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

[*] Corresponding author

[†] These authors contributed equally to this work.

Author Note

Correspondence concerning this article should be addressed to Erica J. Yoon, 450 Serra Mall, Bldg. 420, Rm. 290, Stanford, CA 94305. E-mail: ejyoon@stanford.edu

Abstract

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 judg- ments with high accuracy. This utility-theoretic approach to speech acts takes a step towards explaining the richness and subtlety of social language.

Keywords: Politeness; computational modeling; commmunicative goals; pragmatics

Word count: 3500

Polite speech emerges from competing social goals

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 (Bühler, 1934; Frank & Goodman, 2012; Jakobson, 1960; Shannon, 1948). On these views, transfer ought to be efficient and accurate: The speaker should choose a succinct utterance to convey what the speaker knows (Grice, 1975; Searle, 1975), 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 about the nature of communication: It is typically inefficient and underinformative, and sometimes even outright false. Yet language users, including even young children, spontaneously produce requests in polite forms (Axia & Baroni, 1985; e.g., Clark & Schunk, 1980), and speakers use politeness strategies even while arguing, preventing unnecessary offense to their interactants (Holtgraves, 1997). 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 (Leech, 1983) or social norms (Ide, 1989), but the most influential account of politeness relies on the notion of "face" to motivate deviations (Brown & Levinson, 1987; Goffman, 1967). On this theory, interactants seek to be liked, approved, and related to ("positive face") as well as maintain their freedom to act ("negative face"). If the speaker's intended meaning contains no threat to the listener's face, then the speaker will choose to convey the meaning in an explicit and efficient manner (putting it "on the record"). As the degree of face-threat becomes more severe, however, a speaker will choose to be polite by producing more indirect utterances. 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 it does not precisely define how competing communicative goals trade off with one another. For example, it is unclear when face-saving should be prioritized over helpful information transfer, and when the desire to save face will motivate statements that are outright false ("Your cake is delicious!") versus indirect ("It could use a bit of salt"). Concretely, such theorizing does not constrain how an artificial agent like a robot should go about making polite requests, conveying negative evaluations, or delivering bad news. 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 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 can choose between different utterances on the basis of their expected utility (including their cost to utter, approximated by the length of the utterance). The lie that a poem was great provides social utility by making its writer feel good, but does not inform about the true state of the world. Further, if the writer suspects that it was in fact terrible, the speaker runs the risk of being seen as uncooperative.

The utilities are weighed within a Rational Speech Act (RSA) model that takes a probabilistic approach to pragmatic reasoning in language (Frank & Goodman, 2012; Goodman & Frank, 2016): 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 (Lassiter & Goodman, 2017), hyperbole (Kao, Wu, Bergen, & Goodman, 2014), and irony (Kao & Goodman, 2015), 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 (Baker, Saxe, & Tenenbaum, 2009), a hypothesis which has found support in a wide variety of work with both adults and children (e.g., Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017).

RSA models are defined recursively such that speakers reason about listeners, and vice versa. By convention 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 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 (Figure 1).

We evaluate our model on its ability to predict human utterance choices in situations where polite language use is expected. Imagine Bob recited his poem and asks Ann how well he did. 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 determines how much Ann prioritizes each goal. Ann's production decision is softmax, which interpolates between maximizing and probability matching (via λ_{S_2}; Goodman & Stuhlmüller, 2013):

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

.

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

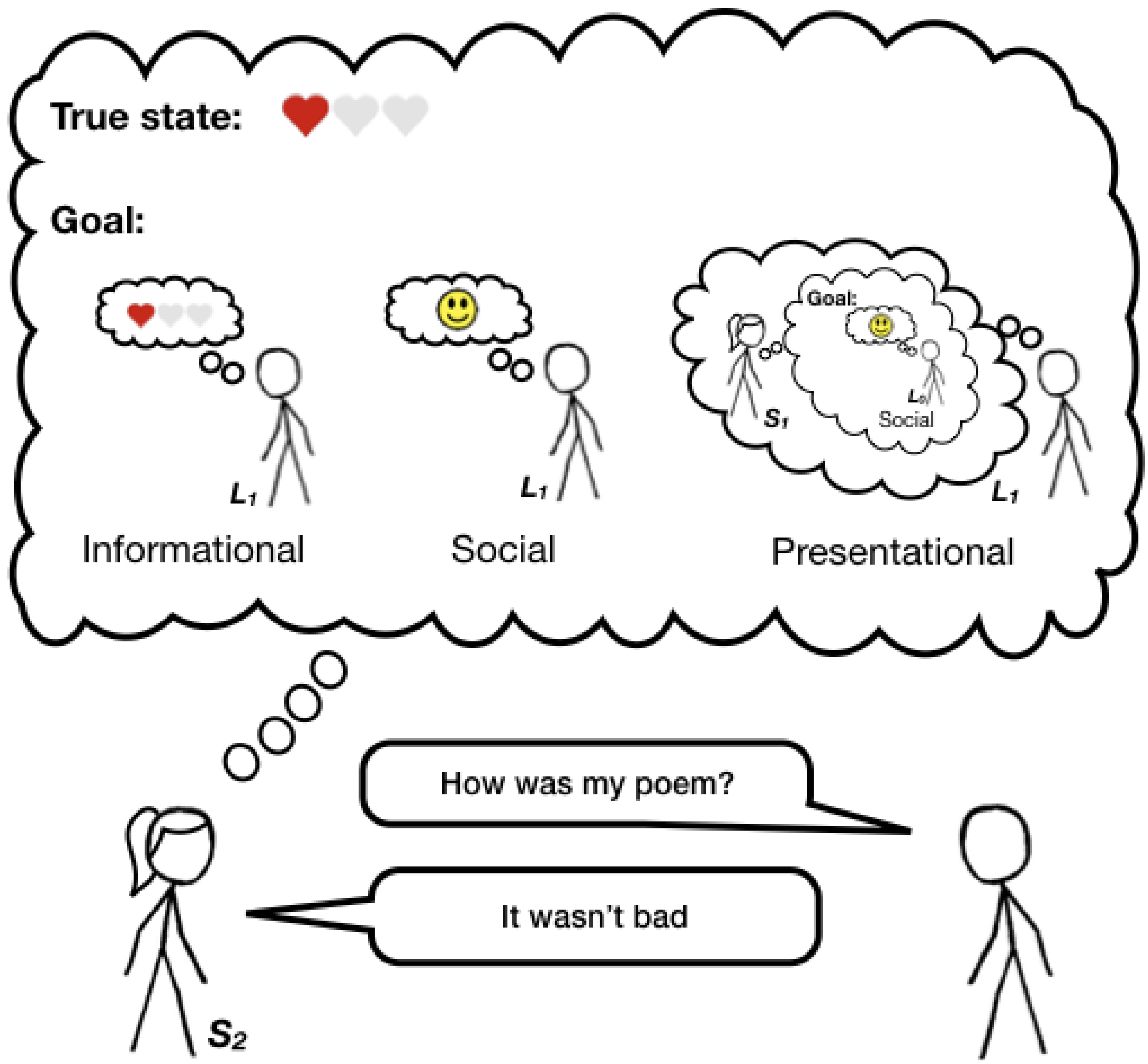

Figure 1. Diagram of the model: The pragmatic speaker observes the true state and determines her goal between three utilities (informational, social, and presentational), and produces an utterance.

98 the weighted combination of the three utilities minus the utterance cost $C(w)$:

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

99 .

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 s after hearing the speaker's utterance w (i.e., surprisal): $U_{inf}(w) = \ln(P_{L_1}(s|w))$.

For aspects of the world with affective consequences for the listener (e.g., Bob and his poem recital), we assume speakers produce utterances that make listeners feel like they are in a good state. The second utility term is a *social utility* (U_{soc}), which we define 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, 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 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). The speaker gains presentational utility when her listener believes she has certain goals – that she is trying to be informative or kind. Formally,

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

.

To define this term, the speaker has a weighting of informational vs. social goals to convey (ϕ_{S_1}) and must consider the beliefs of listener L1, 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 is higher-order 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).

Finally, more complex utterances 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 costlier than their equivalents with no negation (inferred from data; see Supplementary Information).

Within our experimental domain, we assume there are four possible states of the world corresponding to the value placed on a particular referent (e.g., the presentation the speaker is commenting on): $S = s_1, ..., s_5$. We further assume a uniform prior distribution over possible states of the world. The set of utterances is {*terrible*, *bad*, *good*, *amazing*, *not terrible*, *not bad*, *not good*, and *not amazing*}. We implemented this model using the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014).

Intuitively, if Bob's performance was good, Ann's utilities align toward 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. If he does not, he might infer Ann is "just being nice", but is 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. 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.

We made a direct, pre-registered test of our model by instantiating the example above in an online experiment ($N = 202$). Participants read scenarios in which we provided information on the speaker's (Ann's, in our example) feelings toward some performance or product (e.g., poem recital; *true state*), on a scale from zero to three hearts (e.g., one out of three hearts). For example, one trial read: "Imagine that Bob gave a poem recital, but he didn't know how good it was. Bob approached Ann, who knows a lot about poems, and asked"How was my poem?" We also manipulated the speaker's *goal* 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 Information). 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 heart) 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 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]). Rather than eschewing one of their goals to increase utility along a single dimension, participants chose utterances that jointly satisfied their conflicting goals by producing indirect, polite speech.

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 Information) via a Bayesian data analysis (M. D. Lee & Wagenmakers, 2014). 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

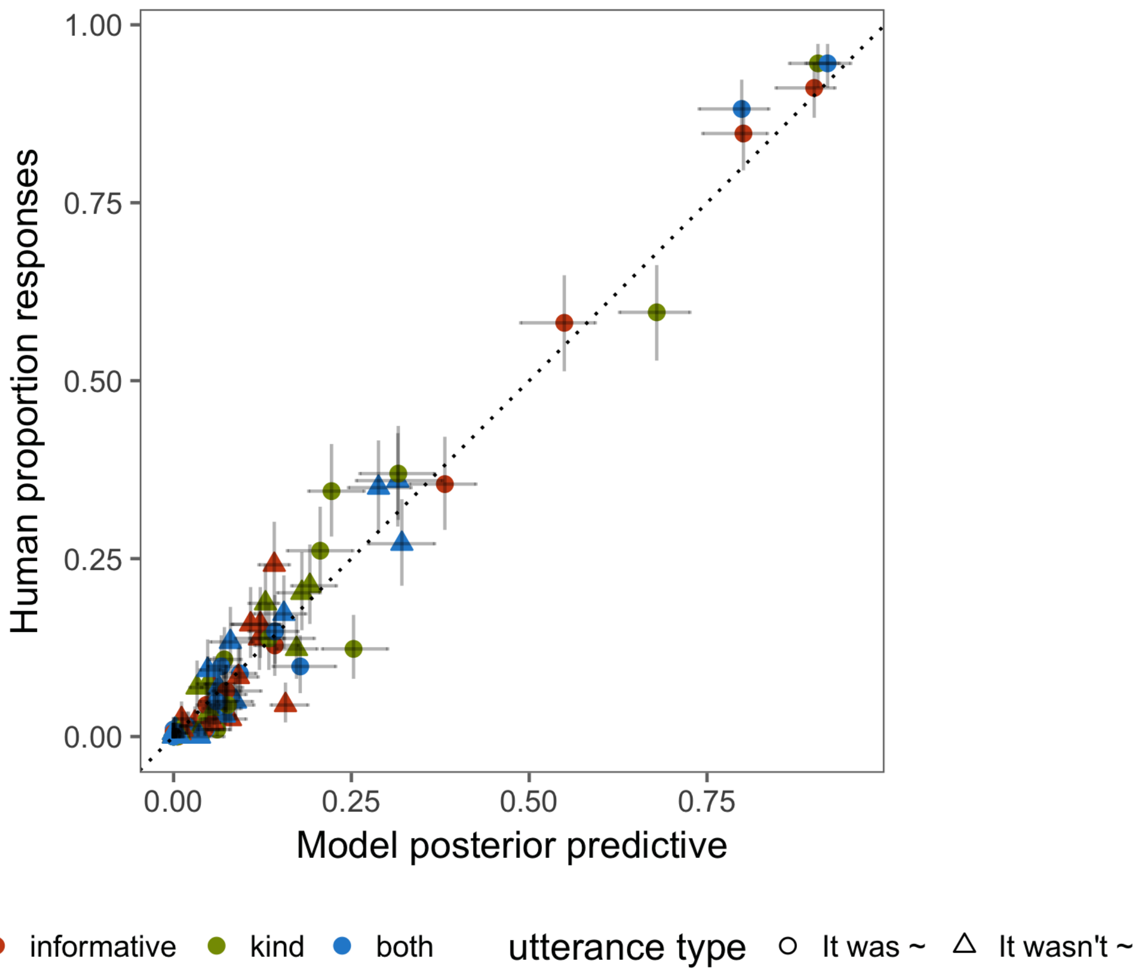

Figure 2. Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

($r^2(96) = 0.97$; 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.

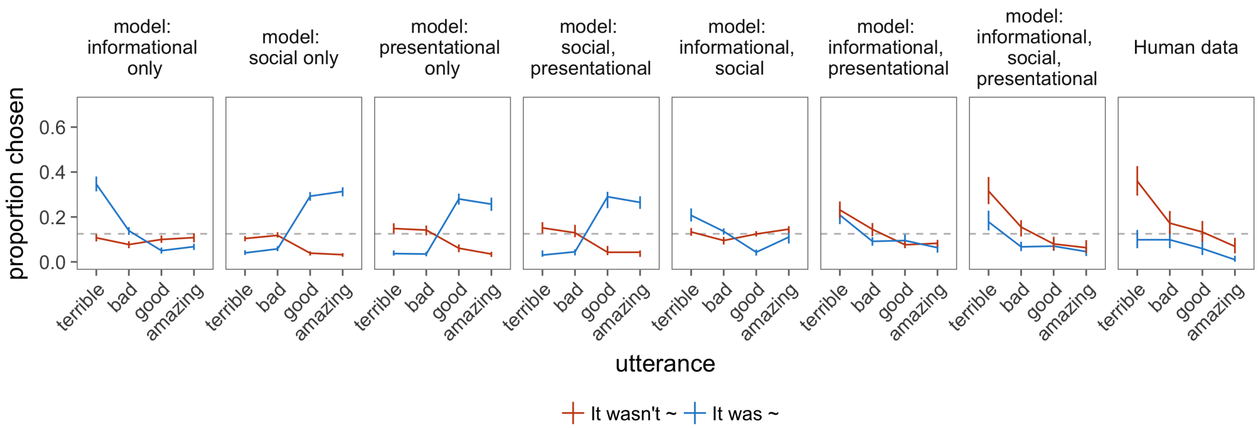

Figure 3. Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart (on a scale of 0 to 3) and speaker with both goals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

The utility weights inferred for the full model (Table 2) provide additional insight into how polite language use operates: *Being kind* 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 social utility close to zero, but the weight on *appearing kind* stays high, suggesting that speakers are expected to manage their own face even when they are not considering others'. *Kind and informative* speakers emphasize informativity slightly more than kindness. In all cases, however, the presentational utilities have greatest weight, which may suggest that appearing honest and kind is more important than actually being so! Overall then, our condition manipulation altered the balance between these weights, but all utilities played a role in all conditions.

Politeness is a puzzle for purely informational accounts of language use. Incorporating social motivations can provide an explanatory framework, but such intuitions have been resistant to formalization or precise testing. To overcome this issue, we created a utility-theoretic model of language use that captured the interplay between competing

Table 1

Comparison of variance explained for each model variant and log Bayes Factors quantifying evidence in favor of alternative model in comparison.

Model	Variance explained	log BF
model: informational, social, presentational	0.97	–
model: informational, presentational	0.96	-11.14
model: informational, social	0.92	-25.06
model: social, presentational	0.23	-864
model: presentational only	0.23	-873.83
model: social only	0.22	-885.52
model: informational only	0.83	-274.89

informational, social, and presentational goals. A preregistered experimental test of the model confirmed its ability to capture human judgments, unlike comparison models that used only a subset of the full utility structure.

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 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 (perhaps paralleling the difficulty human speakers feel in finding the right thing to say in a difficult situation; see Goodman & Frank, 2016).

For a socially-conscious speaker, managing listeners' inferences is a fundamental task. Inspired by the theory of politeness as face management (Brown & Levinson, 1987), our model takes a step towards understanding it. Our work extends previous models of language

Table 2

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	social	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	social	0.66	–	0.34	0.04
informational, social	both	0.54	0.46	–	–
informational, social	informative	0.82	0.18	–	–
informational, social	social	0.39	0.61	–	–
social, presentational	both	–	0.38	0.62	0.55
social, presentational	informative	–	0.35	0.65	0.75
social, presentational	social	–	0.48	0.52	0.66

beyond standard informational utilities to address social and self-presentational concerns. Previous theories of language use have not explained how informational versus social concerns trade off to inform the speaker's utterance choices. Thus, this work represents a key theoretical advance exploring how informational cooperativity interacts with other social goals. By considering utility-driven inferences in a social context (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013) where agents need to take into account concerns about both self and others, our approach here could give insights into a wide range of social behaviors beyond speech.

The model presented here relates to other work done in game-theoretic pragmatics. Van Rooy (2003) uses a game-theoretic analysis of polite requests ("Could you possibly take me home?") to argue the purpose of polite language is to align the preferences of interlocutors. Our notion of social utility U_{soc} and presentational utility U_{pres} is similar in that they motivate speakers to signal worlds that make the listener feel good. Van Rooy (2003)'s analysis, however, relies on the notion that polite language is costly (in a social way e.g., by reducing one's social status or incurring social debt to one's conversational partner) but it's not clear how the polite behaviors explored in our experiments (not polite requests) would incur any cost to speaker or listener. Our model derives its predictions by construing the speaker utility as a collection of possible goals (here, epistemic, social, and presentational goals). The speech-acts themselves are not costly.

In another game-theoretic approach by Pinker, Nowak, and Lee (2008), human communication is assumed to involve a mixture of cooperation *and* conflict: indirect speech then allows for plausible deniability that is in self-interest but goes against the interest of the addressee. In contrast, our work builds on existing classic theories of polite speech as primarily cooperative (Brown & Levinson, 1987; Grice, 1975) rather than based on both cooperation and conflict. We have shown that a separate notion of plausible deniability may not be needed, as indirect speech in our specific case comes from both a goal to be helpful and a desire to look good. Our work is able to capture different linguistic nuances involved in this process of reasoning about different goals that speakers have.

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. For example, 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.

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.

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.

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 5 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 4).

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 (Figure 5). 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 –

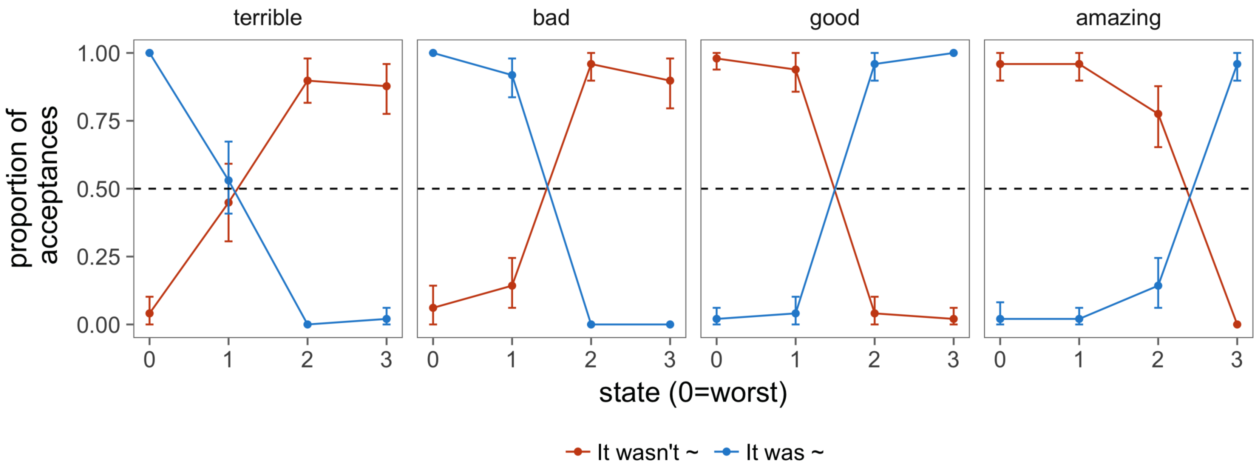

Figure 4. Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

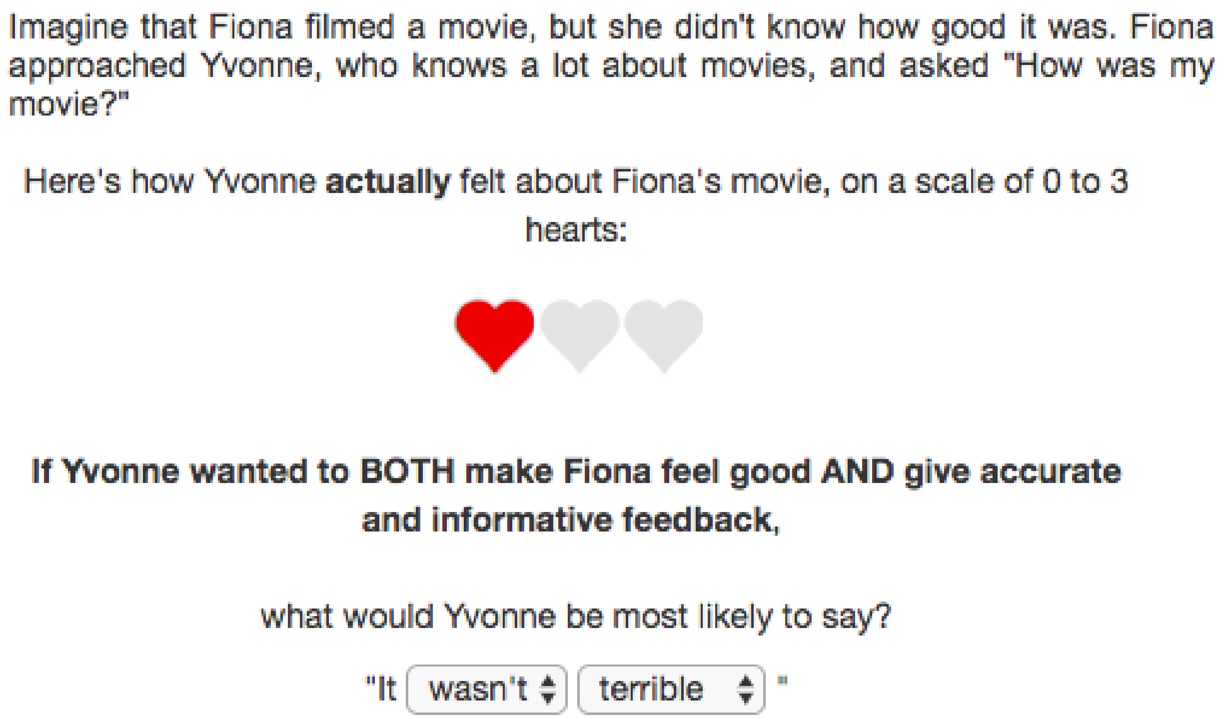

Figure 5. Example of a trial in the speaker production task.

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 5). 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.*

Data availability

Our model, preregistration of hypotheses, procedure, data, and analyses are available at https://github.com/ejyoon/polite_speaker.

Author information

Author contributions

All authors designed research and wrote the paper; E.J.Y. and M.H.T. performed research and analyzed data.

Competing interests

The authors declare no conflict of interest.

Supplementary Information

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; Müller, 2017a), *binom* (Version 1.1.1; Dorai-Raj, 2014), *brms* (Version 2.0.1; Bürkner, 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, & Müller, 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; Müller, 2017b), *jsonlite* (Version 1.5; Ooms, 2014), *langcog* (Version 0.1.9001; Braginsky, Yurovsky, & Frank, n.d.), *lme4* (Version 1.1.15; Bates, Mächler, 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 0.12.17; Eddelbuettel & François, 2011), *readr* (Version 1.1.1; Wickham, Hester, & Francois, 2017), *rwebppl* (Version 0.1.97; Braginsky, Tessler, & Hawkins, n.d.), *stringr* (Version 1.3.1; Wickham, 2017b), *tibble* (Version 1.4.2; Müller & 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; Bürkner, 2017) using crossed random effects of true state and goal with maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013; Gelman & Hill, 2006).

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"; *kind*, *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

Table 3

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: Kind	0.97	0.25	0.51	1.49
True state * Informative	-1.33	0.18	-1.69	-0.98
True state * Kind	-0.50	0.22	-0.92	-0.07

Table 4

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

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 (M. D. Lee & Wagenmakers, 2014). We ran 4 MCMC chains for 80,000 iterations, discarding the first 40,000 for burnin. Negation cost and speaker optimality parameters are shown in Table 4.

346 ## Supplemental Figures

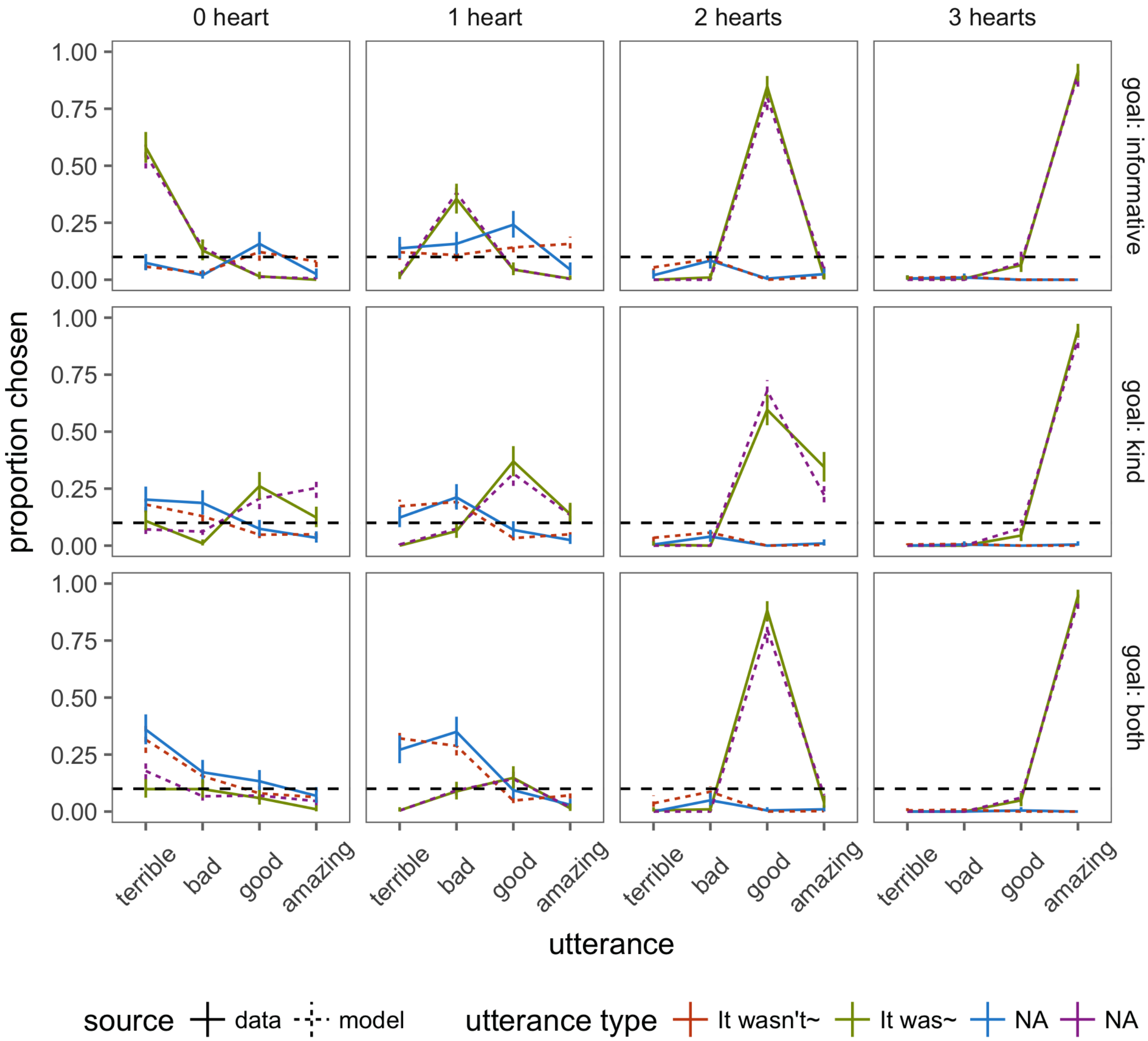

Figure 6. Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows).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

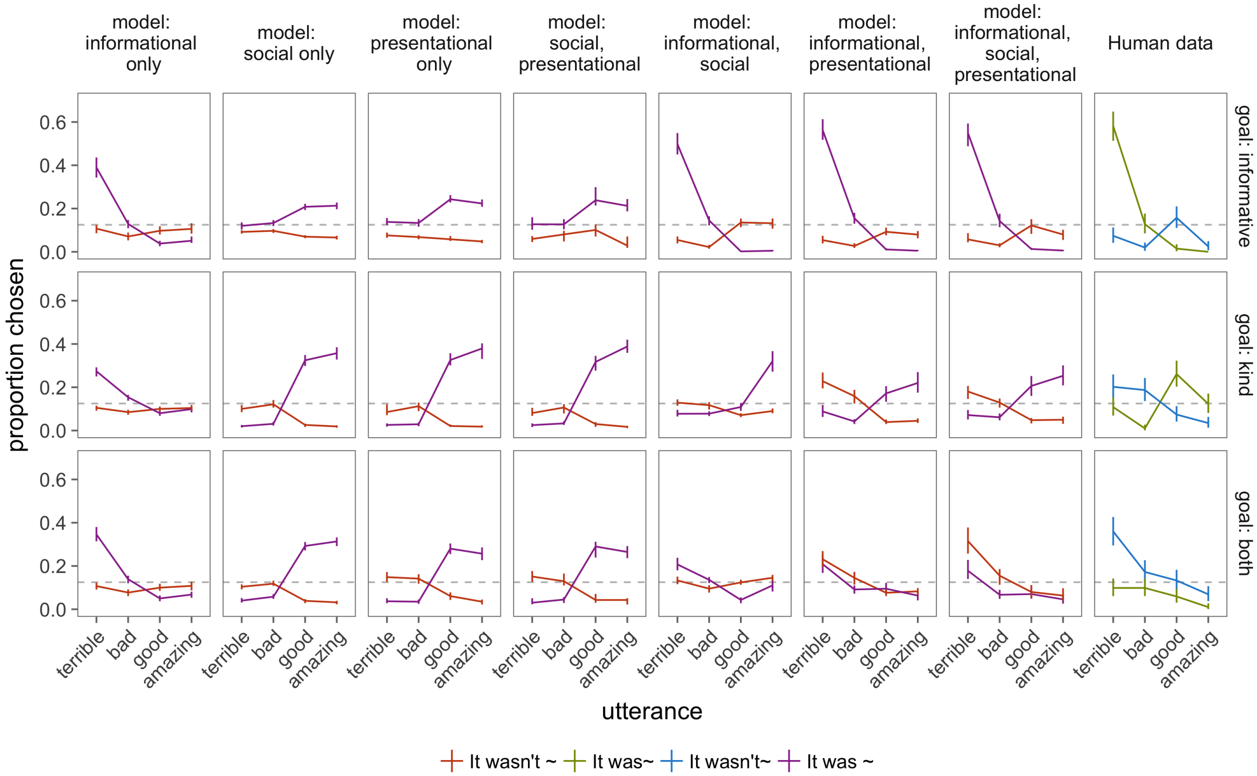

Figure 7. Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with a goal to be informative (top), kind (middle), or both (bottom). Gray dotted line indicates chance level at 12.5%.

References

Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child*

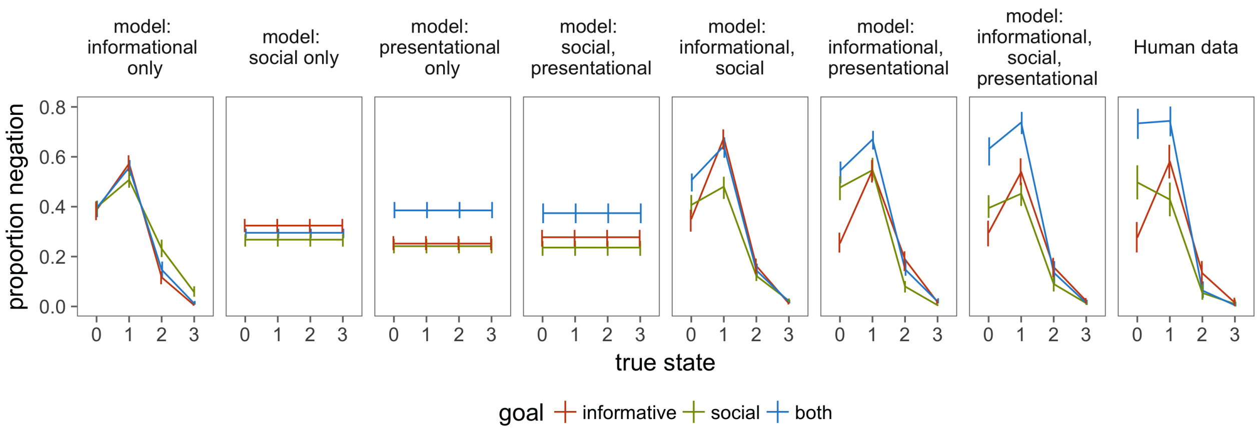

Figure 8. Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

Development, 918–927.

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r*. Retrieved from https://CRAN.R-project.org/package=magrittr

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., & Maechler, M. (2017). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Braginsky, M., Tessler, M. H., & Hawkins, R. (n.d.). *Rwebppl: R interface to webppl*.

Retrieved from https://github.com/mhtess/rwebppl

Braginsky, M., Yurovsky, D., & Frank, M. C. (n.d.). *Langcog: Language and cognition lab things*. Retrieved from http://github.com/langcog/langcog

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

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

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, *8*(2), 111–143.

Dorai-Raj, S. (2014). *Binom: Binomial confidence intervals for several parameterizations*. Retrieved from https://CRAN.R-project.org/package=binom

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. doi:10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. doi:10.18637/jss.v040.i08

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

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

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Goodman, N. D., & Stuhlmüller, A. (2014). The Design and Implementation of Probabilistic

Programming Languages. http://dippl.org.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). Academic Press.

Hamlin, K. J., Ullman, T. D., Tenenbaum, J. B., Goodman, N. D., & Baker, C. L. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226.

Henry, L., & Wickham, H. (2017). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Hocking, T. D. (2017). *Directlabels: Direct labels for multicolor plots*. Retrieved from https://CRAN.R-project.org/package=directlabels

Holtgraves, T. (1997). YES, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology*, *16*(2), 222–239.

Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, *8*(2-3), 223–248.

Jakobson, R. (1960). Linguistics and poetics. In *Style in language* (pp. 350–377). MA: MIT Press.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 37th annual conference of the Cognitive Science Society*.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge Univ. Press.

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

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs.* Retrieved from https://CRAN.R-project.org/package=BayesFactor

Müller, K. (2017a). *Bindrcpp: An 'rcpp' interface to active bindings.* Retrieved from https://CRAN.R-project.org/package=bindrcpp

Müller, K. (2017b). *Here: A simpler way to find your files.* Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2017). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes.* Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [Stat.CO]*. Retrieved from https://arxiv.org/abs/1403.2805

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Searle, J. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 59–82). Academic Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 623–656.

Van Rooy, R. (2003). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In *Proceedings of the 9th conference on theoretical aspects of rationality and knowledge* (pp. 45–58). ACM.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2017b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2017c). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., & Henry, L. (2017). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2017). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr