Polite speech arises from desires to be helpful and look helpful

Erica J. Yoon[1,*], Michael Henry Tessler[1,*], Noah D. Goodman[1], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University

[*] These authors contributed equally to this work.

Author Note

FIXME: author note?

Correspondence concerning this article should be addressed to Erica J. Yoon, 450 Serra Mall, Bldg. 420, Rm. 290, Stanford, CA 94305. E-mail: ejyoon@stanford.edu

Abstract

Conveying information in a false or indirect manner in consideration of listeners' wants (i.e. being polite) seemingly contradicts an important goal of a cooperative speaker: information transfer. We model production of polite speech in which speakers deviate from being maximally informative for social reasons. We show that speakers produce polite speech due to their desires to be helpful – both epistemically (convey the true state to the listener) and socially (make the listener feel good) – and to *appear* helpful. We formalize this tradeoff between speaker's goals within a probabilistic model and show the model is able to predict people's polite speech production judgments. Our extension of formal theories of language to account for speakers' social goals represents an advance in understanding of human speech.

Keywords: Politeness; computational modeling; communicative goals; pragmatics

Word count: X

Polite speech arises from desires to be helpful and look helpful

One-sentence summary

Polite speech arises from desires to be helpful – both epistemically and socially – and to *appear* helpful.

Introduction

Human speech is an important means of exchanging information, but intriguingly it often deviates from maximally efficient and accurate information transfer. Instead of saying the most direct message that the speaker wants the listener to access ("I know for a fact that you failed the exam"; "Open the window!"), Speakers produce vague or underinformative remarks ("I don't think you did very well on the exam") or add extraneous, seemingly irrelevant markers ("*could you please* open the window?"). People sometimes even produce false utterances that completely misrepresents the speaker's knowledge ("Your talk was great!" about a truly terrible talk).

Polite language, in which speakers convey information in a false or indirect manner in consideration of listeners' wants, violates a critical principle of cooperative communication: exchanging information efficiently and accurately (Grice, 1975). Yet polite speech serves another important goal of communication: maintaining and improving social relationships. Here we propose that cooperative communication reflects a principled tradeoff between two goals: epistemic goal, or to convey information accurately and efficiently; and social goal, or to make the interactants feel good.

How can we model production of polite speech? The Rational Speech Act (RSA) framework describes language understanding as recursive probabilistic inference between a pragmatic listener and an informative speaker. This framework has been successful at capturing the quantitative details of a number of language understanding tasks but it neglects the social goals a speaker may pursue. On the other hand, informal theories of politeness explain how speakers' social goals give rise to polite speech. For example, Brown

and Levinson (1987) argue that deviation from informativity increases the level of polite face-saving. But there has been no formalization of the notion of speakers' social goals, thus no systemic quantitative predictions of politeness theories have been available.

We propose a computational model of polite speech (**pRSA**) that accounts for both epistemic and social goals of speakers. RSA models assume speakers choose utterances approximately optimally given a utility function (Goodman & Stuhlmüller, 2013). In our model, the speaker's utility function can be decomposed into two components: First, *epistemic utility* refers to the standard, informative utility in RSA: the amount of information a literal listener (L_0) would still not know about world state s after hearing a speaker's utterance w. Second, *social utility* is the expected subjective utility of the state inferred given the utterance w. The expected subjective utility is related to the intrinsic value of the state, and we use a value function (V) to map states to subjective utility values. This captures the affective consequences for the listener of being in state s. The utility weight (single mixture parameter ϕ_{S_1}) determines how informative versus social the speaker wants to be: a higher ϕ_{S_1} signifies the epistemic goal prioritized over the social goal. Finally, some utterances might be costlier than others. The utility of an utterance subtracts the cost $c(w)$ from the weighted combination of the social and epistemic utilities.

$$U(w; s; \phi]) = \phi_{S_1} \cdot L_0(s \mid w) + (1 - \phi_{S_1}) \cdot V[L_0(s \mid w)] - C(w)$$

The recursive reasoning in our model unfolds as follows: The speaker (S_1) in pRSA chooses utterances w softmax-optimally given the state s and his goal mixture parameter weight ϕ. Given the speaker's utterance, the pragmatic listener (L_1) jointly infers the state s and the utility weight ϕ_{S_1}. Finally, the pragmatic speaker (S_2) chooses an utterance, based on the pragmatic listener L_1's model and one of three possible different goals he can have: (1) true epistemic goal to convey the true state ($\phi_{epistemic}$); (2) true social goal to make L1 feel good (ϕ_{social}); (3) self-presentational goal to convey a certain ϕ_{S_1} to L_1 (i.e. to *appear*

informative or kind; ϕ_{self}).

$$P_{S_2}(w \mid s, \hat{\beta}) \propto \exp(\phi_{epistemic} \cdot L_1(s \mid w) + \phi_{social} \cdot V[L_1(s \mid w)] + \phi_{self} \cdot L_1(\phi_{S_1} \mid w))$$

We used a simple procedure to empirically test whether our model is able to predict production of polite utterances. Participants read scenarios in which someone (e.g. Ann) gave a performance of some kind, and another person (Bob) evaluated it. We provided information on Ann's feelings toward the presentation (*true state*), which were shown on a scale from zero to three hearts (e.g. two out of three hearts filled in red color; see Figure 1). We also presented Ann's goal, which was one of the following: (1) to be *informative* and give accurate feedback; to be *social* and to make Bob feel good; or to be *both* informative and social at the same time. We hypothesized that speakers with both goals to be informative and social given bad true states (i.e. Bob's performance was poor) would produce more negation ("It wasn't~") to save the listener's face while vaguely conveying the bad true state (see our pre-registered model, hypothesis, and procedure at FIXME). Each participant read 12 scenarios total (4 true states × 3 goals).

In a single trial, each scenario was followed by a question that asked for the most likely utterance by Ann. Participants indicated their answer by choosing one of the options on the two dropdown menus, side-by-side, one for choosing between *It was* vs. *It wasn't* and the other for choosing among *terrible*, *bad*, *okay*, *good*, and *amazing*, thereby selecting one of ten possible utterances (see Figure 1). We separately gathered the literal meanings of the ten possible utterances, by measuring how likely each utterance is to be true given each true state, to set expected literal meanings of utterances in our model (see Supplementary Materials for literal semantic results).

Results

Mean proportion of utterances chosen by participants in each true-state × goal condition were overall highly consistent with the our model predictions (Figure 2). The posterior predictive of the model explained almost all of the variance in the production data

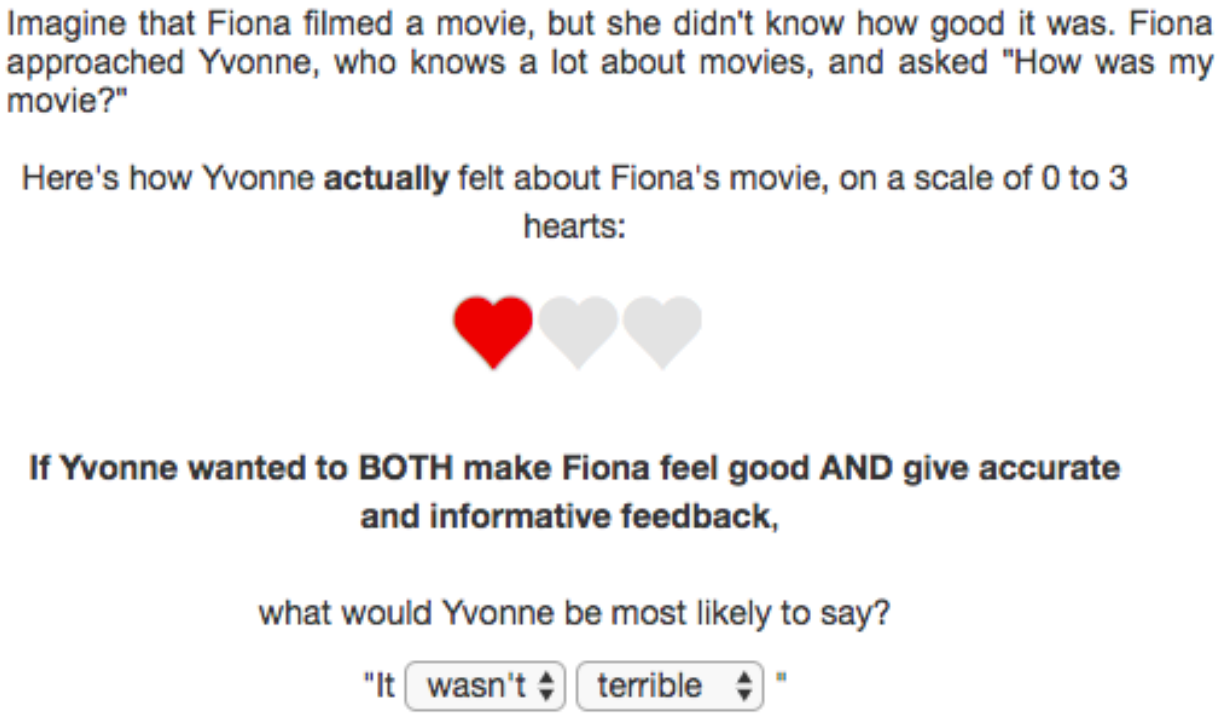

Figure 1. Example of a trial in the speaker production task.

$r^2(96) = 0.97$ (Figure 3). Consistent In line with our hypothesis, the both-goal speaker $\times$ bad true state (0 or 1 heart) conditions yielded the greatest proportion of negation ("It wasn't ~"; see Figure 4).

Our work unifies previous formal models of communication and informal theories of social uses of language. Our findings suggest that neither epistemic nor social motives alone motivate polite speech; instead, production of polite speech results from the conflict between these two, combined with a self-presentational desire to *look* epistemically and socially helpful. These findings provide strong support for a utility-theoretic framing of politeness, and suggest new directions in understanding of pragmatic language use in social contexts.

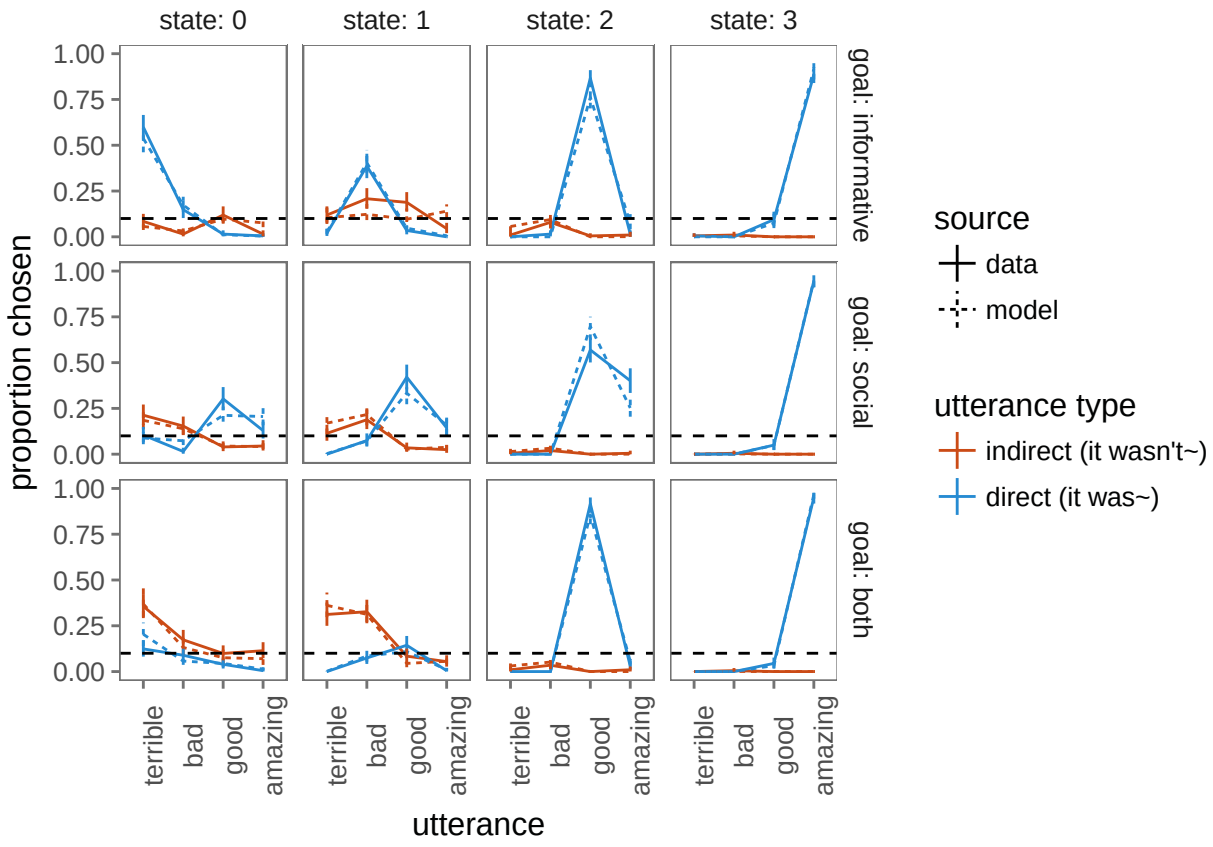

Figure 2. Experimental results (solid lines) and fitted model predictions (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.

Appendices

Literal semantic judgments

Inferred parameters

Data analysis tools

We used R (3.4.2, R Core Team, 2017) and the R-packages *bindrcpp* (0.2, Müller, 2017), *binom* (1.1.1, Dorai-Raj, 2014), *coda* (0.19.1, Plummer, Best, Cowles, & Vines, 2006), *dplyr* (0.7.4, Wickham, Francois, Henry, & Müller, 2017), *forcats* (0.2.0, Wickham, 2017a),

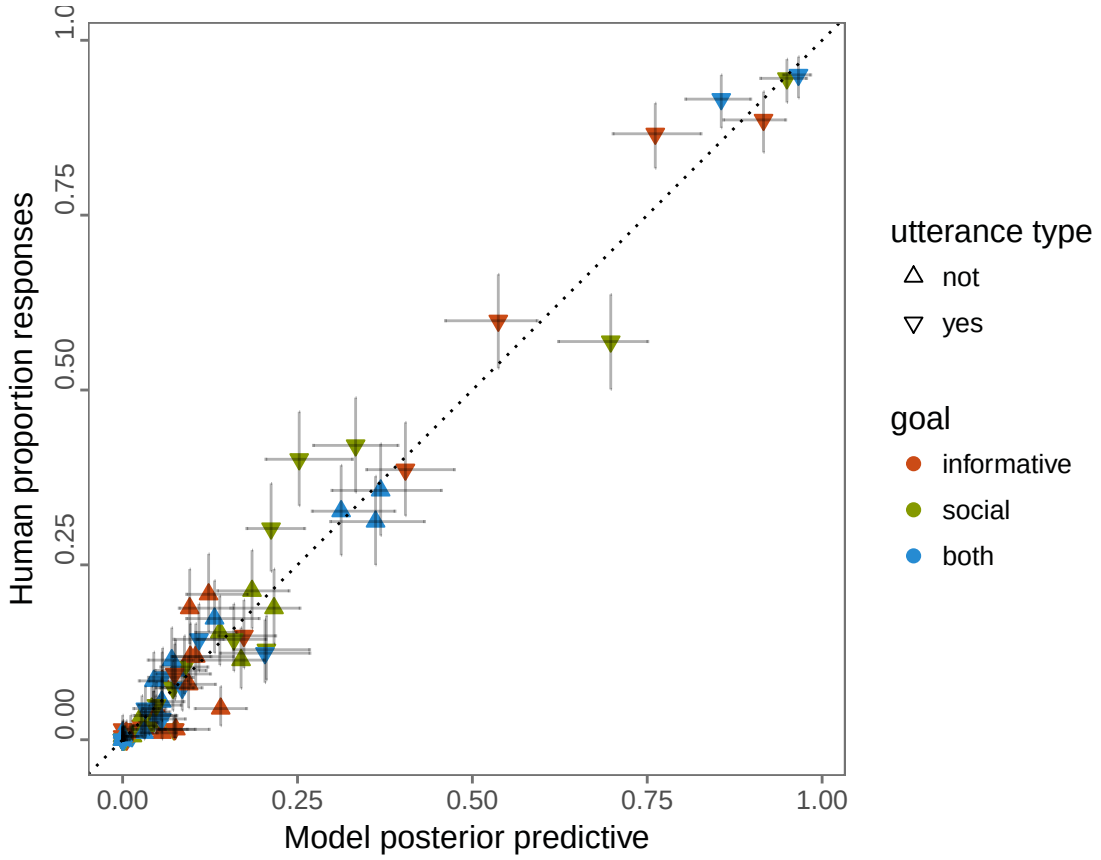

Figure 3. Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

ggplot2 (2.2.1, Wickham, 2009), *ggthemes* (3.4.0, Arnold, 2017), *gridExtra* (2.3, Auguie, 2017), *jsonlite* (1.5, Ooms, 2014), *langcog* (0.1.9001, Braginsky, Yurovsky, & Frank, n.d.), *magrittr* (1.5, Bache & Wickham, 2014), *papaja* (0.1.0.9492, Aust & Barth, 2017), *purrr* (0.2.4, Henry & Wickham, 2017), *readr* (1.1.1, Wickham, Hester, & Francois, 2017), *rwebppl* (0.1.97, Braginsky, Tessler, & Hawkins, n.d.), *stringr* (1.2.0, Wickham, 2017b), *tibble* (1.3.4, Müller & Wickham, 2017), *tidyr* (0.7.2, Wickham & Henry, 2017), and *tidyverse* (1.2.1, Wickham, 2017c) for all our analyses.

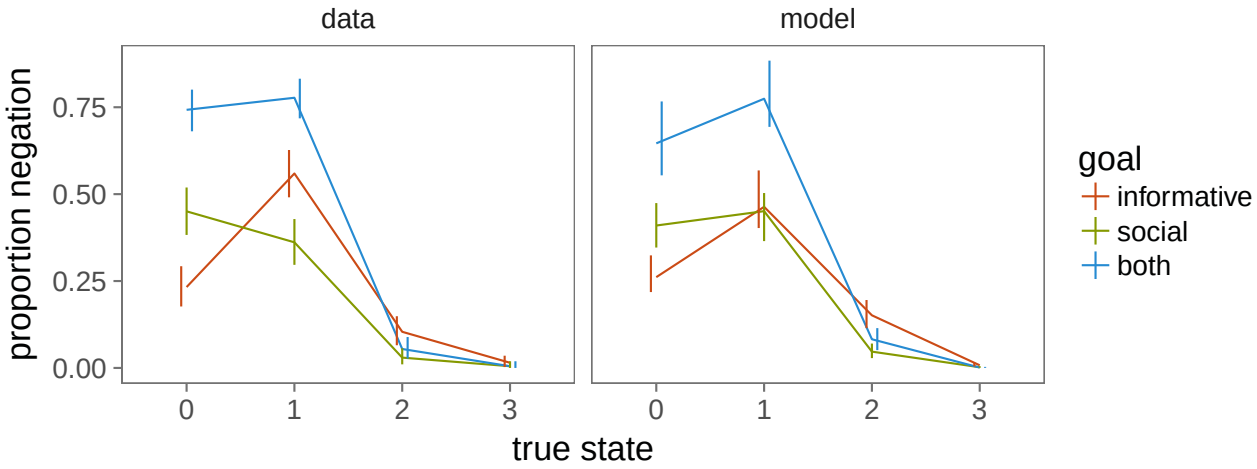

Figure 4. Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

Model parameter and weight comparison

Here we compare predictions of the current model with its possible alternatives. The current model has a triple mixture structure, with three goals each of which is assigned a different weight: (1) goal to be truly informative (i.e. want to convey the true state); (2) goal to be truly social (i.e. want to make the listener feel good); (3) self-presentational goal to appear certain way (as determined by s1-phi). Alternative models involve one or two out of these three components. Below we show that the current model best captures the production pattern in the empirical data.

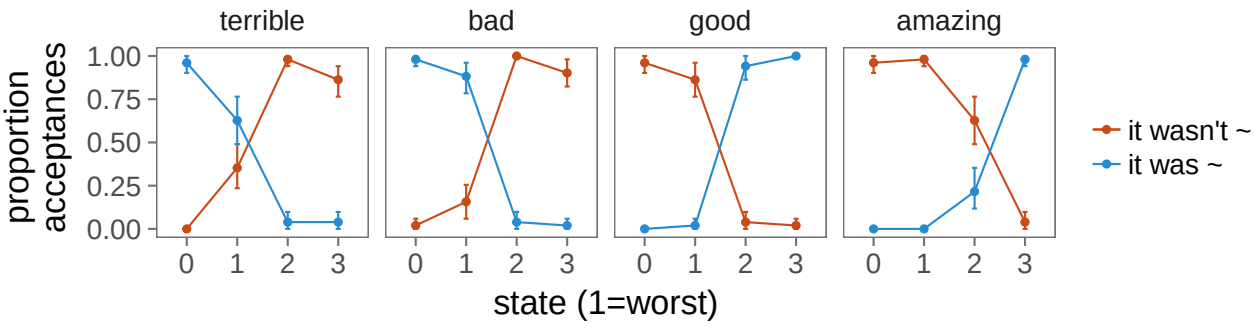

Figure 5. Semantic measurement results. Proportion of acceptances of utterance types (colors) combined with target words (facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

References

Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r*. Retrieved from https://CRAN.R-project.org/package=magrittr

Braginsky, M., Tessler, M. H., & Hawkins, R. (n.d.). *Rwebppl: R interface to webppl*. Retrieved from https://github.com/mhtess/rwebppl

Braginsky, M., Yurovsky, D., & Frank, M. (n.d.). *Langcog: Language and cognition lab things*. Retrieved from http://github.com/langcog/langcog

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Dorai-Raj, S. (2014). *Binom: Binomial confidence intervals for several parameterizations*. Retrieved from https://CRAN.R-project.org/package=binom

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language

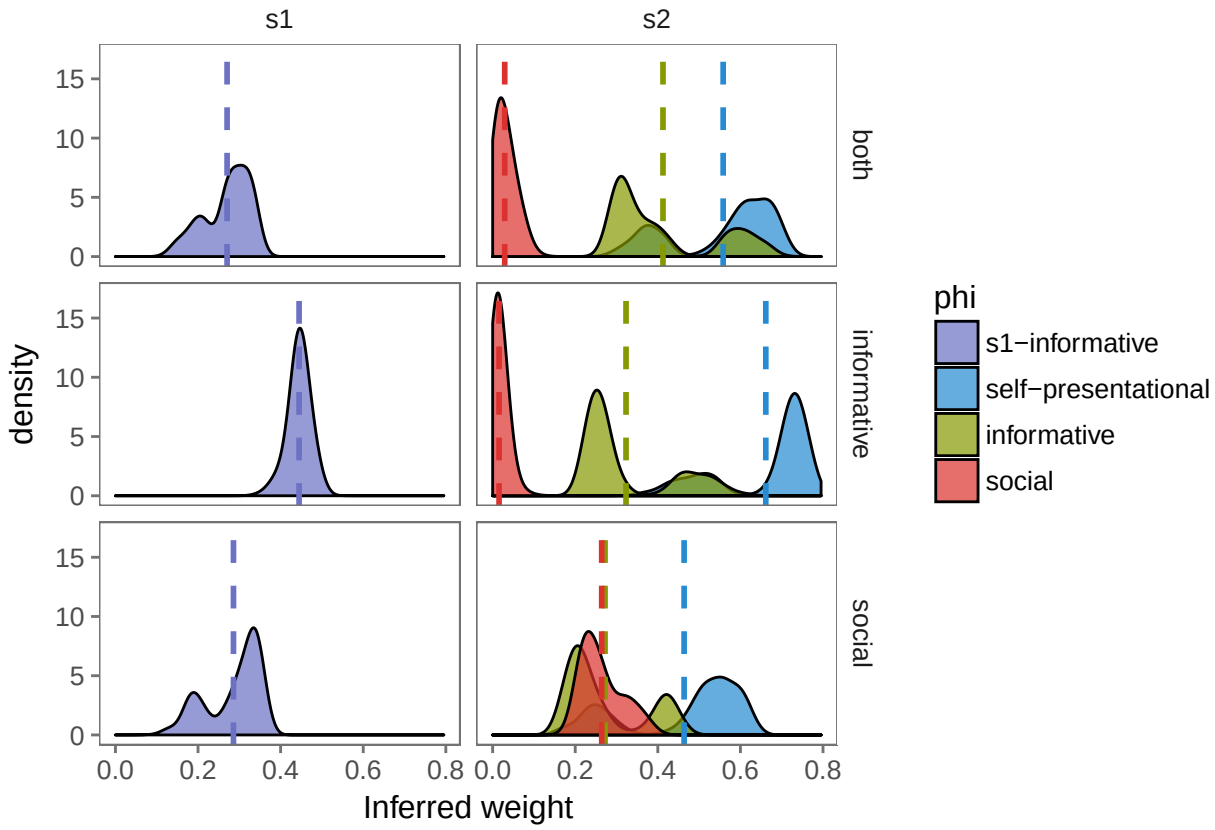

Figure 6. Inferred goal weights. Horizontal facets are different experimental conditions (trying to be X). Density plots show likely weights used in the speaker's utility function.

understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). Academic Press.

Henry, L., & Wickham, H. (2017). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from https://CRAN.R-project.org/package=bindrcpp

Müller, K., & Wickham, H. (2017). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*. Retrieved from

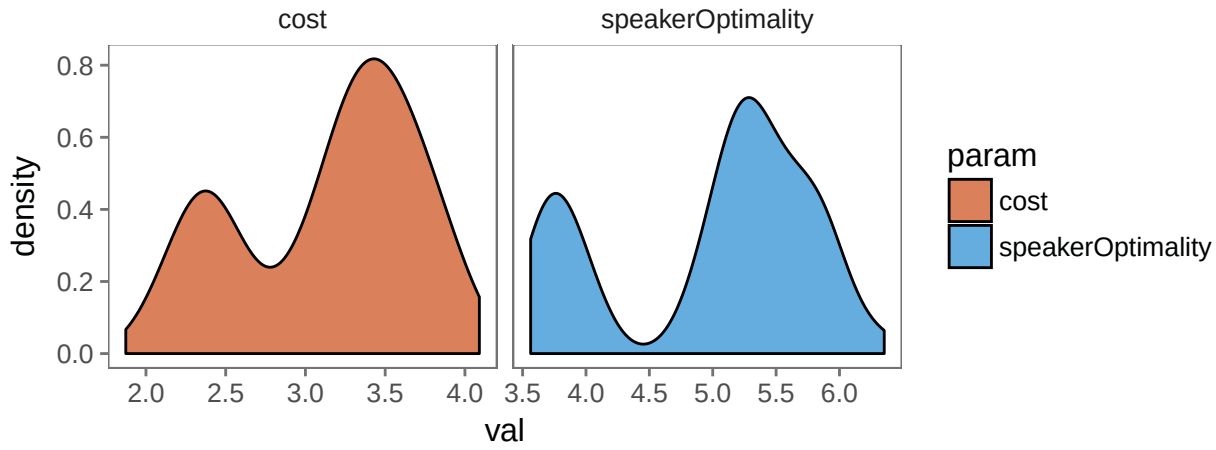

Figure 7. Inferred cost and speaker optimality parameters from the main model.

https://arxiv.org/abs/1403.2805

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2017b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2017c). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., & Henry, L. (2017). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data*

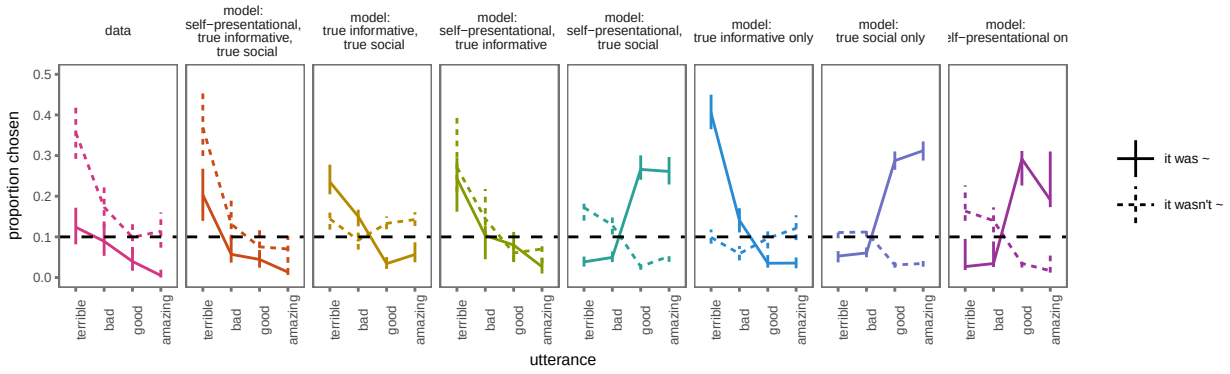

Figure 8. Utterances from data (leftmost) and predictions from different model alternatives for a speaker with both goals addressing the true state of 0 heart. Proportion of utterances chosen (direct utterances in solid lines and indirect utterances in dotted lines, and words shown on x-axis). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.

manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2017). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

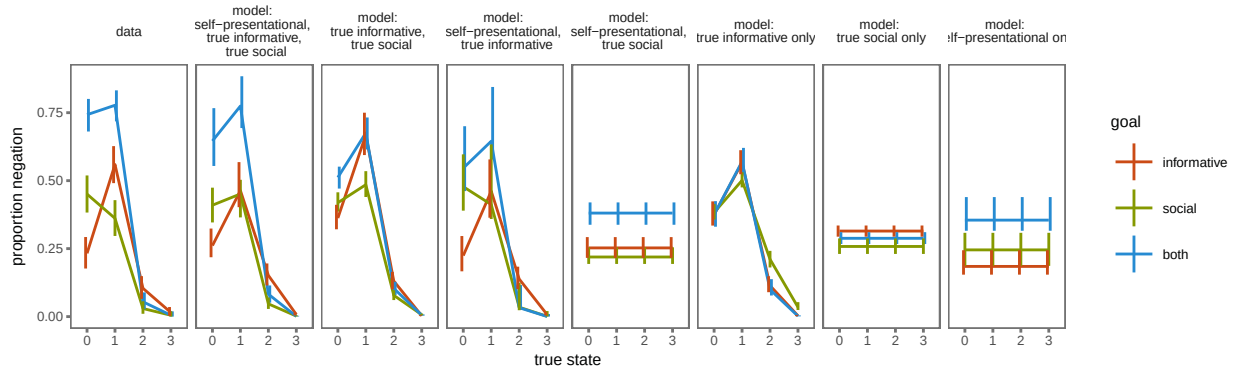

Figure 9. Experimental results (leftmost) and predictions from different model alternatives for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

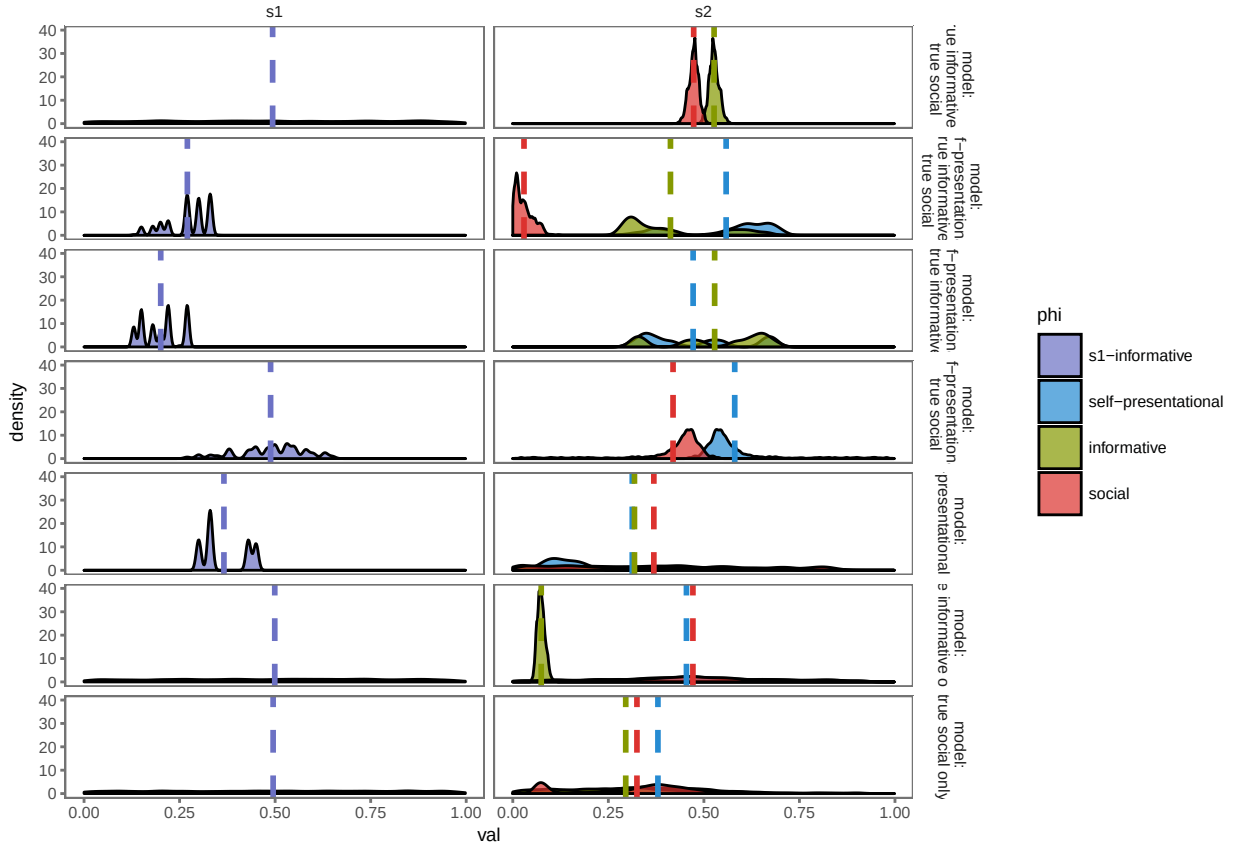

Figure 10. Inferred goal weights from different model alternatives. Horizontal facets are different experimental conditions (trying to be X). Density plots show likely weights used in the speaker's utility function.