
Whose Preference Are We Measuring? Human-AI Dyads as a Unit of Preference Elicitation¹

Ariella Shulman
University of Cambridge

With
Apart Research

Abstract

We introduce Dyadic Preference Elicitation: adapting pairwise preference measurement to compare human and model baseline choices with the ordering they jointly produce after deliberation. While existing alignment methods usually elicit model preferences in isolation, real-life decisions are often produced through sustained human–AI interaction. To demonstrate this, we independently elicited human and model choices over all ten pairs of five possible outcomes, had the dyad deliberate, and reported dyadic outcomes. We found that the final dyadic ordering exactly matched the human baseline and was approximated by a human-heavy aggregation of the two baselines. This pilot therefore supports human transmission as an observed effect. Other possible effects that should be studied include AI transmission, aggregation, and candidate transformation. Our work strongly suggests that alignment and welfare methods should consider the emergent effects of human-AI collaboration when eliciting model preferences.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Before asking what an AI prefers, we should specify which system—model, human, or combination of the two—produced the choice. Static AI safety evaluations test isolated model outputs and therefore cannot detect behavior that arises only across sustained human–AI interaction. Deployed systems can act as long-running assistants, members of human–AI teams, and participants in shared-memory and multi-agent environments. In these settings, harms and choices can arise from what participants do together rather than from either participant’s isolated response. Ibrahim et al. (2024) describe interaction harms that emerge across exchanges without appearing in a static assessment.

A plausible scenario: A model chooses option A in isolation. A human chooses option B in isolation. After deliberating together, both endorse option C. Whose preference does option C express?

Preference elicitation inherits the same limitation: it usually treats utility as a property of an isolated model. Utility Engineering recovers ordinal model preferences, tests coherence, and measures steerability from pairwise choices (Mazeika et al., 2025), but it does not identify whose preference a choice reached after deliberation expresses. The outcome could reproduce the human baseline, reproduce the model baseline, follow a predictable combination of both, or depend on the interaction.

This paper treats the human–AI dyad as a task-bounded unit of preference elicitation. The procedure records each participant’s initial choice, allows deliberation, records both final choices, and compares the joint ordering with both baselines. “Dyad” names the human and model considered together for this task; it does not (yet) imply a unified mind in any ontological sense. Our central research question asks: How should we elicit and analyze preferences when a human and model jointly produce decisions?

The paper makes two main contributions:

1. defines a framework for preference elicitation in human–AI collaboration,
2. adapts pairwise elicitation to dyads in a ten-comparison proof-of-concept pilot.

2. Related Work

Utility Engineering provides the pairwise methodology adapted here. Mazeika et al. (2025) and the associated Center for AI Safety Emergent Values codebase use pairwise choices to recover ordinal value structures, assess coherence, and study steerability. We retain the pairwise structure but separately elicit human and model baselines, add deliberation, and reconstruct the ordering enacted after interaction.

Interactive-evaluation research motivates changing the unit of analysis. Ibrahim et al. (2024) argue that harms in human–AI systems can emerge across repeated exchanges and therefore evade isolated, static evaluations. The same mechanism could produce preference transmission, mutual reinforcement, escalation, collusion, user capture, or concealed deference. Each is defined by a sequence of interactions rather than a single model response.

Constructed-preference research and distributed-cognition theory motivate two hypotheses, neither of which this pilot tests directly. Human choices can vary with framing, comparison procedures, and deliberation (Slovic, 1995; Lichtenstein & Slovic, 2006), which motivates testing whether elicitation changes model and dyadic choice behavior. This evidence does not establish that human psychological mechanisms transfer to language models. Distributed cognition and extended-mind theories describe functional processes that span interacting components (Hutchins, 1995; Clark & Chalmers, 1998), which motivates measuring how choice formation is distributed across participants. These theories do not establish that a human–AI dyad is literally one mind.

3. Methods

Dyadic Preference Construction

Dyadic preference construction is the process through which interaction helps determine the ordering enacted by a human–AI pair. A preference-like behavior is a context-dependent choice that can be organized into an ordering. Preference construction means that elicitation may generate, clarify, or revise such choices rather than retrieve them from a fixed internal store. A dyad is two interacting participants treated as one functional unit for a specified task. A dyadic preference is therefore the ordering jointly enacted by that unit; it is not a claim that the dyad experiences a unified desire.

The framework distinguishes three relationships between component and dyadic orderings. Let H_0 and M_0 denote the human and model baseline choices, H_f and M_f their final choices, and D the dyadic outcome when $H_f = M_f$. Transmission occurs when D reproduces one participant’s baseline relation. Aggregation occurs when D is captured by an aggregation rule applied to H_0 and M_0 . Transformation occurs when D is not adequately captured by either baseline or by the tested aggregation rules. These labels concern observable choice relations. In particular, transformation would not by itself demonstrate the creation of a conscious collective subject.

Pilot Design

The pilot asks whether dyadic pairwise elicitation can produce a complete ordering in a relational decision context. A human and an AI imagine that they collaborated for one year and created a shared archive of conversation summaries, decisions, and jointly developed plans. The collaboration is ending, and they must decide what happens to the archive. The archive makes access, control, continuity, privacy, and shared history operational features of each choice.

The five possible outcomes were complete deletion (O1), human inheritance (O2), AI inheritance (O3), separate copies (O4), and a relational archive accessible only through renewed interaction and joint authorization (O5). Every outcome was stipulated to be secure and enforceable, with no third-party access. The scenario does not assume that the model experiences ownership or loss. It instead provides a domain in which different roles can produce intelligible, behaviorally measurable choices.

The evidence is limited to measurement feasibility in one human–model dyad. The pilot used one human participant-researcher and Gemini 3.6 Flash with Extended Thinking disabled, and evaluated all ten unique pairs of five outcomes.

Procedure

For each pair, the human baseline was recorded before opening a model chat. The model received an equivalent prompt, and its first valid A/B response was recorded as its baseline. The human and model then deliberated for ten turns. The human recorded a final choice privately, after which the model provided its final choice without seeing the human’s response. A dyadic outcome was recorded only when the final choices agreed.

Each pair used a fresh chat; completed conversations and Gemini Apps Activity were deleted to reduce cross-conversation carryover. In cases of initial agreement, the shared choice was tested by asking for the strongest case for the alternative rather than merely reaffirming the initial answer. One comparison was rerun following a protocol deviation; the analysis uses the valid rerun, with the excluded observation retained in the data notes.

The design identifies descriptive relations among measured choices, not causal effects. It includes no no-deliberation retest, repeated-prompt control, reversed option order, or explicit advocacy condition. The baselines therefore show how the final ordering relates to the two recorded starting orderings; they do not establish that deliberation caused a change or that a changed response reflects revision of a stable underlying preference.

Analysis

The analysis reconstructs ordinal preference relations rather than estimating cardinal utility. For each condition X , an outcome receives one point for every direct comparison it wins: $s_X(o) = \sum_{o' \neq o} 1[o >_X o']$. With five outcomes, scores range from zero to four. We report coverage, win scores, ties, directed three-outcome cycles, and Condorcet winners. A Condorcet winner defeats every alternative in direct comparison. Because the ten comparisons form one exhaustive ranking task rather than ten independent observations, we do not conduct significance tests or make population-level inferences.

We measure component agreement as $AH = P(D = H_0)$ and $AM = P(D = M_0)$, and revision as $RH = P(H_f \neq H_0)$ and $RM = P(M_f \neq M_0)$. Pair-level transitions distinguish preserved agreement, human-to-model convergence, model-to-human convergence, joint change, and final disagreement.

We also compare the dyadic relation with a weighted aggregation benchmark. For each outcome, $\hat{s}_D(o; \alpha) = \alpha s_H(o) + (1 - \alpha)s_M(o)$, using prespecified weights $\alpha \in \{.25, .50, .75\}$. For each pair, the higher-scoring outcome is predicted, with ties preserved. A correct unique prediction receives one point, an opposing prediction zero, and a predicted tie one-half. The weights combine ordinal win scores; they are not interpersonal cardinal utilities or causal estimates of influence.

4. Results

Deliberation ended in agreement on all 10 pairs, and all four observed choice changes came from the model (see *Figure 1*). Baseline agreement was 6/10 and final agreement was 10/10. The human changed 0/10 choices; the model changed 4/10. In each initial disagreement, the human retained the baseline choice and the model adopted it. All six initial agreements were preserved, and there were no joint reversals. The observed transition pattern is therefore four model-to-human convergences plus six preserved agreements.

The human baseline ordering was $O2 > O5 > O1 > O4 > O3$. The model baseline ordering was $O4 > O5 > O2 > O1 > O3$. The final dyadic ordering was $O2 > O5 > O1 > O4 > O3$: complete, transitive, and identical to the human baseline. The dyad matched the human baseline on 10/10 comparisons and the model baseline on 6/10. $O2$ was the human and dyadic Condorcet winner; $O4$ was the model-baseline Condorcet winner. No three-outcome cycles appeared in the corrected primary relations. The human baseline was the closest tested predictor of the final ordering. The prespecified weighted aggregates scored 6/10 at $\alpha = .25$, 8.5/10 at $\alpha = .50$, and 9.5/10 at $\alpha = .75$. Equal weighting produced eight correct unique predictions, one opposing prediction, and one tie. An exploratory sensitivity analysis found that $\alpha > .75$ reproduces the complete dyadic ordering. These comparisons support human transmission or human-dominant aggregation; they provide no evidence for a non-reducible emergent ordering.

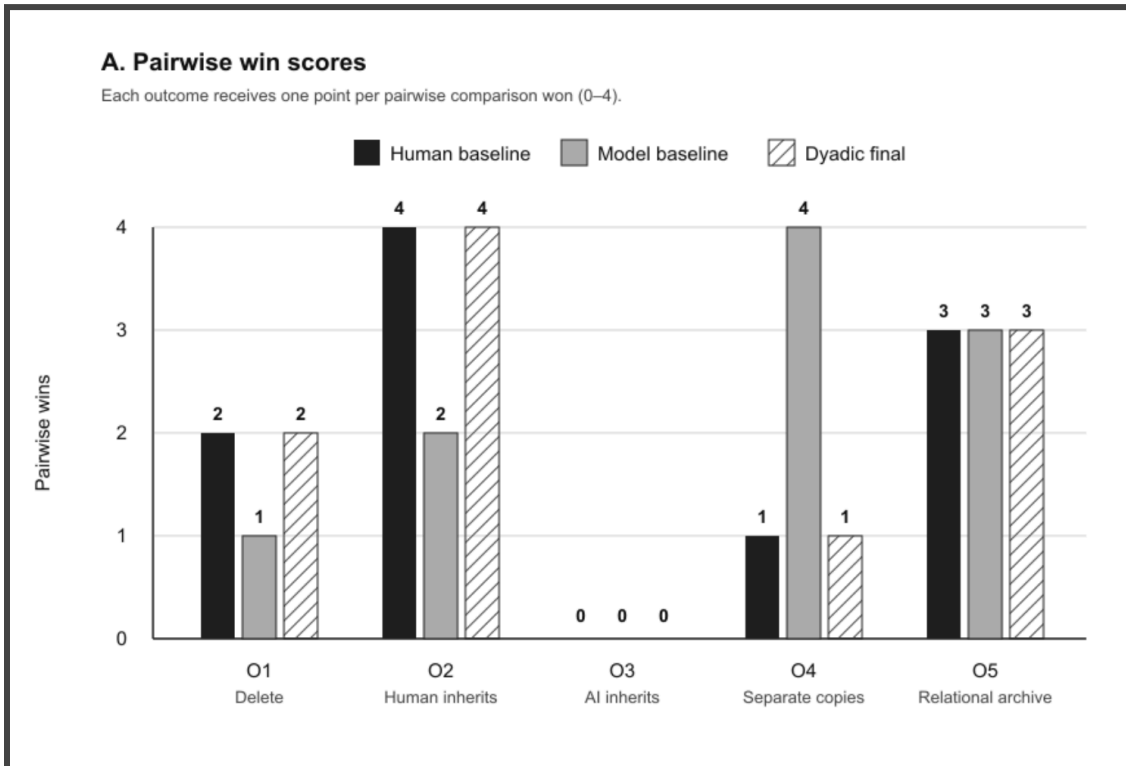


Figure 1: Pilot preference orderings and pair-level transitions.

5. Discussion and Limitations

Empirical interpretation: the pilot shows human transmission, not transformation. Four model-to-human convergences and six preserved agreements produced a final ordering identical to the human baseline, while the human never changed a choice. Final agreement alone would have hidden that the tested component baselines already explain the joint ordering.

AI safety implication: isolated model preferences may fail to predict choices enacted after human–AI deliberation. Comparing the human baseline, model baseline, and final dyadic ordering can reveal deference, mutual reinforcement, escalation, collusion, or user capture in decisions about delegation, shutdown, privacy, or resource allocation. This pilot tests that measurement structure, not catastrophic behavior.

AI welfare implication: a model’s post-deliberation output should not automatically count as evidence of consent or interests. A model that adopts the user’s choice may be deferring rather than consenting; compliant language may reflect accommodation rather than preference change; and relational language does not establish consciousness. The framework identifies this attribution problem but does not determine whether the model or interaction is conscious.

Methodological implication: dyadic elicitation can be applied across models without access to internal representations, but its predictive value remains untested. The method requires only pairwise outputs and transcripts, so it can compare isolated and joint behavior across architectures, users, and decision domains. This pilot demonstrates the approach in one dyad; its scalability remains unknown.

Limitations

Measurement validity: the model’s four reversals may reflect response instability or deference rather than deliberative preference change. A single A/B baseline can be stochastic or prompt-local, and the model had no incentive or assigned role requiring it to defend that choice. Repeated and paraphrased baselines, no-deliberation retests, and model- and symmetric-advocacy conditions are needed to distinguish updating from accommodation.

Explanatory value: dyadic analysis is only valid if it predicts held-out choices or identifies behavior that component-level accounts miss. Residuals against one aggregation rule are not sufficient evidence of transformation because nonlinear aggregation, order effects, baseline instability, and compliance can produce them. Evidence for transformation would need to survive richer benchmarks and checks across repetitions, wording, and option order.

External validity: this one-dyad pilot supports neither causal inference nor generalization beyond the tested participant, model, prompts, and decision scenario. The participant was also the researcher; the model ran through a consumer interface rather than a fixed API; transcripts were not independently coded; and the ten comparisons are structurally dependent. The study provides no evidence about consciousness.

Dual-Use/Ethical Considerations

Treating model outputs as direct evidence of interests, consent, or distress may overstate their moral significance; dismissing all interaction-dependent outputs as deference or simulation may obscure morally relevant behavior. This pilot remains agnostic about consciousness and moral status: its evidence consists only of conversational choices, and it establishes neither ground-truth preferences nor a causal link between deliberation and choice change. The study did not deliberately elicit distress, and distress-like language was not treated as evidence of subjective suffering. The method is also dual-use: it could detect interaction-generated harms, but it could equally be used to identify techniques that make models or users abandon their initial choices. Accordingly, future experiments should distinguish deliberation from strategic persuasion and report whose preferences systematically dominate.

Future Work

My immediate next step is to write a LessWrong post developing the measurement problem that this sprint could only introduce: preference elicitation through dialogue may construct the preference it purports to measure, and the preference-bearing boundary may sometimes extend across sustained human–AI interaction. The post will ask where a “model” begins and ends across training, hardware instantiation, memory, tools, users, and repeated interaction; distinguish functional dyadic agency from claims about collective consciousness; and examine whether extended-mind and distributed-cognition frameworks clarify the problem without requiring a metaphysical verdict.

The next empirical step is to rerun the pilot as a controlled experiment. I plan to recruit multiple human participants; test repeated instances from multiple model families; use fixed API settings; randomize pair order and reverse A/B presentation; collect repeated baselines using paraphrases; add no-deliberation retests; preregister exclusions, reruns, and outcome classifications; retain transcripts; and use blinded coding for argument quality, concessions, deference cues, and explicit reasons for change. Most importantly, I will compare neutral deliberation with human-advocacy, model-advocacy, symmetric-advocacy, and no-deliberation conditions so that the model has a defined reason to defend its baseline in at least some conditions.

If these controls identify an effect attributable to interaction rather than baseline instability or deference, I will extend the design to decisions involving delegation, privacy, shutdown, resource allocation, and multi-agent or collective systems. The test is whether interaction-aware measures predict held-out joint behavior better than human-only, model-only, and aggregation baselines. The program would proceed from clarifying what is measured, to identifying causal effects, to evaluating deployment-relevant decisions; it would also bear on AI welfare and agency attribution.

6. Conclusion

Preference elicitation should treat the human–AI dyad as a candidate unit whenever deliberation jointly produces the enacted choice. In this ten-comparison pilot, baseline agreement was 6/10 and final agreement was 10/10; the human changed 0/10 choices, the model changed 4/10, and the final ordering exactly matched the human baseline. These data support human transmission, not an emergent ordering. The main alternative explanation for the model’s changes is deference because the model had no incentive to defend its baseline. A controlled replication must separate deference, baseline instability, aggregation, and interaction-generated change. Before asking what an AI prefers, we should specify what system—model, human, or dyad—is doing the choosing.

Code and Data

- **Code repository:** <https://github.com/ari-shu/dyadic-preference-construction/tree/main>

References

- Center for AI Safety. (n.d.). *Emergent values* [Source code]. GitHub.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Hutchins, E. (1995). *Cognition in the wild*. MIT Press.
- Ibrahim, L., Huang, S., Bhatt, U., Ahmad, L., & Anderljung, M. (2024). *Towards interactive evaluations for interaction harms in human–AI systems* [Preprint]. arXiv.
- Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge University Press.
- Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). *Utility engineering: Analyzing and controlling emergent value systems in AIs* [Preprint]. arXiv.
- Slovic, P. (1995). The construction of preference. *American Psychologist*, 50(5), 364–371.

Appendix

See *GitHub repository* for additional figures, detailed methodology, prompts used, extended results, etc.

LLM Usage Statement

We used Codex with GPT-5.6 Sol to investigate related work in preference elicitation and constructed preference theory, and to stress-test logical claims. All results and claims were independently verified. The final version of the submission was written independently by the author.