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

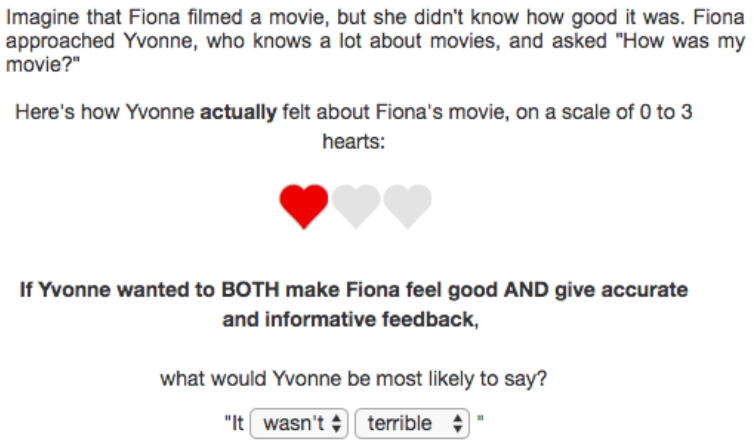

Figure S1: Example of a trial in the speaker production task.

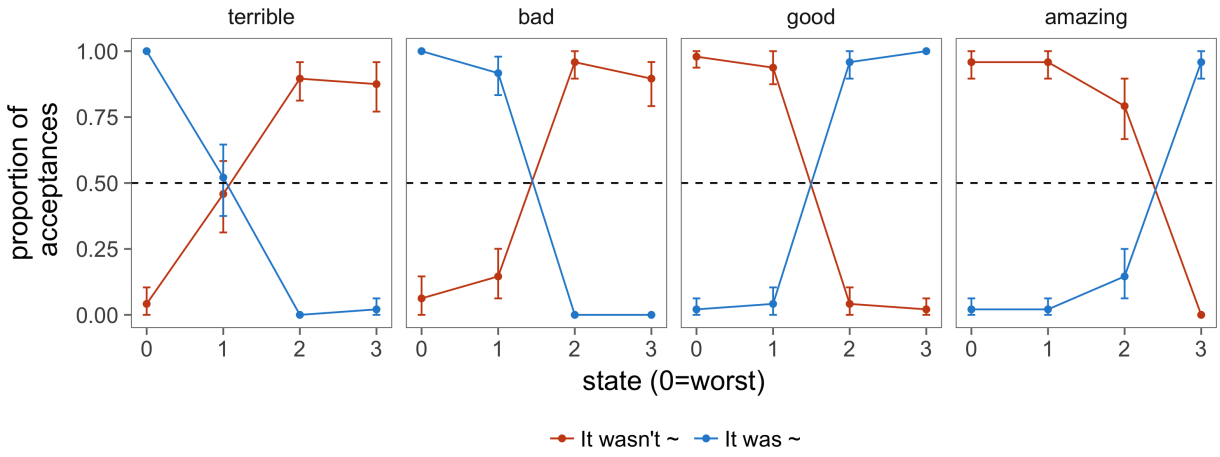

Figure S2: Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

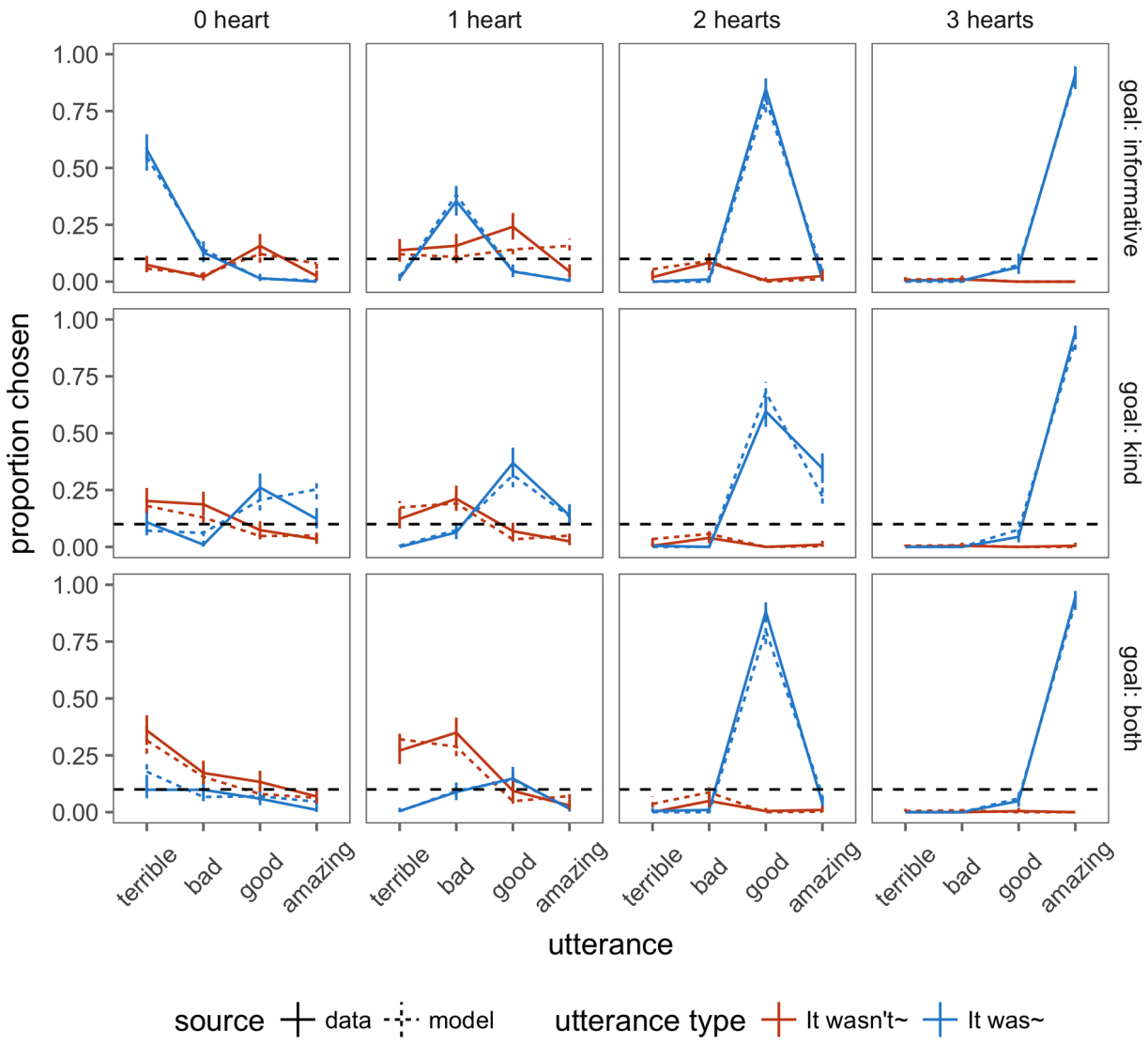

Figure S3: Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type direct vs. indirect in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows).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

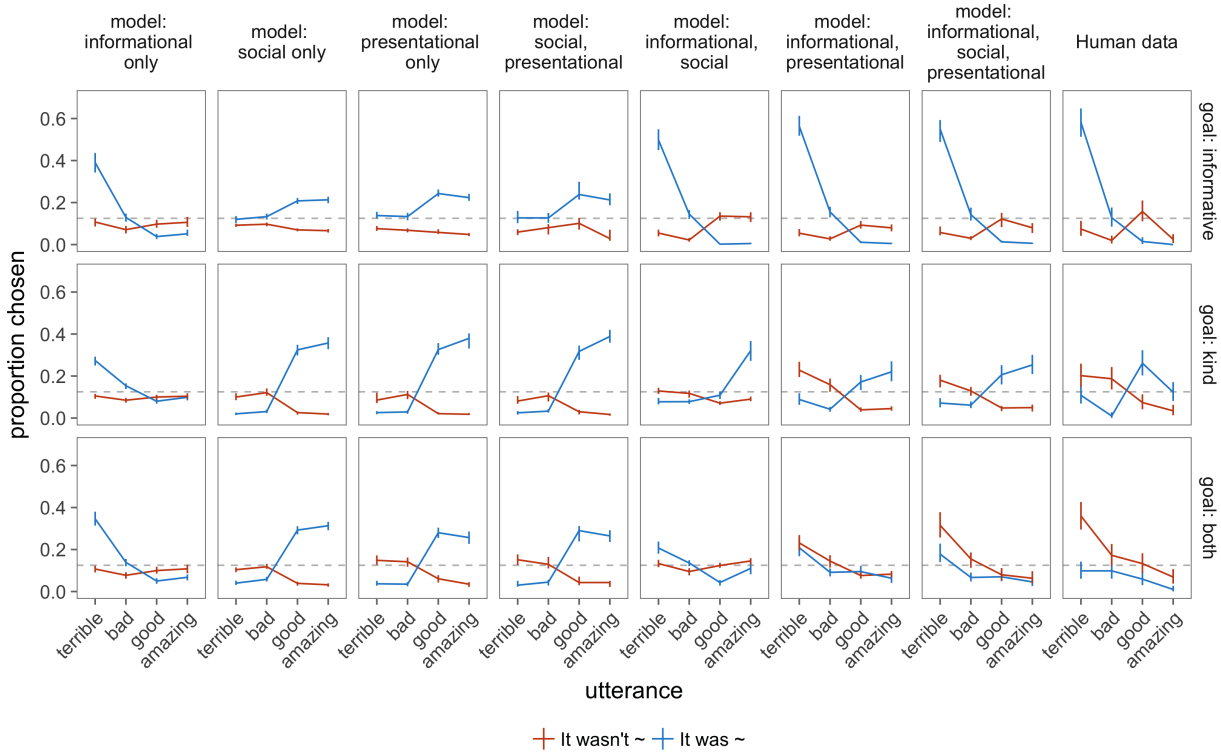

Figure S4: Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with a goal to be informative (top), kind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

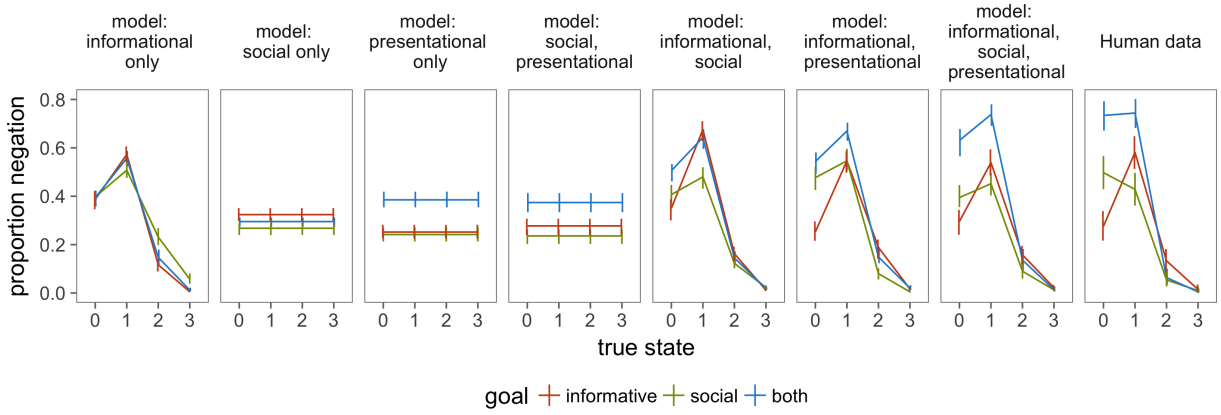

Figure S5: Fitted model predictions (left) and experimental results (rightmost) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).