Polite speech emerges from competing social goals

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.

Language is a remarkably efficient tool for information transfer. Yet to be polite, speakers often behave in ways that are at odds with this goal, making statements that are inefficient, imprecise, or even outright false. Why? We show that polite speech emerges from competing goals: to be informative, to be kind, and to *appear* to be both of these. We formalize this tradeoff using a probabilistic model of speakers' utterance choice, which predicts human judgments with high accuracy. This utility-theoretic approach to speech acts takes a step towards explaining the richness and subtlety of social language.

We don't always say what we're thinking. "Close the window!" could be sufficient, but instead we add "can you please…?" or "would you mind…?" Rather than tell an uncomfortable truth, we lie ("Your dress looks great!") and prevaricate ("Your poem was so appropriate to the occasion"). Such utterances are puzzling for standard views of language use, which see communication as the transfer of information from a sender to a receiver (*1–4*). On these views, transfer ought to be efficient and accurate: The speaker should choose a succinct utterance to convey what the speaker knows (*5, 6*), and the information transferred should be accurate and truthful to the extent of the speaker's knowledge. Polite speech – like the examples above – vio-

lates these basic expectations: It is inefficient, underinformative, and sometimes outright false. Why are we polite?

Theories of politeness explain deviations from optimal information transfer by assuming that speakers take into account informational *and* social concerns. These concerns have been described as polite maxims (*7*) or social norms (*8*), but the most influential account relies on the notion of *face* (*9, 10*). Under the face-based framework for polite language, interactants seek to be liked, approved, and related to (*positive face*) as well as to maintain their freedom to act (*negative face*). Though intuitively appealing, the theory does not describe when face should be prioritized over other concerns (e.g., information transfer) nor when face-saving should yield indirect (e.g., "Your cake could use a bit of salt") vs. false ("It's tasty") statements. Further, a mutually-understood notion of face introduces additional complexity: Speakers sometimes may not want to preserve the listener's face genuinely but only to be *seen* as doing so, hence appearing to be socially apt and saving their own face, which may lead to a different decision from that based on genuine desires to be kind or informative. What is needed is a precise theory of these goals and how they trade off.

To address this challenge, we develop a utility-theoretic model, quantifying tradeoffs between the different communicative goals. In our model, speakers attempt to maximize a set of competing utilities: an informational utility, derived via effective information transmission; a social utility, derived by being kind (giving the listener utility); and a self-presentational utility, derived by being perceived as valuing information transfer or kindness. Speakers then choose utterances on the basis of their expected utility.

Utilities are weighed within a Rational Speech Act (RSA) model that takes a probabilistic approach to pragmatic reasoning in language (*4, 11*). Speakers are modeled as agents who choose utterances by reasoning about their effects on a listener relative to their cost, while listeners are modeled as inferring interpretations by reasoning about speakers and their goals.

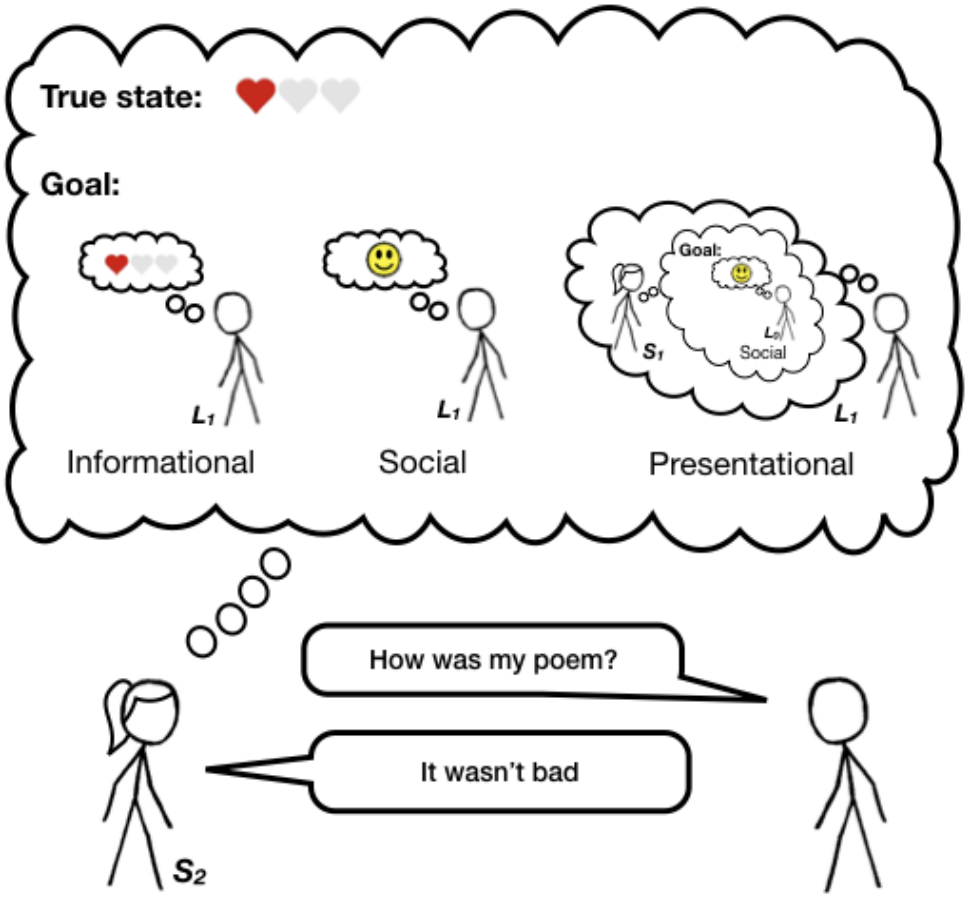

Figure 1: Diagram of the model: The pragmatic speaker observes the true state and determines her goal between three utilities (informational, social, and presentational), and produces an utterance.

RSA models have been used to understand a wide variety of complex linguistic behaviors (*12–14*), and are part of a broader class of models instantiating the idea that human social cognition can be approximated via reasoning about others as rational agents who act to maximize their subjective utility (*15–17*).

RSA models are defined recursively such that speakers reason about listeners, and vice versa. By convention, we index this recursion such that a pragmatic listener L_1 reasons about the intended meaning and goals a speaker S_1 would have had in order to produce a particular utterance. S_1 produces utterances by reasoning about a "literal listener" L_0, modeled as attending

only to the literal meanings of words (rather than their pragmatic implications), hence grounding the recursion. Our current target is a model of a polite speaker S_2, who reasons about what utterance to say to L_1 by considering the set of utilities described above (Figure 1).

We evaluate our model on its ability to predict human utterance choices in situations where polite language use is expected. Imagine Bob recites a poem he wrote and asks Ann for her opinion. Ann (S_2) produces an utterance w based on the true state of the world s (the rating Bob's recital deserved) and a set of goal weights $\hat{\phi}$ that determine how Ann prioritizes each goal. Ann's production decision is softmax, which interpolates between maximizing and probability matching (via parameter λ_{S_2}; (*18*)):

$$P_{S_2}(w|s,\hat{\phi}) \propto \exp(\lambda_{S_2} \cdot U_{total}(w;s;\hat{\phi};\phi_{S_1}))$$

We consider three goals that the speaker weighs to arrive at a polite utterance: informational, social, and presentational. The total utility of an utterance is the weighted combination of the three utilities minus the utterance cost $C(w)$, which captures the general pressure towards economy in speech (*19*):

$$U_{total}(w;s;\hat{\phi};\phi_{S_1}) = \phi_{inf} \cdot U_{inf}(w;s) + \phi_{soc} \cdot U_{soc}(w) + \phi_{pres} \cdot U_{pres}(w;\phi_{S_1}) - C(w)$$

where $\hat{\phi}$ indicates the relative importance of the speaker's three goals (ϕ_{inf}, ϕ_{soc}, and ϕ_{pres}). Utterances with negation (e.g., "not terrible") are longer and thus assumed to be slightly costlier than their unnegated equivalents.

First, the *informational utility* (U_{inf}), represents the speaker's desire to be epistemically helpful, and captures how well the utterance w leads the literal listener (L_0) to infer the true state of the world s: $U_{inf}(w;s) = \ln(P_{L_1}(s|w))$. Second, the *social utility* (U_{soc}) is the expected subjective utility $V(s)$ of the state implied to the listener by the utterance: $U_{soc}(w) = \mathbb{E}_{P_{L_1}(s|w)}[V(s)]$. In our experimental domain, states are explicit ratings, so we use a positive

linear value function V to capture the idea that listeners want to hear that they are in a good state of the world (e.g., Bob prefers that his poem was good).

If listeners try to infer the goals that a speaker is entertaining (e.g., social vs. informational), speakers may choose utterances in order to convey that they had certain goals in mind. The third and the most novel aspect of our model, *presentational utility* (U_{pres}), captures the extent to which the speaker appears to the listener to have a particular goal in mind. Formally,

$$U_{pres}(w; \phi_{S_1}) = \ln(P_{L_1}(\phi_{S_1} \mid w)) = \ln \int_s P_{L_1}(s, \phi_{S_1} \mid w)$$

To define this term, the speaker has a weighting of informational vs. social goals to convey (ϕ_{S_1}) and must consider that the listener L_1 reasons about the speaker's goal-weights together with the true state of the world:

$$P_{L_1}(s, \phi_{S_1}|w) \propto P_{S_1}(w|s, \phi_{S_1}) \cdot P(s) \cdot P(\phi_{S_1})$$

The presentational utility U_{pres} is of a higher order than the other utilities: it can be defined only for S_2, but not S_1.

Intuitively, if Bob's poem was good, Ann's utilities align: By saying "[Your poem] was amazing," Ann is being truthful and kind while also (accurately) appearing to be both. If Bob's performance was poor, however, Ann is in a bind: she could be kind and say "It was great", but at the cost of conveying the wrong information to Bob if he believes her. If he does not, he might infer Ann is "just being nice", but is uninformative. Alternatively, Ann could tell the truth ("It was bad"), but then Bob might think Ann didn't care about him. In this dilemma, our model predicts that indirect speech – like "It wasn't bad" – is the best solution. Her statement is sufficiently vague to leave open the possibility that the poem was good, but her avoidance of the simpler and less costly "It was good" provides both an inference that the performance was mediocre and a signal that she cares about Bob's feelings.

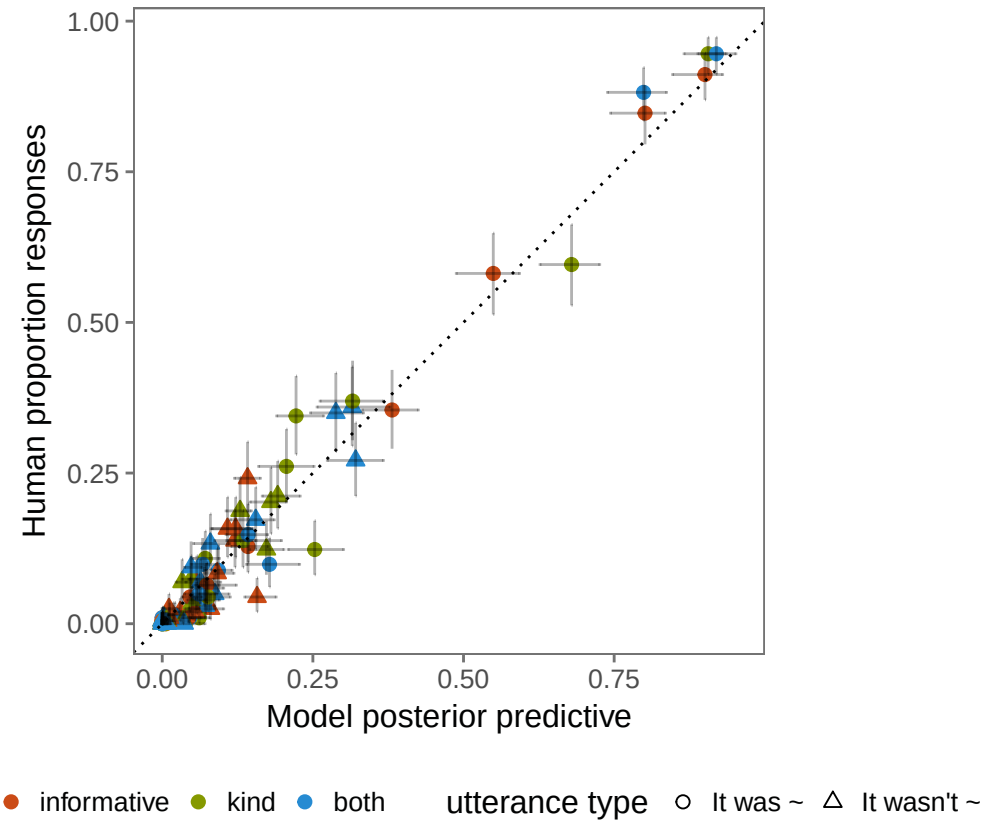

Figure 2: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

We tested our model in a pre-registered online experiment (N = 202). Participants read scenarios with information about the speaker's feelings toward some performance or product (e.g., poem recital; *true state*), on a scale from zero to three hearts. The speaker's *goals* varied across trials: to be *informative* ("give accurate and informative feedback"); to be *kind* ("make the listener feel good"); or to be *both* informative and kind simultaneously. We hypothesized that each of the three goals represented a tradeoff between the three utilities in our model (see Supplementary Materials). In a single trial, each scenario was followed by a question asking for the speaker's most likely utterance. Participants selected one of eight possible utterances, by choosing between *It was* vs. *It wasn't* and then among *terrible*, *bad*, *good*, and *amazing*.

We hypothesized that speakers describing bad states (e.g., Bob's performance deserved 0 hearts) while trying to be both informative and kind would produce more indirect, negative utterances (e.g., "It wasn't bad"). This prediction was confirmed: a Bayesian mixed-effects model predicting negation as a function of true state and goal yielded an interaction such that a speaker with both goals to be informative and kind produced more negation in worse states compared to a speaker with only the goal to be informative (M = -1.33, [-1.69, -0.98]) and goal to be kind (M = -0.50, [-0.92, -0.07]).

Next, to connect the behavioral data to our model, we inferred the parameters of the RSA model (e.g., the speaker's utility weights in each goal condition; see Supplementary Materials) via a Bayesian data analysis (*20*). To approximate the semantics of the words as interpreted by the literal listener L_0, we obtained literal meaning judgments from an independent group of participants (N=51). Predictions from the full polite speaker model showed a very strong fit to participants' utterance choices ($r^2(96) = 0.97$; Figure 2).

We also compared the predictions of our model to its variants containing subsets of the three utilities in the full model. Both the variance explained and the marginal likelihood of the observed data were the highest for the full model (Table 1). Only the full model captured the participants' preference for negation in the condition in which the speaker had both goals to be informative and kind about truly bad states, as hypothesized (Figure 3). All three utilities – informational, social, and presentational – were required to fully explain participants' utterance choices. The utility weights inferred for the full model (Table S2) provide additional insight into how polite language use operates: our condition manipulation altered the balance between these weights, but all utilities played a role in all conditions.

To precisely estimate choice behavior, our experiment abstracted away from natural interactions in a number of ways. Real speakers have access to a potentially infinite range of utterances to manage the tradeoffs in our experiment ("It's hard to write a good poem", "That metaphor

Figure 3: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost), 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%.

Table 1: Comparison of variance explained for each model variant and log Bayes Factors quantifying evidence in favor of the full model, in comparison to each of the alternatives.

Model	Variance explained	log BF
model: informational only	0.83	274.89
model: social only	0.22	885.52
model: presentational only	0.23	873.83
model: social, presentational	0.23	864.00
model: informational, social	0.92	25.06
model: informational, presentational	0.96	11.14
model: informational, social, presentational	0.97	1.00

was so relatable!"). Under our framework, each utterance will have strengths and weaknesses relative to the speaker's goals. Computation in an unbounded model presents technical challenges (*11*) (perhaps paralleling the difficulty human speakers feel in finding the right thing to say in a difficult situation).

For a socially-conscious speaker, managing listeners' inferences is a fundamental task. Inspired by the theory of politeness as face management (*9*), our model takes a step towards understanding it. By considering utility-driven inferences in a social context (*21, 22*), our approach here could give insights into a wide range of social behaviors beyond speech. And by experimenting with different utility weights and value functions, our model could provide a framework for understanding systematic cross-cultural differences in what counts as polite.

Politeness is only one of the ways that language use deviates from pure information transfer. When we flirt, insult, boast, and empathize, we also balance being informative with goals to affect others' feelings and present particular views of ourselves. Our work shows how social and self-presentational motives can be integrated with other concerns more generally, opening up the possibility for a broader theory of social language. Further, a formal account of politeness moves us closer to courteous computation – to computers that can 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. M. C. Frank, N. D. Goodman, *Science* **336**, 998 (2012).

5. H. P. Grice, *Syntax and Semantics*, P. Cole, J. L. Morgan, eds. (Academic Press, 1975), vol. 3, pp. 41–58.

6. J. Searle, *Syntax and Semantics*, P. Cole, J. L. Morgan, eds. (Academic Press, 1975), vol. 3, pp. 59–82.

7. G. Leech, *Principles of pragmatics* (London, New York: Longman Group Ltd., 1983).

8. S. Ide, *Multilingua-journal of cross-cultural and interlanguage communication* **8**, 223 (1989).

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

10. E. Goffman, *Interaction ritual: essays on face-to-face interaction* (Aldine, 1967).

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

12. D. Lassiter, N. D. Goodman, *Synthese* **194**, 3801 (2017).

13. J. T. Kao, J. Y. Wu, L. Bergen, N. D. Goodman, *Proceedings of the National Academy of Sciences* **111**, 12002 (2014).

14. J. T. Kao, N. D. Goodman, *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (2015).

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

16. J. Jara-Ettinger, H. Gweon, L. E. Schulz, J. B. Tenenbaum, *Trends in cognitive sciences* **20**, 589 (2016).

17. S. Liu, T. D. Ullman, J. B. Tenenbaum, E. S. Spelke, *Science* **358**, 1038 (2017).

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

19. N. D. Goodman, D. Lassiter, *Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought* (Wiley-Blackwell, 2015).

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

21. C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, *Nature Human Behaviour* **1**, 0064 (2017).

22. K. J. Hamlin, T. D. Ullman, J. B. Tenenbaum, N. D. Goodman, C. L. Baker, *Developmental science* **16**, 209 (2013).

Acknowledgments

All authors designed research and wrote the paper; E.J.Y. and M.H.T. performed research and analyzed data. Our model, preregistration of hypotheses, procedure, data, and analyses are available at `https://github.com/ejyoon/polite_speaker`. The authors declare no conflict of interest. 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 our main experiment. 51 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. We used thirteen different context items in which a speaker 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; see Figure S1 for an example of the heart scale). 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 four possible words: *terrible*, *bad*, *good*, and *amazing*, giving rise to eight 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 speaker production 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 (i.e., a standard Beta-Binomial model). Meanings of the words as judged by participants were as one would expect (Figure S2).

Speaker production task

202 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. As in the literal semantic task 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 (see

Fig. 2). 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; Figure S1). Each participant read twelve scenarios, depicting every possible combination of the three goals and four 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

Data analysis

We used R (Version 3.4.3; R Core Team, 2017) and the R-packages *BayesFactor* (Version 0.9.12.2; Morey & Rouder, 2015), *bindrcpp* (Version 0.2; Mller, 2017a), *binom* (Version 1.1.1; Dorai-Raj, 2014), *brms* (Version 2.0.1; Brkner, 2017), *coda* (Version 0.19.1; Plummer, Best, Cowles, & Vines, 2006), *directlabels* (Version 2017.3.31; Hocking, 2017), *dplyr* (Version 0.7.4; Wickham, Francois, Henry, & Mller, 2017), *forcats* (Version 0.2.0; Wickham, 2017a), *ggplot2* (Version 2.2.1; Wickham, 2009), *ggthemes* (Version 3.4.0; Arnold, 2017), *gridExtra* (Version 2.3; Auguie, 2017), *here* (Version 0.1; Mller, 2017b), *jsonlite* (Version 1.5; Ooms, 2014), *langcog* (Version 0.1.9001; Braginsky, Yurovsky, & Frank, n.d.), *lme4* (Version 1.1.15; Bates, Mchler, Bolker, & Walker, 2015), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.12; Bates & Maechler, 2017), *papaja* (Version 0.1.0.9655; Aust & Barth, 2017),

Table S1: Predictor mean estimates with standard deviation and 95% credible interval information for a Bayesian linear mixed-effects model predicting negation production based on true state and speaker goal (with both-goal as the reference level).

Predictor	Mean	SD	95% CI-Lower	95% CI-Upper
Intercept	0.88	0.13	0.63	1.12
True state	2.18	0.17	1.86	2.53
Goal: Informative	0.47	0.17	0.14	0.80
Goal: Social	0.97	0.25	0.51	1.49
True state * Informative	-1.33	0.18	-1.69	-0.98
True state * Social	-0.50	0.22	-0.92	-0.07

purrr (Version 0.2.4; Henry & Wickham, 2017), *RColorBrewer* (Version 1.1.2; Neuwirth, 2014), *Rcpp* (Eddelbuettel & Balamuta, 2017; Version 0.12.14; Eddelbuettel & Franois, 2011), *readr* (Version 1.1.1; Wickham, Hester, & Francois, 2017), *rwebppl* (Version 0.1.97; Braginsky, Tessler, & Hawkins, n.d.), *stringr* (Version 1.2.0; Wickham, 2017b), *tibble* (Version 1.3.4; Mller & Wickham, 2017), *tidyr* (Version 0.7.2; Wickham & Henry, 2017), and *tidyverse* (Version 1.2.1; Wickham, 2017c) for all our analyses.

Full statistics on human data

We used Bayesian linear mixed-effects models (`brms` package in R; Brkner, 2017) using crossed random effects of true state and goal with maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013).

Polite RSA model fitting and inferred parameters

In the speaker production task, participants were told the speakers' intentions (e.g., wanted to make Bob feel good). We assume that the intention descriptions conveyed some mixture of weights ϕ_{epi}, ϕ_{soc}, ϕ_{pres}, and ϕ_{S_1} that the speaker was using. We put uninformative priors on the unnormalized mixture weights ($\phi \sim Uniform(0, 1)$) separately for each goal condition

Table S2: Inferred phi parameters from all model variants with more than one utility.

Model	goal	ϕ_{inf}	ϕ_{soc}	ϕ_{pres}	ϕ_{S_1}
informational, social, presentational	both	0.36	0.11	0.54	0.36
informational, social, presentational	informative	0.36	0.02	0.62	0.49
informational, social, presentational	kind	0.25	0.31	0.44	0.37
informational, presentational	both	0.64	-	0.36	0.17
informational, presentational	informative	0.77	-	0.23	0.33
informational, presentational	kind	0.66	-	0.34	0.04
informational, social	both	0.54	0.46	-	-
informational, social	informative	0.82	0.18	-	-
informational, social	kind	0.39	0.61	-	-
social, presentational	both	-	0.38	0.62	0.55
social, presentational	informative	-	0.35	0.65	0.75
social, presentational	kind	-	0.48	0.52	0.66

Table S3: Inferred negation cost and speaker optimality parameters for all model variants.

Model	Cost of negation	Speaker optimality
informational only	1.58	8.58
informational, presentational	1.89	2.93
informational, social	1.11	3.07
informational, social, presentational	2.64	4.47
presentational only	2.58	9.58
social only	1.73	7.23
social, presentational	2.49	5.29

("wanted to be X"; *informative*, *kind*, or *both*). In addition, the full model has two global parameters: the speaker's soft-max parameter λ_{S_2} and soft-max paramater of the hypothetical speaker that the pragmatic listener reasons about λ_{S_1}. λ_{S_1} was 1, and λ_{S_2} was inferred from the data: We put a prior that was consistent with those used for similar models in this model class: $\lambda_{S_2} \sim Uniform(0, 20)$. Finally, we incorporate the literal semantics data into the RSA model by maintaining uncertainty about the semantic weight of utterance w for state s, for each of the states and utterances, and assuming a Beta-Binomial linking function between these weights and the literal semantics data (see *Literal semantics task* above). We infer the posterior distribution over all of the model parameters and generate model predictions based on this posterior distribution using Bayesian data analysis (Lee & Wagenmakers, 2014). We ran 4 MCMC chains for 80,000 iterations, discarding the first 40,000 for burnin. The inferred values of weight mixtures for each model variant (with different ϕ components) and other parameters are shown in Table S2 and Table S3, respectively.

Figs. S1 to S5

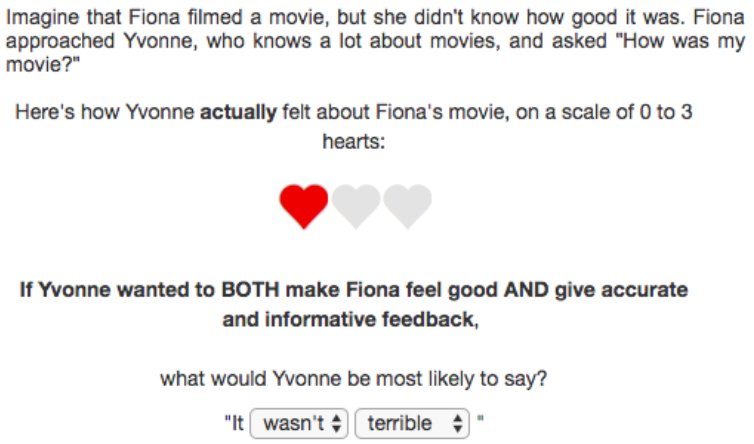

Figure S1: Example of a trial in the speaker production task.

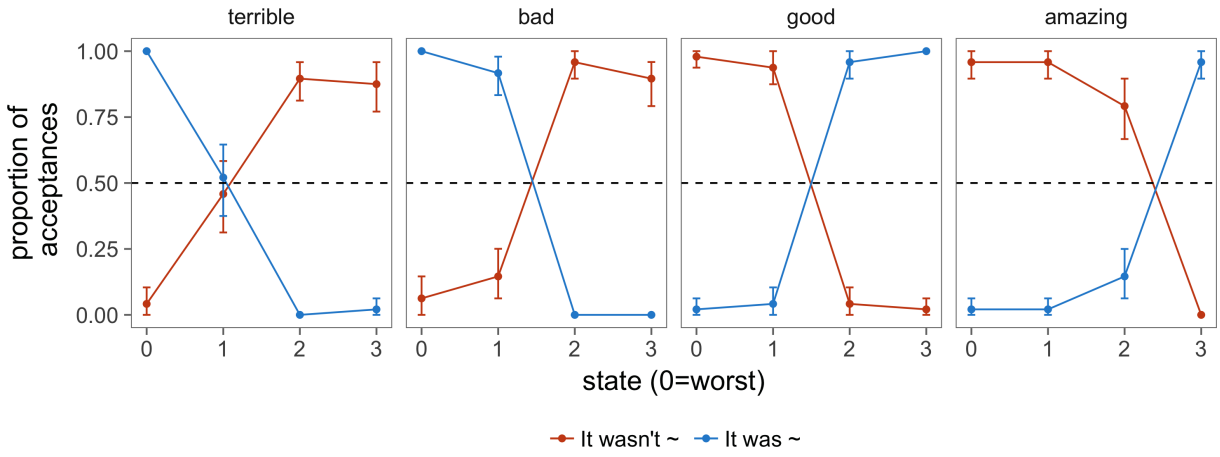

Figure S2: Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

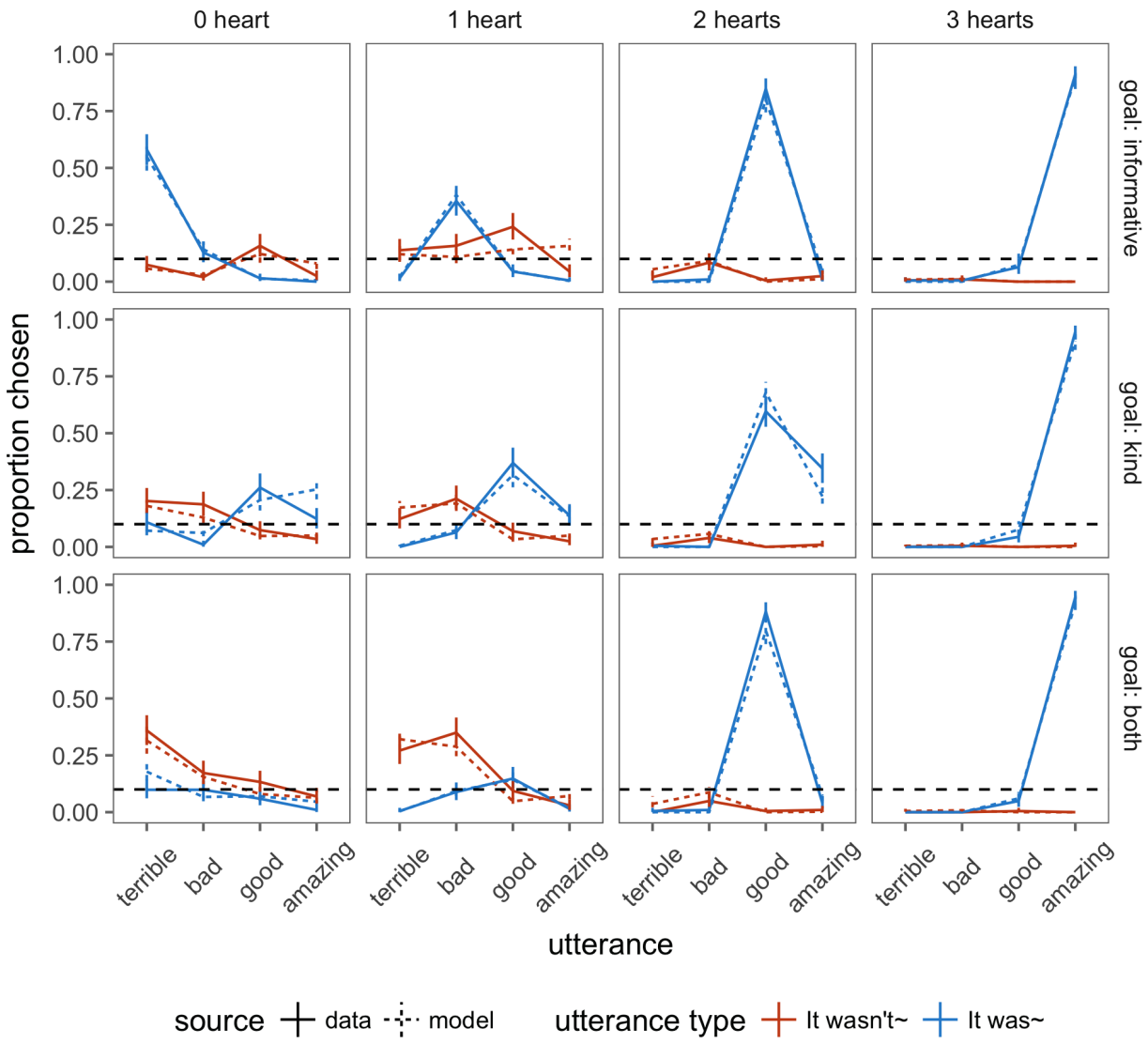

Figure S3: Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type direct vs. indirect in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.

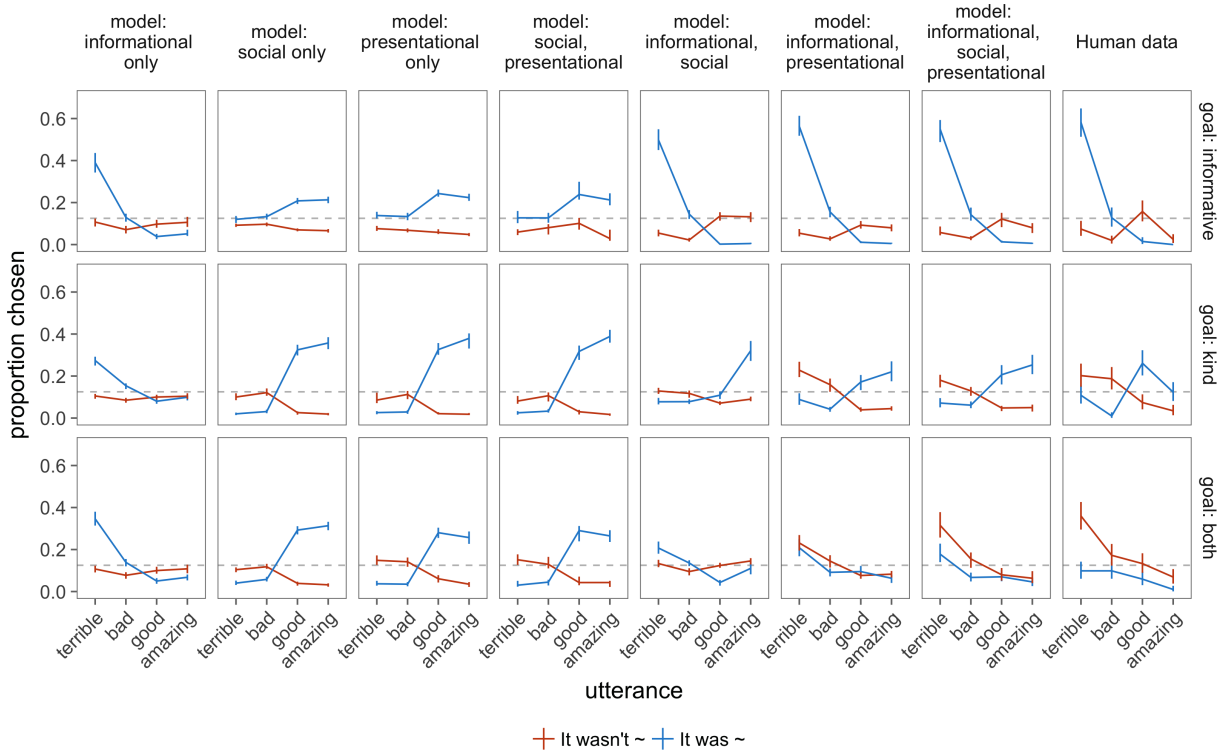

Figure S4: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with a goal to be informative (top), kind (middle), or both (bottom). Gray dotted line indicates chance level at 12.5%.

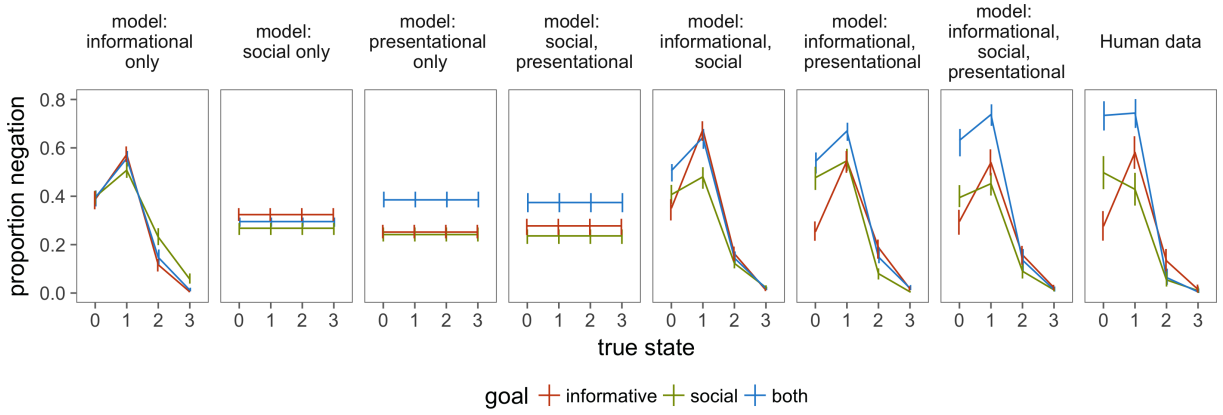

Figure S5: Fitted model predictions (left) and experimental results (rightmost) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).