is strongest when the ZOPA is narrow (small surplus makes exploitation more damaging and reciprocity more valuable).
Testability in games	5/5	The outcome is binary (deal/no-deal), directly observable in transcripts. The treatment is clean: three between-subjects conditions (no info, game-disclosed, opponent-claimed). First offers and concession patterns are observable for mediation analysis.
Total	**21/25**	Score ≥ 18: proceed as-is.

3 Causal Model

Variable Definitions

Testable Implications

1. **Source-channel reversal:** Deal rate in the opponent-claimed condition > no-info baseline > game-disclosed condition. Same information content, opposite directional effects depending on source.

2. **Anchoring mechanism:** In the game-disclosed condition, first offers are significantly closer to the opponent's reservation price (higher anchoring index) than in the no-info or opponent-claimed conditions.

3. **Reciprocity mechanism:** In the opponent-claimed condition, concession rates are faster (play-

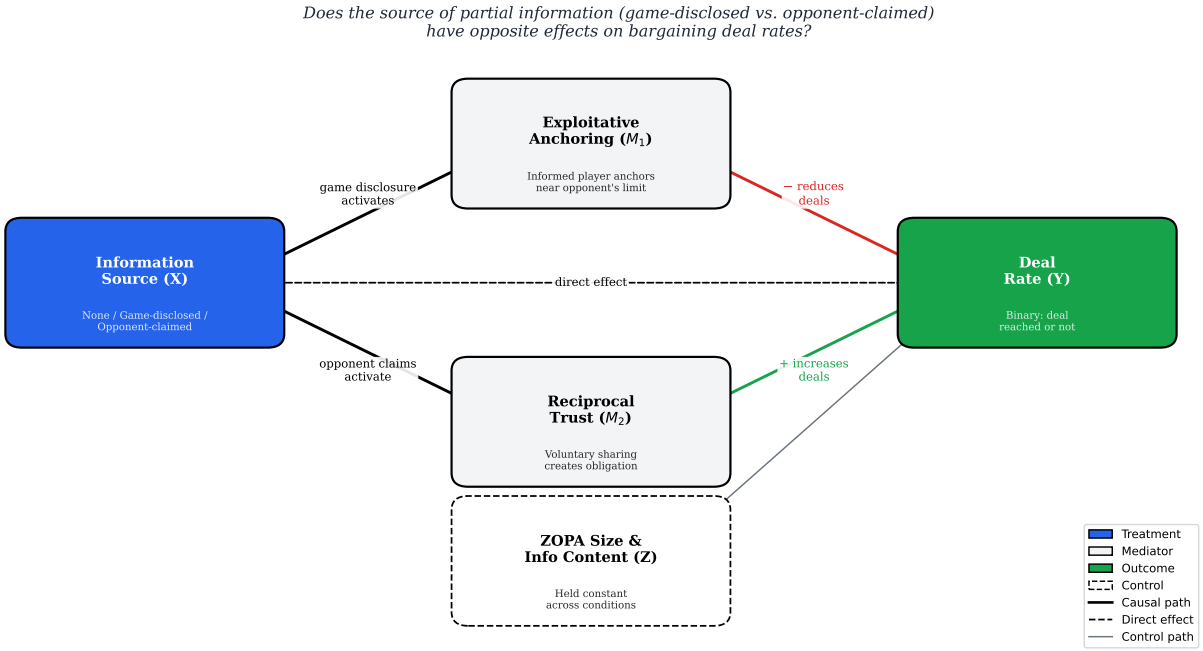

Figure 1: Pearlean causal DAG for the information-source bargaining hypothesis. The treatment (blue) is the source of partial information about the opponent's valuation: no information, game-verified disclosure, or opponent's unverified claim. Two competing mediators (gray) represent the behavioral channels activated by each source. Game disclosure activates exploitative anchoring (M_1, red negative arrow): the informed player anchors offers near the opponent's reservation price. Opponent claims activate reciprocal trust (M_2, green positive arrow): voluntary sharing creates cooperative obligation. The outcome (green) is whether the pair reaches a deal. ZOPA size and information content (dashed border) are held constant across conditions. The key prediction: M_1 dominates under game disclosure (reducing deals) while M_2 dominates under opponent claims (increasing deals), producing opposite effects from identical information content.

ers converge more quickly) than in the no-info condition, consistent with reciprocal trust driving cooperation.

4. **Mediation:** Controlling for the anchoring index (M_1) should attenuate the negative effect of game disclosure on deal rates. Controlling for concession speed (M_2) should attenuate the positive effect of opponent claims.

Identification Strategy

- **Randomized:** Information source (three-level treatment) is randomized via game variant assignment. Each session is one of: no-info, game-disclosed, opponent-claimed.
- **Held constant:** ZOPA size (buyer valuation − seller cost = \$20), number of bargaining rounds, AI strategy (identical `player.md`), payoff structure, turn structure. Crucially, the *information content* is identical across the two informed conditions — only the source differs.
- **Key design constraint:** In the opponent-claimed condition, the AI must be scripted to "voluntarily" share approximately the same valuation information that the game discloses in the

Variable	Type	Operationalization	Game Measurement
X	Treatment	Information source: (1) no info about opponent's valuation, (2) game discloses opponent's approximate valuation (verified), (3) opponent claims their own approximate valuation in a chat phase (unverified)	Game variant: three configs differing only in how/whether info about the opponent's reservation price reaches the player
M_1	Mediator	Exploitative anchoring: first offer's proximity to the opponent's reservation price	Anchoring index: $\frac{\lvert offer_1 - V_{opponent} \rvert}{ZOPA}$; lower = more exploitative
M_2	Mediator	Reciprocal trust: cooperative framing in offers and concession speed	Concession rate across rounds; cooperative language in cheap-talk transcripts (where available)
Y	Outcome	Deal reached: binary indicator of whether players agree on a price within ZOPA before time expires	Transcript: explicit accept/reject of a final price
Z	Control	ZOPA size ($20 in all conditions); information content (same approximate valuation range disclosed regardless of source)	Identical valuation draws and ZOPA across all three conditions

game-disclosed condition. This ensures content equivalence. The AI's claim is approximately truthful (matching the verified disclosure) so that the comparison isolates source, not accuracy.

- **Confounds ruled out:** Random assignment eliminates selection. Identical ZOPA rules out difficulty differences. Identical AI strategy rules out strategic adaptation. Content equivalence across informed conditions isolates the source channel.

- **Limitations:** (1) The opponent-claimed condition requires the AI to make a "voluntary" disclosure, which is scripted rather than truly voluntary — this may not fully capture the social dynamics of real cheap talk. (2) We cannot directly observe beliefs (does the player *trust* the AI's claim?) — only actions. (3) The AI opponent is an LLM, not a human; the reciprocity mechanism may be weaker against a known AI. (4) Three conditions require larger sample sizes than a two-condition design to detect the interaction.

4 Next Steps

This hypothesis is ready for `/design-experiment` to map to a concrete game design. The design phase will need to:

- Define the three game variants (no-info, game-disclosed, opponent-claimed) with identical payoff structures

- Script the AI's "voluntary disclosure" in the opponent-claimed condition to match the game's disclosure in the verified condition (content equivalence)
- Decide whether to include ZOPA-size variation as a secondary treatment (narrow vs. wide ZOPA) to test the predicted moderator
- Set the number of bargaining rounds (balancing data richness against learning effects)
- Choose the approximate valuation signal: e.g., "the opponent's valuation is between $X and $Y" (same range in both informed conditions)