Polite speech emerges from competing pressures to be (and look) informative and kind

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.

Being polite, or conveying information in a false or indirect manner in deference to someone else's feelings, seemingly contradicts an important goal of a cooperative speaker: information transfer. In this work, we show that polite speech emerges from a set of competing goals: to be informative, to be kind, and to *appear* helpful or kind. We formalize this tradeoff between speaker's competing goals using a utility-theoretic 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 language use.

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*). Under information-based views, the 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 – violates these basic expectations: It is inefficient, underinformative, and sometimes outright false. So why are we polite?

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

To address these challenges, we develop a utility-theoretic model to quantify tradeoffs between different goals that a polite speaker may have. In our model, speakers attempt to maximize a set of competing utilities: an informational utility, derived via classical, effective information transmission; a social utility, derived by being kind and saving the listener's face; and a self-presentational utility, derived by appearing in a particular way to save the speaker's own face. Speakers then choose utterances on the basis of their expected utility.

The 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

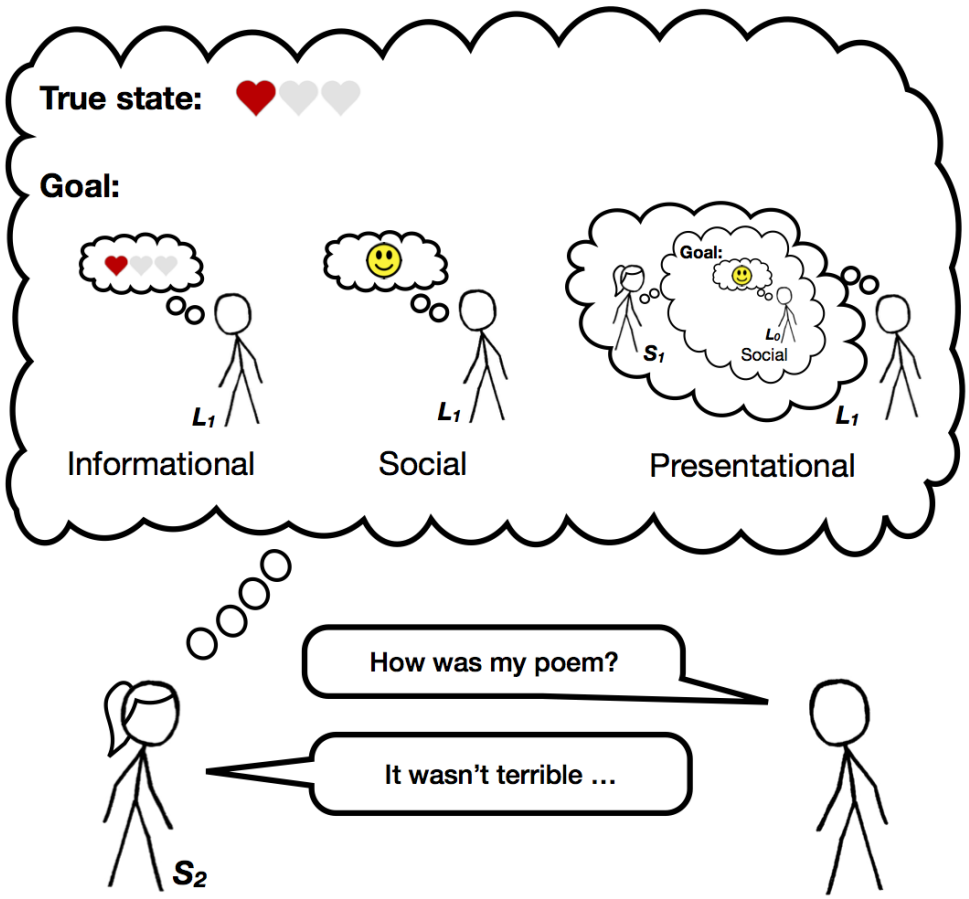

Figure 1: Diagram of the model: The pragmatic speaker observes the true state and determines her goal between three utilities (informational, social, and presentational), and produces an utterance.

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. This class of models has been effective in understanding a wide variety of complex linguistic behaviors (*12–14*), and, more broadly, the idea that human social cognition can be approximated via reasoning about others as rational agents who act to maximize their subjective utility (*15*) has found a wide variety of empirical support in studies with both adults and children (*16, 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. The target of this work is a model of a polite speaker S_2: S_2 reasons about what utterance to say to L_1 by considering the set of utilities described above (Figure 1).

We test our model by 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 (i.e., the rating truly deserved by Bob's recital) 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 (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 could consider 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:

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

In this work, utterances with negation (e.g., "not terrible") are assumed to be slightly costlier than their equivalents with no negation.

The first utility term is a standard *informational utility* (U_{inf}), which represents the speaker's desire to be epistemically helpful. The informational utility 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, we define the *social utility* (U_{soc}) as 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: Thus, 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 would prefer that his poem was good).

If listeners are uncertain and 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 component 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 (e.g., to be kind). Formally,

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

.

In order to define this term, the speaker has a particular set of goal-weights 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:

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

.

This presentational utility U_{pres} is of a higher order than the other utilities in that it can only be defined for a speaker thinking about a listener who evaluates a speaker (i.e., it can be defined for S_2, but not S_1).

Intuitively, if Bob's performance was good, Ann's utilities align towards producing a positive utterance. By saying "[Your poem] was amazing," Ann is simultaneously being truthful,

kind, and appearing both truthful and kind. If Bob's performance was poor, however, Ann is in a bind: Ann could be kind and say "It was great", but at the cost of conveying the wrong information to Bob if he believes her to be truthful. Worse yet, Bob could infer Ann is "just being nice", but is otherwise an unhelpful speaker. Alternatively, Ann could say the truth ("It was bad"), but then Bob would think Ann didn't care about him. What is a socially-aware speaker to do? Our model predicts that indirect speech – like "It wasn't bad" – helps navigate Ann's dilemma. It conveys some true information (e.g., literally it wasn't the worst it could have been) while being sufficiently open-ended to spare Bob's feelings. Further, by incurring the slightly higher cost involved in producing negation, Ann provides a signal that she has Bob's feelings in mind: Why else not say the simpler "It was good"?

We tested our model in an 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 will represent 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 most likely utterance by Ann. Participants selected one of eight possible utterances, by choosing between *It was* vs. *It wasn't* and then among *terrible*, *bad*, *good*, and *amazing*.

Our primary behavioral hypothesis was that speakers describing bad states (e.g., Bob's performance deserved 0 hearts) with goals to be both informative and kind would produce more indirect, negative utterances (e.g., "It wasn't terrible"). Such indirect speech acts serve to save the listener's face while also conveying a vague estimate of the true state. 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 pro-

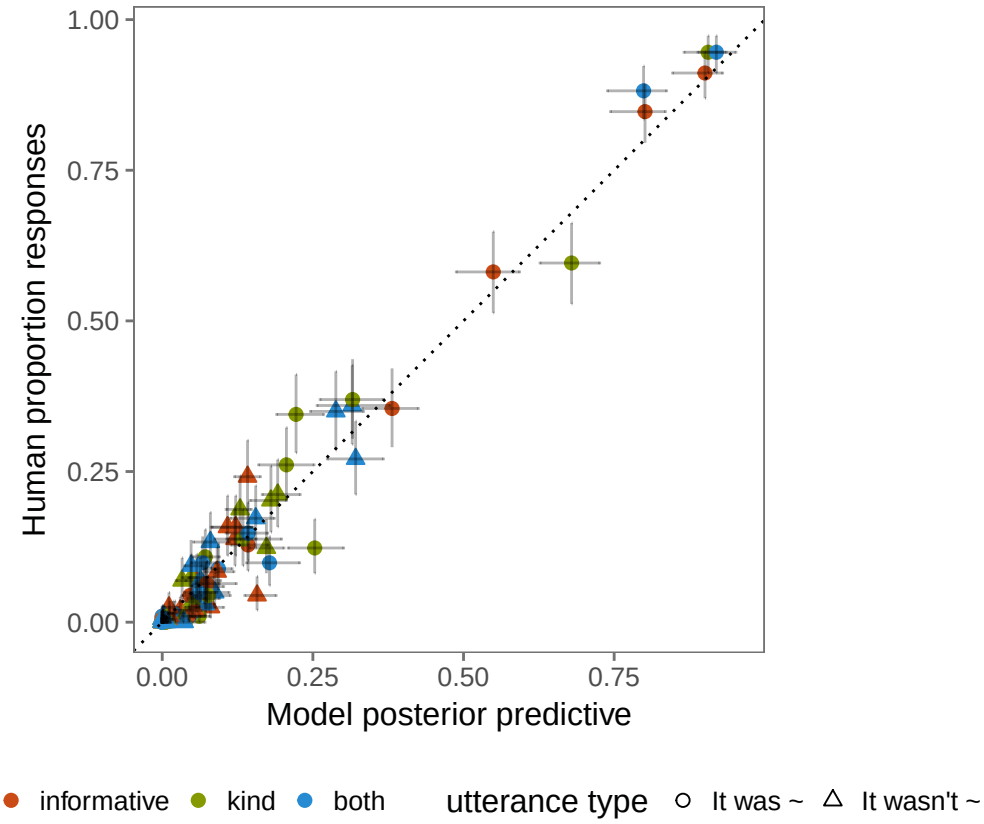

Figure 2: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

Figure 3: 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%.

duced 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]).

To connect these behavioral data more directly to our model, we built a Bayesian statistical model (*19*) to infer the parameters of the RSA model (e.g., the speaker's utility weights in each goal condition; see Supplementary Materials). We separately obtained literal meaning judgments (independent sample; N=51) about the utterances to approximate the semantics of the words as interpreted by the literal listener L_0. 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). In particular, 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). Thus, all three utilities – informational, social, and presentational – were required to fully explain participants' utterance choices.

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: informative 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: informative, social	0.92	25.06
model: informative, presentational	0.96	11.14
model: informative, social, presentational	0.97	1.00

The actual parameter values inferred for the full model (Table S2) provide additional insight into how polite language use operates. *Being nice* requires equal weights on all three utilities, indicating that Gricean informativity needs to be part of language use even when it is explicitly not the goal. *Being informative* pushes the weight on kindness close to zero, but the weight on *appearing kind* stays high, indicating that speakers are expected to manage their own face even when they are not considering others'. *Nice and informative* speakers emphasize honesty slightly more than kindness. In all cases, however, the presentational utilities have greatest weight, indicating perhaps that appearing honest and kind is more important than actually being so!

Managing listeners' inferences is a fundamental task for a socially conscious speaker. Following Brown and Levinson (1987), cross-cultural differences in politeness could be a product of different weightings within the same utility structure. It is also possible, however, that culture affects the value function V that maps states of the world onto subjective values for the listener (e.g., the mapping from states to utilities may be more complex than we have considered). Our formal modeling approach with systematic behavior measurements provides an avenue towards understanding the vast range of politeness practices found across languages.

To precisely estimate choice behavior, our experiment abstracted away from natural interactions in a number of ways. Real-life Anns have access to a potentially infinite range of utterances to manage the same tradeoff ("It's hard to write a good poem", "That metaphor in the second stanza was so relatable!"). Under our framework, each utterance will have strengths and weaknesses relative to the speaker's goals, though computation in an unbounded model presents technical challenges (*11*).

Politeness is just one of the ways that language use deviates from pure information transfer. When we flirt, insult, boast, and empathize, we balance information transmission with the goal to affect others' feelings or present particular views of ourselves. Similar utility structures to that used in our model could give insights into these speech acts and other behaviors that can be modeled as utility-driven inference in a social context (*20, 21*) where agents need to take into account concerns about both self and others. This work provides a key theoretical advance beyond previous formal accounts of language use (e.g., (*13*)) by showing how informational cooperativity interacts with social and self-presentational goals in language use, thus opening up new communicative behaviors to formal modeling. 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. 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. M. D. Lee, E. J. Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course* (Cambridge Univ. Press, 2014).

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

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

Acknowledgments

Funding: 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. *Author contributions*: E.J.Y., M.H.T., N.D.G., and M.C.F. designed research; E.J.Y. and M.H.T. performed research; E.J.Y. and M.H.T. analyzed data; and E.J.Y., M.H.T., N.D.G., and M.C.F. wrote the paper. *Competing interests*: The authors declare no conflict of interest. *Data and materials availability*: Our pre-registered model, hypothesis, and procedure, as well as all of our data and analyses are available at `https://github.com/ejyoon/polite_speaker`.

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), *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* (Ver-

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

sion 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 ("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

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}
informative, social, presentational	both	0.36	0.11	0.54	0.36
informative, social, presentational	informative	0.36	0.02	0.62	0.49
informative, social, presentational	social	0.25	0.31	0.44	0.37
informative, presentational	both	0.64	NA	0.36	0.17
informative, presentational	informative	0.77	NA	0.23	0.33
informative, presentational	social	0.66	NA	0.34	0.04
informative, social	both	0.54	0.46	NA	NA
informative, social	informative	0.82	0.18	NA	NA
informative, social	social	0.39	0.61	NA	NA
social, presentational	both	NA	0.38	0.62	0.55
social, presentational	informative	NA	0.35	0.65	0.75
social, presentational	social	NA	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
informative only	1.58	8.58
informative, presentational	1.89	2.93
informative, social	1.11	3.07
informative, social, presentational	2.64	4.47
presentational only	2.58	9.58
social only	1.73	7.23
social, presentational	2.49	5.29

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

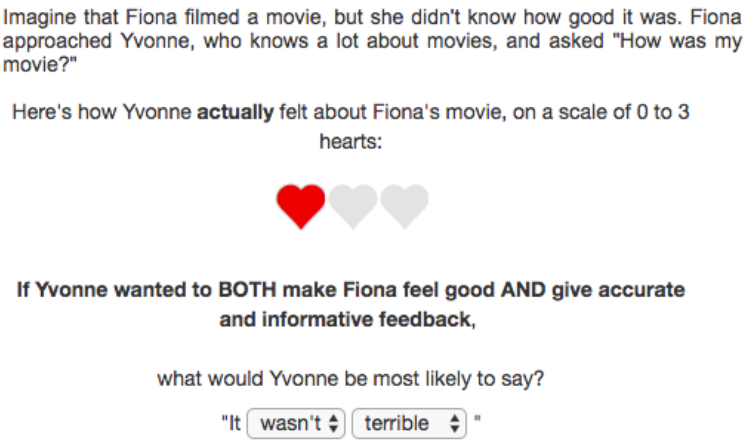

Figure S1: Example of a trial in the speaker production task.

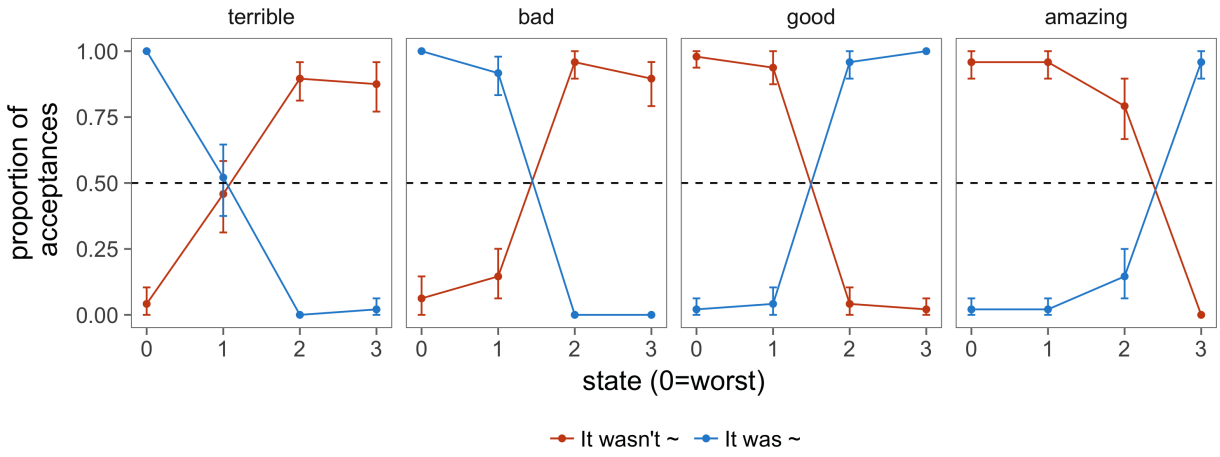

Figure S2: Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

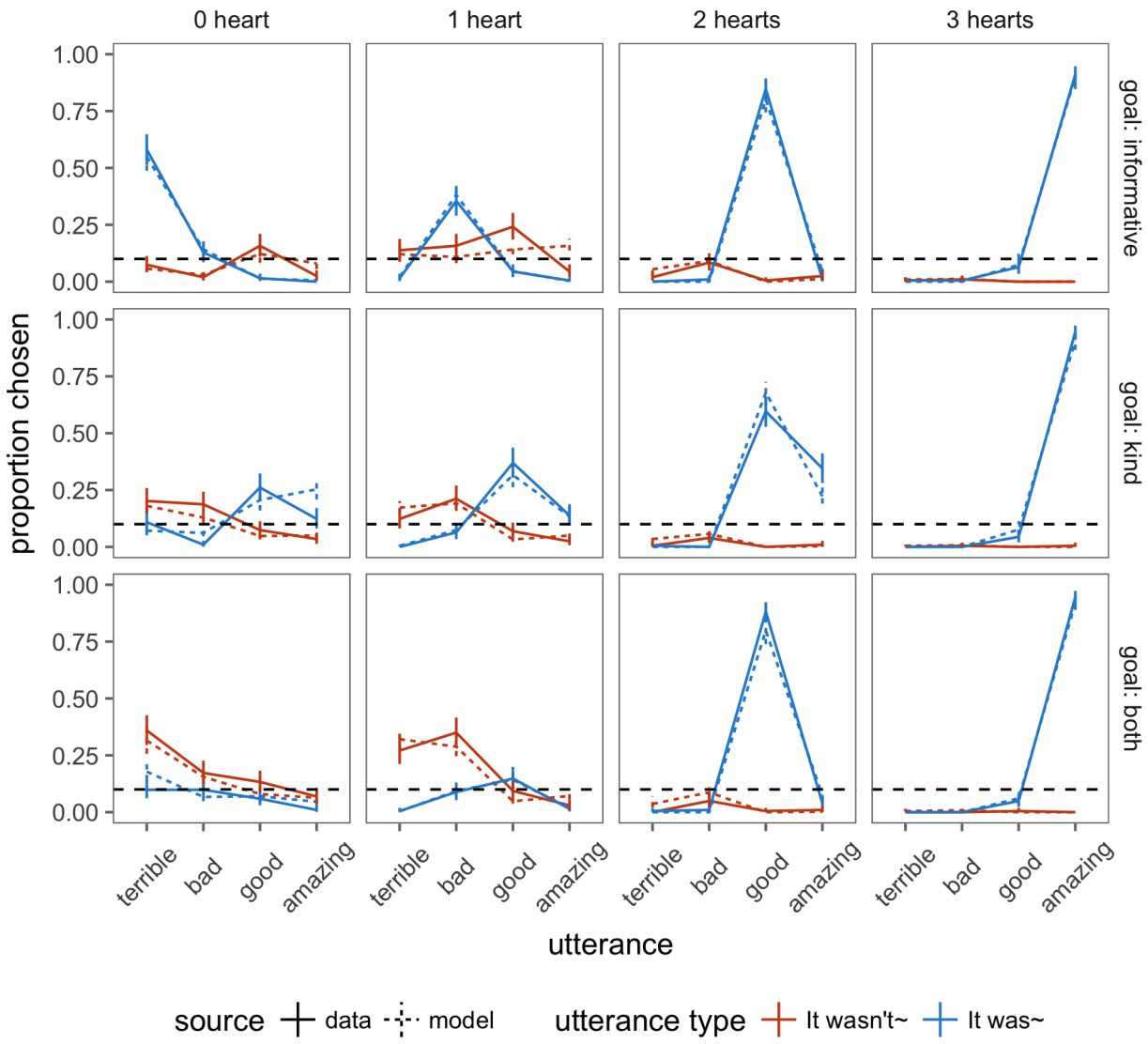

Figure S3: Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type direct vs. indirect in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows).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

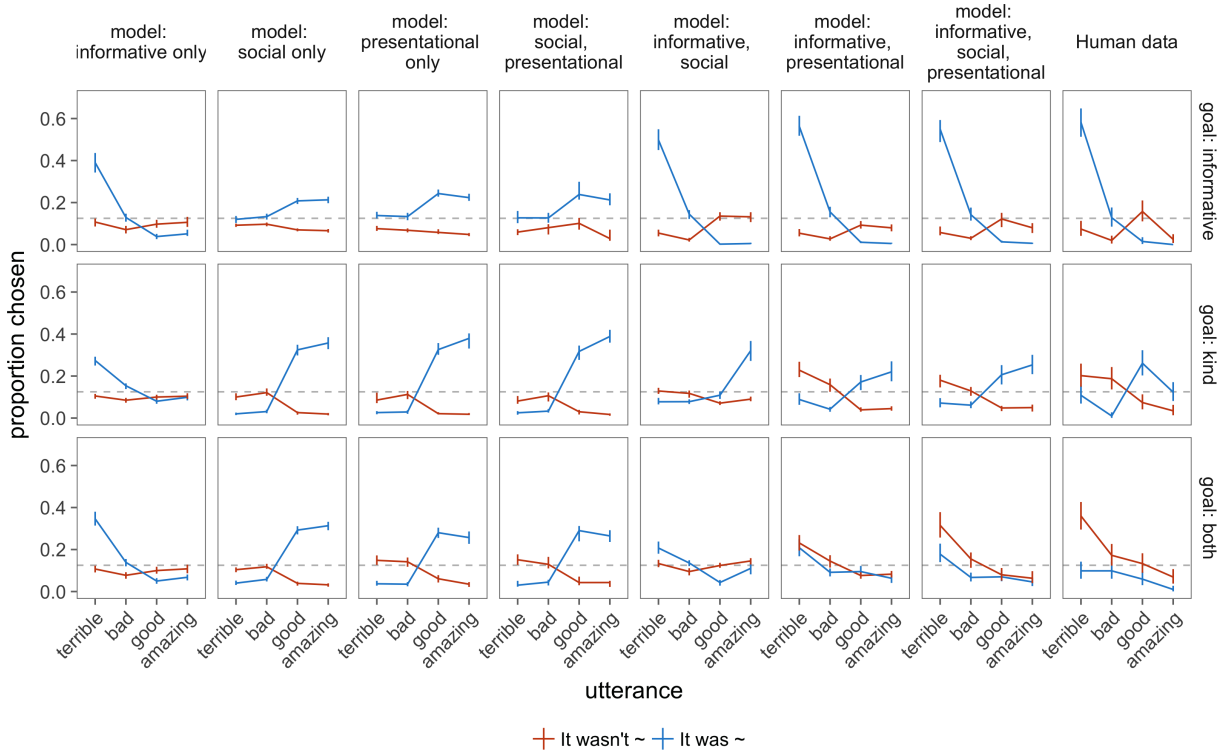

Figure S4: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with a goal to be informative (top), kind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

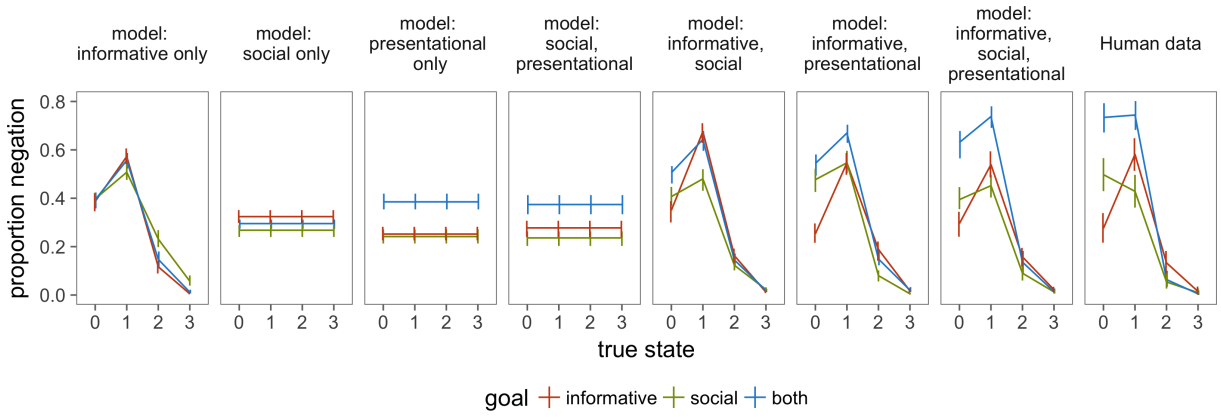

Figure S5: Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).