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 provide positive value to others, and to be self-presentational and *appear* helpful. 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 speech.

We don't always say what we're thinking. Although "close the window!" could be sufficient, we say "can you please…?" or "would you mind…?" Rather than telling 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 from which the listener can recover their intended meaning (*5, 6*), and the information transferred should be accurate and truthful to the extent that the speaker knows or believes to be true. Polite speech – like the examples above – violates these basic expectations about the nature of communication: It is typically inefficient and underinformative, and sometimes even 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. These concerns are sometimes expressed as sets of polite maxims (*7*) or social norms (*8*), but the most influential account of politeness relies on the notion of "face" to motivate deviations (*9, 10*). On this theory, speakers seek to be liked, approved, and related to ("positive face") as well as maintain both their and the listeners' freedom to act ("negative face"). Both inefficient indirect speech and untruthful lies in communication are then the result of speakers' strategic choices relative to possible face threats.

The face-based framework for polite language use provides an intuitive and appealing explanation of many types of polite speech, but theorizing at level of the abstract notions like "face" does not make quantitative predictions in any individual circumstance nor constrain how an artificial agent should go about making polite requests, conveying negative evaluations, or delivering bad news. It is not obvious how to quantify a face threat in a given situation (e.g., how much of the listener's positive face will be damaged by hearing "your poem was terrible"), or how social and informational motivations will trade off in the mind of a speaker (given that the poem recital was terrible, should the speaker say that the listener's poem was "okay," "not bad," or "marvelous"?). Further, the recursive nature of reasoning about face has not been formally addressed: 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.

To address these challenges, we develop a utility-theoretic quantitative model for understanding polite speech, in a unified framework to quantify tradeoffs between different goals that a 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 providing positive affect to others, thereby saving the listener's face; and a self-presentational utility, derived by appearing in a particular way to other agents and saving the speaker's own face. Speakers then can choose between different utterances on the basis of their expected utility. The lie that a poem "was good" provides social utility by making its writer feel good, but does not inform about the true state of the world.

Informational, social, and self-presentational utilities are weighed within a Rational Speech Act (RSA) model. RSA models take 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. This class of models has been effective in understanding a wide variety of complex linguistic behaviors, including vagueness (*12*), hyperbole (*13*), and irony (*14*), among others. More broadly, RSA models provide a instantiation for language of the idea that human social cognition can be approximated via reasoning about others as rational agents who act to maximize their subjective utility (*15*), a hypothesis which has found support in a wide variety of work with both adults and children (*16, 17*).

RSA models are defined recursively such that speakers reason about listeners, and vice versa. By convention the level of this recursion is indexed such that a pragmatic listener L_1 reasons about what intended meaning and goals would have led a speaker S_1 to produce a particular utterance. Then S_1 reasons about a "literal listener" L_0, modeled as attending only to the literal meanings of words (rather than their pragmatic implications), and hence grounds

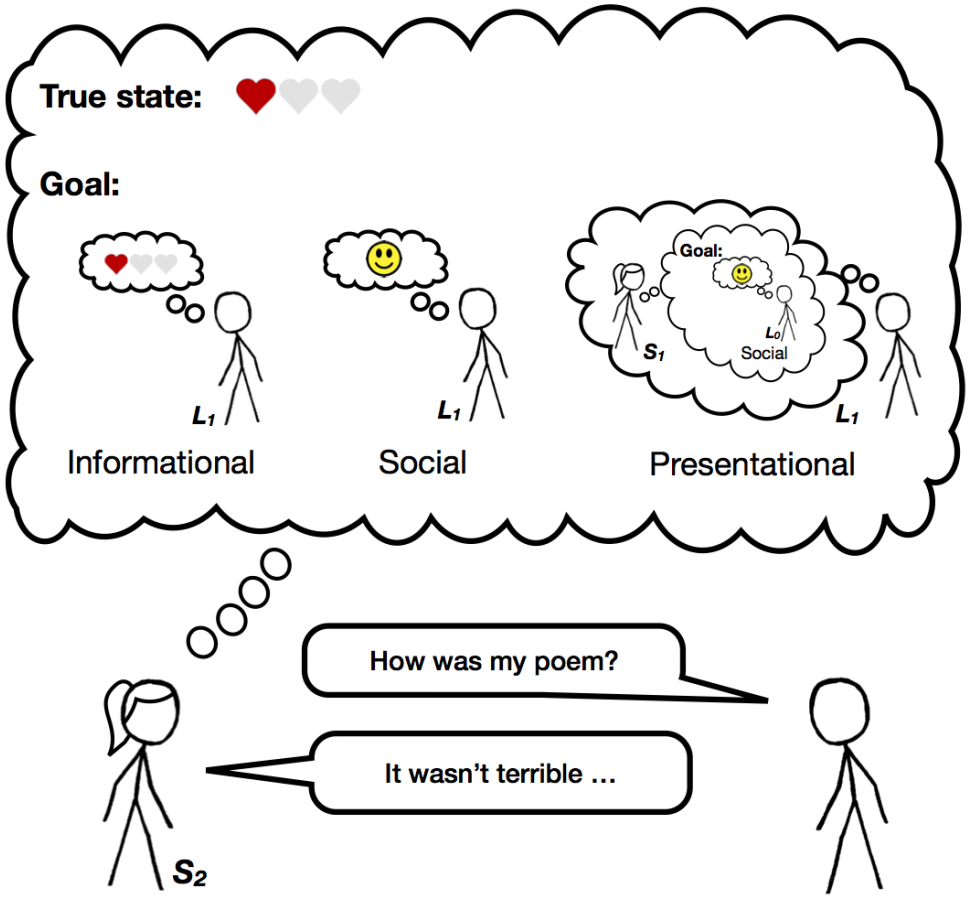

Figure 1: Diagram of the model: The pragmatic speaker observes the true state and determines her goal between three utilities (informational, social, and presentational), and produces an utterance.

the recursion. The target of our current 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: namely, whether an utterance results in L_1 gaining information, feeling positively, or judging S_2 to be either informative or kind (Figure 1).

We evaluate our model by predicting human utterance choices in situations where polite language use is expected. Imagine Bob recited his poem and is ignorant of the quality of his poem recital; he asks Ann how well he did. Ann (the pragmatic speaker 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 $\hat{\phi}$, each of which determines how much speaker Ann prioritizes a particular goal compared to other possible goals. Speaker Ann then chooses utterances depending on their expected utility, specifically as a softmax which interpolates between maximizing and probability matching (via the parameter λ_{S_2}; (*18*):

$$P_{S_2}(w|s,\hat{\phi}) \propto \exp(\lambda_{S_2} \cdot \mathbb{E}[U_{total}(w;s;\hat{\phi})])$$

.

What goals must the speaker consider to arrive at a polite utterance? We consider three utilities: informational, social, and presentational. The total utility of an utterance is the weighted combination of the three utilities minus the cost of the utterance $C(w)$, approximated by the length of the utterance:

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

.

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 the amount of information

a literal listener (L_0) would still not know about the world state after hearing the speaker's utterance: $U_{inf}(w) = \ln(P_{L_1}(s|w))$.

For aspects of the world with affective consequences for the listener (e.g., Bob and his poem recital), we define the *social utility* (U_{soc}) as the value $V(s)$, or expected subjective utility, to the listener of the state inferred given the utterance: $U_{soc}(w) = \mathbb{E}_{P_{L_1}(s|w)}[V(s)]$. This value captures 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 recital was good). We use a positive linear value function (V) to map states to subjective values: better ratings are more positively valued.

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 component, *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). The speaker gains presentational utility when her listener believes she has certain goals – that she is trying to be informative or kind. Formally,

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

.

The speaker considers the beliefs of listener L_1, who hears an utterance and jointly infers both the speaker's utilities and the true state of the world:

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

.

This presentational utility, which is the most novel aspect of our model, is higher-order in that it can only be defined for a speaker thinking about a listener who evaluates a speaker. (That is, it can be defined for S_2, but not S_1.)

Finally, utterances that are more complex incur a greater cost, $C(w)$ – capturing the general pressure towards economy in speech. In our work, utterances with negation (e.g., "not terrible") are assumed to be slightly more costly than their equivalents with no negation (given by a parameter inferred from data; see Supplemental Materials).

Intuitively, if Bob's performance was good, Ann's utilities align to lead her to say something positive. By saying "[Your poem] was amazing," Ann is simultaneously being truthful, kind, and appearing both and 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 she does so at the cost of conveying the wrong information to Bob (e.g., if he mistakenly infers Ann's goal to be truthful and hence believes that his recital was actually good). Worse yet, Bob could infer that Ann is "just being nice," inferring her goal to be social, and discount her comment as uninformative. Alternatively, she 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 was 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 another word, Bob could reason that Ann had reasons for not saying a simpler alternative like "It was good" and and that she must have taken his feelings into account in her utterance.

We made a direct test of our model by instantiating the example above in an online experiment (N = 202; see our pre-registered model, hypothesis, and procedure at `https://github.com/ejyoon/polite_speaker`). Participants read scenarios in which we provided information about the speaker's (Ann's, in our example) feelings toward some performance or product (e.g., poem recital; *true state*), which were shown on a scale from zero to three hearts (e.g., one out of three hearts). We manipulated the speaker's *goal* across trials: to be *informative* ("give accurate and informative feedback"); to be *social* ("to make the listener

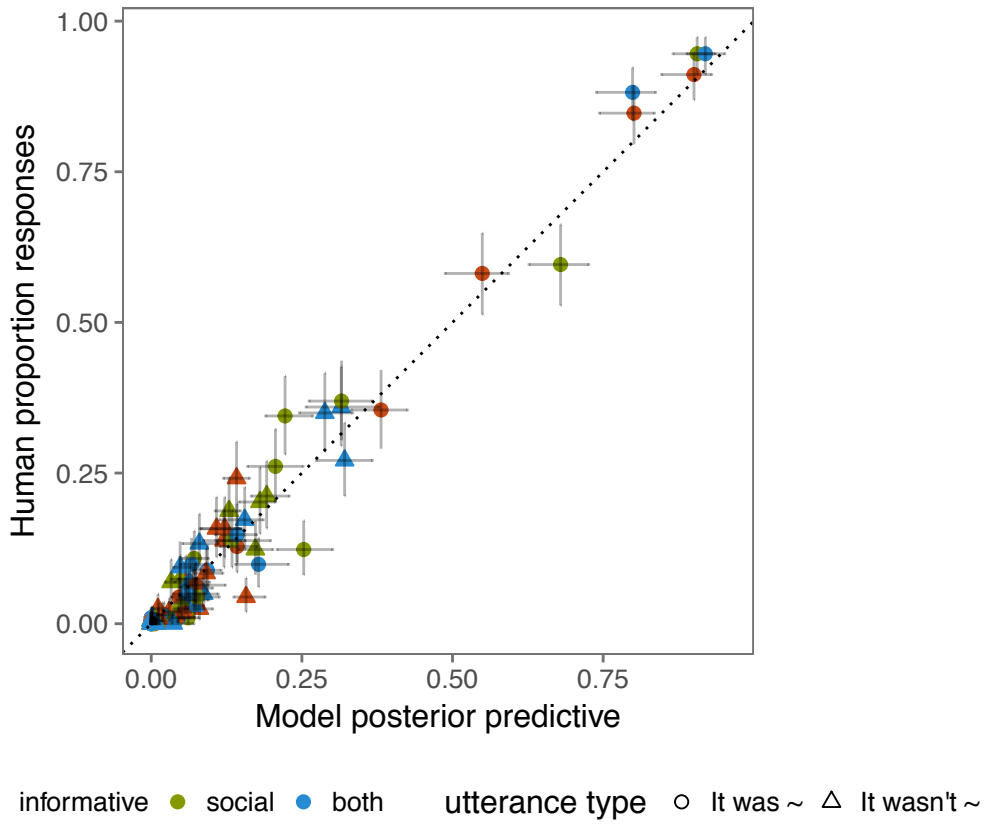

Figure 2: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

feel good"); or to be *both* informative and social at the same time. We hypothesized that each of the three goals will represent a tradeoff between the three utilities in our model described above (their inferred values are available in the Supplementary Materials). In a single trial, each scenario was followed by a question that asked 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 who found themselves describing bad states (e.g., Bob's performance was bad) and who had goals to be both informative and social would produce more indirect, negative utterances (e.g., "It wasn't terrible"). Such indirect

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%.

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 informational and social goals produced more negation in worse states compared to a speaker with only the informational goal (M = -1.33, [-1.69, -0.98]) and social goal (M = -0.50, [-0.92, -0.07]).

To connect these behavioral data more directly to our model, we separately obtained literal meaning judgments about the utterances and incorporated them into our model's predictions. Using an independent sample of N=51 participants, we measured judgments of how well different utterances apply to each of the levels on the heart scale (e.g., to what extent is "terrible" true of 2 out of 3 hearts?). These measurements were used in the Bayesian data analysis to approximate the semantics of the words as interpreted by the literal listener agent L_0 (see Supplementary Materials for literal semantic results). Then we used a Bayesian analytic technique (*19*) to infer the parameters of the model (e.g., the speaker's utility weights in each goal condition; see Supplementary Materials for inferred parameters).

Predictions from the full polite speaker model showed a very strong fit to participants' ut-

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

terance choices ($r^2(96) = 0.97$; Figure 2). We also compared the predictions of our model with model variants containing different subsets of the three utilities in the full model (Figure 3; see Supplemental Materials: Model Comparison). Both the variance explained and the marginal likelihood of the observed data were the highest for the full model (see Table 1 and Figure 3). 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 social about truly bad states, as hypothesized. The full model was superior to: the model with social and presentational utilities, which predicted outright false statements ("It was good"); the model with informational and social utilities, which predicted truthful statements "It was terrible" and "It wasn't amazing" (that is semantically true when the poem was terrible); and to the model with informative and presentational utilities, which predicted that the speaker was equally likely to be truthful ("It was terrible") or presentational ("It wasn't terrible"). Thus, all three utilities – informational, social, and presentational – were required to fully explain participants' choices.

To better measure choice behavior, our experiment abstracted away from natural interactions in a number of ways. Real-life Anns will have access to a potentially infinite range of utterances to manage the same tradeoff ("It wasn't my cup of tea", "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 (see (*11*)).

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., pointing out bad states could be considered prosocial in certain cultural contexts; more generally, 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. Further, 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. A similar utility structure to the one we employed here could give insights into these behaviors as well.

The formalization of the presentational utility is especially meaningful in that it begins to precisely define self-oriented motivations behind polite speech and other related behaviors. To the best of our knowledge, previous theories of politeness have not explained how the motivations of the other- vs. self-oriented concerns are related or how they trade off to inform the speaker's utterance choices. In our current model, the self-oriented concern stems from an other-oriented concern, as the speaker wants to appear to care about the other person's face or access to knowledge. The model then makes precise predictions about how the speaker considering both of these concerns will choose her utterances. This work then can be extended to not only other speech acts, but also a wide range of 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.

In sum, 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. 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

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}
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

("wanted to X"; *social*, *informative*, 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: λ_{S_2} ~ $Uniform(0, 20)$. 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

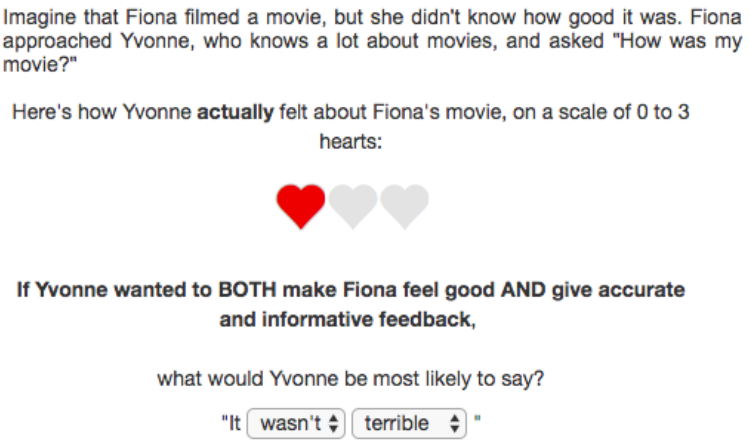

Figure S1: Example of a trial in the speaker production task.

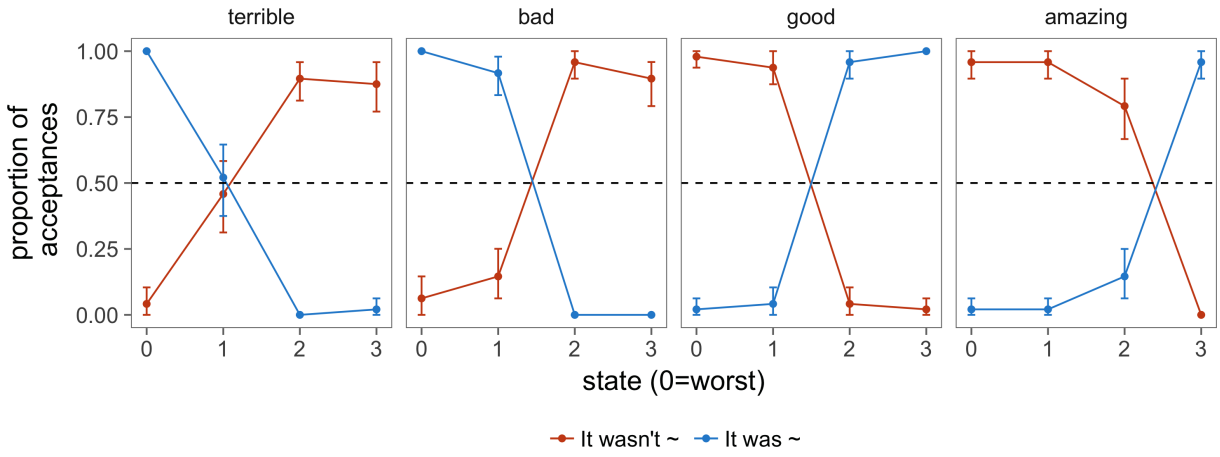

Figure S2: Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

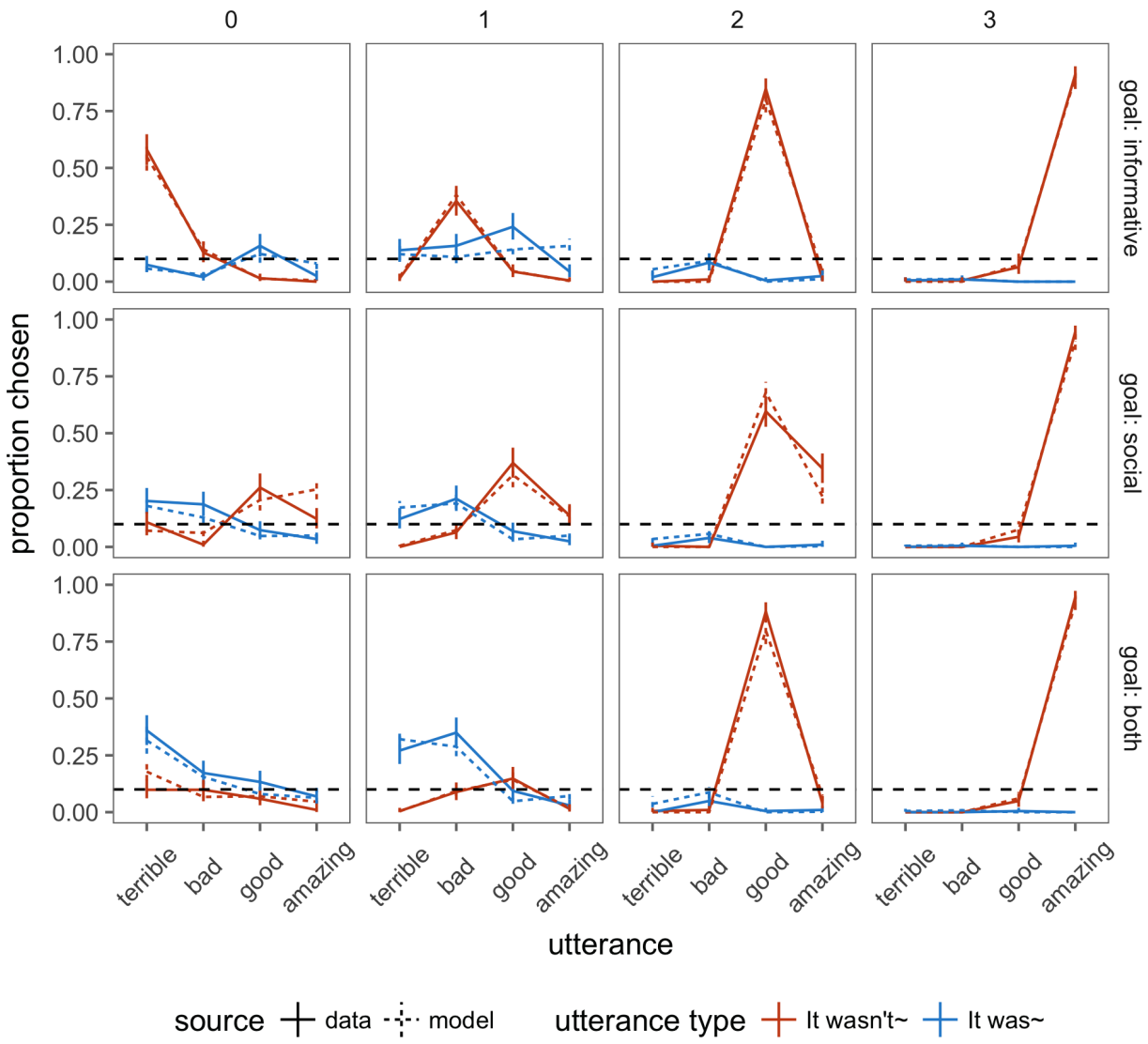

Figure S3: Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type direct vs. indirect in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows).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

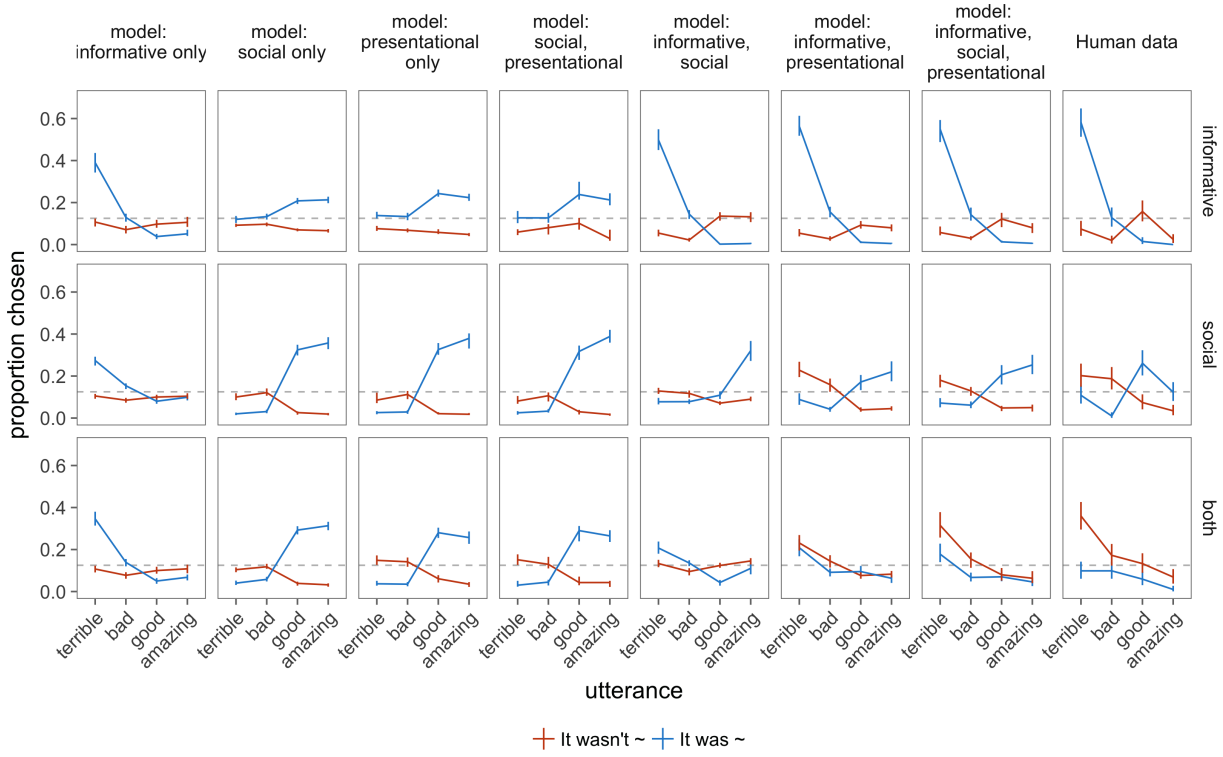

Figure S4: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with informative (top), social (middle),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

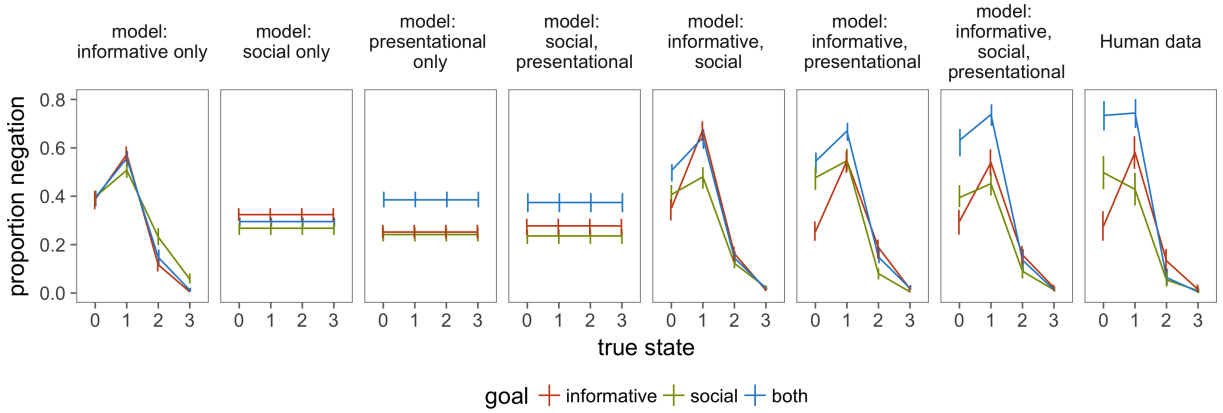

Figure S5: Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).