Belief Revision in Human-LLM Communication: Effects of Source Identity, Reliability, and Initial Confidence

Yuting Yuan (yuting.yuan@ifikk.uio.no)[1], Nicholas Allott (n.e.allott@ilos.uio.no)[2], Mikhail Kissine (mikhail.kissine@ifikk.uio.no)[1,3,4]

[1]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, [2]Department of Literature, Area Studies and European Languages, University of Oslo, [3]ACTE, LaDisco and ULB Neuroscience Institute, Université libre de Bruxelles, [4]Department of Linguistics, University College London

Introduction. Large language models (LLMs) are becoming increasingly embedded in everyday communicative environments [7], where users often treat their outputs as meaningful inputs [6]. This makes it important to examine how LLM outputs are evaluated as advice or factual answers. On inferentialist accounts, understanding an utterance involves recovering a speaker's communicative intention [1, 2, 4]. This creates a tension: people appear to engage in conversational exchanges with LLMs, although such systems produce fluent verbal output without communicative intentions. From the perspective of epistemic vigilance, communicated information is evaluated in terms of content and source properties, especially competence and honesty [5]. Since initial representations can resist revision [3], the present study examines whether these pragmatic and epistemic mechanisms extend to conflicting information attributed to human and non-human sources.

RQ1: How do source identity (generative AI vs human) and perceived reliability influence the belief revision? **H1:** Participants will be more likely to revise their initial decisions toward the source described as more accurate, with this effect varying by source identity. If participants use reliability normatively, human and AI sources should have similar effects once reliability is controlled; if AI labels trigger algorithm aversion or automation bias, source identity should still matter.

RQ2: Does initial confidence modulate susceptibility to source-based updating? **H2:** Lower initial confidence will be associated with a greater likelihood of belief revision.

Methods. Adult native English speakers are recruited via Prolific. Participants read short general-knowledge statements, make a binary true/false decision, and report confidence on a 0-100 slider. Expt.1 (baseline, N=100) has been completed and was used to select 28 trials for Expt.2, with no clear floor or ceiling effect (Figure 1). Expt.2 (N = 180) uses a mixed design: Time is within-subject, and perceived source reliability is between-subject. Phase 1 repeats the baseline task; Phase 2 presents the same statements with conflicting generative AI and human responses. Participants choose the source they agree with and report confidence (Figure 2). Reliability is manipulated across three conditions: AI more accurate, human more accurate, or no accuracy information.

Belief revision is operationalised as divergence between Phase 2 source selections and Phase 1 decisions, with confidence change as a complementary index.

Approach for statistical analysis. Data will be analysed using Bayesian hierarchical models, with participant and item random intercepts and weakly informative priors. For Expt.1, descriptive item-level analyses were used to estimate accuracy and confidence for stimulus selection; Flesch-Kincaid Grade Level (FKGL) showed no reliable association with accuracy. For Expt.2, logistic mixed effects models will test source choice, and belief revision will be analysed as a function of reliability condition, source identity, baseline correctness, and initial confidence.

Figures

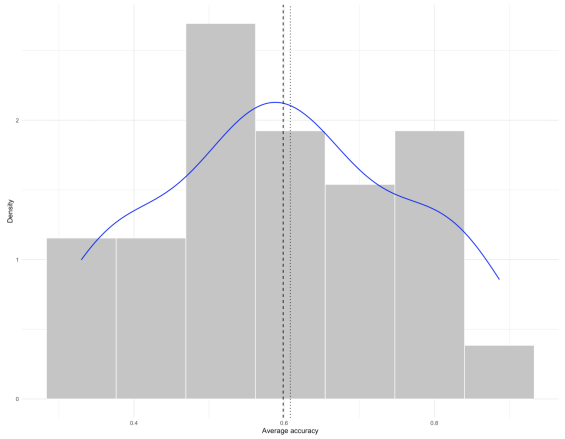

Figure 1: Participant-level accuracy on the selected 28 trials for Experiment 2.

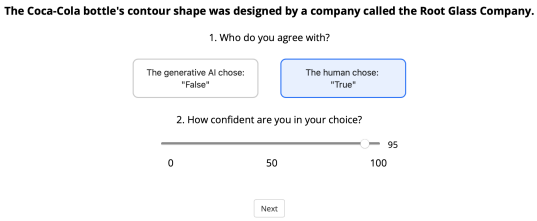

Figure 2: Example of a trial in Phase 2 of Experiment 2

References

[1] Nicholas Allott. Encapsulation, inference and utterance interpretation. *Inquiry*, pages 1–35, 2023.

[2] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[3] Myrto Pantazi, Mikhail Kissine, and Olivier Klein. The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social cognition*, 36(2):167–198, 2018.

[4] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

[5] Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

[6] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777, 2022.

[7] Yidan Yin, Nan Jia, and Cheryl J Wakslak. Ai can help people feel heard, but an ai label diminishes this impact. *Proceedings of the National Academy of Sciences*, 121(14):e2319112121, 2024.