

# Polite speech emerges from competing social goals

Erica J. Yoon<sup>a,1,2</sup>, Michael Henry Tessler<sup>a,1</sup>, Noah D. Goodman<sup>a</sup>, and Michael C. Frank<sup>a</sup>

<sup>a</sup>Department of Psychology, Stanford University

This manuscript was compiled on October 24, 2018

**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: convey information, be kind, and present oneself in a good light. We formalize this tradeoff using a probabilistic model of utterance production, which predicts human utterance choices in socially-sensitive situations with high quantitative accuracy. This utility-theoretic approach to speech acts takes a step towards explaining the richness and subtlety of social language use.**

politeness | computational modeling | communicative goals | pragmatics

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 (1–4). On the classical view, transfer ought to be efficient and accurate: Speakers are expected to choose succinct utterances to convey their beliefs (5, 6), 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 of a language spontaneously produce requests in polite forms (7), and adults use politeness strategies while arguing (8).

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 (9), social norms (10), or aspects of a speaker and/or listener’s identity, known as *face* (11, 12). 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. How this socially-aware calculation unfolds, however, is not well understood. When should a speaker decide to say something false (*Your poem was great!*) 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 sav-

ing the listener’s face; and self-presentational utility—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 (2, 13). 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 model class has provided a quantitative understanding of a wide variety of complex linguistic behaviors, including vagueness (14), hyperbole (15), and irony (16), 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 (17), a hypothesis which has found empirical support in a wide variety of work with both adults and children (e.g., (18, 19)).

RSA models are defined recursively such that speakers  $S$  reason about listeners  $L$ , and 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

## Significance Statement

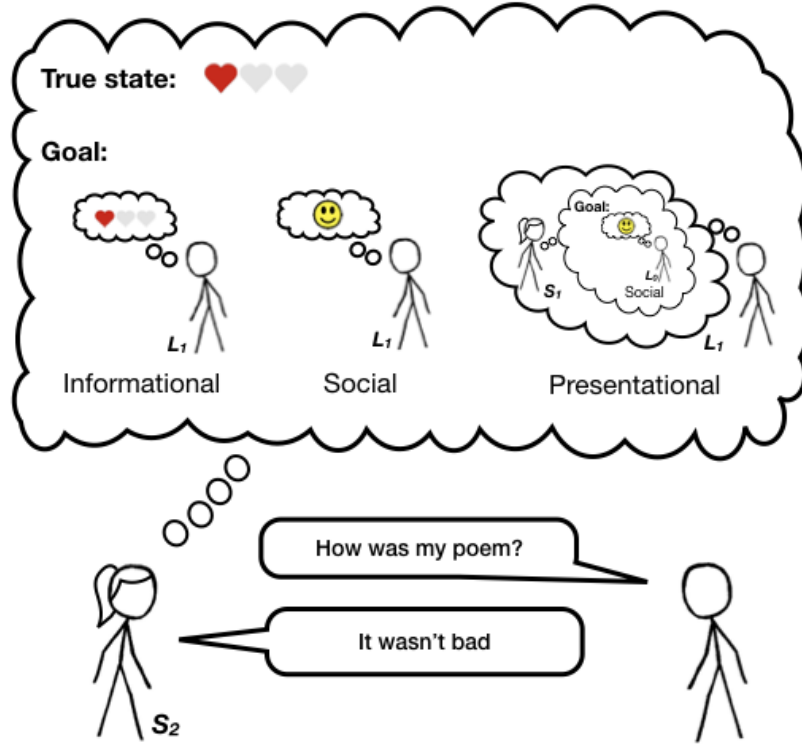
Standard views of language use emphasize its role in information transfer. On these views, the ubiquity of politeness is a puzzle. We present a new quantitative viewpoint on social language use that resolves the puzzle by assuming that speakers balance three potentially conflicting goals: an informational goal (*be informative*), a social goal (*be nice*), and a self-presentational goal (*look informative/nice*). This formal work provides a framework for developing a quantitative understanding of social language use more broadly. Further, by providing a theory of how speakers consider the feelings of others in formulating their utterances, it paves the way for the development of courteous computational agents.

Please provide details of author contributions here.

Please declare any conflict of interest here.

<sup>1</sup>E.J.Y. and M.H.T. contributed equally to this work.

<sup>2</sup>To whom correspondence should be addressed. E-mail: ejyoon@stanford.edu



**Fig. 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.

to the literal meanings of words (rather than their pragmatic implications), and hence grounds the recursion (see Materials and Methods for definitions of  $L_0$  and  $S_1$ ). 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 on its ability to predict human utterance choices in situations where polite language use is expected to varying degrees. 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  $\lambda_{S_2}$ ; (20)):

$$P_{S_2}(w|s, \hat{\phi}) \propto \exp(\lambda_{S_2} \cdot \mathbb{E}[U_{total}(w; s; \hat{\phi}; \phi_{S_1})]). \quad [1]$$

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 component utilities minus the utterance cost  $C(w)$ :

$$U_{total}(w; s; \hat{\phi}; \phi_{S_1}) = \phi_{inf} \cdot U_{inf}(w; s) + \phi_{soc} \cdot U_{soc}(w) + \phi_{pres} \cdot U_{pres}(w; \phi_{S_1}) - C(w). \quad [2]$$

We define *social utility*  $U_{soc}(w)$  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), they might try to infer the weighting (e.g., "was she just being nice?"). But then a sophisticated speaker can produce utterances in order to appear *as if* they had certain goals in mind (i.e., a self-presentational goal). 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* ( $U_{pres}$ ) and is the most novel component of our model. The speaker gains presentational utility when her listener believes she has particular goals, represented by a mixture weighting  $\phi_{S_1}$  between trying to be genuinely informative vs. kind. Formally,

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

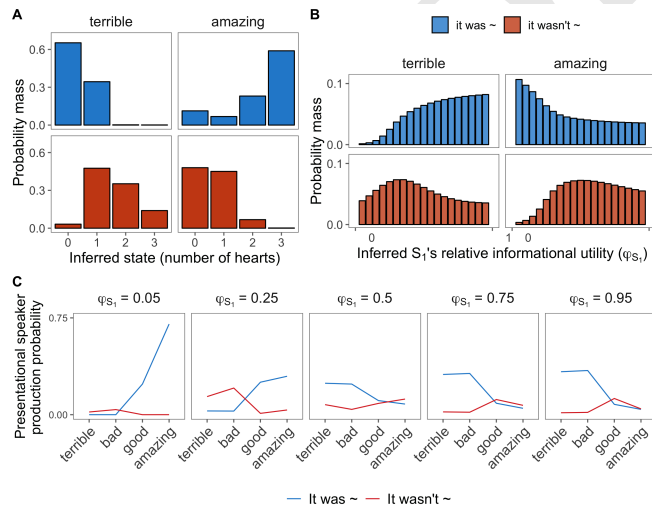
The speaker conveys a particular weighting of informational vs. social goals ( $\phi_{S_1}$ ) by considering the beliefs of listener  $L_1$ , who hears an utterance and jointly infers the speaker's utilities and the true state of the world:

$$P_{L_1}(s, \phi_{S_1} | w) \propto P_{S_1}(w | s, \phi_{S_1}) \cdot P(s) \cdot P(\phi_{S_1}). \quad [4]$$

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  $\phi_{S_1}$ . Definitions for  $S_1$  and  $L_0$  otherwise mirror those of  $S_2$  and  $L_1$  and can be found in the Materials and Methods 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 (and this cost differential is 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 poem the speaker is commenting on), represented in terms of numbers of hearts (Figure 1):  $S = s_0, \dots, s_3$ . Since the rating scale is relatively abstract, we assume a uniform prior distribution over possible states of the world. The set of utterances is  $\{\text{terrible}, \text{bad}, \text{good}, \text{amazing}, \text{not terrible}, \text{not bad}, \text{not good}, \text{and not amazing}\}$ . We implemented this model using the probabilistic programming language WebPPL (21) and a demo can be found at <http://forestdb.org/models/politeness.html>.



**Fig. 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 ( $\phi_{S_1}$  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 vs. social as represented by  $\phi_{S_1}$ ).

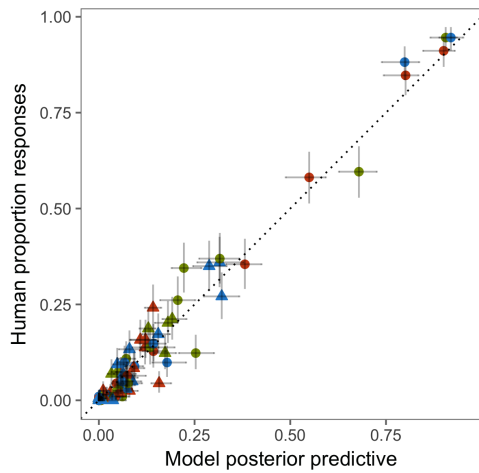
## Results

**Model behavior.** The pragmatic listener model  $L_1$  draws complex inferences about both the true state of the world (Figure 2A) and the speaker's goals (Figure 2B). Upon hearing *[Your poem] was terrible* (Fig. 2 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 (Fig. 2 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 (Fig. 2 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 (Fig. 2 bottom-left).

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  $\phi_{S_1}$  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.,  $\phi_{S_1}$  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.

**Experimental findings.** 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 ( $N = 202$ ). Participants read scenarios with information on the speaker's (Ann's, in our example) 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?"*

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 include a self-presentational component (i.e., some mixture of the three utilities; see Supplementary Information). 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*.



goal ● informative ● kind ● both utterance type ○ It was ~ △ It wasn't ~

**Fig. 3.** 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).

**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 |

Our primary behavioral hypothesis was that speakers describing bad states (e.g., Bob's poem deserved 0 hearts) with goals to be both informative and kind would produce more indirect, negative utterances (e.g., *It wasn't terrible*). Such indirect speech acts both save the listener's face and provide some information about the true state, and thus, are the kind of thing that a socially-conscious speaker would say (Figure 2). This prediction was confirmed: a Bayesian mixed-effects model predicts more negation as a function of true state and goal via 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.

The model parameters (softmax parameters and each goal condition's utility weights) can be inferred from the behavioral data using a Bayesian data analysis model ((22); see Supplementary Information) 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 ( $N=51$ ). The posterior predictions from the the three-utility polite speaker model

**Table 2. Maximum A-Posteriori (MAP) estimates of phi parameters from all model variants with more than one utility.**

| Model (utilities) | goal        | $\phi_{inf}$ | $\phi_{soc}$ | $\phi_{pres}$ | $\phi_{S1}$ |
|-------------------|-------------|--------------|--------------|---------------|-------------|
| inf, soc, pres    | both        | 0.36         | 0.11         | 0.54          | 0.36        |
| inf, soc, pres    | informative | 0.36         | 0.02         | 0.62          | 0.49        |
| inf, soc, pres    | social      | 0.25         | 0.31         | 0.44          | 0.37        |
| inf, pres         | both        | 0.64         | —            | 0.36          | 0.17        |
| inf, pres         | informative | 0.77         | —            | 0.23          | 0.33        |
| inf, pres         | social      | 0.66         | —            | 0.34          | 0.04        |
| inf, soc          | both        | 0.54         | 0.46         | —             | —           |
| inf, soc          | informative | 0.82         | 0.18         | —             | —           |
| inf, soc          | social      | 0.39         | 0.61         | —             | —           |
| soc, pres         | both        | —            | 0.38         | 0.62          | 0.55        |
| soc, pres         | informative | —            | 0.35         | 0.65          | 0.75        |
| soc, pres         | social      | —            | 0.48         | 0.52          | 0.66        |

(informational, social, presentational) showed a very strong fit to participants' actual utterance choices ( $r^2(96) = 0.97$ ; Figure 3). We compared these to six model 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 participants' preference for negation when the speaker wanted to be informative and kind about truly bad states, as hypothesized (Figure 4). In sum, the full collection of informational, social, and presentational were required to fully explain participants' utterance choices.

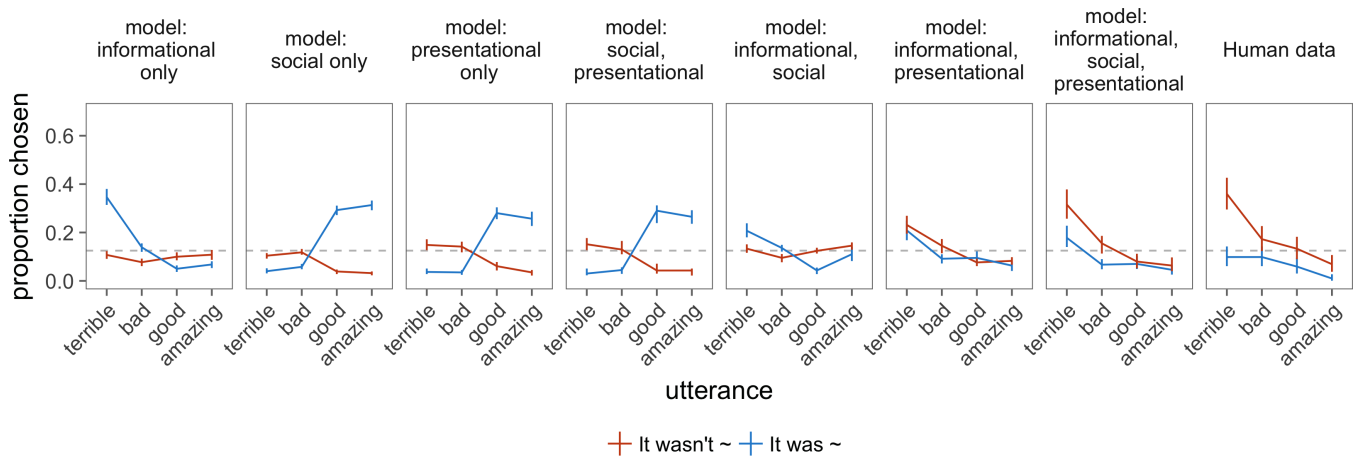
The utility weights inferred for the three-utility model (Table 2) provide additional insight into how polite language use operates in our experimental context and possibly beyond: *Being kind* requires equal weights on all three utilities, indicating that Gricean informativity is a part of language 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, suggesting that managing the listener's inferences about oneself was integral to participants' 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 that used 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





**Fig. 4.** 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 dashed line indicates chance level at 12.5%.

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 (13)).

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 (11) 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 (23, 24) 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; (25)) or a separately-motivated notion of plausible deniability (26). 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 (11), though our account is not inconsistent with the invocation of plausible deniability accounts in cases of conflict.

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 that 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.

## Materials and Methods

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

$$P_{L_0}(s|w) \propto [w](s) * P(s). \quad [5]$$

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  $P(s)$ .

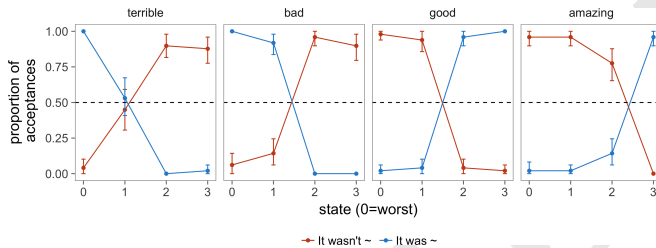
The *speaker*  $S_1$  chooses utterances approximately optimally given a utility function, which can be decomposed into two components. First, informational utility ( $U_{inf}$ ) is the amount of information a literal listener  $L_0$  would still not know about world state  $s$  after hearing a speaker's 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.

$$U(w; s; \phi_{S_1}) = \phi_{S_1} \cdot \ln(P_{L_0}(s | w)) + (1 - \phi_{S_1}) \cdot \mathbb{E}_{P_{L_0}(s|w)}[V(s)] - C(w). \quad [6]$$

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

$$P_{S_1}(w | s, \phi_{S_1}) \propto \exp(\lambda_1 \cdot \mathbb{E}[U(w; s; \phi_{S_1})]) \quad [7]$$

**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 6 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 5).



**Fig. 5.** 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.

Imagine that Fiona filmed a movie, but she didn't know how good it was. Fiona approached Yvonne, who knows a lot about movies, and asked "How was my movie?"

Here's how Yvonne **actually** felt about Fiona's movie, on a scale of 0 to 3 hearts:



If Yvonne wanted to **BOTH** make Fiona feel good **AND** give accurate and informative feedback,

what would Yvonne be most likely to say?

"It wasn't terrible"

**Fig. 6.** Example of a trial in the speaker production task.

**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, 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 6). 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; Figure 6). Each participant read 12 scenarios, depicting every possible combination of the 3 goals and 4 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, procedure, data, analyses and their preregistration are available at [https://github.com/ejyoon/polite\\_speaker](https://github.com/ejyoon/polite_speaker).

**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.

- Bühler K (1934) *Sprachtheorie*. (Oxford, England: Fischer).
- Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. *Science* 336(6084):998–998.
- Jakobson R (1960) Linguistics and poetics in *Style in language*. (MA: MIT Press), pp. 350–377.
- Shannon CE (1948) A mathematical theory of communication. *Bell Syst. Tech. J.* 27:623–656.
- Grice HP (1975) *Logic and conversation*, eds. Cole P, Morgan JL. (Academic Press) Vol. 3, pp. 41–58.
- Searle J (1975) *Indirect Speech Acts*, eds. Cole P, Morgan JL. (Academic Press) Vol. 3, pp. 59–82.
- Axia G, Baroni MR (1985) Linguistic politeness at different age levels. *Child Development* pp. 918–927.
- Holtgraves T (1997) Yes, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology* 16(2):222–239.
- Leech G (1983) *Principles of pragmatics*. (London, New York: Longman Group Ltd.).
- Lee 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.
- Brown P, Levinson SC (1987) *Politeness: Some universals in language usage*. (Cambridge university press) Vol. 4.
- Goffman E (1967) *Interaction ritual: essays on face-to-face interaction*. (Aldine).
- Goodman ND, Frank MC (2016) Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11):818–829.
- Lassiter D, Goodman ND (2017) Adjectival vagueness in a bayesian model of interpretation. *Synthese* 194(10):3801–3836.
- Kao JT, Wu JY, Bergen L, Goodman ND (2014) Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33):12002–12007.
- Kao JT, Goodman ND (2015) Let's talk (ironically) about the weather: Modeling verbal irony in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
- Baker CL, Saxe R, Tenenbaum JB (2009) Action understanding as inverse planning. *Cognition* 113(3):329–349.
- Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB (2016) The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20(8):589–604.
- Liu S, Ullman TD, Tenenbaum JB, Spelke ES (2017) Ten-month-old infants infer the value of goals from the costs of actions. *Science* 358(6366):1038–1041.
- Goodman ND, Stuhlmüller A (2013) Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* 5(1):173–184.
- Goodman ND, Stuhlmüller A (2014) The Design and Implementation of Probabilistic Programming Languages (<http://dippl.org>).
- Lee MD, Wagenmakers EJ (2014) *Bayesian Cognitive Modeling: A Practical Course*. (Cambridge Univ. Press).
- Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4):0064.
- Hamlin KJ, Ullman TD, Tenenbaum JB, Goodman ND, Baker CL (2013) The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science* 16(2):209–226.
- 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*. (ACM), pp. 45–58.
- Pinker S, Nowak MA, Lee JJ (2008) The logic of indirect speech. *Proceedings of the National Academy of sciences* 105(3):833–838.