Polite speech emerges from competing social goals

Abstract

Language is a remarkably efficient tool for transmitting information. Yet human speakers make statements that are inefficient, imprecise, or even contrary to their own beliefs, all in the service of being polite. What rational machinery underlies polite language use? Here, we show that polite speech emerges from the competition of three communicative goals: to convey information, to be kind, and to present oneself in a good light. We formalize this goal tradeoff using a probabilistic model of utterance production, which predicts human utterance choices in socially-sensitive situations with high quantitative accuracy, and we show that our full model is superior to its variants with subsets of the three goals. This utility-theoretic approach to speech acts takes a step towards explaining the richness and subtlety of social language use.

Keywords: politeness, computational modeling, communicative goals, pragmatics

Word count: 3816

Polite speech emerges from competing social goals

Introduction

We rarely say exactly what's on our mind. Although "close the window!" could be an effective message, we dawdle by adding "can you please. . . ?" or "would you mind. . . ?" Rather than tell an uncomfortable truth, socially-aware speakers lie ("Your dress looks great!") and prevaricate ("Your poem was so appropriate to the occasion"). Such language use is puzzling for classical views of language as information transfer (Bühler, 1934; Frank & Goodman, 2012; Jakobson, 1960; Shannon, 1948). On the classical view, transfer ought to be efficient and accurate: Speakers are expected to choose succinct utterances to convey their beliefs (Grice, 1975; Searle, 1975), and the information conveyed is ideally truthful to the extent of a speaker's knowledge. Polite speech violates these basic expectations about the nature of communication: It is typically inefficient and underinformative, and sometimes even outright false. Yet even young speakers spontaneously produce requests in polite forms (Axia & Baroni, 1985), and adults use politeness strategies while arguing (Holtgraves, 1997), even though polite utterances may risk high-stakes misunderstandings (Bonnefon, Feeney, & De Neys, 2011).

If politeness only gets in the way of effective information transfer, why be polite? Clearly, there are social concerns, and most linguistic theories assume utterance choices are motivated by these concerns, couched as either polite maxims (Leech, 1983), social norms (Ide, 1989), or aspects of a speaker and/or listener's identity, known as *face* (Brown & Levinson, 1987; Goffman, 1967). Face-based theories predict that when a speaker's intended meaning contains a threat to the listener's face or self-image (and potentially the speaker's face), her messages will be less direct, less efficient, and possibly untruthful. Indeed, listeners readily assume speakers' intentions to be polite when interpreting utterances in face-threatening situations (Bonnefon, Feeney, & Villejoubert, 2009). How this socially-aware calculation unfolds, however, is not well understood. When should a speaker decide to say something false ("Your poem was great!" based on an example from Bonnefon et al. (2009)) rather than just be indirect (*Some of the metaphors were tricky to understand.*)? How does a speaker's own self-image enter into the calculation?

We propose a utility-theoretic solution to the problem of polite language use by quantifying the tradeoff between competing communicative goals. In our model, speakers attempt to maximize utilities that represent their communicative goals: informational utility—derived via classical, effective information transmission; social utility—derived by being kind and saving the listener's face; and self-presentational utility—the most novel component of our model, derived by appearing in a particular way to save the speaker's own face. Speakers then produce an utterance on the basis of its expected utility (including their cost to speak). The lie that a poem was great provides social utility by making the writer feel good, but does not provide information about the true state of the world. Further, if the writer suspects that the poem was in fact terrible, the speaker runs the risk of being seen as uncooperative.

We assume that speakers' utilities are weighed within a probabilistic model of pragmatic reasoning: the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016). Speakers are modeled as agents who choose utterances by reasoning about their potential effects on a listener, while listeners infer the meaning of an utterance by reasoning about speakers and what goals could have led them to produce their utterances. 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. In this framework, language use builds on 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 S reason about listeners L, and

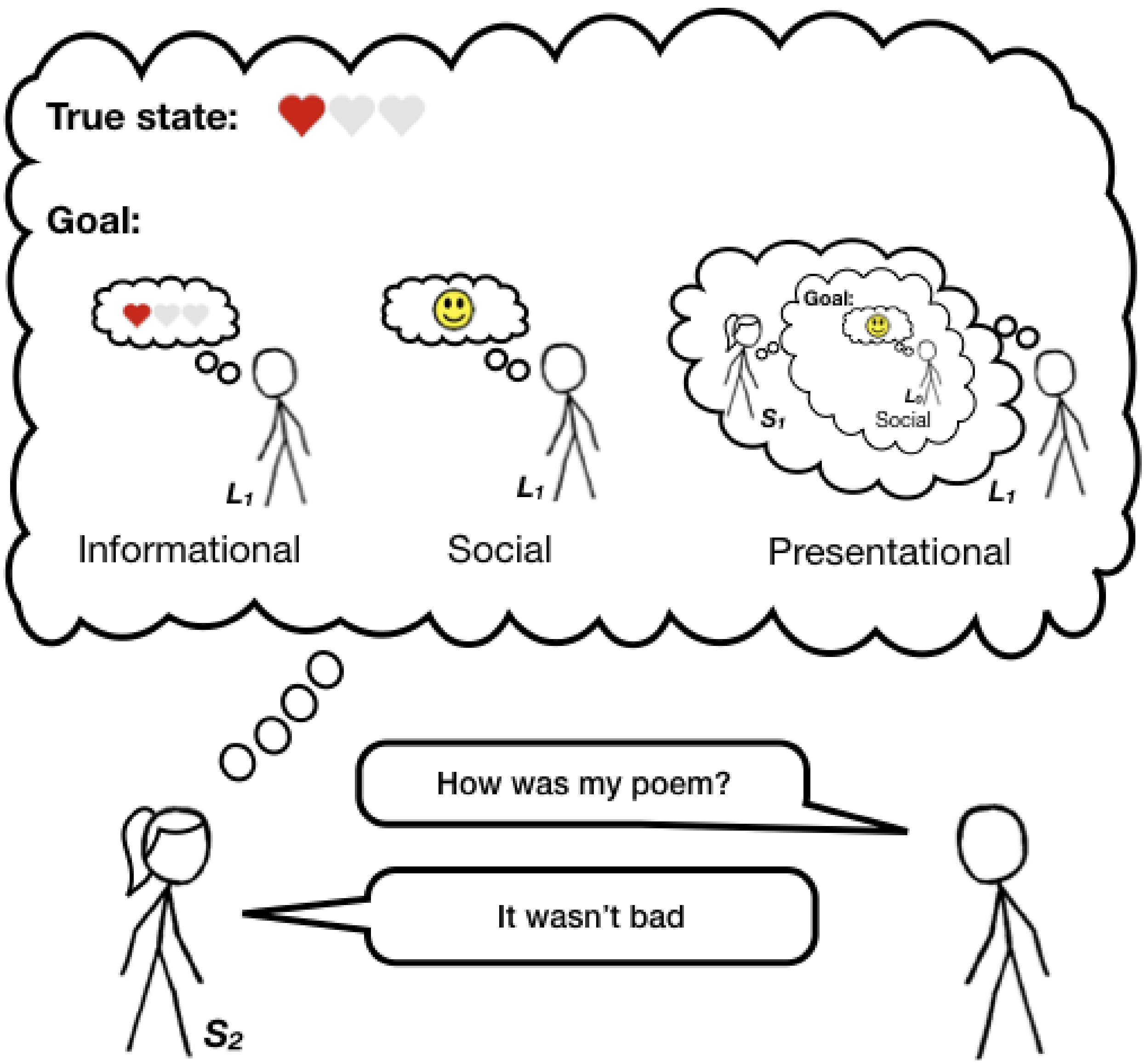

Figure 1. Diagram of the model: The polite speaker observes the true state and determines her goal between three utilities (informational, social, and presentational), and produces an utterance.

vice versa. We use a standard convention in indexing and say 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, who is 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 who reasons about what to say to L_1 by considering informational, social, and self-presentational goals (Figure 1).

We evaluate our model's ability to predict human utterance choices in situations where polite language use is expected. Imagine Bob recited a poem and asked Ann how good it was. Ann (S_2) produces an utterance w based on the true state of the world s (i.e., the rating, in her mind, truly deserved by Bob's poem) and a set of goal weights $\hat{\phi}$, that determines how much Ann prioritizes each of the three possible goals. Ann's production decision is softmax, which interpolates between maximizing and probability matching (via λ_{S_2}; Goodman & Stuhlmüller, 2013):

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

We posit that a speaker's utility contains three distinct components: informational, social, and presentational. The total utility U_{total} of an utterance is thus the weighted combination of the three utilities minus the utterance cost $C(w)$:

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

We define *social utility* (U_{soc}) as the expected subjective utility of the state $V(s)$ implied to the pragmatic listener by the utterance: $U_{soc}(w) = \mathbb{E}_{P_{L_1}(s|w)}[V(s)]$. The subjective utility function $V(s)$ could vary by culture and context; we test our model when states are explicit ratings (e.g., on a 4-point scale) and we assume a positive linear value relationship between states and values V to model a listener's preference to be in a highly rated state (e.g., Bob would prefer to have written a poem deserving 4 points rather than 1 point).

At the same time, a speaker may desire to be epistemically helpful, modeled as standard *informational utility* (U_{inf}). The informational utility indexes the utterance's *surprisal*, or amount of information the listener (L_1) would still not know about the state of the world s after hearing the speaker's utterance w (e.g., how likely is Bob to guess Ann's actual opinion of the poem): $U_{inf}(w) = \ln(P_{L_1}(s|w))$. Speakers who optimize for informational utility produce accurate and informative utterances while those who optimize for social utility produce utterances that make the listener feel good.

If a listener is uncertain how their particular speaker is weighing the competing goals to be honest vs. kind (informational vs. social utilities), he might try to infer the weighting (e.g., "was she just being nice?"). But a sophisticated speaker can produce utterances in order to appear *as if* she had certain goals in mind, for example making the listener think that the speaker was being both kind and informative ("she wanted me to know the truth but without hurting my feelings"). The extent to which the speaker *appears* to the listener to have a particular goal in mind (e.g., to be kind) is the utterance's *presentational utility* (U_{pres}). The speaker gains presentational utility when her listener believes she has particular goals, represented by a mixture weighting ϕ_{S_1} between trying to be genuinely informative vs. kind. Formally,

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

The speaker conveys a particular weighting of informational vs. social goals (ϕ_{S_1}) by considering the beliefs of listener L_1, who hears an utterance and jointly infers the speaker's utilities and the true state of the world:

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

The presentational utility is the highest-order term of the model, defined only for a speaker thinking about a listener who evaluates a speaker (i.e., defined for S_2, but not S_1). Only the social and informational utilities are defined for the S_1 speaker (via reasoning about L_0); thus, S_1's utility weightings can be represented by a single number, the mixture parameter ϕ_{S_1}. Definitions for S_1 and L_0 otherwise mirror those of S_2 and L_1 and can be found in the Supplmentary Materials: Model details section.

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 (this cost is inferred from data; see Supplementary Materials).

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 poem the speaker is commenting on), represented in terms of numbers of hearts (Figure 1): $S = s_0, ..., s_3$. Since the rating scale is relatively abstract, we 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) and a demo can be found at http://forestdb.org/models/politeness.html.

Model predictions

The pragmatic listener model L_1 draws complex inferences about both the true state of the world (Fig. 2A) and the speaker's goals (Figure 2B). Upon hearing *[Your poem] was terrible* (Figure 2A and 2B top-left), the listener infers the poem is probably truly terrible (i.e., worthy of zero hearts) and that the speaker has strong informational goals. *It was amazing* is more ambiguous (Figure 2A and 2B top-right): The poem could indeed be worthy of three hearts, but it is also plausible the speaker had strong social goals and the poem was mediocre. Negation makes the meanings less precise and introduces more uncertainty into the inference about the state: A listener who hears *It wasn't amazing* sees it as a relatively kind way of saying that the poem was quite bad (0 or 1 hearts), inferring a balance of social and informational goals for the speaker (Figure 2A and 2B bottom-right). *It wasn't terrible* is the most open-ended, leaving open the possibility that the poem was worthy of 0 hearts (i.e., *it was terrible*) but conveying to the listener that the speaker cares about both informational and social goals, with a slight preference of towards being social (Figure 2A and 2B bottom-left).

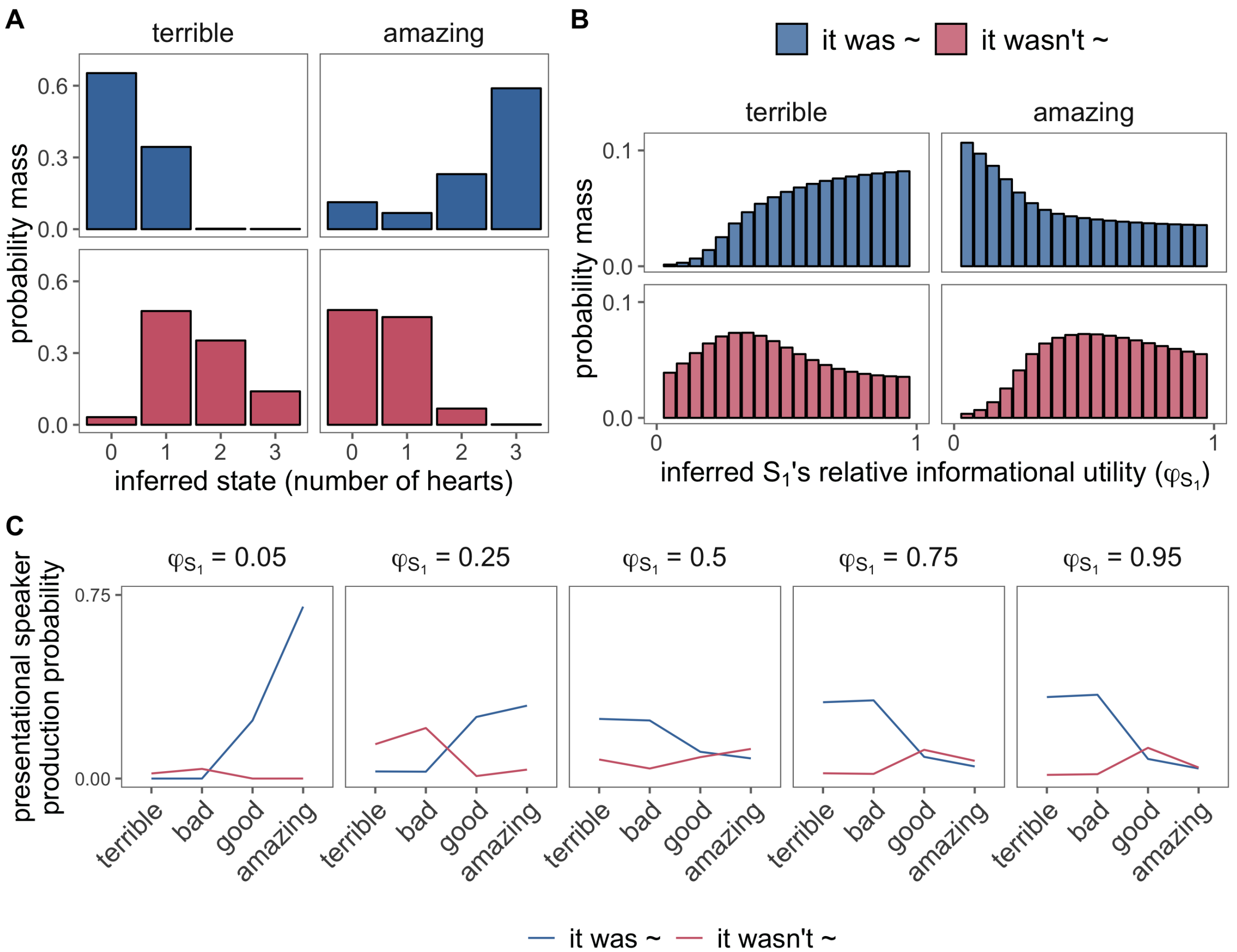

Figure 2. Model behavior. Listener inferences about the true state (e.g., the rating truly deserved by the poem; A) and the speaker's utility weighting (ϕ_{S_1} or how informational vs. social the speaker is, where $\phi_{S_1} = 0$ is fully social, and $\phi_{S_1} = 1$ is fully informational; B) as a function of the utterance heard (facets). C: Purely self-presentational speaker production behavior as a function of the kind of speaker they wish to present themselves as (facets; relatively more informational, e.g., $\phi_{S_1} = 0.05$, vs. social as represented, e.g., $\phi_{S_1} = 0.95$).

The self-presentational utility guides the speaker S_2 to care about how she will be viewed in the eyes of the listener L_1 (Figure 2C). If the speaker wants to present herself as someone who is socially-minded (e.g., informational mixture or ϕ_{S_1} of 0.05), she should produce direct, positive utterances (e.g., *amazing*). The best way to appear honest (e.g., informational mixture of 0.95) is to say direct, negative utterances (e.g., *terrible*). The desire to appear as someone concerned with telling the truth while also caring about the listener's feelings (e.g., ϕ_{S_1} of 0.25) leads the speaker to produce indirect utterances (e.g., *not terrible*). Such indirect speech acts are sufficiently open-ended to include the possibility that the poem was good, but the avoidance of a more direct utterance (e.g., *good*) provides the listener with a way to recover the true state (e.g., the poem was mediocre) by way of reasoning that the speaker cares about his feelings by not saying the blunt truth.

Experiment: Speaker production task

We made a direct, fully pre-registered test of our speaker production model and its performance in comparison to a range of alternative models, by instantiating our running example in an online experiment.

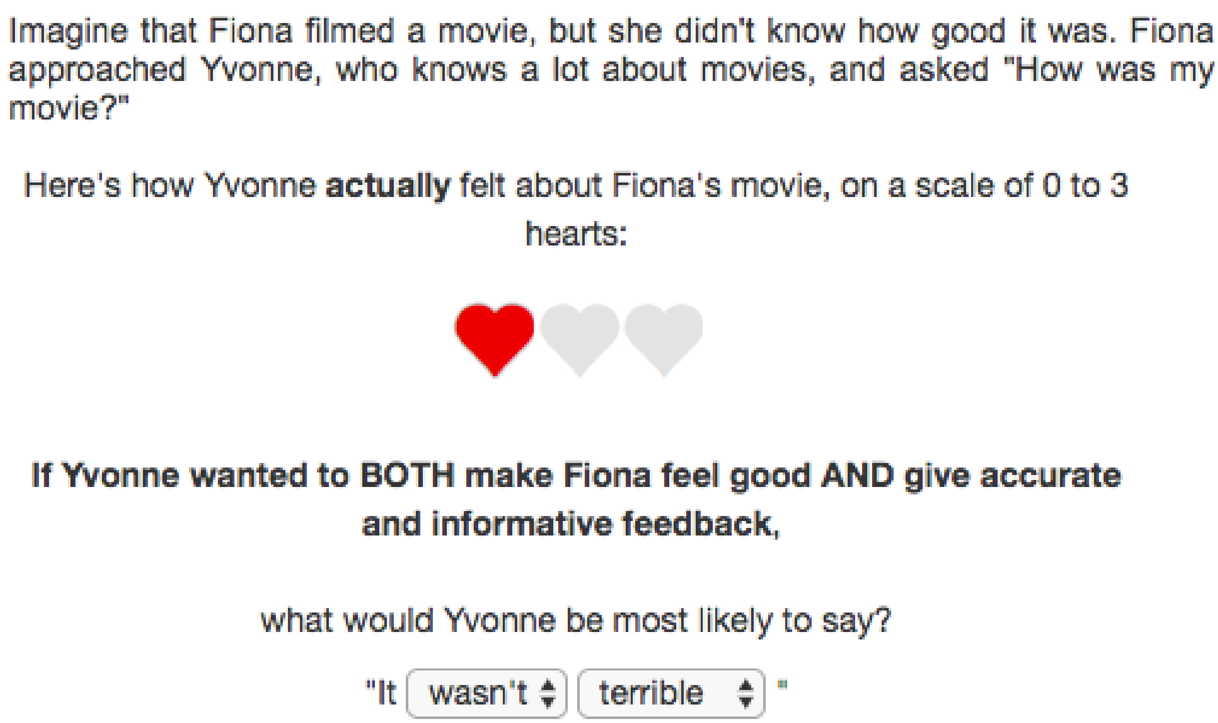

Figure 3. Example of a trial in the speaker production task.

Participants

202 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Design and Methods

Participants read scenarios with information on the speaker's feelings toward some performance or product (e.g., a 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?" Additionally, we manipulated the speaker's goals 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 experimentally-induced goals would induce a different tradeoff between social and informational utilities in our model, as well as modulating the self-presentational component. In a single trial, each scenario was followed by a question asking for the most likely produced 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.*

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.*

Behavioral results

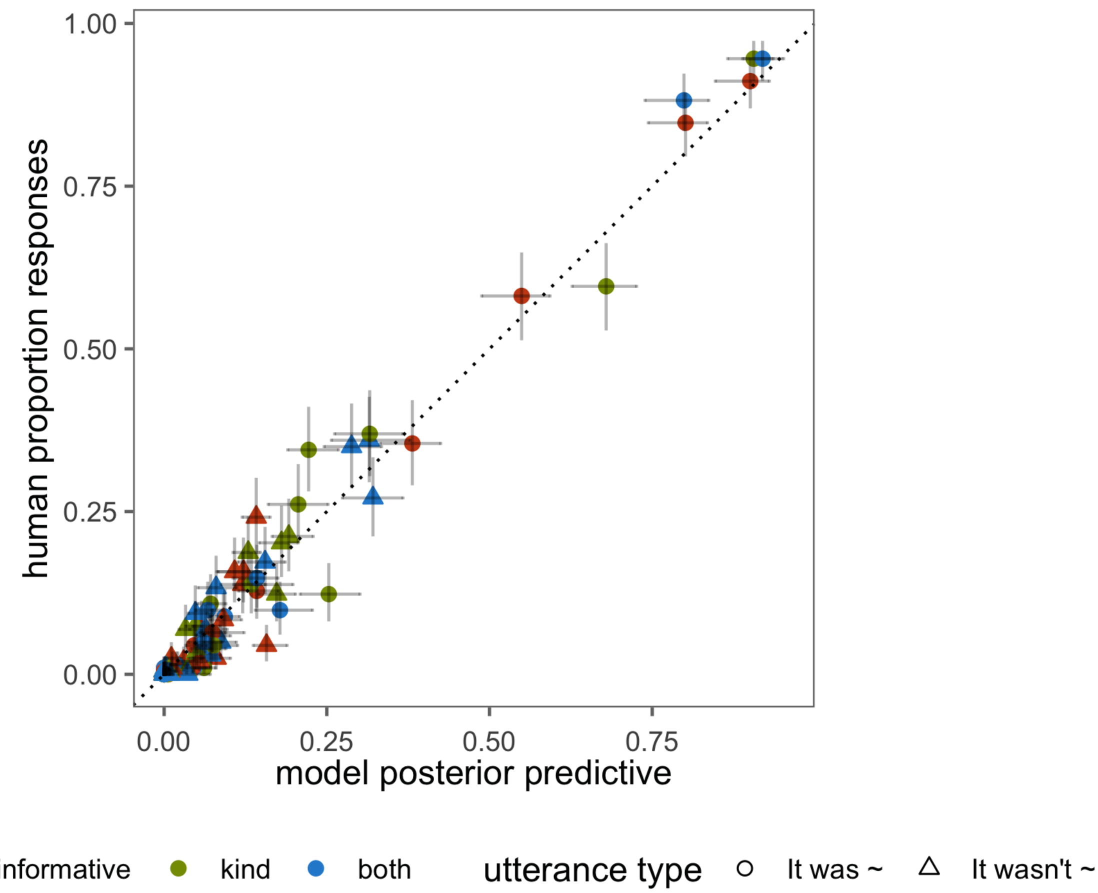

Figure 4. Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

Our primary behavioral hypothesis was that speakers describing bad states (e.g., poem deserving 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 both save the listener's face and provide some information about the true state, and thus, are what a socially-conscious speaker would say (Figure 2). This prediction was confirmed, as a Bayesian mixed-effects model predicts more negation as a function of true state and goal via an interaction: 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 speech.

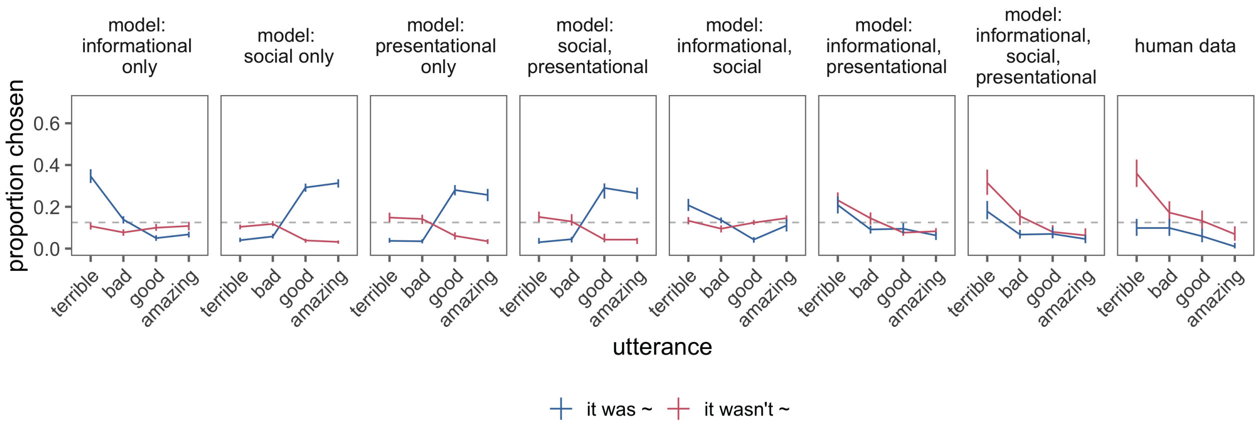

Figure 5. Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart (on a scale of 0 to 3) and speaker with both goals to be informative and kind. Gray dotted line indicates chance level at 12.5%.

Model results

The model parameters (softmax parameters and each goal condition's utility weights) can be inferred from the behavioral data using Bayesian data analysis (M. D. Lee & Wagenmakers, 2014). To approximate the literal meanings (i.e., the semantics) of the words as interpreted by the literal listener L_0, we obtained literal meaning judgments from an independent group of participants (See Supplmentary Materials: Literal semantic task section). The posterior predictions from the the three-utility polite speaker model (informational, social, presentational) showed a very strong fit to participants' actual utterance choices ($r^2(96)$ = 0.97; Figure 4). We compared these to six model variants containing subsets of the three utilities in the full model. Both the variance explained and marginal likelihood of the observed data were the highest for the full model (Table 1). Only the full model captured

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
informational, social, presentational	0.97	–
informational, presentational	0.96	-11.14
informational, social	0.92	-25.06
social, presentational	0.23	-864
presentational only	0.23	-873.83
social only	0.22	-885.52
informational only	0.83	-274.89

participants' preference for negation when the speaker wanted to be informative and kind about truly bad states, as hypothesized (Figure 5). In sum, the full set of informational, social, and presentational were required to fully explain participants' utterance choices.

The utility weights inferred for the three-utility model (Table 2) provide additional insight into how polite language use operates in our experimental context and possibly beyond: *Being kind* ("social") requires not only weights on social and presentational utilities but equal weights on all three utilities, indicating that informativity is a part of language use even when it is explicitly not the goal. *Being informative* ("informative") pushes the weight on social utility (ϕ_{soc}) close to zero, but the weight on *appearing kind* (ϕ_{pres}) stays high, suggesting that speakers are expected to manage their own face even when they are not considering others'. *Kind and informative* ("both") speakers emphasize informativity slightly more than kindness. In all cases, however, the presentational utilities have greatest weight, suggesting that managing the listener's inferences about oneself was integral to participants'

Table 2

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

model (utilities)	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

decisions in the context of our communicative task. Overall then, our condition manipulation altered the balance between these weights, but all utilities played a role in all conditions.

Discussion

Politeness is puzzling from an information-theoretic perspective. Incorporating social motivations adds a level of explanation, but so far such intuitions and observations have resisted both formalization and precise testing. We present a utility-theoretic model of language use that captures the interplay between competing informational, social, and presentational goals, and provide preregistered experimental evidence that confirmed its ability to capture human judgments, unlike comparison models with only a subset of the full utility structure.

To estimate precisely choice behavior in the experiment, it was required to abstract away from natural interactions in a number of ways. Human speakers have access to a potentially infinite set of utterances to select from in order to manage the three-utility tradeoff (*It's hard to write a good poem*, *That metaphor in the second stanza was so relatable!*). In theory, 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. Our work extends previous models of language beyond standard informational utilities to address social and self-presentational concerns. Further, our model builds upon the theory of politeness as face management (Brown & Levinson, 1987) and takes a step towards understanding the complex set of social concerns involved in face management. Our approach can provide insight into a wide range of social behaviors beyond speech 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.

Previous game-theoretic analyses of politeness have either required some social cost to an utterance (e.g., by reducing one's social status or incurring social debt to one's conversational partner; Van Rooy, 2003) or a separately-motivated notion of plausible deniability (Pinker, Nowak, & Lee, 2008). The kind of utterance cost for the first type of account would necessarily involve higher-order reasoning about other agents, and may be able to be defined in terms of the more basic social and self-presentational goals we formalize here. A separate notion of plausible deniability may not be needed to explain most politeness behavior, either. Maintaining plausible deniability is in one's own self-interest (e.g., due to controversial viewpoints or covert deception) and goes against the interests of the addressee; some amount of utility dis-alignment is presumed by these accounts. Politeness behavior appears present even in the absence of obvious conflict, however: In fact, you might be even more motivated to be polite to someone whose utilities are more aligned with yours (e.g., a friend). In our work here, we show that such behaviors can in fact arise from purely cooperative goals (Brown & Levinson, 1987), though in cases of genuine conflict, plausible deniability likely plays a more central role in communication.

Utility weights and value functions in our model could provide a framework for a quantitative understanding of systematic cross-cultural differences in what counts as polite. Cross-cultural differences in politeness could be a product of different weightings within the same utility structure. Alternatively, culture could affect 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 nonlinear and involve reasoning about the future). 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 language use deviates from purely informational transmission. We flirt, insult, boast, and empathize by balancing informative transmissions with goals to affect others' feelings or present particular views of ourselves. Our work shows how social and self-presentational motives are integrated with informational concerns more generally, opening up the possibility for a broader theory of social language. In addition, a formal account of politeness moves us closer to courteous computation – to machines that can talk with tact.

Supplementary Materials

Model details

The *literal listener* L_0 is a simple Bayesian agent that takes the utterance to be true:

$$P_{L_0}(s|w) \propto [\![w]\!](s) * P(s).$$

where $[\![w]\!](s)$ is the truth-functional denotation of the utterance w (i.e. the utterance's literal meaning): It is a function that maps world-states s to Boolean truth values. The literal meaning is used to update the literal listener's prior beliefs over world states $P(s)$.

The *speaker* S_1 chooses utterances approximately optimally given a utility function, which can be decomposed into two components. First, informational utility (U_{inf}) is the amount of information a literal listener L_0 would still not know about world state s after hearing a speaker's utterance w. Second, social utility (U_{soc}) is the expected subjective utility of the state inferred given the utterance w. The utility of an utterance subtracts the cost $c(w)$ from the weighted combination of the social and epistemic utilities.

$$U(w; s; \phi_{S_1}) = \phi_{S_1} \cdot \ln(P_{L_0}(s \mid w)) + (1 - \phi_{S_1}) \cdot \mathbb{E}_{P_{L_0}(s|w)}[V(s)] - C(w).$$

The speaker then chooses utterances w softmax-optimally given the state s and his goal weight mixture ϕ_{S_1}:

$$P_{S_1}(w \mid s, \phi_{S_1}) \propto \exp(\lambda_1 \cdot \mathbb{E}[U(w; s; \phi_{S_1})]).$$

Literal semantic task

We probed judgments of literal meanings of the target words assumed by our model and used in our main experiment.

Participants. 51 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Design and Methods. 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 3 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.

Behavioral results. 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 6).

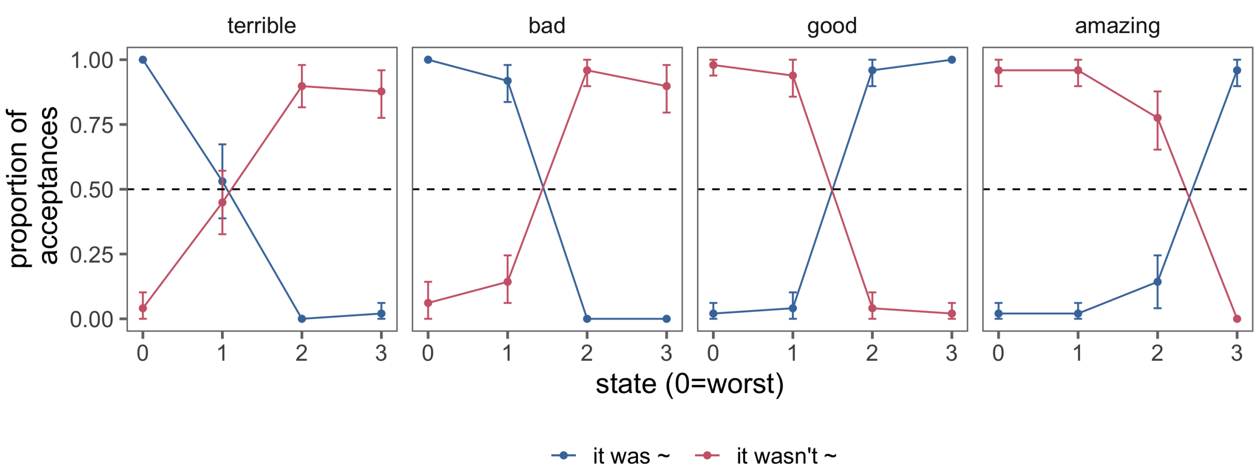

Figure 6. Semantic measurement results. Proportion of acceptances of utterance types (shown in different colors) combined with target words (shown in different facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

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 0.2.2; 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 0.7.7; Wickham, Francois, Henry, & Müller, 2017), *forcats* (Version 0.2.0; Wickham, 2017a), *ggplot2* (Version 3.0.0; Wickham, 2009), *ggthemes* (Version 3.4.0; Arnold, 2017), *gridExtra* (Version 2.3; Auguie, 2017), *here* (Version 0.1; Müller, 2017b), *jsonlite* (Version 1.6; 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 0.2.5; Henry & Wickham, 2017), *RColorBrewer* (Version 1.1.2; Neuwirth, 2014), *Rcpp* (Eddelbuettel & Balamuta, 2017; Version 0.12.19; 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). The full statistics are shown in Table 3.

Model fitting and inferred parameters

Other than speaker goal mixture weights explained in the main text (shown in Table 2), the full model has two global parameters: the speaker's soft-max parameter λ_{S_2} and soft-max

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
ninformational only	1.58	8.58
ninformational, presentational	1.89	2.93
ninformational, social	1.11	3.07
ninformational, social, presentational	2.64	4.47
presentational only	2.58	9.58
social only	1.73	7.23
social, presentational	2.49	5.29

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 (M. D. Lee & Wagenmakers, 2014). We ran 4 MCMC chains for 80,000 iterations, discarding the first 40,000 for burnin. The inferred values of parameters are shown in Table 4.

Data Availability

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

Supplemental Figures

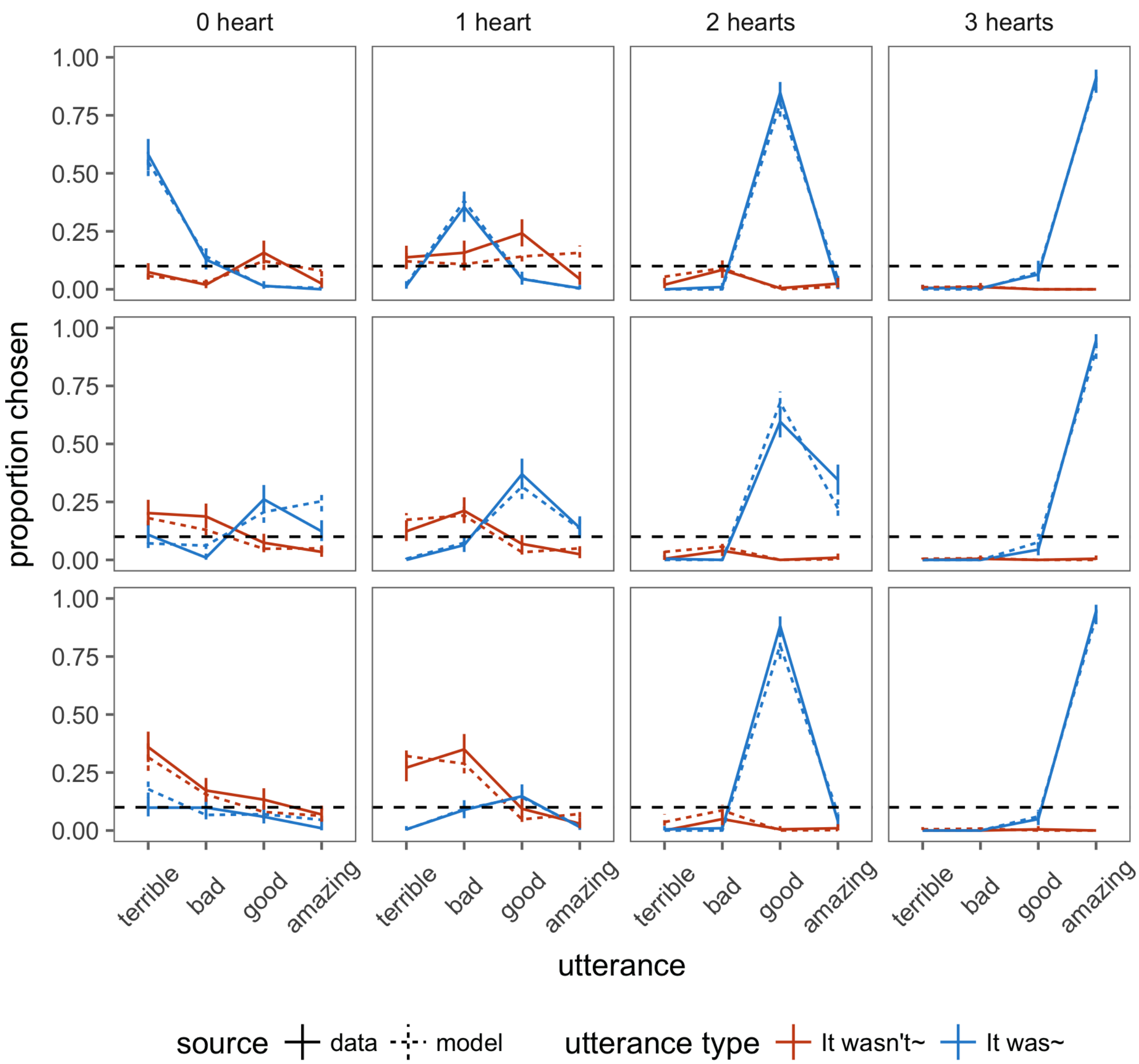

Figure 7. Experimental results (solid lines) and fitted predictions from the full model (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model. Black dotted line represents the chance level.

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

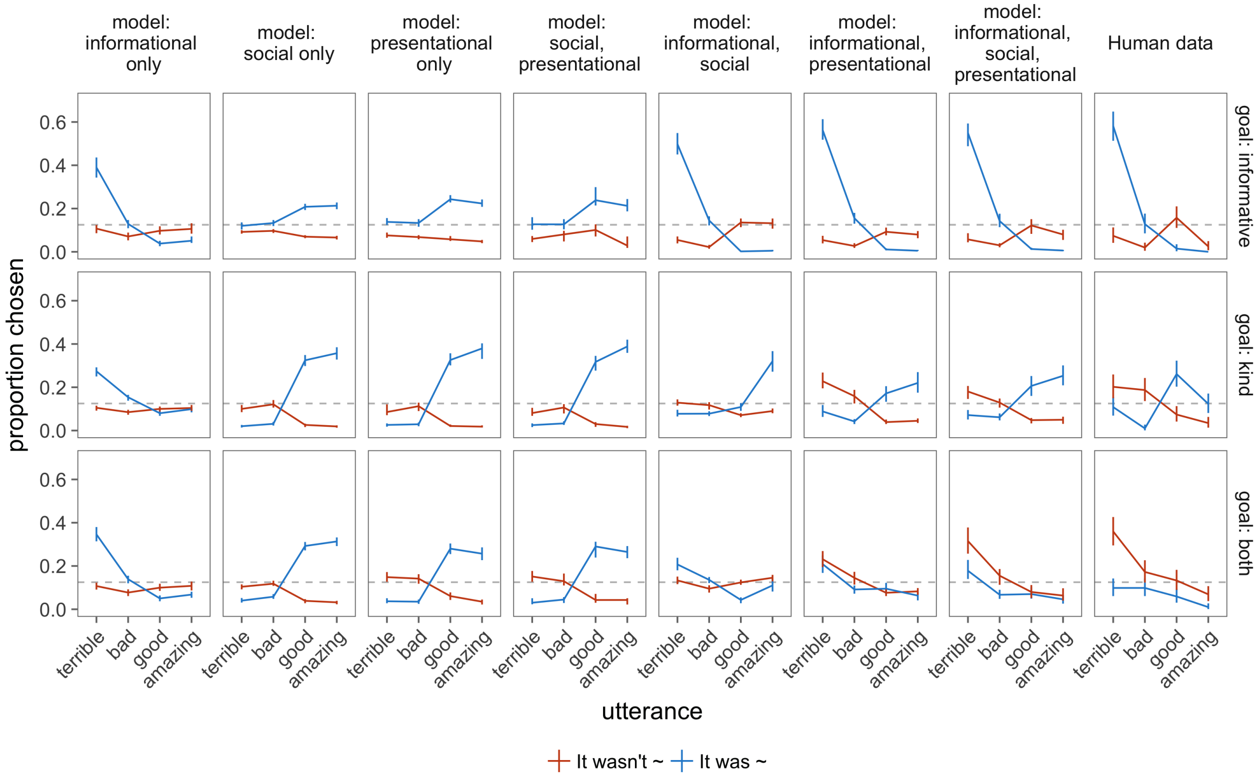

Figure 8. Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart and speaker with a goal to be informative (top), kind (middle), or both (bottom). Gray dotted line indicates chance level at 12.5%.

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 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*

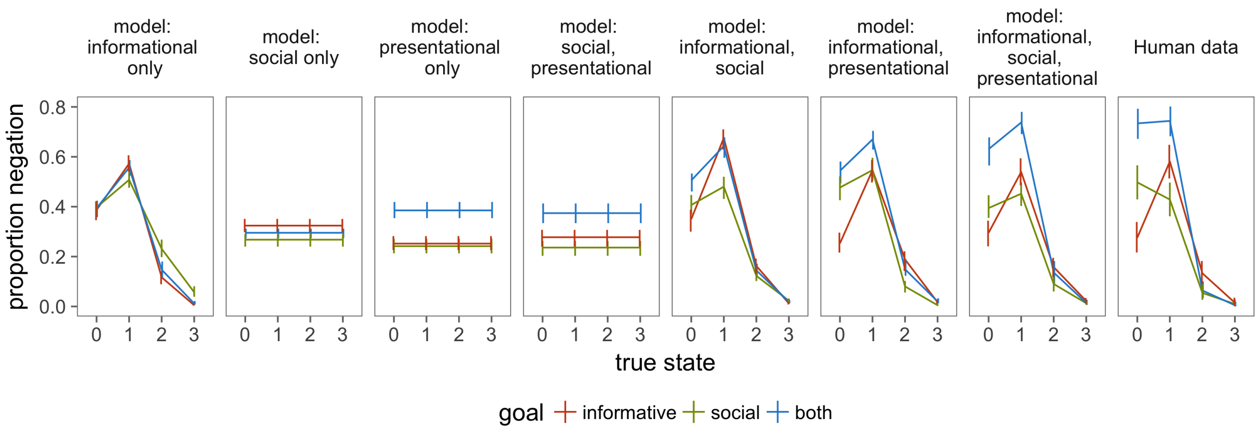

Figure 9. Experimental results (left) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

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

Bonnefon, J.-F., Feeney, A., & De Neys, W. (2011). The risk of polite misunderstandings. *Current Directions in Psychological Science*, *20*(5), 321–324.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

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

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