

Not as rational as RSA predicts: Failure to reason about alternative messages

Alexandra Mayn, John Duff, Natalia Bila and Vera Demberg

Department of Language Science and Technology, Saarland University

Introduction

The Rational Speech Act framework (RSA; Frank & Goodman, 2012) formalizes cooperative communication and has been used for modeling various pragmatic phenomena. While the model most commonly used in the RSA framework is L_2 (listener who reasons about a *pragmatic* speaker), a simpler listener model L_1 (listener who reasons about a *literal* speaker) has also been used. Franke (2011) showed that the L_1 model is sufficient for computing quantity implicatures, and Franke & Degen (2016) found that most of their participants were best described by the L_1 model, rather than L_2 . L_1 models highlight that rational agents can make inferences even when the speaker is literal and equally likely to produce any true message, by reasoning over the number of alternative messages available to the speaker.

Recent work by Mayn, Loy and Demberg (2024) suggests that while people readily reason about the rationality of the speaker, they may not reason about alternative messages. When the speaker is believed to be a 4-year-old child, who is presumably not a sophisticated reasoner, there is a very high proportion of literal interpretations, contrary to the prediction that people should favor a pragmatic interpretation even when the speaker is literal.

In this study, we investigate whether people make pragmatic inferences when reasoning about a speaker who is explicitly presented as literal. We find that participants overwhelmingly fail to consider alternative messages: only 5 out of 79 participants make pragmatic inferences based on alternative messages as the L_1 model would predict, which suggests that L_1 is not a plausible model of pragmatic reasoning.

Experiment

Participants


94 native English speakers recruited on Prolific.



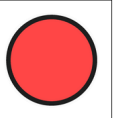
Materials

Participants were told that they would play a communication game with a simple computer program called *basic_message_picker*, which randomly selected any true message to refer to an object. Participants briefly practiced selecting messages as if they were the computer program and received feedback to ensure they understood the program's expected behavior.





On each trial, participants saw three objects and a message that *basic_message_picker* purportedly selected to refer to one of them. Participants then indicated how likely they believed each object to be the intended referent by distributing 100 points between the objects.

On critical trials, as shown in Figure 1, the message is true of two objects. However, for one of the objects, the target, one of its features is inexpressible (in our example, there is no message for blue), whereas the other object could also be referred to with another message (green paint). RSA predicts that the listener will assign a probability of $2/3$ to the blue triangle because for the green triangle, the speaker's probability mass is split equally between the messages "triangle" and "green", whereas for the blue triangle, the whole probability mass is on "triangle".


basic_message_picker selected the following message: 


These were the messages basic_message_picker could select from:


How likely do you think it is that basic_message_picker selected the message above to refer to each of the following objects?
Use the sliders below to respond. Remember that slider values need to sum up to 100.



0



0



0

Sum: 0

Figure 1. Example trial in the critical condition.

Procedure

The experiment consisted of three phases: Block 1, training, and Block 2. Participants were randomly assigned to one of two conditions, No training or Training.

Block 1 was the same for all participants. They completed 8 critical trials, 16 unambiguous trials, and 4 ambiguous filler trials.

At the end of the first block, participants saw one critical trial again and were prompted to briefly explain their response in a textbox. Participants' explanations were annotated based on Mayn et al. (2024). Explanations indicating belief that the two fitting objects were equally probable were labeled *guess*, while explanations reflecting reasoning

about alternative messages predicted by RSA were labeled *correct_reasoning*. Explanations revealing expectations of rational behavior from the program were labeled *ascribe_rationality*. Unclear responses were labeled *unclear*.

After Block 1, participants were assigned to one of the two conditions. In the No training condition, participants completed an unrelated task (a 10-question version of Raven's matrices) instead of training, and Block 2 was identical to Block 1. In the Training condition, participants received training aimed at making them aware of alternative messages which could be used to refer to each object. During training and in Block 2, before interpreting the message using sliders, participants had to indicate, for each object, which messages *basic_message_picker* could have sent to refer to it.

Results

15 participants were excluded for accuracy below 80% on unambiguous fillers, misunderstanding instructions, or changing their mind upon reflection. The remaining 79 participants (39 in the Training condition, 40 in No training) were analyzed.

Participants' responses in each block were categorized into three classes based on likelihood of coming from normal distributions centered around 50 (literal interpretation), 66 (pragmatic interpretation) and 100 (ascribing rationality to the speaker) with $sd=2$. Participants whose likelihood for all classes was below a threshold set based on a pilot study (10^{-30}) were classified as "other".

Figure 2 shows participants' mean target ratings in each block with their assigned class and annotation of their reported strategy. In both conditions, most participants (65% or more) were assigned to the "50", indicating literal responding. Only two people in the No training condition and three in the Training condition were in the "66" class, suggesting pragmatic responding predicted by RSA. Six participants, three in each condition, expected *basic_message_picker* to behave rationally, contrary to instructions. Training did not result in an increase in pragmatic responding.

Participants' explanations were generally consistent with class assignment based on ratings but revealed that sometimes participants gave the "correct" rating for the wrong reason: Of the five participants assigned to the correct class, only one provided an explanation indicating reasoning about alternative messages.

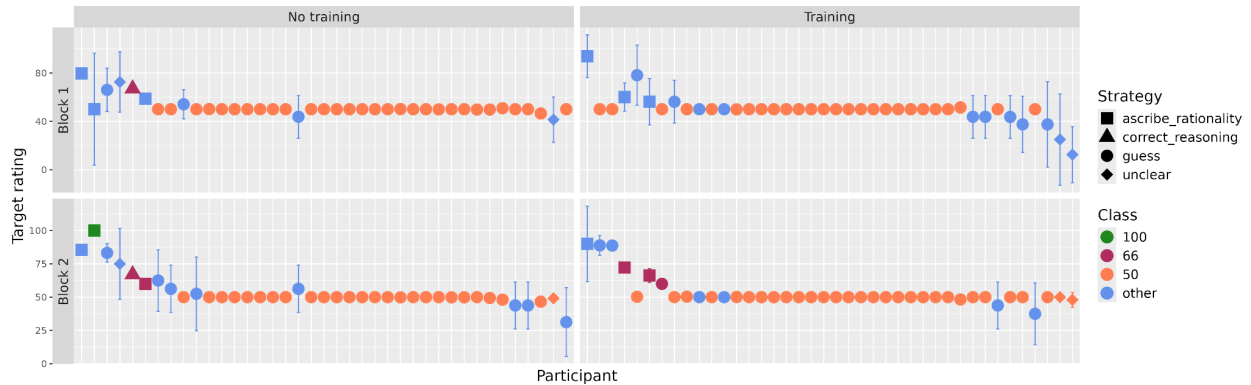


Figure 2. Mean target rating in the critical condition by participant across blocks in the two conditions.

Discussion

We find that participants overwhelmingly fail to consider alternative messages and derive a pragmatic interpretation of a message by a literal speaker. Furthermore, making people aware of alternative messages does not improve performance. This finding is striking but consistent with literature on errors in Bayesian reasoning in other types of problems (Fox & Levav, 2004; Starns et al. 2019). This also suggests that the L_1 listener model, where a pragmatic listener exploits the number of alternative messages to derive inferences from a literal speaker, may not be a plausible model of pragmatic capacity.

It is also possible that people are reasoning about alternative messages but misestimating their probabilities. To test this hypothesis, we plan to conduct a follow-up where participants will be trained to correctly divide the probability space of available messages using a visual aid.

References

- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998.
- Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4, 1-1.
- Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PLoS ONE*, 11(5), e0154854. doi:10.1371/journal.pone.0154854

Fox, C. R., & Levav, J. (2004). Partition-Edit-Count: Naive extentional reasoning in judgment of conditional probability. *Journal of Experimental Psychology: General*, 133(4), 626-642. doi:10.1037/0096-3445.133.4.626

Mayn, A., Loy, J. E. & Demberg, V. (2024). Beliefs about the speaker's reasoning ability influence pragmatic interpretation: Children and adults as speakers. *PsyArXiv preprint*. doi:10.31234/osf.io/n3v69

Starns, J. J., Cohen, A. L., Bosco, C., & Hirst, J. (2019). A visualization technique for Bayesian reasoning. *Applied Cognitive Psychology*, 33, 234-251. doi:10.1002/acp.3470