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,
450 Serra Mall, Stanford, CA 94305.
[*]These authors contributed equally to this work.
[†]To whom correspondence should be addressed; E-mail: ejyoon@stanford.edu.

People often deviate from saying the maximally informative and succinct message (e.g. "Your talk was terrible") and instead speak *politely*: Speakers convey information in a false or indirect manner in consideration of listeners' wants (e.g. "It was a great talk!"; "The talk wasn't bad; it's hard to give a good presentation anyway"). If language is an important means of information transfer, does polite speech contradict cooperative communication? We argue that polite speech reflects a cooperative speaker's desire to balance between two communicative goals: to be *epistemically* helpful and convey the true state to the listener; and to be *socially* helpful and make the listener feel good. We also argue that a polite speaker has a desire to *look* helpful and appear to have these two goals in mind. We formalize the tradeoff between the 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 communication to account for speakers' social goals represents an advance in understanding of human speech.

Human speech is an important means of exchanging information, but intriguingly speech often deviates from maximally efficient and accurate information transfer. Instead of saying the most direct message that the speaker wants the listener to access ("Your presentation was terrible"; "Tell me where Jordan Hall is"), speakers often produce vague or underinformative remarks ("It's hard to give a good presentation") or add extraneous, seemingly irrelevant markers ("*Could you please* tell me where Jordan Hall is?") (*1, 2*). People sometimes even produce false utterances that completely misrepresents the speaker's knowledge ("Your muffins are the best ever!" about truly terrible muffins) (*3*).

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 (*4*). Yet polite speech serves another important goal of communication: maintaining and improving social relationships. Here we propose that polite speech reflects a principled tradeoff between goals to be: *epistemic*, or to convey information accurately and efficiently; *social*, or to make the interactants feel good; and *self-presentational*, or to *appear* to be epistemically and socially helpful to others.

How can we model production of polite speech? Informal theories of politeness explain how speakers' social goals give rise to polite speech. For example, Brown and Levinson (*5*) 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. On the other hand, formal theories of language have accounted for speakers' desires to be informative, but not for their potential social goals. The Rational Speech Act (RSA) framework describes language understanding as recursive probabilistic inference between a pragmatic listener and an informative speaker (*6*). 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.

We propose a computational model of polite speech (**pRSA**) that unifies formal theories of informative communication and informal theories of polite speech, and accounts for both epistemic and social goals of speakers. RSA models assume speakers choose utterances approximately optimally given a utility function (*7*). 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) 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} that the speaker had in mind. Finally, the pragmatic speaker (S_2) chooses an utterance, based on the pragmatic listener L_1's model and weights on three different goals: (1) genuine epistemic goal to convey the true state ($\phi_{epistemic}$); (2) genuine social goal to make L1 feel good (ϕ_{social}); and (3) self-presentational goal to convey certain ϕ_{S_1} to L_1 (i.e. to *appear* informative or kind; ϕ_{self}).

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

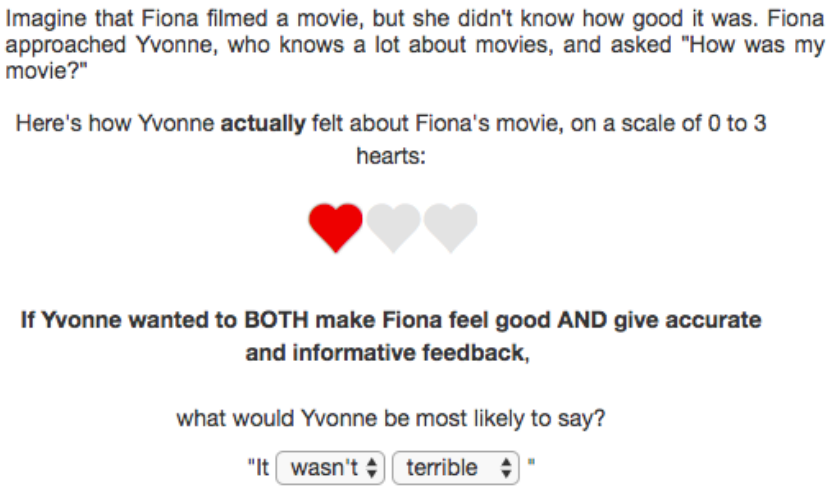

Figure 1: Example of a trial in the speaker production task.

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. one out of three hearts filled in red color; see Figure 1). We also presented Ann's goal, which was one of the following: 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 $\times$ 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

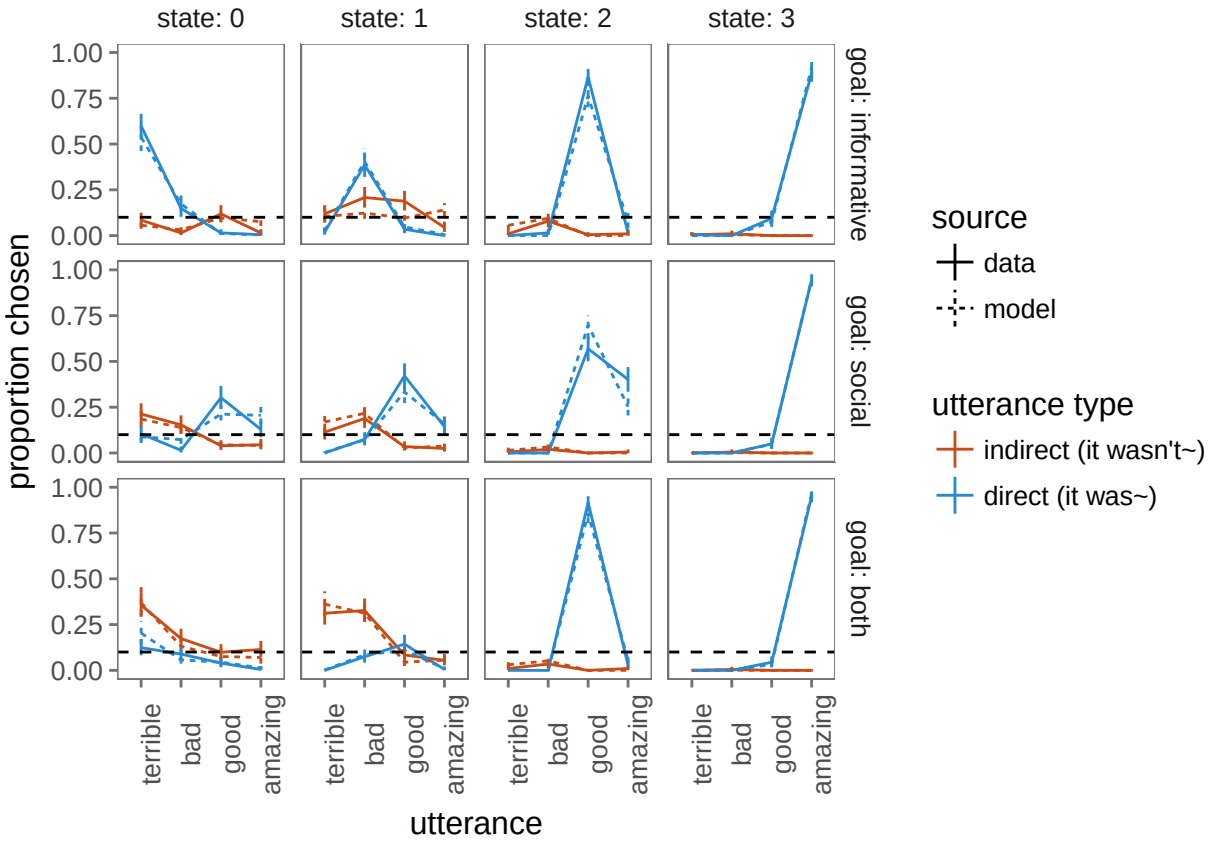

Figure 2: Experimental results (solid lines) and fitted model predictions (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars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

for choosing among *terrible*, *bad*, *good*, and *amazing*, thereby selecting one of eight possible utterances (see Figure 1). We separately gathered the literal meaning judgments for the eight 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).

Mean proportion of utterances chosen by participants in each true-state $\times$ 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 ($r^2(96) = 0.97$;

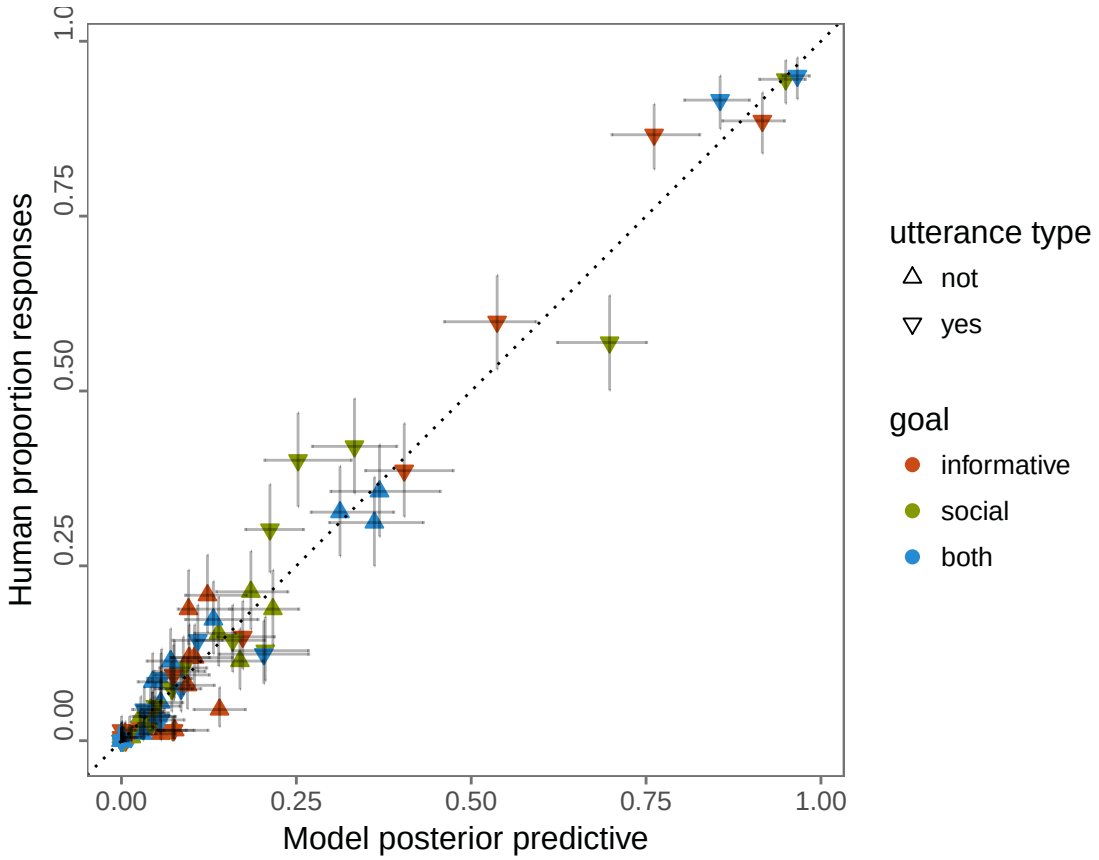

Figure 3: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

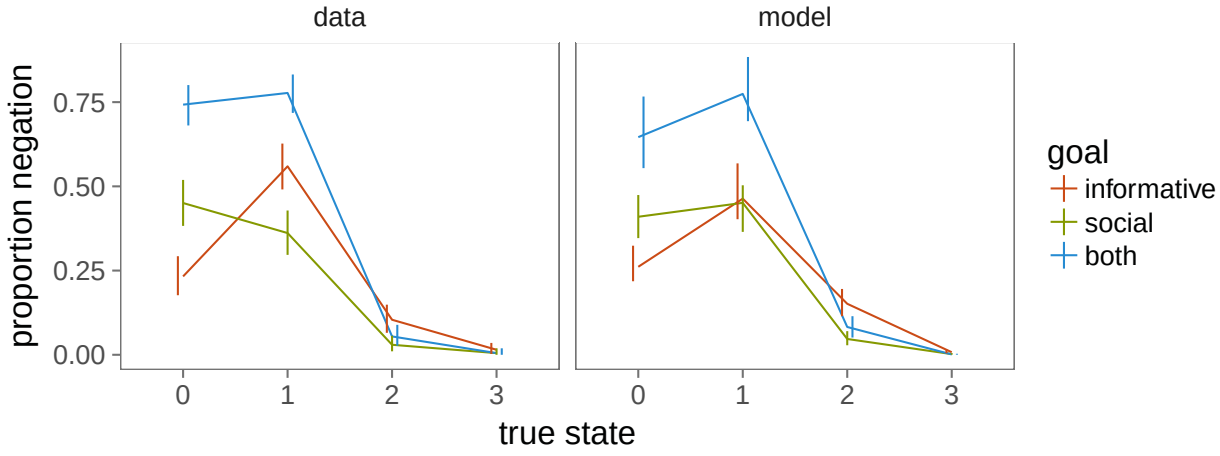

Figure 4: Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

Figure 3). In line with our hypothesis, conditions in which the both-goal speaker tried to convey bad true state (0 or 1 heart) yielded the greatest proportions 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.

References

1. H. H. Clark, D. H. Schunk, *Cognition* **8**, 111 (1980).

2. T. Holtgraves, *Cognitive Psychology* **37**, 1 (1998).

3. B. M. DePaulo, D. A. Kashy, S. E. Kirkendol, M. M. Wyer, J. A. Epstein, *Journal of personality and social psychology* **70**, 979 (1996).

4. H. P. Grice, *Logic and conversation* (Academic Press, 1975), vol. 3, pp. 41–58.

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

6. N. D. Goodman, M. C. Frank, *Trends in Cognitive Sciences* **20**, 818 (2016).

7. N. D. Goodman, A. Stuhlmüller, *Topics in cognitive science* **5**, 173 (2013).

8. M. D. Lee, E. J. Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course* (Cambridge Univ. Press, 2014).

Acknowledgments

This work was supported by NSERC PGS Doctoral scholarship PGSD3-454094-2014 to EJY, NSF Graduate Research Fellowship DGE-114747 to MHT, ONR grant N00014-13-1-0788 to NDG, and NSF grant BCS 1456077 to MCF.

Supplementary materials

Materials and Methods

Literal semantic task

We probed judgments of literal meanings of the target words assumed by our model and used in all our experiments. 51 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. We used 13 different context items in which someone evaluated a performance of some kind. For example, in one of the contexts, Ann saw a presentation, and Ann's feelings toward the presentation (*true state*) were shown on a scale from zero to three hearts (e.g., two out of three hearts filled in red color). The question of interest was "Do you think Ann thought the presentation was / wasn't X?" and participants responded by choosing either "no" or "yes." The target could be one of five possible words: *terrible*, *bad*, *good*, and *amazing*, giving rise to ten different possible utterances (with negation or no negation). Each participant read 32 scenarios, depicting every possible combination of states and utterances. The order of context items was randomized, and there were a maximum of four repeats of each context item per participant. For this and the subsequent experiment, we analyzed the data by collapsing across context items.

For each utterance-state pair, we computed the posterior distribution over the semantic weight (i.e., how consistent X utterance is with Y state) assuming a uniform prior over the weight. Meanings of the words as judged by participants were as one would expect (see Figure

S1). We used the fraction of participants that endorsed utterance w for state s to set informative priors to infer posterior credible values of the literal meanings from data in the speaker production experiment.

Speaker production task

202 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. As in the semantics measurements above, we used scenarios in which a person (e.g., Bob) gave some performance and asked for another person (e.g., Ann)'s opinion on the performance. Additionally, we provided information on the speaker Ann's goal – *to make Bob feel good*, or *to give as accurate and informative feedback as possible*, or *both* – and the true state – how Ann actually felt about Bob's performance (e.g., two out of three hearts, on a scale from zero to three hearts). Each participant read 12 scenarios, depicting every possible combination of goals and states. The order of context items was randomized, and there were a maximum of two repeats of each context item per participant.

Each scenario was followed by a question that read, "If Ann wanted *to make Bob feel good* but not necessarily give informative feedback (or *to give accurate and informative feedback* but not necessarily make Bob feel good, or *BOTH make Bob feel good AND give accurate and informative feedback*), what would Ann be most likely to say?" 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*, *good*, and *amazing*.

Supplementary Text

Model fitting

We ran 4 MCMC chains for 40,000 iterations, discarding the first 20,000 for burnin.

Inferred parameters

In the speaker production task, participants were told what speakers' intentions were (e.g. wanted to make Bob feel good). We assume that the intention descriptions conveyed the weight mixtures $\phi_{epistemic}$, ϕ_{social}, ϕ_{self} and ϕ_{S_1} that the speaker was using. We put uninformative priors on each of these mixtures ($\phi \sim$ Uniform(0,1)) and inferred their credible values separately for each goal condition ("wanted to X") using Bayesian data analytic techniques (*8*). For the "wanted to give informative feedback" (*informative*) condition, FIXME. For the "wanted to make [listener] feel good" (*social*) condition, FIXME. For the "wanted BOTH to make [the listener] feel good and give informative feedback" (*both*) condition, FIXME (see Figure S2).

There were two additional parameters of the model, on which we put uninformative priors: the value scale parameter ($\alpha \sim$Unif(0,10)) in the utility function; and the cost parameter ($C(u) \sim$Unif(1,10)). We inferred their posterior credible values from the data. The Maximum A-Posteriori (MAP) estimates and 95% Highest Probability Density Intervals (HDI) were: FIXME (see Figure S3).

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 ϕ_{S_1}). Alternative models include one or two out of these three components. Below we show that the current model with all three utility components best captures the production pattern in the empirical data.

FIXME: talk about the two figures.

Data analysis tools

We used R (3.4.2, R Core Team, 2017) and the R-packages *bindrcpp* (0.2, Mller, 2017), *binom* (1.1.1, Dorai-Raj, 2014), *coda* (0.19.1, Plummer, Best, Cowles, & Vines, 2006), *dplyr* (0.7.4, Wickham, Francois, Henry, & Mller, 2017), *forcats* (0.2.0, Wickham, 2017a), *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, Mller & Wickham, 2017), *tidyr* (0.7.2, Wickham & Henry, 2017), and *tidyverse* (1.2.1, Wickham, 2017c) for all our analyses.

Figs. S1 to S5

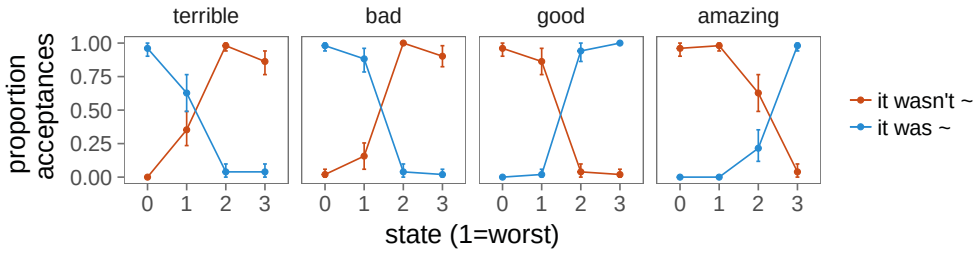

Figure S1: Semantic measurement results. Proportion of acceptances of utterance types (colors) combined with target words (facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

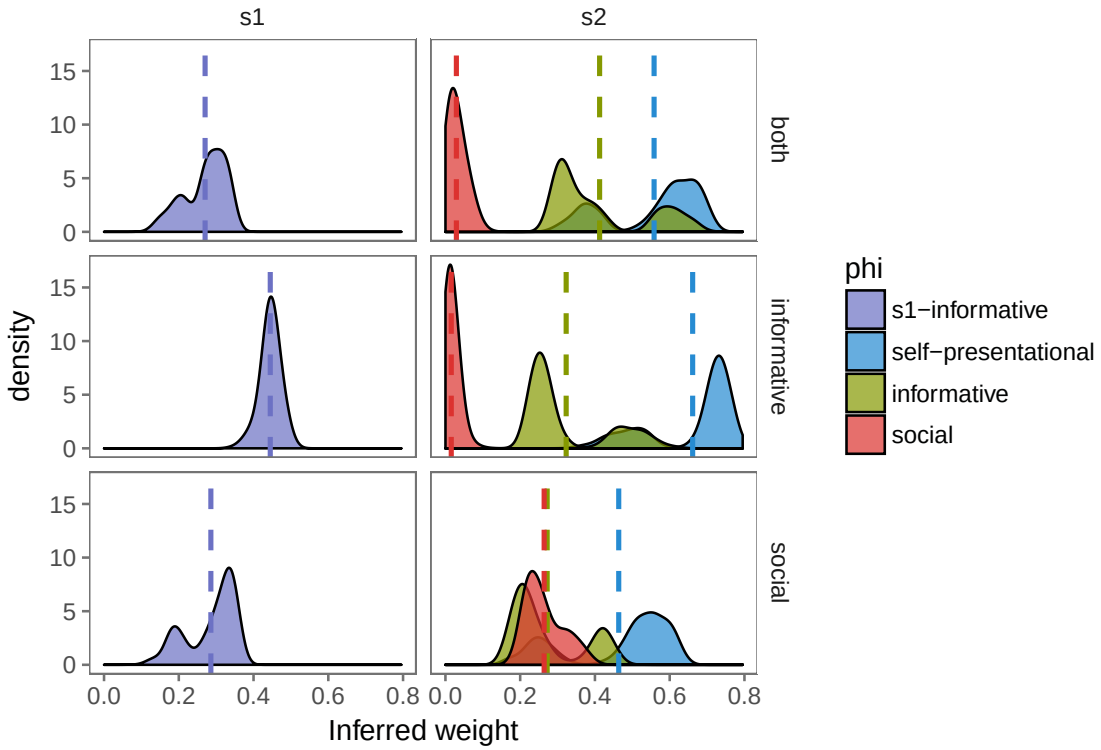

Figure S2: Inferred goal weights for the main model. Horizontal facets are different experimental conditions (trying to be X). Density plots show likely weights used in the speaker's utility function.

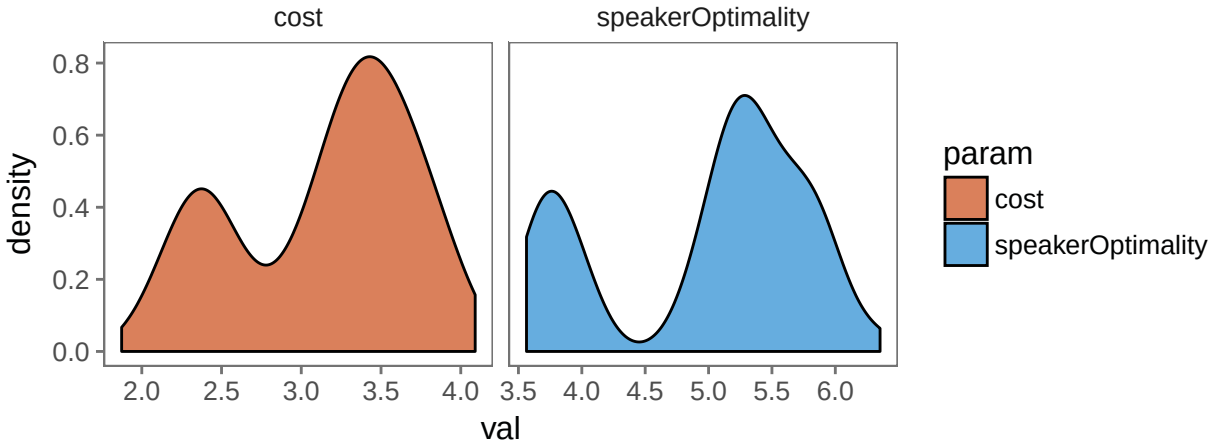

Figure S3: Inferred cost and speaker optimality parameters from the main model.

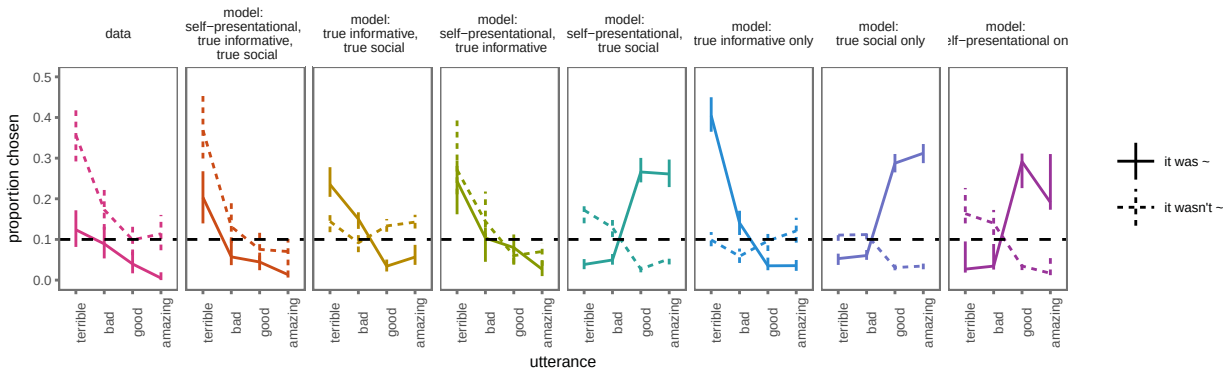

Figure S4: Utterances from data (leftmost) and predictions from different model alternatives for a speaker with both goals addressing the true state of 0 heart. Proportion of utterances chosen (direct utterances in solid lines and indirect utterances in dotted lines, and words shown on x-axis). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.

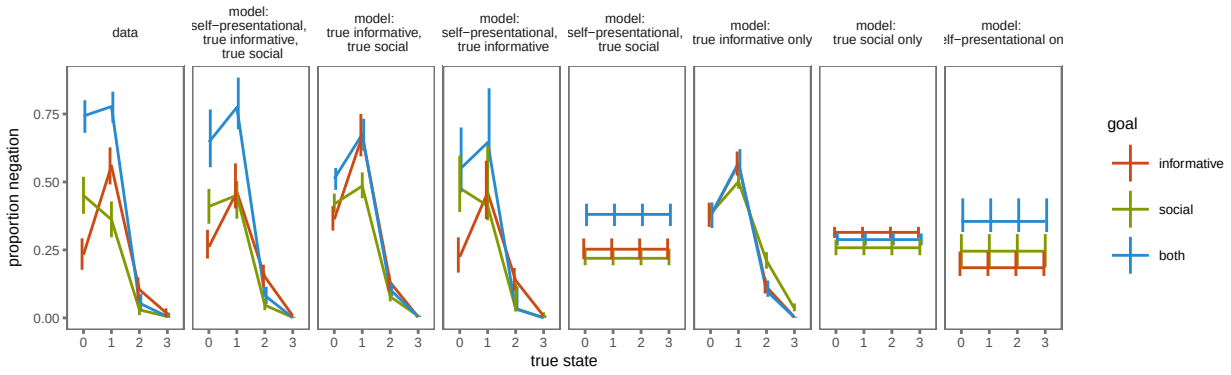

Figure S5: Experimental results (leftmost) and predictions from different model alternatives for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).