The lady doth protest too much! Discourse exepctations affect judgments of argument strength

S.R. Pulapura
University of Edinburgh
s.pulapura@sms.ed.ac.uk

Hannah Rohde
University of Edinburgh
hannah.rohde@ed.ac.uk

Catherine Lai
University of Edinburgh
c.lai@ed.ac.uk

Abstract

Bayesian models of argument strength predict that a rational agent's belief in a claim should increase monotonically when more reasons are provided in favor of it. This study investigates a competing prediction: in natural language dialogues, arguments may feel weaker when *too many* reasons are provided. Extensive prior experimental evidence shows that listeners expect speakers to produce informative utterances, and will make inferences to rationalize perceived redundancies. We propose that pragmatic speakers prefer to produce "informative" reasons that significantly raise a literal listener's credence from a disbelieving prior. If this is so, a pragmatic listener may interpret seemingly superfluous reasons by inferring that the speaker expected him to be skeptical. Using the Rational Speech Act framework, we develop a pragmatic model of argument strength that predicts the latter inference, and present a behavioral experiment which partially confirms its predictions.

1 Introduction

Trustworthiness and explainability are intuitively related notions. The "explanation" relation holds between a primary discourse segment, and a secondary segment from which it is *inferrable* (Kehler, 2002). In argumentative discourses, speakers use explanations to negotiate the acceptance of controversial propositions: the role of a "reason" in these cases is to reduce or eliminate the need for trust among interlocutors by making the target content easier to infer from the listener's accepted beliefs (Mercier, 2020). It seems evident that listeners are more likely to accept– and perhaps, are more epistemically justified in accepting (Sperber et al., 2010)– beliefs that are more thoroughly explained (by a human or an LLM). This analysis is predicted by a number of psychological models of argument strength (Oaksford and Hahn, 2004) and forms the conceptual core of "chain-of-thought" reasoning (Wei et al., 2022).

We present a behavioral study and a probabilistic pragmatic model that may challenge the latter assumption: explanations may actually be judged as less trustworthy when they are redundant ("the lady doth protest too much!"). We suggest that this effect is predicted by rational theories of language use (Frank and Goodman, 2012; Grice, 1989). In order to maximize information exchange under cognitive constraints, pragmatic speakers tend to avoid mentioning predictable information. Listeners expect this production preference, and can adjust their model of the context to "rationalize" perceived redundancies (Kravtchenko and Demberg, 2022). For example, "Alice went to the grocery store... she paid the cashier!" can generate an inference that Alice doesn't reliably pay for groceries.

If speakers are similarly pragmatic when justifying beliefs with reasons, a costly explanation would only be worthwhile when the conclusion is particularly improbable or difficult to reconcile with the listener's other beliefs. Then, if listeners rationalize excess reasons by assuming that the speaker expected a disagreement, reasons that evidentially count in favor of the conclusion could pragmatically count against it.

2 Model

We develop a pragmatic model of over-explanation using the Rational Speech Act framework (Frank and Goodman, 2012). We focus our study on reasons for actions, using stimuli in the form *"I did [action] because [reasons]"*.

2.1 Notation

We model the listener's belief state as a beta distribution on the interval $O \in [0; 1]$, parameterized by mean μ and fixed sample size. Lower values of O are assigned to less favorable or preferable actions, and higher values of O to more favorable or preferable actions. Therefore, a distribution with higher

probability on lower values of O corresponds to an unfavorable opinion, and a distribution with higher probability on higher values of O corresponds to a favorable opinion.

2.2 Literal Listener

The model is standardly grounded in a literal listener L_0 who performs no pragmatic reasoning. He updates his beliefs simply by conditionalizing on the reasons R in a straightforward application of Bayes' theorem:

$$P_{L_0^\mu}(O|R) \propto P_{L_0^\mu}(R|O) \cdot P_{L_0^\mu}(O) \tag{1}$$

We use the notation L_0^μ for a literal listener whose prior has mean μ.

2.3 Pragmatic Speaker

A pragmatic speaker S_1's objective is to maximize the informativity of her utterance, while minimizing its cost (Eq. 3). The informativity is given by the log probability (negative surprisal) that the listener assigns to favorable opinions (i.e. $P_{L_0^\mu}(O > \theta)$, for some threshold θ).

$$P_{L_0^\mu}(O > \theta|R) = \int_\theta^1 P_{L_0^\mu}(O|R)dO \tag{2}$$

$$\mathbb{U}_{S_1}(R;\theta,\mu) = \underbrace{\log P_{L_0^\mu}(O > \theta|R)}_{\textbf{informativity}} - \text{cost(R)} \tag{3}$$

Since longer explanations have higher production cost, the speaker prefers to provide many reasons only when she assumes she is talking to a literal listener with an unfavorable prior (low μ). The reasons would significantly increase the probability of obtaining the threshold in that case. If she thinks the literal listener *already* has a favorable prior, the informativity of the reasons is negligible, so it is more economical for her to omit them.

The speaker selects an utterance by softmax-optimizing her utility:

$$P_{S_1}(R|\theta,\mu) = \text{softmax}(\alpha \cdot \mathbb{U}_{S_1}(R;\theta,\mu)) \tag{4}$$

2.4 Pragmatic Listener

By assuming utility-maximizing behavior from S_1, the pragmatic listener L_0 can infer which distribution S_1 must be treating as L_0's prior. He performs a joint inference (Eq. 5) on θ and μ: the opinion S_1 is trying to induce, and the opinion she expects L_0 to have, respectively. He marginalizes out θ to infer μ (Eq. 6).

$$P_{L_1}(\theta,\mu|R) \propto P_{S_1}(R|\theta,\mu) \cdot P(\theta) \cdot P(\mu) \tag{5}$$

$$P_{L_1}(\mu|R) = \int_0^1 P_{L_1}(\theta,\mu|R)d\theta \tag{6}$$

Since S_1 is more likely to produce a long explanation when she expects an unfavorable prior μ, L_1 tends to infer lower values of μ when he hears longer explanations. He performs the final Bayesian update from an adjusted prior computed as a weighted average over all possible μ:

$$P_{L_1}(O|R) \propto \int_0^1 \underbrace{P(R|O) \cdot P_{L_1^\mu}(O)}_{\textbf{Conditionalization}} \cdot \underbrace{P_{L_1}(\mu|R)}_{\textbf{Prob. of prior}} d\mu \tag{7}$$

Since longer explanations induce lower values of μ, the evidential strength of the reasons may not be sufficient to recover from the reduction of the prior, resulting in the "backfire" effect.

3 Methods

60 participants were recruited for a behavioral study to test the model's predictions. Stimuli consisted of text-message dialogues, in which the speaker either explained (long-4R) or did not explain (null-0R) a typical or atypical action. For the "atypical" items, the listener also prompted the speaker to elaborate, creating a context where multiple reasons would be expected/informative. Using a 7-point Likert scale, participants responded to two questions.

- RQ1: How favorably did the speaker expect the listener to feel about the action? (i.e. μ)
- RQ2: How favorably will the listener actually feel? (i.e. L_1's posterior $P(O|R)$).

In the four-reasons condition, participants were also asked to complete a free-text response, asking why the speaker chose to explain herself. Sample trials can be found in the appendix.

4 Results

4.1 Model selection

Analysis for the Likert rating questions was conducted using a Bayesian cumulative logit model for ordinal data, using the R (R Core Team, 2021) package `brms` (Bürkner, 2017) with default priors. The fixed effects included REASONS (null-0R vs long-4R), ACTIONTYPE (typical vs atypical), and their interaction, treating null-0R and typical as the baseline conditions. To account for participant and item-level variability, random effects for PARTICIPANT and ITEM were also included.

4.2 RQ1: Reduction of the prior

The first Likert rating question targets RQ1, whether listeners rationalize over-explanation by inferring that the speaker expected an unfavorable prior. Responses are plotted in Figure 1.

The model identified a credible main effect of REASONS (β = -0.77, 95% CI -1.09, -0.43), indicating less favorable expected priors when more reasons were given. There was also a credible interaction of REASONS with ACTIONTYPE (β = 0.83, 95% CI 0.35, 1.30) reversing the main effect entirely for atypical actions. There was also no credible effect of ACTIONTYPE, indicating that typical actions were not consistently rated higher or lower in favorability than atypical actions in the null-0R condition. Results of the models are summarized in Table 1.

These results are in keeping with the pragmatic model. Participants rationalized the speaker's decision to provide superfluous reasons by inferring that the speaker expected the listener to have an unfavorable prior (i.e. he would need to be *convinced* to share her opinion).

4.3 RQ2: Reduction of the posterior

The second Likert rating question targets RQ2: whether higher-order inferences about *why the speaker felt an explanation was necessary* would mitigate its effectiveness is persuading the listener. Responses are plotted in Figure 2.

There was no credible main effect of REASONS for typical actions. However, there was a credible interaction of REASONS with ACTIONTYPE (β = 0.75, 95% CI 0.28, 1.22), meaning that the reasons improved the posterior favorability ratings *only for atypical actions*. There was still no credible main effect of ACTIONTYPE, indicating both action types were rated similarly in the null-0R condition. Results of the models are summarized in Table 2.

The listeners' posterior favorability ratings were never credibly *lower* in the long-4R condition than in the null-0R condition. However, these results still support a pragmatic analysis. While typical and atypical actions received similar ratings with no reasons, the reasons significantly *increased* the rating of the atypical actions. Since no reduction of the prior was observed for those items, the evidential strength of the reasons yielded a net improvement in the posterior. By contrast, since a reduction of the prior was observed for typical actions, the evidential strength of the reasons was enough to recover the prior, but not enough to show a net increase in the posterior.

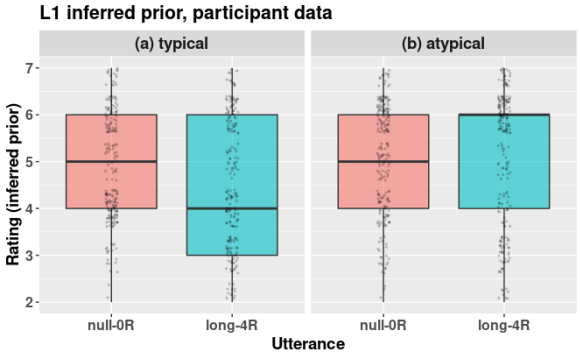


Figure 1: Responses to Q1. Prior favorability ratings reduced when reasons were produced for typical actions.

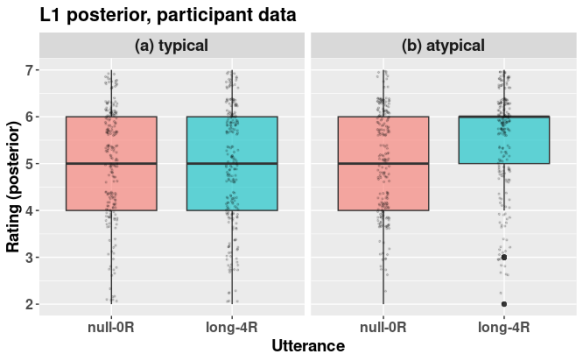


Figure 2: Responses to Q2. Posterior favorability ratings increased only when reasons were provided for atypical actions.

5 Discussion

In this study, we set out to explain the intuition that arguments sometimes seem weaker when more reasons are provided. This intuition seems to contradict basic principles of rational belief revision, which predict that argument strength should increase monotonically as additional favorable evidence becomes available. We proposed that this effect might be produced by social reasoning about the expected behavior of an efficient communicator, which we modeled using the RSA framework.

Our experiment confirmed that listeners do indeed make inferences about speakers' motivations for providing reasons, and that those inferences do diminish the evidential effect of the reasons on listeners' posterior beliefs, compared to cases where the inference is absent. However, the inference we observed from redundant reasons was not strong enough to result in a net *reduction* of the posterior, which the RSA model had predicted.

Limitations

Although our results show that listeners rationalize redundant reasons by inferring that the speaker expected a disagreement, and that this inference can negatively impact argument strength, there is no evidence for true "non-monotonicity" (i.e. the posterior was not credibly lower in the four-reasons condition than the zero-reasons condition). Further, as with most RSA designs, the model's ability to predict the data is highly dependent on the choice of parameter values, in particular the utterance cost and likelihood. Additional norming will be required for empirically motivated parameter selection.

References

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

H. Paul Grice. 1989. Logic and conversation. In *Studies in the Way of Words*, pages 22–40. Harvard University Press, Cambridge, MA.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. Number 104 in CSLI Lecture Notes. CSLI Publications, Stanford, CA.

Ekaterina Kravtchenko and Vera Demberg. 2022. Informationally redundant utterances elicit pragmatic inferences. *Cognition*, 225:105159.

Hugo Mercier. 2020. *Not Born Yesterday*. Princeton University Press, Princeton.

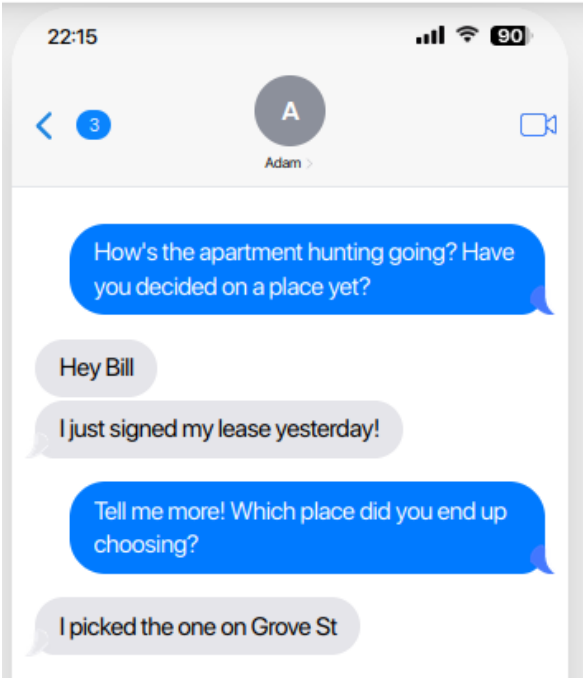
Mike Oaksford and Ulrike Hahn. 2004. A bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58(2):75–85.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

A Appendix



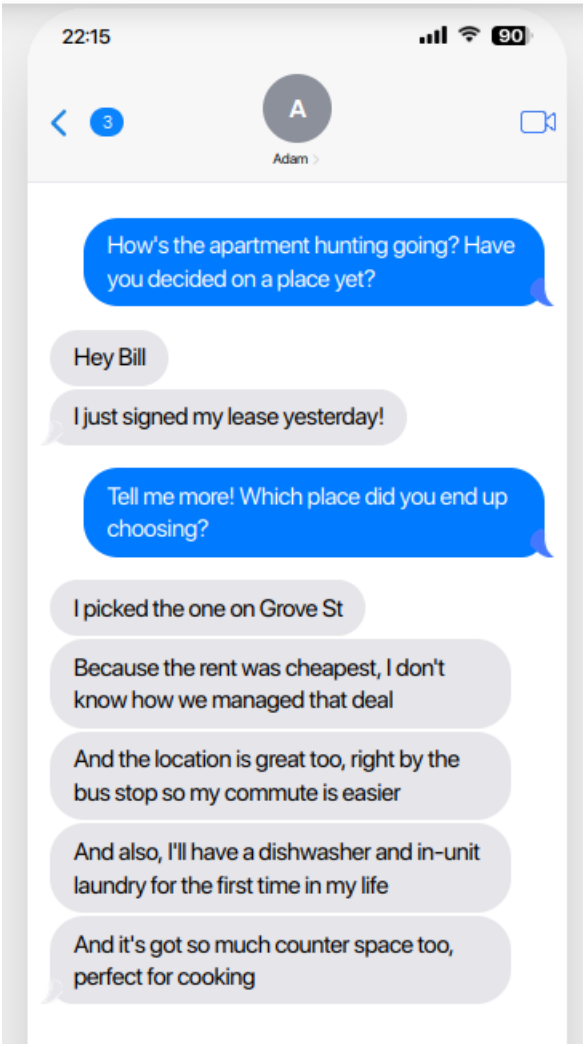


Figure 3: Sample atypical stimuli in the null-0R and long-4R conditions.

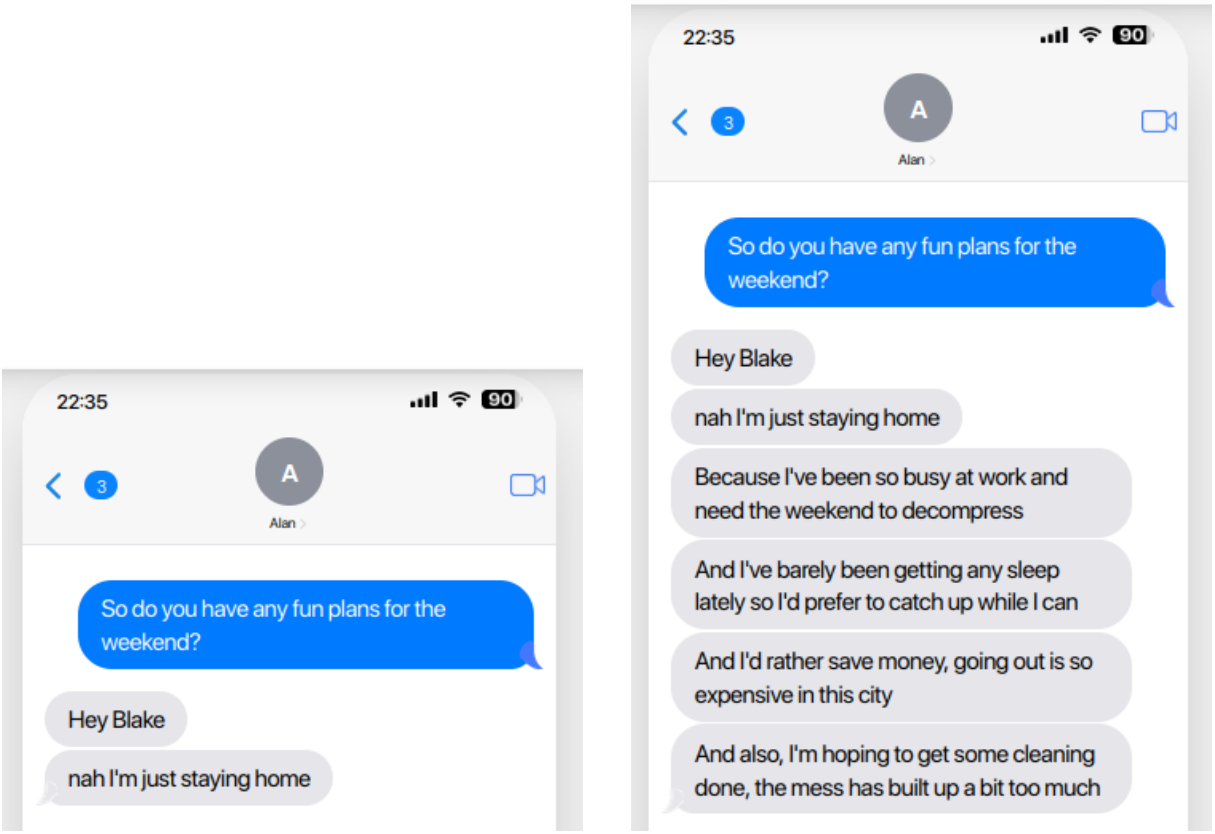


Figure 4: Sample typical stimuli in the null-0R and long-4R conditions.

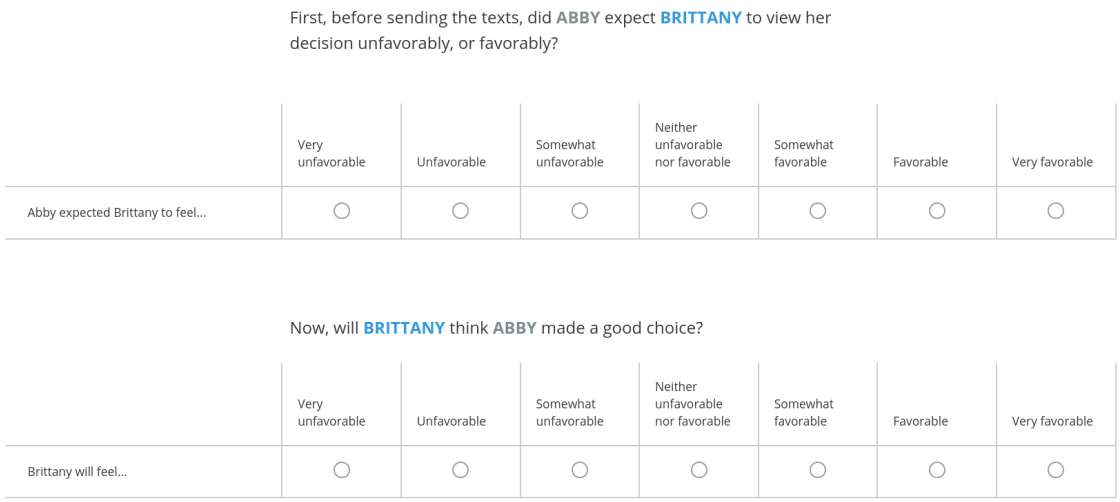


Figure 5: Sample trial.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
LONG4R	**-0.77**	**0.17**	**-1.09**	**-0.43**
ATYPICAL	0.19	0.66	-1.12	1.52
LONG4R:ATYPICAL	**0.83**	**0.24**	**0.35**	**1.30**

Table 1: Log-odds coefficients from the Bayesian ordinal model of RQ1. A **credible** main effect of REASONS, moderated by a **credible** interaction with ACTIONTYPE, is observed.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
LONG4R	0.15	0.17	-0.20	0.50
ATYPICAL	0.25	0.76	-1.25	1.79
LONG4R:ATYPICAL	**0.75**	**0.24**	**0.28**	**1.22**

Table 2: Log-odds coefficients from the Bayesian ordinal model of RQ2. A **credible** interaction of REASONS and ACTIONTYPE is observed.