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

Introduction

The Rational Speech Act framework (RSA; Frank & Goodman, 2012) formalizes cooperative communication and has been used for modeling various pragmatic phenomena. RSA has successfully predicted people's response patterns in reference games, where the listener recovers the speaker's intended meaning based on an ambiguous message. RSA predicts that even when the speaker is literal and is equally likely to produce any true message, the listener will still derive inferences based on alternative messages available to the speaker. Indeed, Franke & Degen (2016) showed that most of their participants were best fit by an RSA model of a listener reasoning about a literal speaker.

However, recent work by Mayn, Loy and Demberg (2024) suggests that people may not reason about alternative messages as RSA predicts. They show that when the speaker is said to be a 4-year-old child, who is presumably believed not to be a sophisticated reasoner, there is a very high proportion of literal interpretations, contrary to RSA's prediction that people should favor a pragmatic interpretation, even when the speaker is a literal speaker. Mayn et al. (2024) hypothesize that while people readily reason about the speaker's intentions or reasoning sophistication, they may often fail to reason about alternative messages.

While Mayn et al. (2024)'s observations are compelling, they may be confounded by people's differing beliefs about children's rationality. In this study, we investigate whether people are able to reason about alternative messages when the speaker is explicitly presented as literal. We find that participants overwhelmingly fail to consider alternative messages: only 2 out of 76 participants make inferences consistent with RSA's predictions.

Experiment

Participants

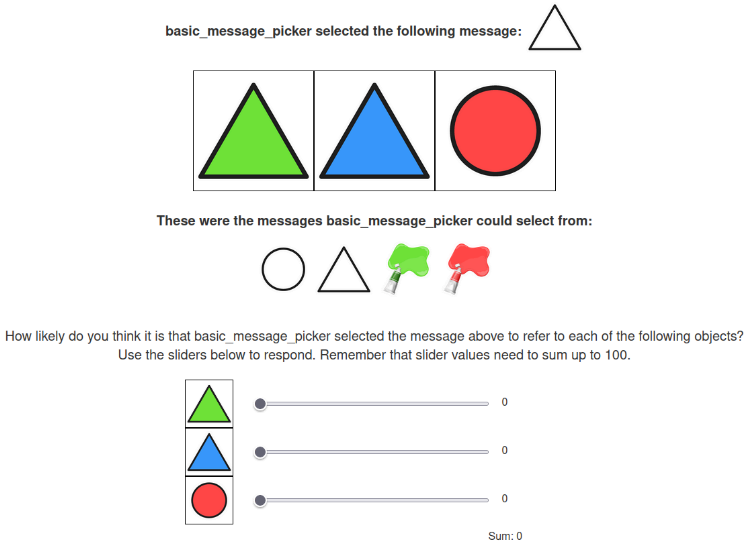
90 native English speakers recruited on Prolific.

Materials and procedure

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 responded how likely they believed each of the objects to be the intended referent by distributing 100 points between the objects.

Participants completed 8 critical trials, 16 unambiguous and 4 ambiguous filler trials. 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.

Figure 1. Example trial in the critical condition.

At the end, participants saw one critical trial again and were then 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.*

Results

14 participants were excluded for accuracy below 80% on unambiguous fillers, misunderstanding instructions or changing their mind upon reflection. The remaining 76 participants were analyzed.

Participants were categorized into three classes based on likelihood of their responses coming from normal distributions centered around 50 (corresponding to 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 with their assigned class and annotation of their explanation. Most participants (63 out of 76) were assigned to the “50” class because they gave a rating of exactly or near 50% to the target every time, consistent with failing to reason about alternative messages. Only two people were assigned to the “66” class, consistent with pragmatic responding predicted by RSA. Three participants expected the computer program to behave rationally, contrary to instructions. Participants’ explanations support this classification: both participants in the “66” class provided explanations involving correct estimations of probability of alternative messages, whereas all but one participant in the “50” class gave explanations consistent with a literal interpretation.

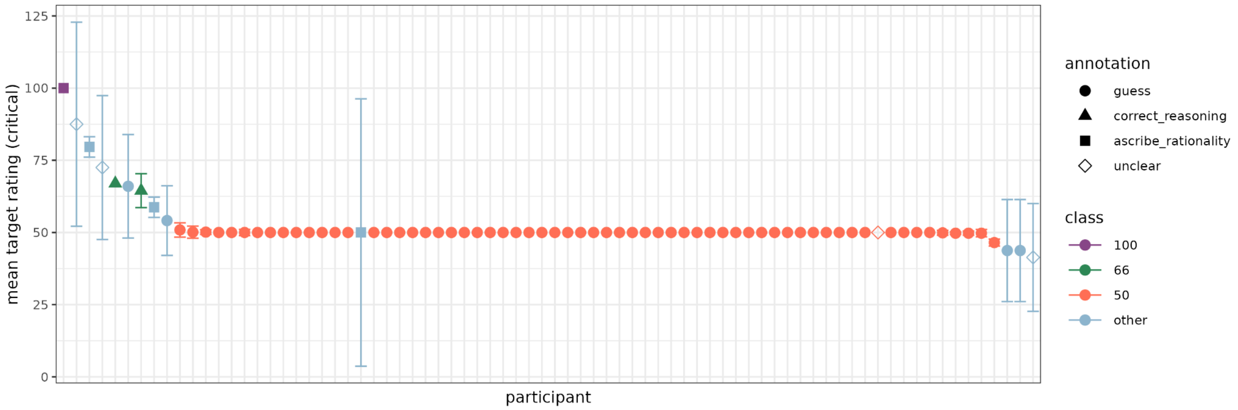


Figure 2. Mean target rating in the critical condition by participant.

Discussion

We find that, contrary to RSA’s predictions, participants overwhelmingly fail to reason about alternative messages when interpreting a message by a literal speaker. 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).

We predicted that prompting participants to consider alternative messages might lead to improved performance. In an additional experiment, participants first played the reference game as described above and then played the same reference game but additionally indicated all possible messages for each referent. Participants' performance did not improve, however, and sometimes even dropped on unambiguous trials, suggesting that the implementation may have been confusing or overwhelming. We plan to conduct a follow-up experiment with a simpler design.

We also consider another explanation: participants may be reasoning about alternative messages but making systematic errors in estimating their probabilities. Fox & Levav (2004) argue that classic fallacies of probabilistic reasoning arise from incorrectly partitioning the probability space. They show that performance improves when participants are nudged to consider all relevant events. To test whether this hypothesis could explain our results, we plan to conduct another follow-up experiment including a training phase, which will teach participants to correctly divide the probability space of available messages using a partitioned circle.

References

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998.

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