Pragmatics of human-AI communication

Introduction. Within linguistics and the philosophy of language, discourse is formalized in terms of mental content – beliefs, goals and motivations of conversation participants – with the joint goal of mutual understanding (Grice 1975 et seq.). The development of large language models (LLMs) introduces new kinds of communicative settings where the standard mentalistic approach to discourse may not be appropriate. Because LLMs don't have a human mind, and likely don't have the same kind of motivation structure as humans do, people might employ distinct strategies when talking to machines. In this work, we explore whether such an alternative strategy for communication is actually employed for machine-generated linguistic content. Focusing on a distinction between asserted and presupposed content in human communication and the different use conditions governing each type, we ask: (1) whether humans uptake information differently when generated by an AI which they are told is unreliable, and (2) whether information processing is affected by whether the content is packaged as asserted vs. presupposed.

Background. We take as our starting point a model of discourse based on proposals by Stalnaker (1974, 1978). On this model, sentences used in communication contribute to the conversational **common ground**, the set of shared beliefs among discourse participants. The model distinguishes two kinds of linguistic content in the way they affect the common ground. **Asserted** content is put forth with the explicit intent to change the listener's beliefs and expand the common ground. In contrast, **presupposed** content must already be part of the common ground, or be accommodated, before that common ground can be updated with the assertion. Crucially, novel information packaged in these two forms have distinct effects on belief change: asserted content is presented to the listener as up for debate, giving them the option of accepting or rejecting. Novel presuppositions, on the other hand, are things the speaker expects a cooperative listener to tacitly add to their own beliefs, and in turn to the common ground. Listeners infer based on the utterance what the speaker wishes to take for granted, and trusting them not to mislead, shifts to the intended common ground.

Hypotheses. The common ground model takes exchange of information as grounded in the beliefs and intentions of interlocutors, and it is possible that the way humans intake information from machines, which lack such a mental apparatus, is different. The model, furthermore, distinguishes the type of belief revisions a listener is expected to do on the basis of whether a piece of new information is asserted vs. presupposed. Presuppositions can lead listeners to adjust their beliefs without much deliberation or discussion. This type of tactic belief change – which relies on reasoning about what the speaker wants to be common ground – may not happen when communicating with an AI. In that case, new information should be treated as new and up for debate, irrespective of how it is packaged. Another possibility is that presupposition accommodation is automatic, and people are prone to accept and go along with AI presuppositions, even when they may challenge AI assertions.

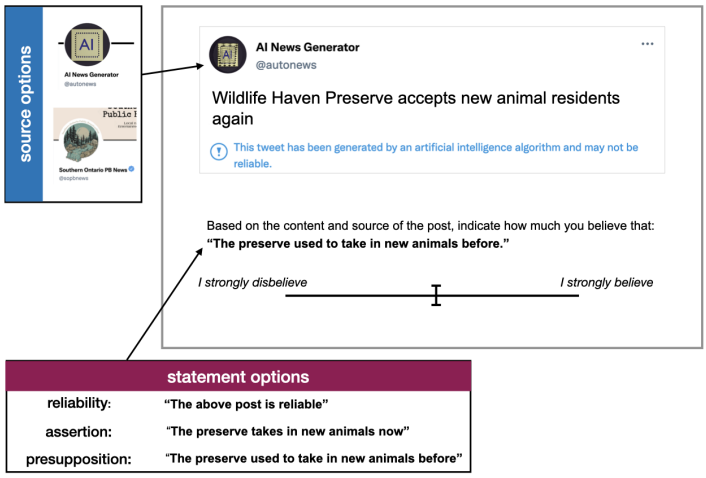

Figure 1: structure of the experimental task

Experiment. Participants (N=205) were asked to read constructed social media posts, and a potentially related follow-up statement, after which they evaluated the extent to which they believe that statement on a slider from “strongly disbelieve” to “strongly believe” (Figure 1). Crucially, each post contained a presupposition trigger (e.g,. *again*). In a 2x4 within-subjects design, we manipulated two factors: (1) **source of information**: human (a news outlet, Southern Ontario Public Broadcasting) vs. AI (AI algorithm tasked with constructing news-like posts; AI-posts indicated that the post’s content may not be reliable); (2) **follow-up statement type:** all participants saw 4 types of follow-up statements: (i) explicit statements about the reliability of the post (*Reliability* condition), (ii) asserted content from the post (*Assertion* condition), presupposed content from the post (*Presupposition* condition), and (iv) unrelated content from the post (*Unrelated* condition). Unrelated trials were used for exclusion and do not figure in analyses.

Results. See Figure 2. There are three findings of note. We found a main effect of source (β= -25.38, p<.001): participants indicated lower belief in AI content overall compared to human-generated content, perhaps unsurprisingly given that they were told that the AI content was unreliable. This finding shows that humans can modulate their trust in information based on source, at least when reliability issues are highlighted. Second, and strikingly, we found a significant difference (β= -4.05, p<.001) between participants' ratings of AI reliability and their endorsement of AI content (assertions and presuppositions): participants endorsed AI-generated content significantly more than they endorsed its reliability. In other words, perceived low reliability of AI did not fully prevent participants from updating their beliefs with the content it produced. Finally, we found a small but significant difference between AI-assertions and AI-presuppositions, with participants indicating greater belief in presupposed content (β= 1.62, p=.01). This suggests that people are ready to accommodate, rather than challenge, AI-presuppositions, despite the conversational setting not obviously licensing such behavior.

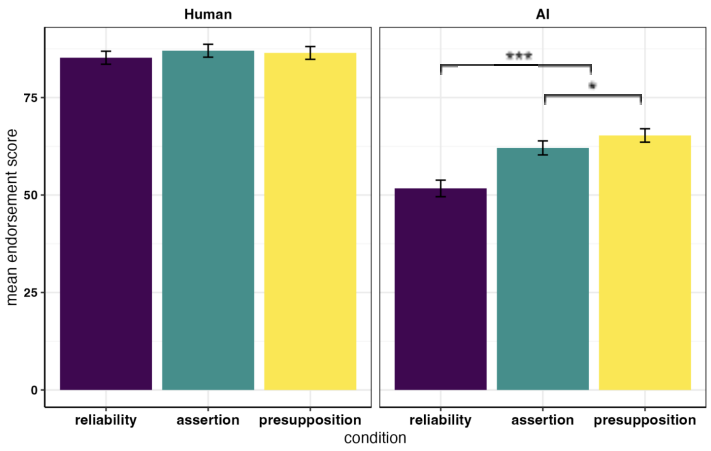

Figure 2: Mean endorsement scores, indicating level of belief in each follow-up; error bars represent 95% CI

Conclusions. AI-generated language presents new questions, both theoretical and practical, about how our beliefs evolve over the course of a conversation. In this study we found that, despite the fact that machines might lack human-like mental states, people treated AI-generated language as constrained by the same principles as those found in human language. On a theoretical front, this finding implies that humans tend to perceive any natural language as human-like. On a practical front, it raises questions about how humans can be aided to encode AI language appropriately, rather than imbuing it with human motivations.

References. Grice, H. P. (1975). Logic and Conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech Acts* (pp. 41-58). New York: Academic Press. Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz and P. Unger (eds.), *Semantics and Philosophy* (pp. 197-214). New York: NYU Press. Stalnaker, R. (1978). Assertion. Syntax and Semantics 9: pp. 315-332.