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: 5500

Polite speech emerges from competing social goals

Introduction

We don't always say what's on our minds. Although "close the window!" could be sufficient, we dawdle, adding "can you please...?" or "would you mind...?" Rather than tell an uncomfortable truth, socially-aware speakers exaggerate ("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 (Buhler, 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 pervasively – even while arguing (Holtgraves, 1997), and 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 speaker behavior is 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, when interpreting utterances in face-threatening situations, listeners readily assume that speakers intend to be polite (Bonnefon, Feeney, & Villejoubert, 2009). How this socially-aware calculation unfolds, however, is not well understood. Adopting an example from Bonnefon et al. (2009), when should a speaker decide to say something false ("Your poem was great!"

said of an actually-mediocre poem) rather than to tell the truth ("Your poem was bad") or to be indirect ("Some of the metaphors were tricky to understand.")? How do the speaker's goals enter into the calculation?

We propose a utility-theoretic solution to the problem of understanding polite language, in which speakers choose their utterance by attempting to maximize utilities that represent competing communicative goals. Under the classic pragmatic view of language production, speakers want to be informative and convey accurate information as efficiently as possible (Goodman & Frank, 2016; Grice, 1975); this desire for informative and efficient communication we call *informational utility*. In addition, speakers may want to be kind and make the listener feel good (i.e., save the listener's face), for example, by stating positive remarks about the listener. We call this goal to be *prosocial utility*.

If a speaker wanted to be informative and kind, then she would ideally produce utterances that satisfy both goals. The nuances of reality, however, can make it difficult to satisfy both goals. In particular, when the true state of the world is of low value to the listener (e.g., the listener's poem was terrible), informational and prosocial goals pull in opposite directions. Informational utility could be maximized by stating the blunt truth ("your poem was terrible.") but that would very likely hurt the listener's feelings and threaten the listener's self-image (low prosocial utility); prosocial utility could be maximized through a white lie ("your poem was amazing"), but it would be misleading (low informational utility). In such situations, it seems impossible to be both truthful and kind. A first contribution of our work here is to formalize the details of this tradeoff so that it can make contact with experimental data.

A second contribution of our work is to develop and test a new theoretical proposal, namely, that speakers may use indirect speech to present themselves positively when they are caught by the conflict between truthfulness and kindness. We propose that speakers may navigate their way out of this conflict by signalling to the listener that they care about both of the goals even though they are genuinely unable to fulfill them. We formalize this notion

of *self-presentational utility* and show that it leads speakers to prefer more indirect speech, namely utterances that provide less information relative to their length than alternatives with a similar meaning.

We look at indirect speech in this paper through negated adjectival phrases (e.g., "It wasn't bad"). The relationship between negation and politeness is a topic of long-standing interest to linguists and psychologists (Bolinger, 1972; Horn, 1989; Stern, 1931; Stoffel, 1901). Comprehending negation, as a logical operation, can be psychologically more complex than comprehending an unnegated assertion; negating assertions can result in difficulty in processing (Clark & Chase, 1972; see Nordmeyer & Frank, 2014 for an underlying pragmatic explanation) as well as failure to recognize or recall the asserted content (Lea & Mulligan, 2002; MacDonald & Just, 1989). Our interest in negation, however, is for its information-theoretic properties: Negating an assertion that has a specific meaning results in a meaning that is less precise and lower in informativity (e.g., negating "Alex has blue eyes" results in the statement that Alex has eyes that are some color other than blue"). In our paradigm, we use negation as a way of turning a relatively direct statement ("It was terrible") into an indirect statement ("It wasn't terrible") whose interpretation includes some possibilities that are consistent with or close to the unnegated statement (i.e., the poem was not terrible, but it was still pretty bad).

Multifactorial, verbal theories – like previous proposals regarding politeness – are very difficult to relate directly to behavioral data. Therefore, to test our hypotheses about the factors underlying the production of polite language (what we refer to as its utility structure), we take a model comparison approach. We do this by formalizing the trade-off between different combinations of speakers' utilities in a class of probabilistic models of language use (the Rational Speech Act (RSA) framework; Frank and Goodman (2012); Goodman and Frank (2016)), with a particular focus on models with and without the self-presentational utility. In this framework, 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. These models build 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). Indeed, this class of pragmatic language models has been productively applied to understand 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.

Model

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. 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 (Figure 1, top). 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 some combination of informational, social, and self-presentational goals (Figure 1, bottom).

We evaluate our model's ability to predict human speaker production behavior in situations where polite language use is expected. Our experimental context involves a speaker ("Ann") responding to the request of their listener ("Bob") to evaluate the listener's (Bob's) creative product. For instance, 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 $\boldsymbol{\omega}$, that determines how much Ann prioritizes each of the three possible goals, as well as a goal weight to convey to the listener ϕ; more details below). Following standard practice in RSA models, Ann's

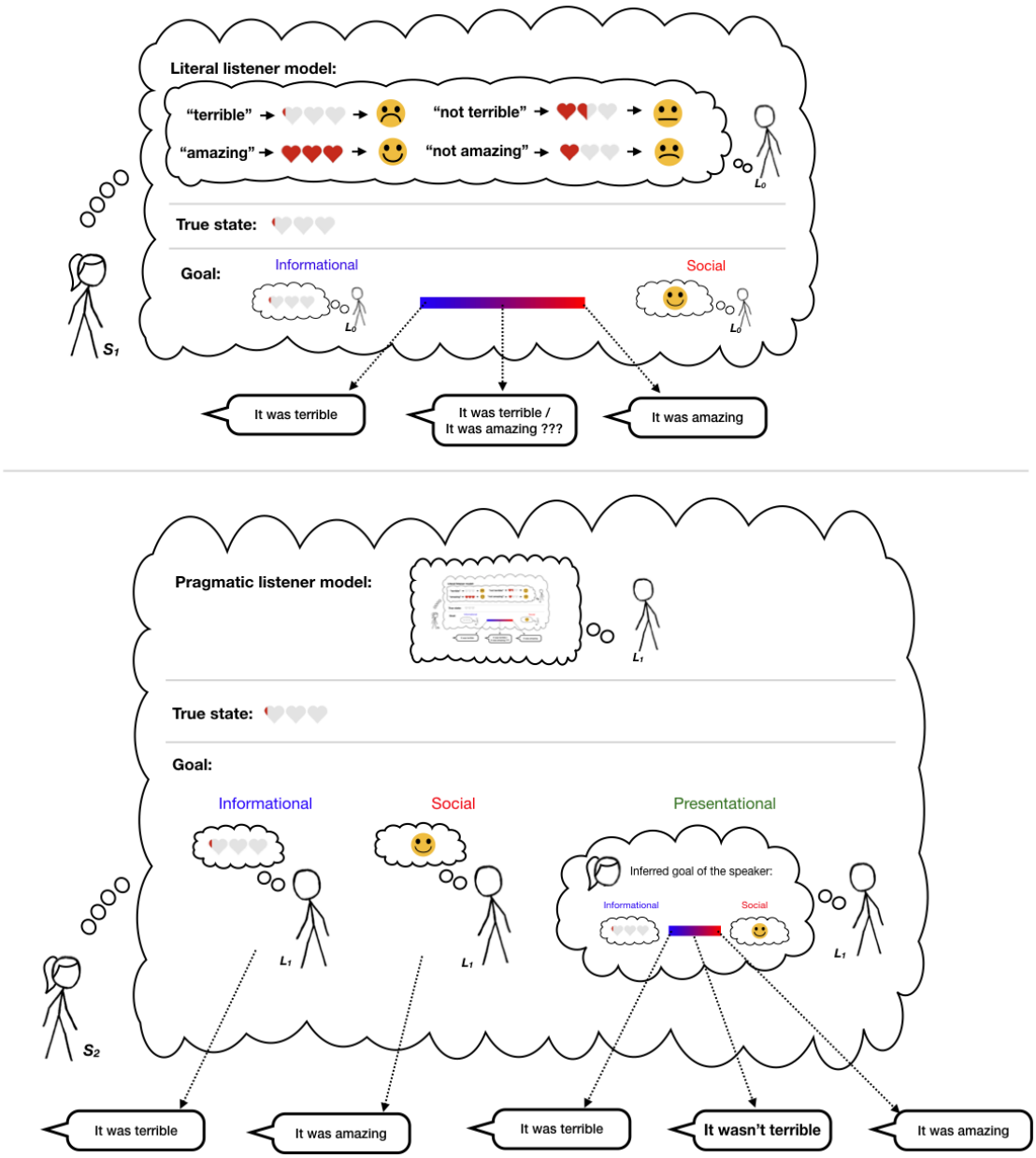

Figure 1. Diagram of the model. Top: First-order polite speaker (S_1) produces an utterance by thinking about: (1) the true state of the world (i.e., how good a given performance was); (2) the reasoning of literal listener who updates his beliefs about the true state via the literal meanings of utterances (e.g., "not terrible" means approximately 1.5 heart out of 3 hearts) and their affective consequences for the listener; and (3) her goal of balancing informational and social utilities. Bottom: Second-order polite speaker (S_2) produces an utterance by thinking about (1) the true state; (2) the pragmatic listener who updates his beliefs about the true state and the first-order speaker S_1's goal (via reasoning about the S_1 model); and (3) her goal of balancing informational, prosocial, and self-presentational utilities. Different utterances shown correspond to different weightings of the utility components.

production decision is softmax, which interpolates between choosing the maximum-utility utterance and probability matching (via speaker optimality parameter α; Goodman & Stuhlmüller, 2013):

$$P_{S_2}(w|s, \boldsymbol{\omega}) \propto \exp(\alpha \cdot \mathbb{E}[U_{total}(w; s; \boldsymbol{\omega}; \phi)]). \quad (1)$$

We posit that a speaker's utility contains distinct components that represent three possible goals that speakers may entertain: informational, social, and presentational. These components were determined based on multiple iterations of preliminary experiments, after which we conducted the preregistered test of our specified model with the specific utilities that we report below.

We take the total utility U_{total} of an utterance to be the weighted combination of the three utilities minus the utterance cost $C(w)$, simply used to capture the general pressure towards economy in speech (e.g., longer utterances are more costly):

$$U_{total}(w; s; \boldsymbol{\omega}; \phi) = \omega_{inf} \cdot U_{inf}(w; s) + \omega_{soc} \cdot U_{soc}(w) + \omega_{pres} \cdot U_{pres}(w; \phi) - C(w). \quad (2)$$

First, 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. We define *social utility* (U_{soc}) to be 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 $V(s)$ is the mapping from states of the world to subjective values, which likely varies by culture and context; we test our model when states are explicit

ratings (e.g., numbers on a 4-point scale) and we assume the simplest positive linear relationship between states and values V, where the subjective value is the numerical value of the state (i.e., the number of hearts). For example, Bob would prefer to have written a poem deserving 4 hearts rather than 1 heart and the strength of that preference is 4-to-1.

Listeners who are aware that speakers can be both kind and honest could try to infer the relative contribution of these two goals to the speaker's behavior (e.g., by asking himself: "was Ann just being nice?"). Thus, we use a pragmatic listener model who has uncertainty about the speaker's goal weight (relative contribution of niceness vs. informativeness) in addition to their uncertainty about the state of the world (number of hearts; Eq. 4). A speaker gains presentational utility when her listener believes she has particular goals, represented by a mixture parameter ϕ weighting the goals to be genuinely informative vs. kind.

A sophisticated speaker can then 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. Such a *self-presentational* goal may be the result of a speaker trying to save their own face (*I want the listener to see that I'm a decent person*) and can result in different speaker behavior depending on the intended, communicated goal of the speaker (e.g., *I want the listener to think I'm being honest* vs. *nice* vs. *both*)[1].

The extent to which the speaker *projects* a particular goal to the listener (e.g., to be kind) is the utterance's *presentational utility* (U_{pres}). Formally,

$$U_{pres}(w;\phi) = \ln(P_{L_1}(\phi \mid w)) = \ln \int_s P_{L_1}(s, \phi \mid w). \tag{3}$$

The speaker projects a particular weighting of informational vs. social goals (ϕ) by considering the beliefs of listener L_1, who hears an utterance and jointly infers the speaker's

[1] In principle, one could define a listener L_2 who reasons about this clever speaker and tries to uncover the goals that the speaker was trying to convey to them; we think such reasoning is reserved for very special relationships and is unlikely to manifest in the more basic acts of polite language use that we study here.

utilities and the true state of the world:

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

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 the second-order speaker S_2, but not the first-order speaker S_1). Only the social and informational utilities are defined for the first-order 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. Definitions for S_1 and L_0 otherwise mirror those of S_2 and L_1 and we use the same speaker optimality parameter for S_1 as for S_2 for simplicity; these sub-models are defined in the next section and appear in more detail in the Supplementary Materials. The complete model specification is shown in Figure 6.

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 1-to-4 numeric rating of the poem the speaker is commenting on), represented in terms of numbers of hearts (Figure 1): $S = s_0, ..., s_3$. In the experiment, participants are told that the listener has no idea about the quality of the product; thus, both listener models L_1 and L_0 assume uniform priors P(s) over the four possible heart states. The pragmatic listener's prior distribution over the first-order speaker's utility weights $P(\phi)$ encodes baseline assumptions about the relative informativeness vs. niceness listener's expect, which plausibly varies by culture and context; for simplicity, we assume this distribution to be uniform over the unit interval (0, 1). The set of utterances for the speaker models S_2 and S_1 is {*terrible*, *bad*, *good*, *amazing*, *not terrible*, *not bad*, *not good*, and *not amazing*} and the cost of an utterance is its length in terms of number of words (i.e., utterances with negation are costlier than those without negation) scaled by a free parameter. 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.

192

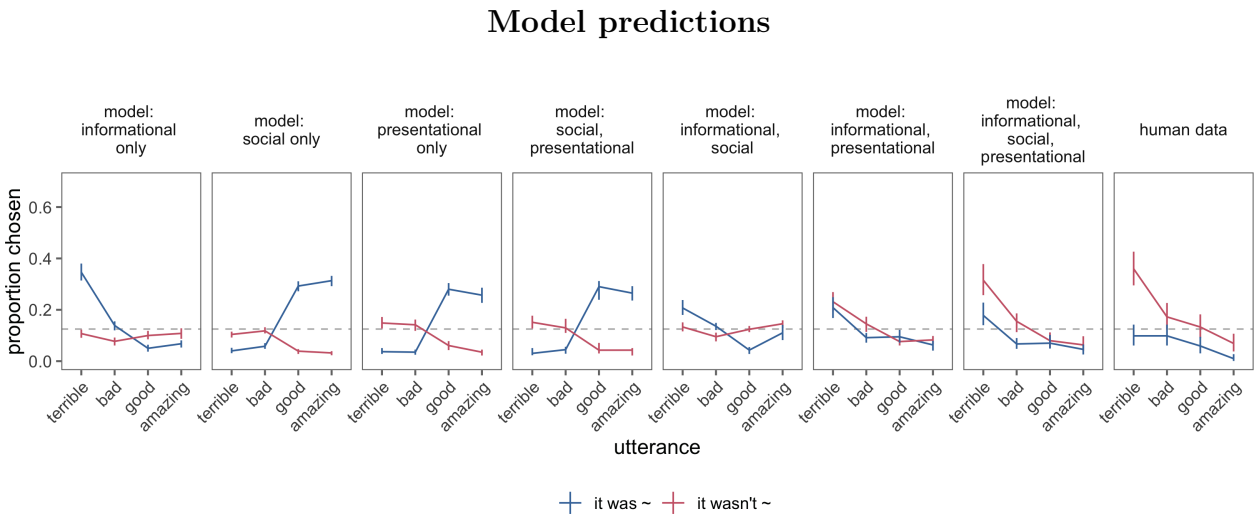

Figure 2. Comparison of predictions for proportion of utterances chosen by pragmatic speaker from possible model variants (left) and human data (rightmost) for average proportion of negation produced among all utterances, given true state of 0 heart (on a scale of 0 to 3) and speaker with both goals to be informative and kind. Gray dotted line indicates chance level at 12.5%.

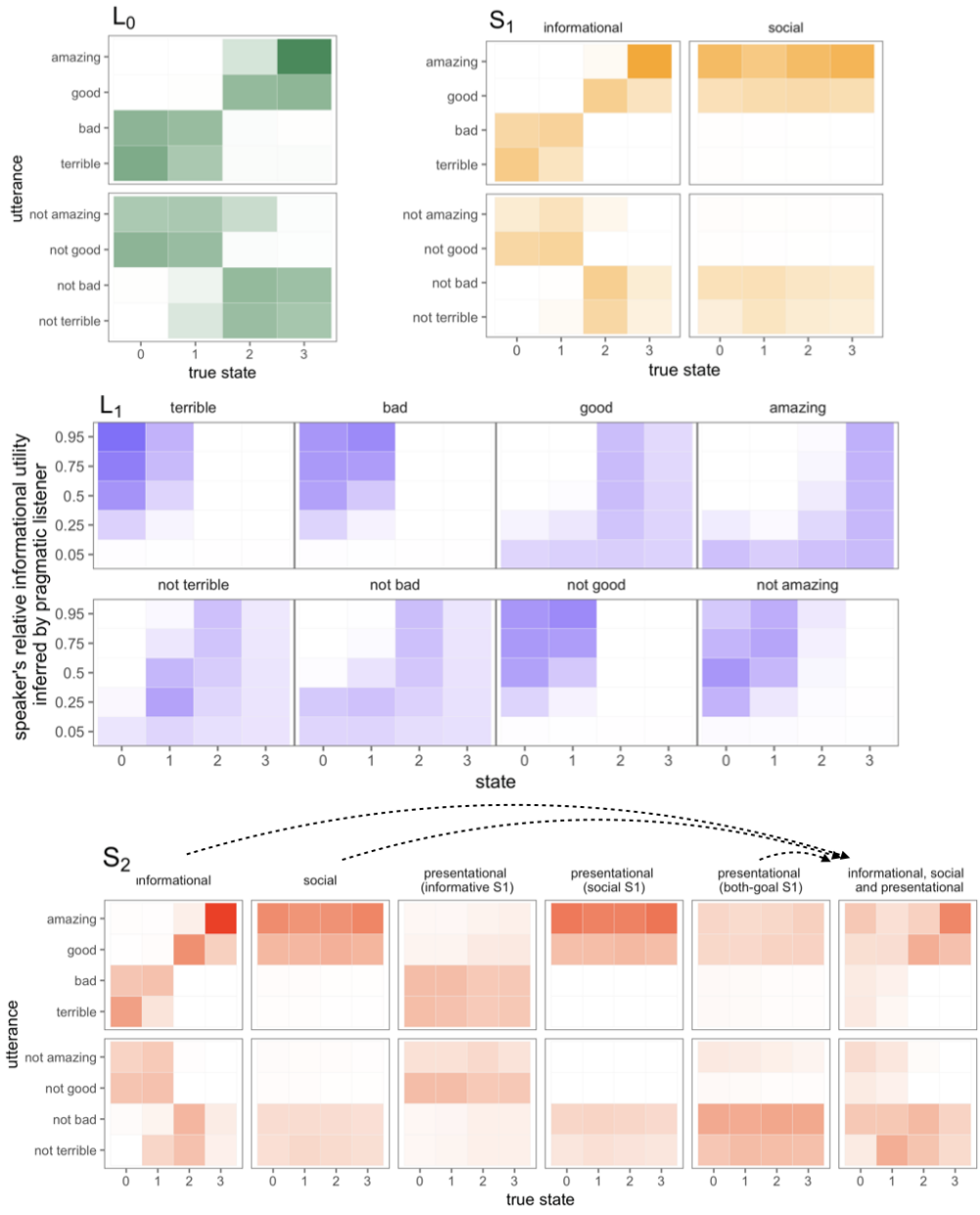

Figure 3. Model overview with schematic predictions. More saturated color indicates high probability (listener models) or high utility (speaker models). Top left: The literal listener L_0 posterior probability distribution over the true state (x-axis) given utterances (y-axis). Top right: Speaker S_1's utility of utterances (y-axis) for different states (x-axis) given either the informational or social goal (facets); shown in color (orange indicates high utility and white indicates low utility). Informational utility tracks the literal meanings and varies by true state; social utility favors utterances than favor higher valued states. Middle: Politeness-aware pragmatic listener L_1's joint posterior distribution over the true state (x-axis) and S_1 utility weighting (y-axis; higher value indicates greater weight on informational utility) given utterances (facets). Bottom: S_2's utility of utterances (y-axis) for different states (x-axis) and different goals (facets). Informational utility tracks the literal meanings and varies by true state; social utility favors utterances that signal high-valued states; three versions of self-presentational utility are shown, corresponding to whether the speaker wants to project informativeness, kindness, and a balance. Only the balanced self-presentational speaker shows a preference for indirect speech. The bottom right-most facet shows S_2's utterance preferences when they want to balance between the three utilities (informational, social, and presentational to project informativeness and kindness).

The behavior of the model can be understood through increasing levels of recursive reasoning. To ground the recursion, we have the literal listener modelL_0: a simple Bayesian agent who updates their prior beliefs over world states P(s) (assumed to be uniform) with the truth-functional denotation of the utterance w: $P_{L_0}(s|w) \propto [\![w]\!](s) * P(s)$ (i.e. the utterance's literal meaning). We assume soft-semantic meanings, which we elicit empirically in a separate experiment (N = 51, see Supplementary Materials). For example, the utterance "good" is compatible with both the 2- and 3-heart states, "not terrible" is also compatible with states 2- and 3-, though also to some extent with the 1-heart state (Figure 3, top left).

The first-order speaker S_1 chooses utterances given a utility function with two components defined in terms of the literal listener: informational and social utility. *Informational utility* (U_{inf}) is the amount of information about the world state conveyed to the literal listener L_0 by the utterance w; for example, the highest information utterance associated with the 2-heart state is "good"; the best way to describe the 0-heart state is "terrible" (Figure 2, top right; left facet). *Social utility* (U_{soc}) is the expected subjective utility of the world state inferred by the literal listener L_0 given the utterance w and does not depend on the true state.[2] For instance, the highest social utility utterance is "amazing", because it strongly implies that the listener is in the 3-heart state; negated negative utterances like "not bad" also have some degree of social utility, because they imply high heart states, albeit less directly (Figure 3, top right; right facet).. The speaker combines these utilities assuming some weighting *phi* and subtracts the cost of the utterance (defined in terms of the length of the utterance) in order to arrive at an overall utility of an utterance for a state and a goal-weighting:
$U(w; s; \phi) = \phi \cdot \ln(P_{L_0}(s \mid w)) + (1 - \phi) \cdot \mathbb{E}_{P_{L_0}(s|w)}[V(s)] - C(w)$. The speaker then chooses utterances w softmax rationally given the state s and his goal weight mixture ϕ.

[2]The independence between true state and social utility stems from the assumption of no shared beliefs between speaker and listener about the true state. This independence is a deliberate feature of our experimental setup, designed to best disambiguate the models proposed. In future work, it would be important to examine how shared beliefs about the true state and the speaker's goals may influence the speaker's utterance choice.

The pragmatic listener model L_1 reasons jointly about both the true state of the world and the speaker's goals (Fig. 3, middle). Upon hearing [Your poem was] "amazing", the listener faces a tough credit-assignment problem: The poem could indeed be worthy of three hearts, but it is also possible that the speaker had strong social goals and then no inference about the quality of the poem is warranted. Hearing [Your poem] was "terrible", the inference is much easier: the poem is probably truly terrible (i.e., worthy of zero hearts) and the speaker probably does not have social goals. Negation makes the interpreted meanings less precise and hence, inferences about goals are also fuzzier: "not amazing" can be seen as a way of saying that the poem was worthy of 0 or 1 hearts, which satisfies some amount of both social and informational goals. "Not bad" is less clear: the speaker could be being nice and the poem was actually worthy of 0- or 1-hearts (i.e., it was bad) or the speaker could be being honest (i.e., it was not bad) and the poem was worth 2-hearts.

The second-order pragmatic speaker model (S_2) reasons about the pragmatic listener L_1 to decide which utterances to produce based on both the true state of the world and the speaker's goals (Figure 3, bottom). The informational and social utilities of the second-order speaker mirror those of the first-order speaker: Direct utterances are more informative than those involving negation and utterances that signal many hearts are more prosocial.[3] The interesting novel behavior of this level of recursion comes from the different flavors of the self-presentational goal (Figure 3, bottom). When the second-order pragmatic speaker wants to *project* kindness (i.e., appear prosocial) they even more strongly display the preference for utterances that signal positive states (i.e., they are over-the-top positive). When the speaker wants to project honesty and informativeness, they take the exact opposite strategy, producing utterances that cannot be explained by virtue of social utility: direct, negative

[3]The second-order speaker informational utilities take into account the listener's pragmatic inferences about the speaker's goals. This only really affects the utility of "not terrible", which has higher information for the 1-heart state because the pragmatic listener strongly infers that the utterance was produced for social reasons. That is, for the second-order speaker, the utterance "not terrible" is loaded in a way that other utterances are not.

utterances (e.g., “it was terrible”). Finally, the speaker may present themselves in more subtle ways (e.g., intending to convey they are both kind and honest): This goal uniquely leads to the indirect, negative utterances (e.g., “not terrible”, “not bad”) having high utility. These utterances are literally incompatible with low-heart states, but are not highly informative; this unique combination is what gives rise to the subtle inference of a speaker who cares about both goals.

Experiment: Speaker production task

We made a direct 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. We developed the preceding model iteratively on the basis of a sequence of similar experiments, but importantly, the current test was fully pre-registered and confirmatory. All data analytic models and our full model comparison approach were registered ahead of time to remove any opportunities for overfitting the behavioral data through changes to the model or the evaluation.

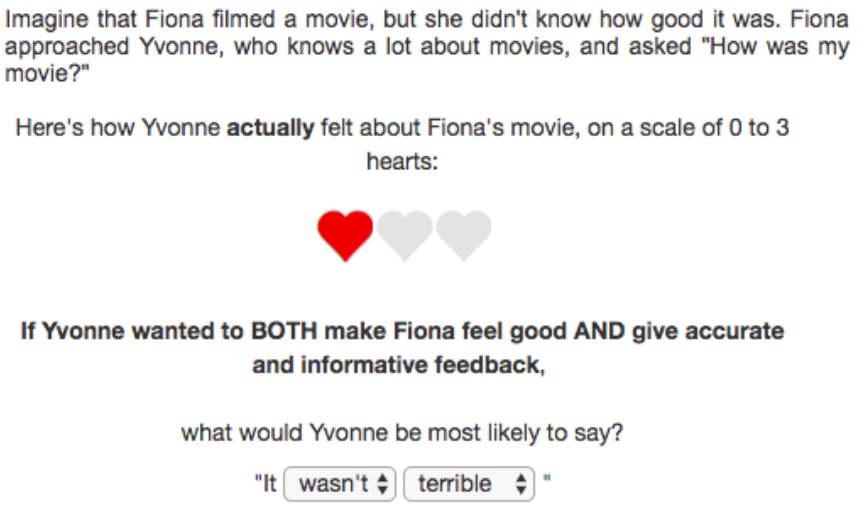

Figure 4. Example of a trial in the speaker production task.

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. Notably, we did not mention a self-presentational goal to participants; rather, we hypothesize this goal can arise spontaneously from a speaker's inability to achieve the first-order goals of niceness and honesty (i.e., if a speaker wants to, but can't, be both honest and nice, they would instead try to signal that they care about both goals). We hypothesized that each of the three experimentally-induced goals (*informative*, *kind*, *both*) would induce a different tradeoff between the informational, prosocial, and self-presentational utilities in our model . 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

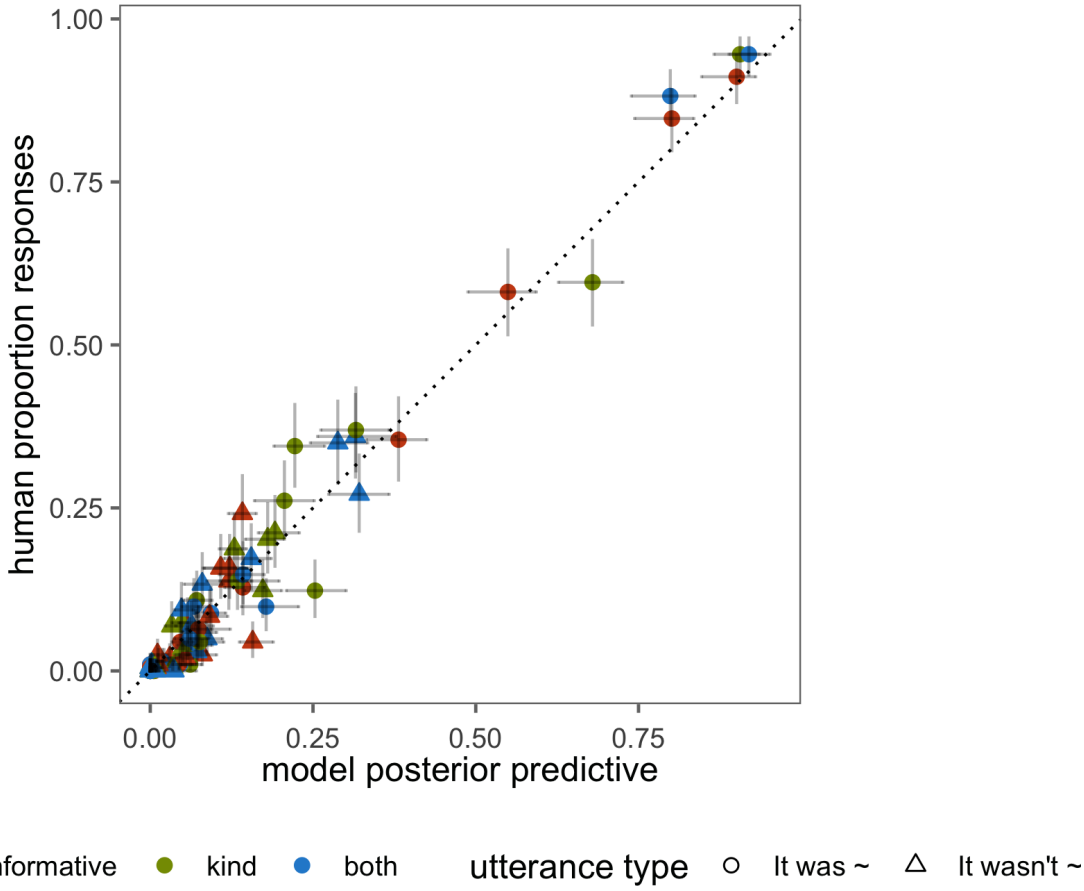

Figure 5. Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data (vertical) and 95% highest density intervals for the model (horizontal).

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 3, bottom). This prediction was confirmed, as a Bayesian mixed-effects model predicts more negation as a function of true state and goal via

Table 1

Inferred goal weight (ω_g) and speaker-projected informativity-niceness weight (ϕ) parameters from all model variants with more than one utility.

model (utilities)	goal	ω_{inf}	ω_{soc}	ω_{pres}	ϕ
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

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.

Model results

We assume our experimental goal conditions (informative vs. kind vs. both) induce a set of weights over the utilities $\boldsymbol{\omega}$ in participants' utterance production model. In addition, the self-presentational utility is defined via a communicated social weight ϕ (i.e., the mixture

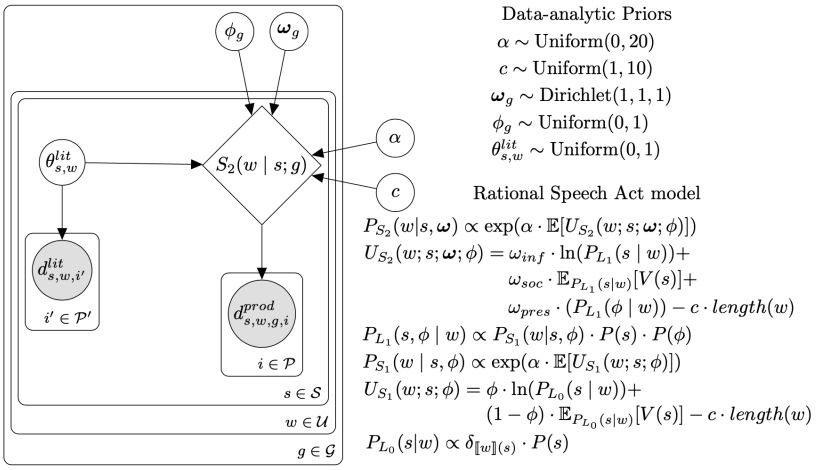

Figure 6. Graphical model representing our Bayesian data analytic approach for the full 3-component model (other models contain subsets of the parameters shown). S_2 represents the RSA speaker model defined by Eq. 1, which is used to predict the production responses d^{prod} of each participant i, for each state s (number of hearts), for each utterance w, in each goal condition g. The RSA speaker model takes as input the literal meaning variables θ, which additionally are used to predict the literal meaning judgments d^{lit} assuming a Bernoulli linking function. Additionally, the RSA model takes the speaker's goal weights ω and intended presentational goal weight ϕ, which are inferred separately for each goal condition g. Finally, the RSA model uses two global free parameters: the cost of negation c and the speaker's rationality parameter α.

of informative vs. social that the speaker is trying to project). The mapping from social situations into utility weights and communicated social weight is a complex mapping, which we do not attempt to model here; instead, we infer these parameters for each goal condition from the data. We additionally infer the literal meanings (i.e., the semantics) of the words as interpreted by the literal listener L_0 with the additional constraint of the literal meaning judgments from an independent group of participants (See Supplementary Materials: Literal semantic task section). Finally, the RSA model has two global free parameters: the softmax speaker optimality and utterance cost of negation from the data (Figure 6). We implement this data analytic model for each of the alternative models and infer the parameters using Bayesian statistical inference (M. D. Lee & Wagenmakers, 2014). We use uninformative priors over ranges consistent with the prior literature on RSA models: $\theta^{lit}_{s,w} \sim \text{Uniform}(0, 1)$, $\phi_g \sim \text{Uniform}(0, 1)$, $\boldsymbol{\omega}_g \sim \text{Dirichlet}(1, 1, 1)$, $\alpha \sim \text{Uniform}(0, 20)$, $c \sim \text{Uniform}(1, 10)$. This analysis tells us which, if any, of these models can accomodate all of the patterns in the

Table 2

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

empirical data. The posterior predictions from 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 5). Other models (e.g., informational + presentational), however, show comparably high correlations to the full data set; correlations can be inflated through the presence of many 0s (or 1s) in the data set, which our data contains since certain utterance choices are implausible given a particular state and goal condition. Thus, we compare model variants using a bonafide model comparison technique, Bayes Factors, which balance predictive accuracy with model complexity in quantifying the goodness of fit of a model.

Bayes Factors compare the likelihood of the data under each model, averaging over the prior distribution of the model parameters; by averaging over the prior distribution over parameters, Bayes Factors penalize models with extra flexibility because increasing the flexibility of the model to fit more data sets decreases the average fit of the model to a particular data set (M. D. Lee & Wagenmakers, 2014), capturing the intuition that a theory that can predict anything predicts nothing. That is, simply because a model has more

parameters and can explain more of the variance in the data set does not entail that it will assign the highest marginal likelihood to the actual data. Here, however, both the variance explained and marginal likelihood of the observed data were the highest for the full model: The full model was at least 5 x 10^4 times better at explaining the data than the next best model (Table 2). 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 2). In sum, the full set of informational, social, and presentational utilities were required to fully explain participants' utterance choices.

The utility weights inferred for the three-utility model (Table 1) 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 (ω_{soc}) close to zero, but the weight on *projecting kindness* (ω_{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' 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 into theories of language use adds a level of explanation, but so far such intuitions and observations have resisted both formalization and precise testing. We presented a set of utility-theoretic models of language use that captured different proposals about the interplay between competing informational, social, and presentational goals. Our

full model instantiated a novel theoretical proposal, namely that indirect speech is a response to the conflict between informational and social utilities that preserves speakers' self-presentation. Our confirmatory test of the comparison between these models then provided experimental evidence that the full model best fit participants' judgments, even accounting for differences in model complexity.

The most substantial innovation in our full model is the formalization of a self-presentational utility, defined only for a speaker who reasons about a listener who reasons about a speaker. We hypothesized that a speaker who prioritizes presentational utility will tend to produce more indirect speech (negation in our experimental paradigm). Indeed, this is consistent with previous work showing that people prefer to use negation ("that's not true" as opposed to "that's false") when prompted to speak more "politely" (Giora, Balaban, Fein, & Alkabets, 2005) and that utterances involving negation tend to be interpreted in a more mitigated and hedged manner compared to direct utterances (H. L. Colston, 1999). It also may help explain the phenomenon of negative strengthening, where negation of a positive adjective can be interpreted in a rather negative manner (e.g., "He's not brilliant" meaning "he is rather unintelligent"; Gotzner, Solt, & Benz, 2018). Our work builds on this previous work that shows a preference for negation by elucidating the goal-directed underpinnings of this behavior and possible contextual modulation of this preference. An interesting open question is whether other negation-related politeness phenomena (e.g., indirect questions such as "You couldn't possibly tell me the time, could you?"; Brown & Levinson, 1987) can be derived from the basic information-theoretic goals we formalize.

In order to conduct quantitative model comparisons, we needed to create an experiment with repeated trials and a restricted range of choices. Thus, we had to abstract away from the richness of natural interactions. These choices decrease the validity of our experiment. Despite these abstractions, we showed that behavior in the experiment reflected social and informational pressures described in previous theories of polite language, providing some face validity to the responses we collected. With a formal model in hand, it

now will be possible to consider relaxing some of the experimental simplifications we put into place in future work. Most importantly, human speakers have access to a potentially infinite set of utterances to select from in order to manage the politeness-related tradeoffs (e.g., *It's hard to write a good poem*, *That metaphor in the second stanza was so relatable!*). Each utterance will have strengths and weaknesses relative to the speaker's goals. 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), and addressing these challenges is an important future direction.

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. This latter point illustrates a general feature of why explicit computational models provide value: only by formalizing the factors in Brown and Levinson (1987)'s theory were we able to recognize that they were an insufficient description of the data we were collecting in previous versions of the current experiment. Those failures allowed us to explore models with a broader range of utilities, such as the one reported here.

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. Cultures may place value on satisfying different communicative goals, and speakers in these cultures may pursue those goals more strongly than speakers from other cultures. For example, we found in our model that a speaker who wants to appear informative should speak more negatively than a truly informative speaker; one could imagine run-away effects where a group becomes overly critical from individuals' desires to appear informative. Culture could also affect the value function V that maps states of the world onto subjective values for the listener. For example, the mapping from states to utilities may be nonlinear and involve reasoning about the future; a social utility that takes into account reasoning about the future could help explain why it can often be nice to be informative. Our formal modeling approach, with systematic behavior measurements, provides an avenue towards understanding the vast range of politeness practices found across languages and contexts (A. N. Katz, Colston, & Katz, 2005).

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 can be integrated with informational concerns more generally, opening up the possibility for a broader theory of social language. A formal account of politeness may also move us closer to courteous computation – to machines that can talk with tact.

References

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

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

Bolinger, D. (1972). *Degree words* (Vol. 53). Walter de Gruyter.

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.

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

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

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

Colston, H. L. (1999). ?Not good? Is ?bad,? But ?not bad? Is not ?good?: An analysis of three accounts of negation asymmetry. *Discourse Processes*, *28*(3), 237–256.

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

Giora, R., Balaban, N., Fein, O., & Alkabets, I. (2005). Negation as positivity in disguise. *Figurative Language Comprehension: Social and Cultural Influences*, 233–258.

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.

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*, 1659.

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

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

Horn, L. (1989). A natural history of negation.

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 naive 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.

Katz, A. N., Colston, H., & Katz, A. (2005). Discourse and sociocultural factors in understanding nonliteral language. *Figurative Language Comprehension: Social and Cultural Influences*, 183–207.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of

interpretation. *Synthese*, *194*(10), 3801–3836.

Lea, R. B., & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 303.

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.

MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 633.

Nordmeyer, A., & Frank, M. C. (2014). A pragmatic account of the processing of negative sentences. In *Proceedings of the thirty-sixth annual meeting of the cognitive science society* (Vol. 36).

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.

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.

Stern, G. (1931). Meaning and change of meaning; with special reference to the english language.

Stoffel, C. (1901). *Intensives and down-toners: A study in english adverbs*. Carl Winters Universitätsbuchhandlung.

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.