"I won't lie, it wasn't amazing": Modeling polite indirect speech

Erica J. Yoon, **Michael Henry Tessler**, **Noah D. Goodman** and **Michael C. Frank**
{ejyoon, mtessler, ngoodman, mcfrank} @stanford.edu
Department of Psychology, Stanford University

Abstract

Why are we polite when we talk to one another? One hypothesis is that people expect others to choose what to say based on their goals both to transfer information efficiently (an epistemic goal) and to make the listener feel good (a social goal). In our previous work, we found that when these two goals conflict, they sometimes produce white lies. In the current work, we expand on this theory to consider another prominent case of polite speech: indirect remarks using negation (e.g., "It wasn't amazing"). With minimal extensions from our previous framework, our formal model suggests that a pragmatic speaker will produce more indirect remarks when the speaker wants to be informative and seem considerate at the same time. These predictions were borne out in a language production experiment. These findings suggest that the conflict between social and epistemic goals can account for a broad range of politeness phenomena.

Keywords: Politeness; computational modeling; communicative goals; pragmatics

Introduction

Language users hear and produce *polite speech* on a daily basis. Adults and even young children spontaneously produce requests in polite forms (Axia & Baroni, 1985; Clark & Schunk, 1980), and speakers use politeness strategies even while arguing, preventing unnecessary offense to their interactants (Holtgraves, 1997). But being polite conflicts with one important goal of cooperative communication: exchanging information efficiently and accurately (Grice, 1975). People tell white lies ("Your new dress is gorgeous!") and produce indirect speech that is longer and more nuanced than the simplest form of their intended message ("I don't think that dress looks phenomenal on you" as opposed to "It looks terrible") to make others feel good. Speakers risk potential loss of their intended message (indirect speech), intentionally convey wrong information (lies), and suffer inefficiencies – all in the service of being polite. If information transfer were the only currency in communication, politeness would be both infelicitous and undesirable.

A *cooperative speaker*, however, can be imagined as one with both an epistemic goal to improve the listener's knowledge as well as a social goal to minimize potential damage to the hearer's (and the speaker's own) self-image, called *face* (Brown & Levinson, 1987). If the speaker's intended meaning contains no threat to the speaker or 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.

Inspired by this set of ideas, we have argued that listeners think about polite speech as reflecting a tradeoff between two goals: information transfer (which we called *epistemic utility*) and face-saving (*social utility*; Yoon, Tessler, Goodman, & Frank, 2016). A speaker with a high weight on social utility will try to save her listener's face: She hides or risks losing information in her intended message by making her utterance false to some degree. On the other hand, a speaker with a high weight on epistemic utility prioritizes truthfulness and informativity, and she may risk a loss of the listener's (or the speaker's own) face. These ideas were formalized in a model of pragmatic language understanding, building on the Rational Speech Act (RSA) theory (for a review, see Goodman & Frank, 2016). We tested the polite RSA model (pRSA) by examining white lies. The model captured human participants' inferences about a speaker's goals given her utterance (e.g., saying a *good* talk was "amazing" implies that she is being nice) and about the true state of the world given a speaker's goal (e.g., saying "good" may mean the talk was only *okay* if the speaker wanted to be nice).

In the current work, we extend our framework to another polite speech act: *indirect speech*. White lies are produced when a speaker tries to save the listener's face by stretching the truth. But instead of lying, people sometimes try to be polite by being more indirect. Through indirect speech, a speaker can express meaning that is different from the literal meaning of the utterance (Searle, 1975). In this work, we focus on negation ("not"), which has the potential to be indirect. For instance, "Mark isn't the cleanest person I know" may suggest that the speaker thinks Mark is *unclean* (inferred meaning) rather than not being the person who has the greatest degree of cleanliness (literal meaning). Negation can be used as a hedging or mitigating device to address an undesirable state that is face-threatening to the addressee (Brown & Levinson, 1987; Grice, 1975).

What may lead a speaker to produce indirect remarks? An indirect remark may be motivated by the speaker's goal to convey some face-threatening information, while being seen as a polite person who avoids threatening others' face. In our previous work, we described a pragmatic listener that jointly inferred the true state and the goals of the speaker. Building on this model, we describe here a speaker whose goal is to lead this pragmatic listener to infer the true state *and* attribute to the speaker certain goals (e.g., face-saving). For instance, "It wasn't amazing" does not preclude the possibility that the presentation was bad, and may in fact be pragmatically strengthened to mean that it was actually bad. Yet because the speaker does not choose the more direct "It was bad", the listener will infer a face-saving goal. Thus saying "It wasn't amazing" can accomplish the goal of conveying that the presentation was bad while the speaker is seen as not wanting to make the listener feel bad. On the other hand, if the speaker does not care about being seen as face-saving,

she will produce less indirect speech. Further, if the presentation was actually good, or even decent, the speaker will prefer to produce a directly positive remark ("It was good") in either case. Thus we predict more indirect speech when the true state is bad, and an interaction with the speaker's desire to both be informative and be seen as wanting to save face. In what follows, we derive our hypotheses using our formal model and present an empirical test of the hypotheses.

Computational Model

In the current work, we introduce a minimal extension to our previous RSA model (pRSA; Yoon et al., 2016) to allow for speaker production of indirect remarks using negation.

Polite RSA

RSA models assume speakers choose utterances approximately optimally given a utility function (Goodman & Stuhlmuller, 2013). pRSA posited that the speaker's utility function can be decomposed into two components. First, *epistemic utility* (U_{epi}) refers to the standard, informative utility in RSA: 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 expected subjective utility is related to the intrinsic value of the state, and we use a value function (V) to map states to subjective utility values. This captures the affective consequences for the listener of being in state s. Finally, some utterances might be costlier than others. The utility of an utterance subtracts the cost $c(w)$ from the weighted combination of the social and epistemic utilities.

$$U(w;s;\hat{\beta}) = \beta_{epi} \cdot \ln(P_{L_0}(s \mid w)) + \beta_{soc} \cdot \mathbb{E}_{P_{L_0}(s|w)}[V(s)] - C(w)$$

The speaker (S_1) in pRSA chooses utterances w softmax-optimally given the state s and his goal weights $\hat{\beta}$. The pragmatic listener (L_1) jointly infers the state s and the utility weights of the speaker, β_{epi} and β_{soc} (Goodman & Lassiter, 2015; Kao, Wu, Bergen, & Goodman, 2014).

$$P_{L_1}(s, \hat{\beta} \mid w) \propto P_{S_1}(w \mid s, \hat{\beta}) \cdot P(s) \cdot P(\hat{\beta}) \quad (1)$$

$$P_{S_1}(w \mid s, \hat{\beta}) \propto \exp(\lambda_1 \cdot \mathbb{E}[U(w;s;\hat{\beta})]) \quad (2)$$

$$P_{L_0}(s \mid w) \propto [\![w]\!](s) \cdot P(s) \quad (3)$$

Within our experimental domain, we assumed there were five possible states of the world corresponding to the value placed on a particular referent (e.g., rating deserved by the presentation the speaker is commenting on, akin to a Yelp rating): $S = \{s_1, ..., s_5\}$. We assume a uniform prior distribution over possible states of the world. The states have subjective numerical values $V(s_i) = \alpha \cdot i$, where α is a free parameter. $[\![w]\!](s)$ corresponds to the lexical meaning of the utterance w (e.g., "good") when applied to state s. We gather independent ratings for these literal meanings.

Extensions to pRSA

We build on pRSA by adding negative utterances and modeling a more sophisticated speaker. First, we extend the utterance alternatives to include negation. Previously we considered five possible utterances: {It was *terrible*, *bad*, *okay*, *good*, and *amazing*}, all direct assertions of specific states (e.g., "It was amazing" would be true for the state of 5 but untrue for the states of 1 or 2). Now the speaker may say, {It *wasn't* terrible, bad, okay, good, and amazing}. These utterances indirectly address the referent by negating certain state. We assume that it is more costly to say utterances with negation, which makes the utterance morphemically longer and is harder to process (Clark & Chase, 1972). In our full data analysis, we put a prior on this negation cost parameters and infer its likely values from the data.

Most importantly, we extended the recursive reasoning in the model. For our experiment, we consider the pragmatic speaker (S_2) who chooses an utterance based on the pragmatic listener model (Eq. 1), thinking about the state as well as goal weights that the pragmatic listener will infer.

$$P_{S_2}(w \mid s, \hat{\beta}) \propto \exp(\lambda_2 \cdot \ln(P_{L_1}(s, \hat{\beta} \mid w)) - C(w))$$

This crucially captures the idea that the speaker both wants to convey the state s, and to be seen as someone with goals $\hat{\beta}$. We simplify from the Yoon et al. (2016) model by including only a single mixture parameter ϕ governing the extent to which the speaker is being informative vs. face saving: $\beta_{epi} = \phi$, $\beta_{soc} = 1 - \phi$.

We implemented this model using the probabilistic programming language WebPPL (Goodman & Stuhlmuller, 2014)[1]. In the next section, we explore the model's predictions for speaker productions of indirect speech with negation vs. direct speech with no negation.

Model predictions

Before describing our experimental data, we derive predictions from the pRSA model. In these initial simulations, we use fixed goal weights and parameters – in later fits, we will derive these parameters from the data using Bayesian data analysis. Since the model requires measurements of literal semantics (e.g., what "not good" means on a given dimension), we first describe these measurements and then give model predictions using them.

Semantic measurement

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

Materials, methods, and results 25 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk. We used 13 different context items that were previously used in Yoon et al. (2016), in which someone

[1] A complete implementation of the model, raw data and analyses, and links to the experiments and pre-registration of hypotheses and method can be found at `https://github.com/ejyoon/cogsci2017`.

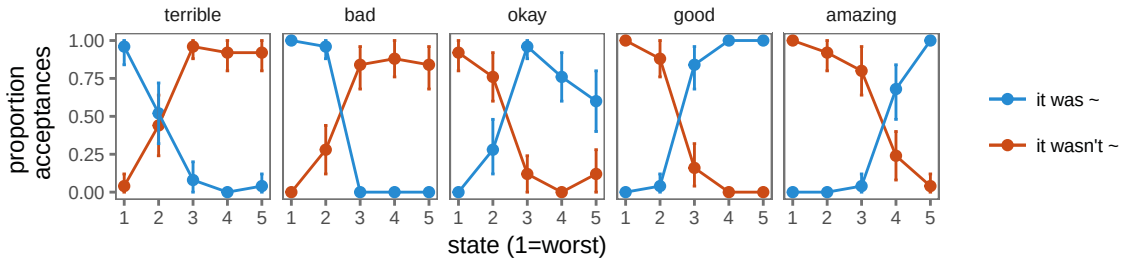

Figure 1: Semantic measurement results. Proportion of acceptances of utterance types (colors) combined with target words (facets) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

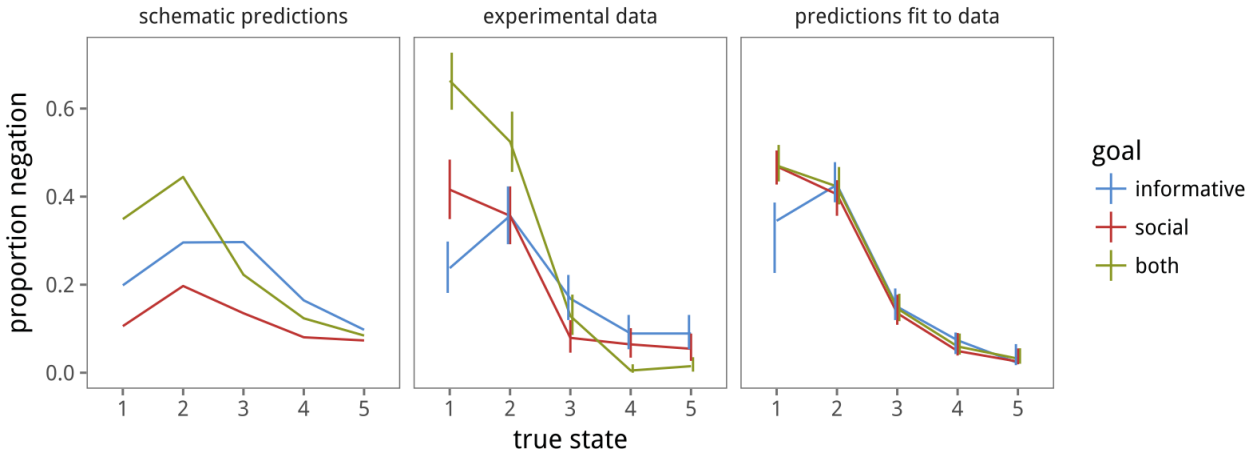

Figure 2: Schematic model predictions (left), experimental results (center) and fitted model predictions (right) for average proportion of negation produced among all utterances, given true states (x-axis) and goals (colors).

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 out of five hearts (e.g., two out of five hearts filled in red color). 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 five possible words: *terrible*, *bad*, *okay*, *good*, and *amazing*, giving rise to ten different possible utterances (with negation or no negation). Each participant read 50 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 subsequent experiments, 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. Meanings of the words as judged by participants were as one would expect (see Figure 1). We used the fraction of participants that endorsed utterance w for state s to set informative priors to infer posterior credible values of the literal meanings from data in the speaker production experiment.

Model parameters and predictions

The S_2 speaker in our model has the goal to convey the state and to be seen as having a particular set of goals. We explore predictions for 3 hypothetical speakers, corresponding to 3 different ϕ mixture parameter weights: (a) an *informative* speaker who wants to convey high epistemic utility (prioritizing information transfer; $\phi = 0.9$) (b) a *social* speaker who wants to convey high social utility (making the listener feel good; $\phi = 0.1$) (c) a *both-goal* speaker who wants to convey a balance between the two utilities ($\phi = 0.5$).[2]

Figure 2 (left) shows the speaker's production probabilities associated with producing an indirect speech act (i.e., an utterance with negation) for the three different speakers as the true state of the world is varied. We see, consistent with our intuition, that indirect speech was relatively more preferred in bad states than in good states. As well, we see higher probability of negation production for the speaker who wants to convey both goals (epistemic and social) relative to each goal

[2]In addition, the model has a few parameters not of theoretical interest. For the purposes of generating model predictions *a priori*, we assign values to these parameters consistent with the previous literature with this class of models: the speaker optimality parameter (λ_1 assigned to 2); the pragmatic speaker optimality parameter (λ_2 to 2); the value scale parameter (α to 1) in the utility function; and the parameter governing the cost of producing a negation ($C(u)$ to 2).

independently. Indirect speech does not convey that much information and so the informative speaker (a) would disprefer it. The social speaker (b) who wants to convey a face-saving goal would tend to signal a better-that-actual state through direct positive remarks. The both-goal speaker produces indirect remarks to avoid direct remarks that are either true but face-threatening, or face-saving but false.

Speaker production experiment

To compare against our model predictions, we measured participants' predictions for the most likely utterance (w) produced by the speaker, given a description of the true state. For example, given that Ann wanted to make Bob feel good but felt that his poem deserved 2 out of 5 hearts, what would she say? We hypothesized that when there was no tradeoff between informativity and face-threat avoidance (i.e., when the addressee's performance was great), speakers would use truthful and face-saving direct remarks ("[Your poem] was amazing") regardless of their described goals. However, when there was a conflict between the epistemic and social goals (i.e., when the addressee's performance was poor), a speaker who tried to convey both goals would use vague indirect remarks ("[Your poem] wasn't terrible") more often than direct face-threatening remarks ("[Your poem] was bad"; preferred by a speaker who only considered the epistemic goal) or direct face-saving remarks ("[Your poem] was good"; preferred by a speaker who wanted to convey only a social goal).

Method

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

Stimuli and Procedure As in the semantics measurements 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. 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., 2 out of 5 hearts). Each participant read 15 scenarios, depicting every possible combination of goals and 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 *was* vs. *wasn't* and the other for choosing among *terrible*, *bad*, *okay*, *good*, and *amazing* (see Figure 3).

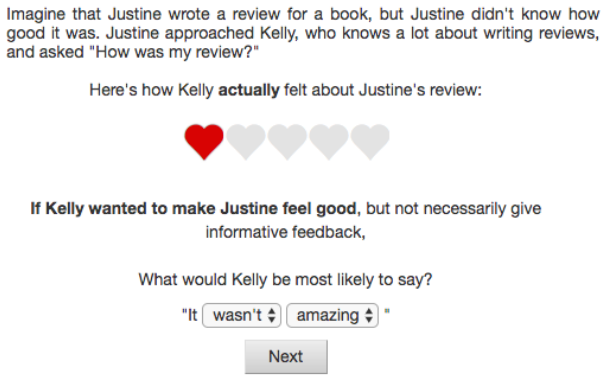

Figure 3: Example of a trial in Experiment 1.

Behavioral results

Our hypotheses for utterance production by speakers with different goals were borne out (see full results in Figure 4).
For good states (4 and 5 hearts), positive direct remarks were judged to be the most likely utterances across all three goal conditions. For less-than-perfect, but still decent states, there was a greater degree of expectation of white lies (e.g., "It was amazing" for 4 hearts) given a social goal. For bad states (1 and 2 hearts), as predicted, there were more instances of expected indirect remarks overall across all goal conditions given bad states. Critically, speakers with both informative and social goals produced more indirect remarks than were observed in the other two goal conditions (Figure 2, center).

Model results

Model fitting In this experiment, participants were told what speakers' intentions were (e.g., wanted to make Bob feel good). We assume that the intention descriptions conveyed the weight mixture ϕ that the speaker was using. We put uninformative priors on this mixture ($\phi \sim \text{Uniform}(0,1)$) and inferred their credible values separately for each goal condition ("wanted to X") using Bayesian data analytic techniques (Lee & Wagenmakers, 2014). We also used the fraction of participants that endorsed utterance w for state s to set informative priors to infer posterior credible values of the literal meanings from data.

There were four additional parameters of the model, on which we put uninformative priors: the speaker optimality parameter ($\lambda_{S_1} \sim \text{Unif}(0,20)$); the pragmatic speaker optimality parameter ($\lambda_{S_2} \sim \text{Unif}(0,5)$); the value scale parameter ($\alpha \sim \text{Unif}(0,5)$) in the utility function; and the cost parameter ($C(u) \sim \text{Unif}(1,10)$). We inferred their posterior credible values from the data. We ran 4 MCMC chains for 80,000 iterations, discarding the first 40,000 for burnin. The Maximum A-Posteriori (MAP) estimate and 95% Highest Probability Density Interval (HDI) were: λ_{S_1}: 2.16 [2.02, 3.61]; λ_{S_2}: 0.91 [0.83, 1.75]; α: 2.71 [0.98, 4.59]; $C(w)$: 2.04 [1.95, 2.25]. To generate utterance predictions, given our model and the inferred parameters, we evaluated the posterior predictive distribution, marginalizing out all parameters.

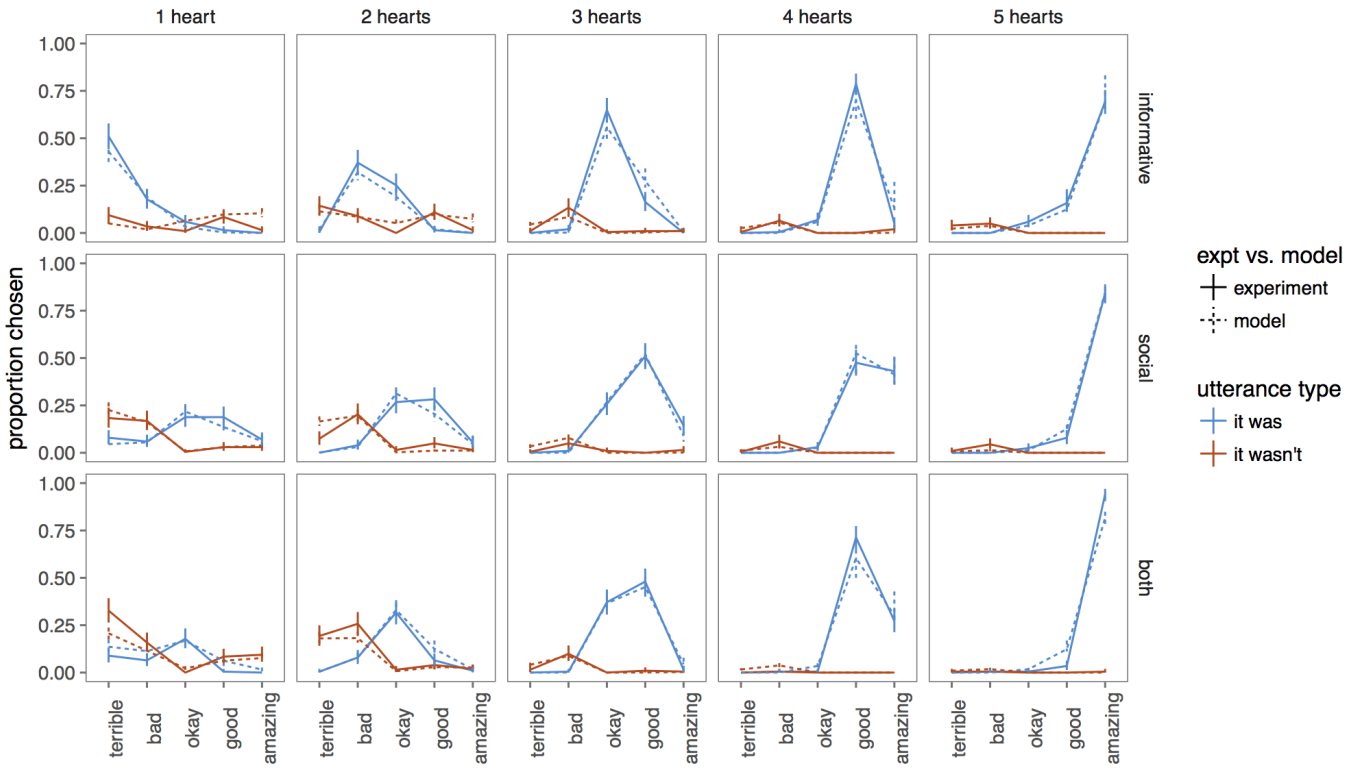

Figure 4: Experimental results (solid lines) and fitted model predictions (dashed lines) for speaker production. Proportion of utterances chosen (utterance type – direct vs. indirect – in different colors and words shown on x-axis) given the true states (columns) and speaker goals (rows). Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model.

Results The inferred weights for each goal condition were largely as expected: For the "wanted to give informative feedback" (*informative*) condition, the model put a relatively high weight on epistemic utility (0.81). For the "wanted to make [listener] feel good" (*social*) condition, the model inferred that the speaker was using a moderate weight on epistemic utility (0.51). For the "wanted BOTH to make [the listener] feel good and give informative feedback" (*both*) condition, the model assigned a weight on epistemic utility between the weights for the other two goal conditions (0.57). Overall, the weights tended to be more biased towards prioritizing the epistemic utility.

The predictions for the speaker's utterance were overall highly consistent with the experimental findings (Figure 4). The posterior predictive of the model explained almost all of the variance in the production data $r^2(150) = 0.962$ (Figure 5). The model successfully predicted distinct patterns for each goal condition. The *informative* speaker produced direct remarks whose literal meanings mapped onto the true states (e.g., "It was terrible" given 1 heart). The *social* speaker produced remarks that were positively biased compared to the true states (e.g., "It was okay" given 2 hearts).

While the model in the *both* condition did produce indirect utterances (e.g., "It wasn't terrible" given 1 heart) it did so slightly less than the empirical data. For this reason, the model did not yield the expected difference for negation production between both-goal and social conditions (Figure 2, right); though the trend was numerically correct, the effect was much smaller in the fit model than the schematic one. There are several possible explanations for this small deviation. In our experimental data, the social speaker placed a higher weight on epistemic utility than in our schematic predictions. Thus, the particular goal descriptions we used in the experiment may have suggested that the social speaker still wanted to be seen as informative, and have led to little differentiation between the social vs both-goal speaker. Another possible cause is that participants preferred a *different kind* of indirect speech than the model – in particular, the both-goal speaker preferred to produce "It wasn't amazing" in the schematic model predictions, whereas participants in our experiment chose "It wasn't terrible." This discrepancy between the two remarks is interesting, because their implied meaning is similar. In a pilot experiment where participants were asked to infer the true state (number of hearts) from an utterance, "It wasn't amazing" and "It wasn't terrible" were very similar (˜2 hearts).

Discussion

Why are we polite? Here we explored a formal instantiation of the hypothesis that two conflicting speaker goals – epis-

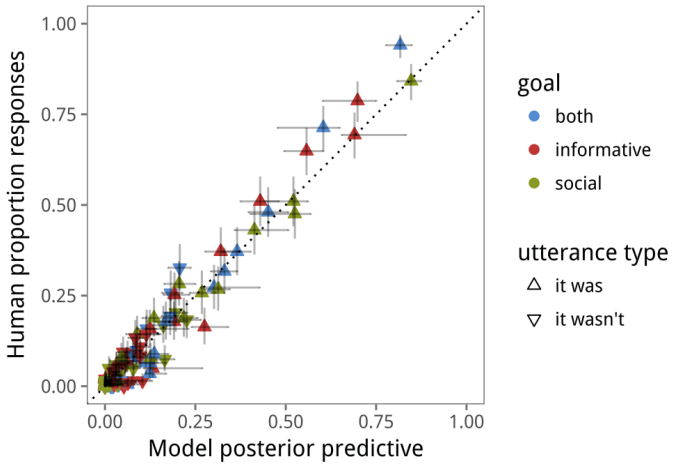

Figure 5: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

temic and social – can be used to explain a range of polite behavior, including white lies and indirect speech acts using negation. Our model predicted that speakers should produce more indirect remarks in cases of greater face threat (given the addressee's poorer performance) and in cases where speakers wanted to be both informative and nice. Our experimental data confirmed these predictions. The model's overall fit to the data was very strong, although it did not show the predicted dominance of indirect speech for the both-goal speaker at low states. Whether this discrepancy between the initial and data-fitted predictions was due to variation in goal weight based on experimental scenarios or a discrepancy in preferences for particular utterances is a question for future work.

An important contribution of this work is in showing the generalizability of our formal model (pRSA) to the case of indirect speech acts. The current work took a step in addressing speakers' self-presentation: Not only do speakers want to save the listener's face, but they also want to save their *own* face, by appearing informative and considerate to the listener. In future work we hope to explore this aspect more and test how our model's utilities can be extended to capture the speaker's desire to appear polite, genuine, and even modest. Using the model to explore other kinds of polite speech such as indirect requests ("Would you mind closing the window?"; Clark & Schunk, 1980) and manifestations of polite speech in different cultures (e.g., Holtgraves & Joong-nam, 1990) are also important future directions.

In sum, our formal model and experimental work represent an advance in polite speech understanding. With a minimal extension to our existing model, we were able to capture subtle patterns in people's inferences about indirect speech production. Our empirical findings suggest that neither epistemic nor social motives alone motivate indirect speech; instead, the need for indirect speech results from the conflict between these two. These findings provide strong support for a utility-theoretic framing of politeness, and suggest new directions in understanding of pragmatic language use in social contexts.

Acknowledgments

This work was supported by NSF grant BCS 1456077 to MCF, ONR grant N00014-13-1-0788 to NDG, NSF Graduate Research Fellowship DGE-114747 to MHT, and NSERC PGS Doctoral scholarship PGSD3-454094-2014 to EJY.

References

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child Development*, 918–927.

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

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517.

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

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

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory, 2nd edition*. Wiley-Blackwell.

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

Goodman, N. D., & Stuhlmuller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. `http://dippl.org`.

Grice, H. P. (1975). Logic and conversation. In *Readings in language and mind*. Blackwell.

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

Holtgraves, T., & Joong-nam, Y. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of Personality and Social Psychology*, *59*(4), 719.

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.

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

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), *Syntax and semantics (vol. 3): Speech acts*. New York: Academic Press.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the thirty-eighth annual conference of the Cognitive Science Society*.