Speakers try to be helpful and look helpful: Modeling epistemic and social goals in speech

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.

Through language, people not only communicate information about the world, but also form, maintain, and improve relationships with others. We propose a computational model of speech in which the speaker's goals reflect these functions of language, and she wants to speak in a way that is: *epistemically* helpful, conveying the true state to the listener; *socially* helpful, making the listener feel good; and *presentational*, having herself appear to be helpful and have these two goals in mind. Tested against a case of simple polite speech, our model is able to predict people's speech production judgments. Our extension of formal theories of communication to account for speakers' social goals represents an advance in understanding of human speech.

Imagine your friend tries on her new dress and excitedly asks you, "How do I look?" Even at the most hideous sight, you might choose to say "You look great!" or mumble after a moment of hesitation, "Well, I don't think it looks amazing on you", rather than answer "You look terrible" to be most informative. [ejy: example at the start? if so which one? analogous to both-goal speaker's "not terrible"? need to elaborate more?]

1

Language is a versatile tool for both information exchange and social rapport: Through language, people can communicate information to improve the listeners' epistemic knowledge, but also form and improve their social relationships with others. On one hand, language can be seen as a transmission device that transfers information that reflects context or the state of affairs from a sender to a receiver (*1–3*). Informativity then is an integral assumption of language use, on which speakers rely to convey more than what their spoken words literally suggest (*4, 5*). On the other hand, language performs important social roles, whereby people make contact with others and form relationships (*6, 7*). Language users tend to speak in a manner that abides by expectations of the community (*8*), promoting a sense of social solidarity. Thus, epistemic and social functions of language translate to speakers' goals to be informative and social in their language use.

But maximizing on both informativity and social goals is difficult, as they often conflict with each other. Courteous language to make the listener feel good ("Your dress is gorgeous" about a truly hideous dress; "It's hard to give a good presentation" after an awful talk) will risk potential loss of their intended message, suffer inefficiencies, or even convey wrong information. Conversely, maximally informative and efficient utterances can seem too blunt or rude ("Be quiet!"; "Read my essay and give feedback now.") and fail to promote or maintain good relationships. How do people then speak to find balance between goals to be informative versus socially apt?

Polite speech, in which people try to preserve other people's feelings by softening their potentially harsh message, reflects the compromise between informative and social goals in conflict. On a daily basis, adults and even young children spontaneously produce polite, indirect requests ("Can you please close the window?") rather than direct orders ("Close the window.") (*9, 10*), and speakers use polite forms of speech even while arguing, preventing unnecessary offense to their interactants (*11*). Based on informativity-focused accounts of lan-

guage, politeness clashes with the idea of cooperative communication as it violates the goal to exchange information efficiently and accurately (*4*); if information transfer were the only currency in communication, politeness would be both infelicitous and undesirable. A polite speaker, however, can be seen to be cooperative and helpful via both an *epistemic* goal to improve the listener's knowledge and a *social* goal to minimize potential damage to the hearer's (and the speaker's own) self-image, called *face* (*12*). If the speaker's intended meaning contains no threat to the speaker or listener's face, then the speaker will choose to convey the meaning in an explicit and efficient manner (putting it "on the record"). As the degree of face-threat becomes more severe, however, a speaker will choose to be polite by producing more indirect utterances.

In this paper, using a case study of simple polite speech, we formalize the idea that language reflects a principled tradeoff between speakers' epistemic and social communicative goals. Previous informal theories of politeness tried to explain how speakers' social goals give rise to polite speech: For example, a prominent theory of politeness (*12*) proposes that deviation from informativity increases the level of polite face-saving. But there has not yet been a formal account of the social goals in speech, thus no systematic, quantitative predictions of polite speech production have been available. On the other hand, formal theories of language have extensively 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 (*13*). 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. Here we propose a computational model that unifies formal theories of informative communication and informal theories of polite speech, and accounts for both epistemic and social goals of speakers.

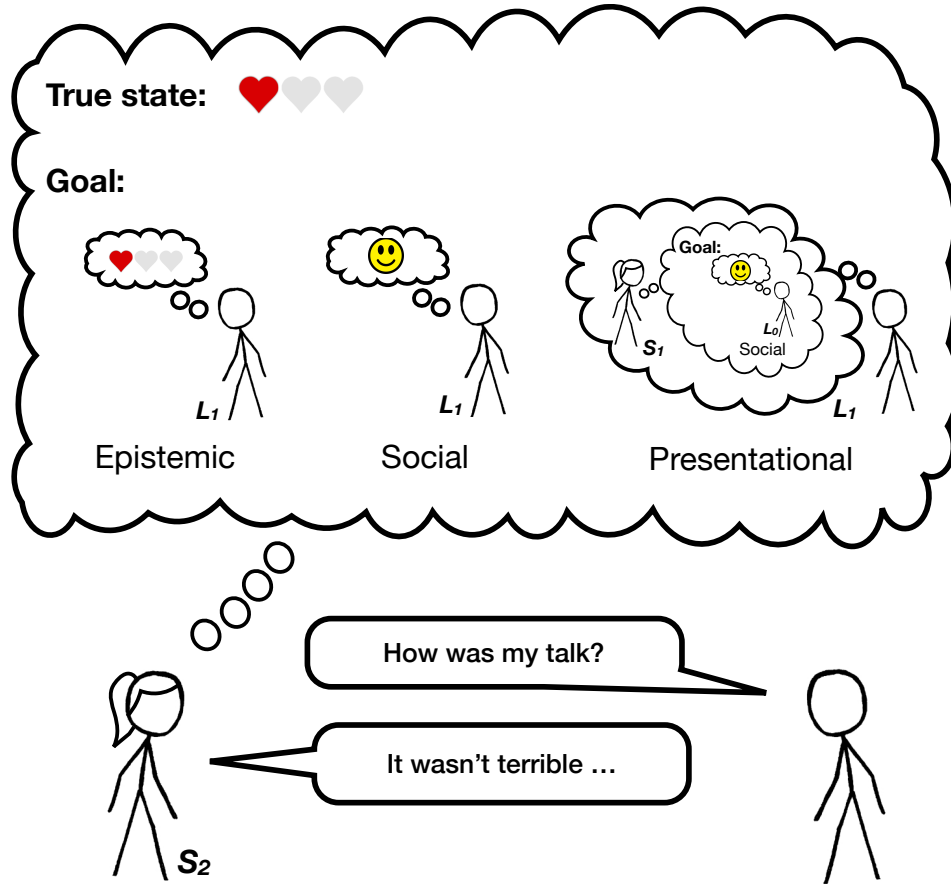Our model addresses contexts in which someone (e.g., Bob the listener) gave a performance

Figure 1: Diagram of the pRSA model: The "pragmatic speaker" observes the true state and determines her goal between three utilities, epistemic, social, and presentational (i.e. desire to appear informative, social, or both to the pragmatic listener; make the pragmatic listener infer a certain set of goals), and produces an utterance.

Figure 2: Example of a trial in the speaker production task.

of some kind, such as a presentation, and asks another person (e.g., Ann the speaker) how well he did. The speaker considers two factors to choose her utterance: the true state, or the rating truly deserved by the performance, and her goal (Figure 1). [ejy: how much detail to go into for the model?] The speaker's goal is represented as a utility function, based on which she chooses utterances approximately optimally (*14*). The speaker's utility function in the current model is made up of three 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. In our model, besides the standard epistemic utility, we add utilities to represent the speaker's social desires that lead her to produce utterances that deviate from informativity but help maintain good relationships and self-reputation: *Social utility* is the expected subjective utility of the state inferred given the utterance w, and is related to the intrinsic value of the state. We use a value function (V) to map states to subjective utility values, which captures the affective consequences for the listener of being in state s. *Presentational utility* concerns conveying a particular utility (ϕ_{S_1}) to the listener, such that the speaker appears to have a particular goal (e.g. to be socially helpful) in mind. A weight is assigned to each utility component to determine which goal is to

5

be prioritized ($\phi_{epistemic}$, ϕ_{social}, $\phi_{presentational}$ respectively). Finally, some utterances might be costlier than others. The total utility of an utterance subtracts the cost $C(w)$ from the weighted combination of the epistemic, social, and presentational utilities).

$$U(w; s; \hat{\phi}) = \phi_{epistemic} \cdot \ln(P_{L_1}(s \mid w)) + \phi_{social} \cdot \mathbb{E}_{P_{L_1}(s|w)}[V(s)]+$$

$$\phi_{presentational} \cdot \ln(P_{L_1}(s, \phi_{S_1} \mid w)) - C(w) \quad (1)$$

Based on the utility function above, the speaker (S_2) chooses utterances w approximately optimally (as per speaker optimality parameter λ_{S_2}) given the state s and his goal weights $\hat{\phi}$.

$$P_{S_2}(w \mid s, \hat{\phi}) \propto \exp(\lambda_{S_2} \cdot \mathbb{E}[U(w; s; \hat{\phi})]) \quad (2)$$

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. Bob) gave a performance of some kind, and another person (Ann) evaluated it (Figure 2). 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). 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 assumed that the goal descriptions conveyed to the participants a particular set of goal weights ($\phi_{epistemic}, \phi_{social}, \phi_{presentational}, \phi_{S_1}$) that the speaker was using. We put uninformative priors on these weights ($\beta \sim Uniform(0, 1)$) and inferred their credible values separately for each goal condition (see Supplementary Materials). 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
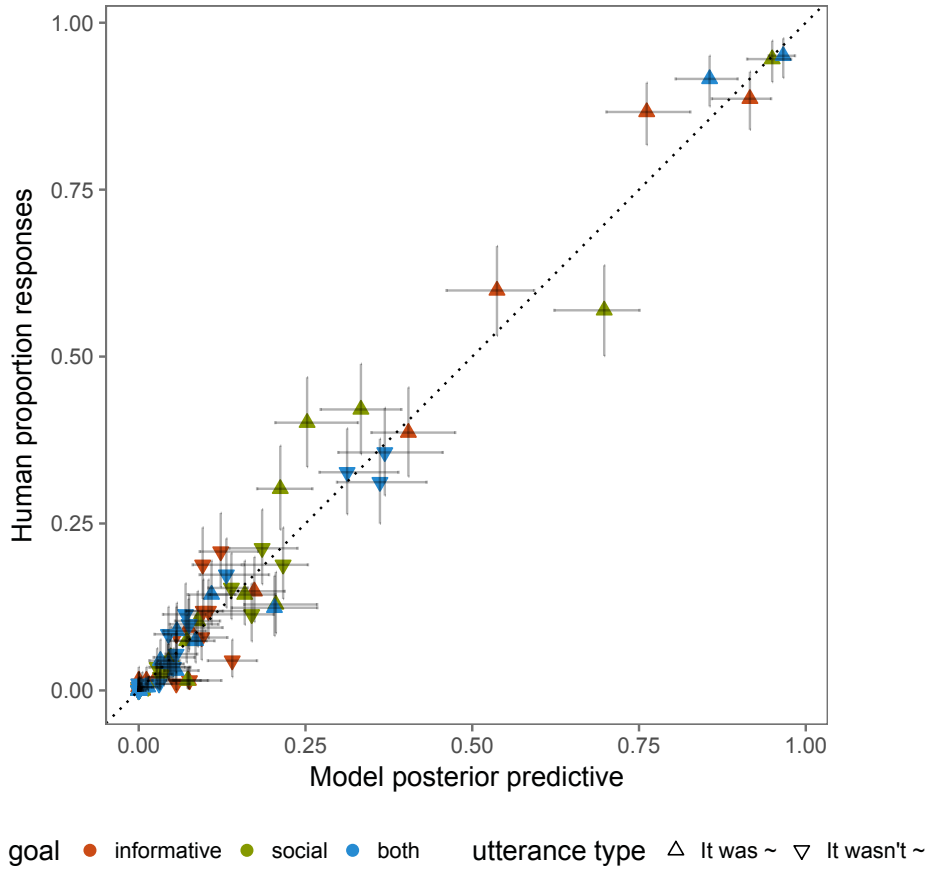
Figure 3: Full distribution of human responses vs. fitted model predictions for pragmatic speaker production. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).
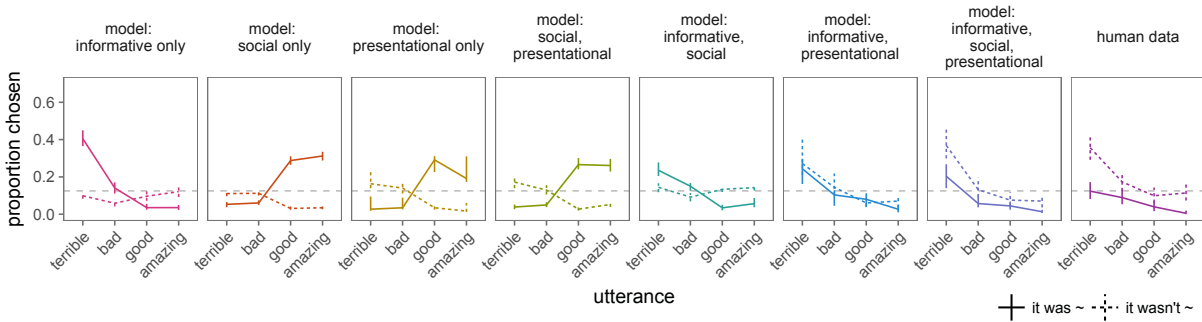

Figure 4: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart (on a scale of 0 to 3) and speaker with both goals. Gray dotted line indicates chance level at 12.5%.

7

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*, *good*, and *amazing*, thereby selecting one of eight possible utterances. 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. The posterior predictive of the model explained almost all of the variance in the production data ($r^2(96) = 0.97$; Figure 3). We compared the predictions of our model with its variants, which had different combinations of utility components: with one or two of the three utilities to be epistemic, social, and pre-sentational (Figure 4). The current model yielded the closest predictions to the human data, especially for the condition in which the speaker was described to have both goals to be in-formative and social. The model fitting for the both-goal condition suggested that the speaker had a high weight on epistemic utility, low weight on social utility, and high weight on presen-tational utility to convey high social utility, meaning that the speaker wanted to be genuinely informative while appearing to care about the listener's feelings (see Supplementary Materials for inferred weights for all goal conditions). Both the variance explained and the Bayes factor (FIXME) were the highest for the current model compared to its alternatives.

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.

Our model casts new light on understanding cognitive mechanisms that involve both in-

formation processing and interactions among agents. Many processes of human learning and inference have been extensively and fruitfully examined under computational modeling approaches similar to our own (*15–18*). [ejy: what references should go here?] However, these processes have been mostly described as purely information-driven, separate or independent of social goals of the agents involved. Our work provides the next step for addressing the issue of how social goals influence cognitive processes that involve more than one agent.

Our formal model also has important implications for building an artificial system that is more human-like. Social interactions require more than merely following simple rules (say "please", "thank you") at the right moments; they require a balance of more intricate goals to be informative and social. We successfully modeled the speaker who considers the tradeoff between communicative informativity, kindness and presentation of herself as helpful. This work takes a concrete step toward quantitative models of the nuances of human speech. And it moves us closer to courteous computation – to computers that communicate with tact.

References

1. K. Bühler, *Sprachtheorie* (Oxford, England: Fischer, 1934).

2. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 623 (1948).

3. R. Jakobson, *Style in language* (MA: MIT Press, 1960), pp. 350–377.

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

5. J. Searle, *Indirect Speech Acts* (Academic Press, 1975), vol. 3, pp. 59–82.

6. M. A. K. Halliday, *Learning How to Mean: Explorations in the Development of Language* (London: Edward Arnold, 1975).

7. S. M. Ervin-Tripp, *Journal of social issues* **23**, 78 (1967).

8. S. M. Ervin-Tripp, *Advances in experimental social psychology* **4**, 91 (1969).

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

10. G. Axia, M. R. Baroni, *Child Development* pp. 918–927 (1985).

11. T. Holtgraves, *Journal of Language and Social Psychology* **16**, 222 (1997).

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

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

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

15. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, *Science* **331**, 1279 (2011).

16. F. Xu, J. B. Tenenbaum, *Psychological review* **114**, 245 (2007).

17. E. Bonawitz, *et al.*, *Cognition* **120**, 322 (2011).

18. C. L. Baker, R. Saxe, J. B. Tenenbaum, *Cognition* **113**, 329 (2009).

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

Acknowledgments

Supplementary materials

[ejy: SM sections and figures should be restructured...]

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.

11

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
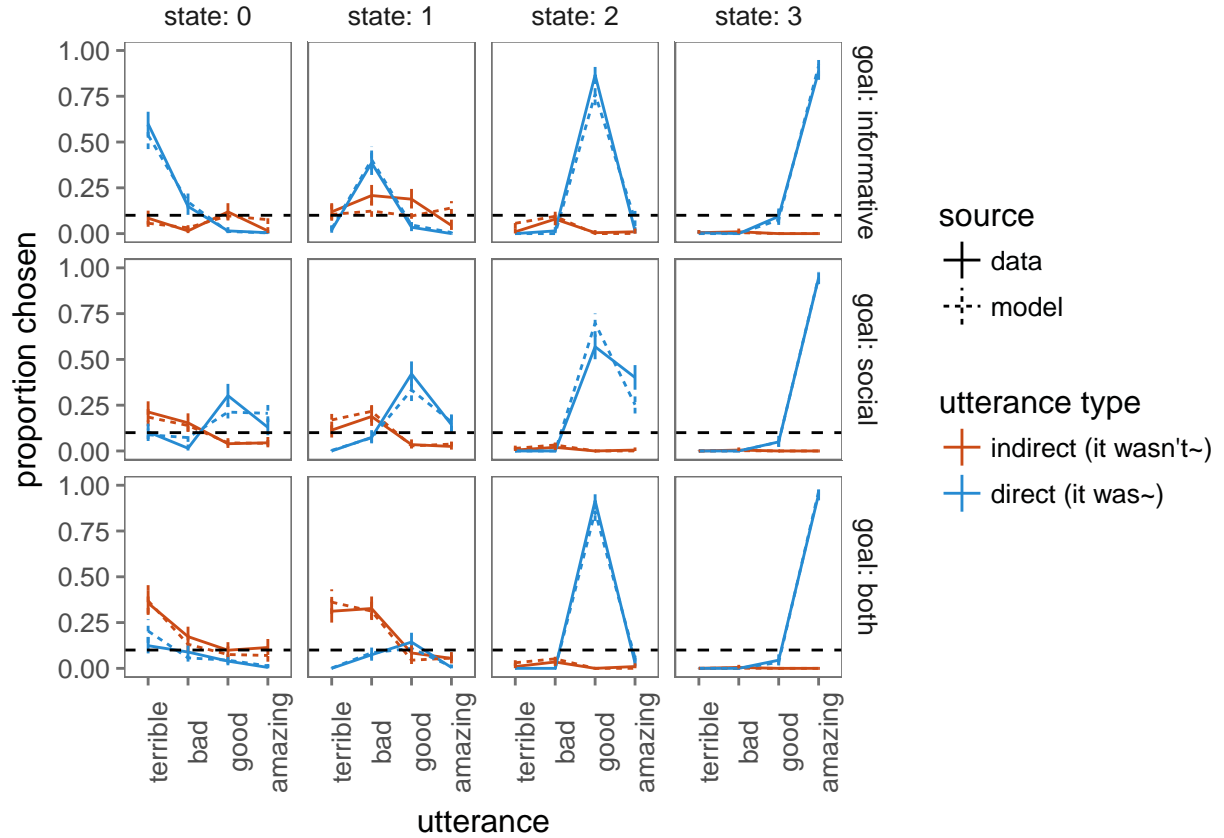
Figure 5: Experimental results (solid lines) and fitted model predictions (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.
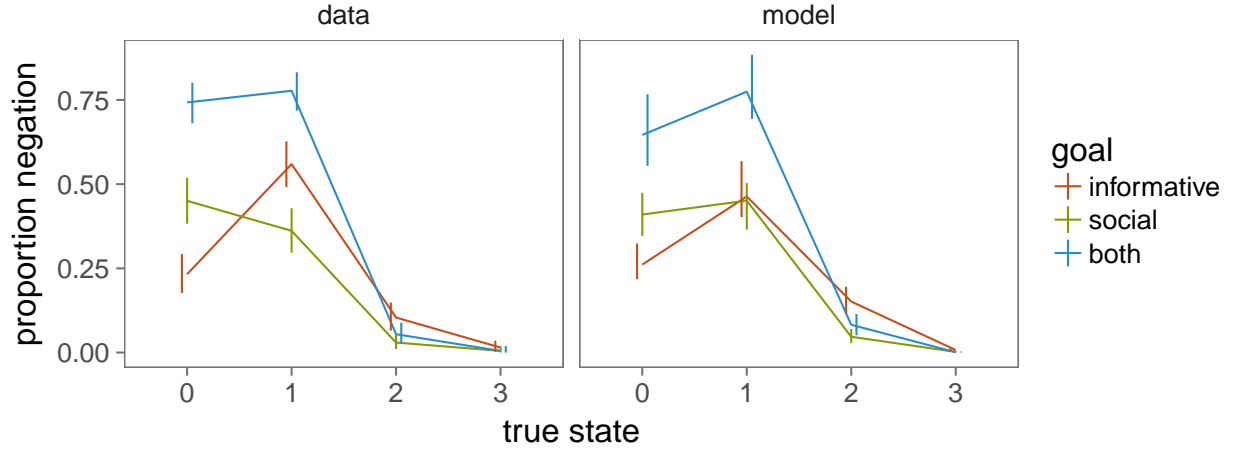
Figure 6: Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

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 (*19*). 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 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

14

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.

[ejy: FIXME: talk about the two figures. include var explained and bayes factor for each model]

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.
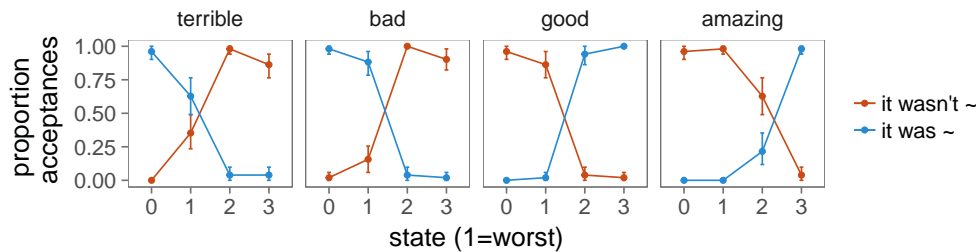
Figs. S1 to S4

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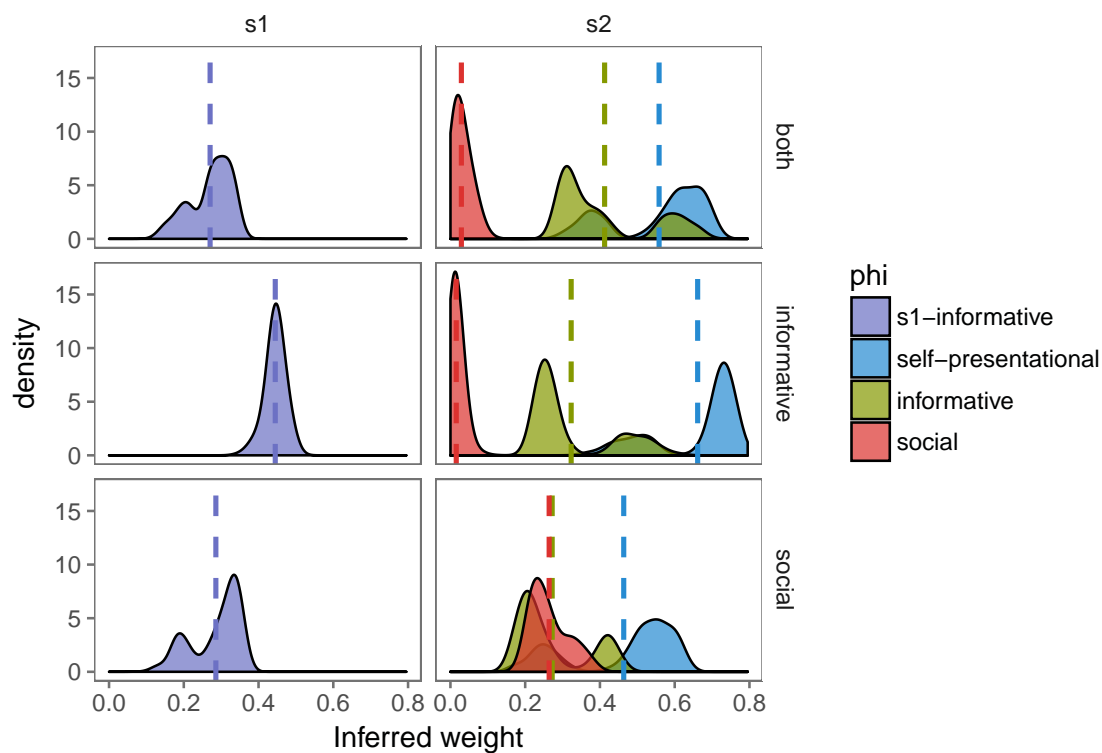
Figure S2: Inferred goal weights for the main model. Horizontal facets are different experimental conditions (trying to be X). Density plots show likely weights used in the speaker's utility function.
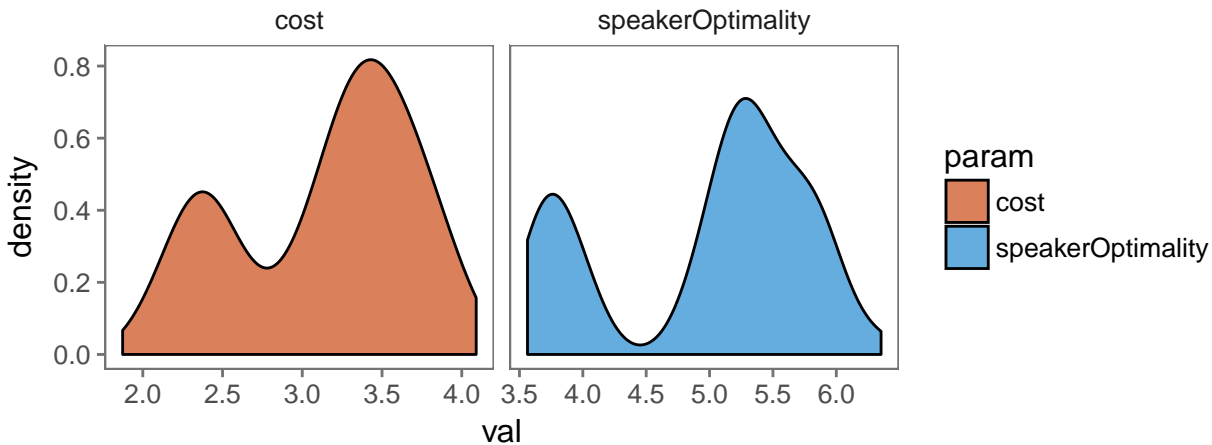

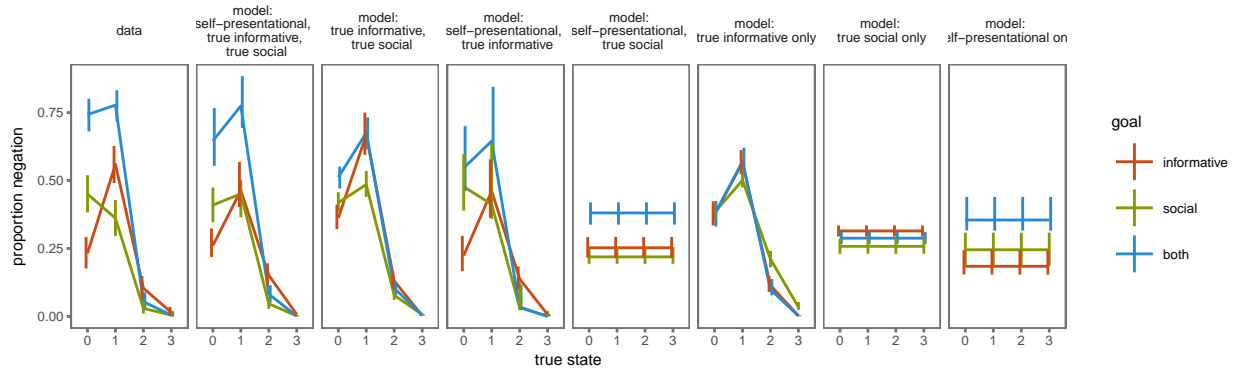Figure S3: Inferred cost and speaker optimality parameters from the main model.

Figure S4: 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).