Tiptoeing around it: Inference from absence in potentially offensive speech

Monica A. Gates[1] (mgates@berkeley.edu), **Tess L. Veuthey**[2] (tess.veuthey@ucsf.edu)
Michael Henry Tessler[3] (mhtessler@stanford.edu), **Kevin A. Smith**[4] (k2smith@mit.edu), **Tobias Gerstenberg**[4] (tger@mit.edu)
Laurie Bayet[5] (laurie.bayet@childrens.harvard.edu) & **Joshua B. Tenenbaum**[4] (jbt@mit.edu)
[1]Psychology, University of California, Berkeley, [2]Neuroscience, University of California, San Francisco,
[3]Psychology, Stanford University, [4]Brain and Cognitive Sciences, Massachusetts Institute of Technology,
[5]Laboratories of Cognitive Neuroscience, Harvard Medical School & Boston Children's Hospital

Abstract

Language that describes people in a concise manner may conflict with social norms (e.g., referring to people by their race), presenting a conflict between transferring information efficiently and avoiding offensive language. When a speaker is describing others, we propose that listeners consider the speaker's use or absence of *potentially offensive language* to reason about the speaker's goals. We formalize this hypothesis in a probabilistic model of polite pragmatic language understanding, and use it to generate predictions about interpretations of utterances in ambiguous contexts, which we test empirically. We find that participants are sensitive to potentially offensive language when resolving ambiguity in reference. These results support the idea that listeners represent conflicts in speakers' goals and use that uncertainty to interpret otherwise underspecified utterances.

Keywords: politeness; social meaning; pragmatics; Bayesian cognitive model; Rational Speech Act model

Introduction

Referring to strangers can be challenging. Without knowing their name, you could describe them by their physical appearance, but not all attributes are equally informative. One problem for speakers is that highly diagnostic attributes can be potentially offensive (e.g., an overweight person's weight).

Grice (1975, p. 46) was aware of this problem: "There are, of course, all sorts of other maxims (aesthetic, social, or moral in character), such as 'Be polite', that are also normally observed by participants in talk exchanges." In a politeness framework, the avoidance of potentially offensive words illustrates how speakers balance being informative with social goals (Brown & Levinson, 1987). Specifically, Brown and Levinson (1987) outline ambiguous speech as a form of indirect or "off-record" politeness. We draw inspiration from these ideas and hypothesize that the use or avoidance of words that carry social meaning prompts listeners to reason about the speaker's social goals. Do listeners hypothesize that speakers are constrained to use inoffensive language, and use this understanding to infer a speaker's intended meaning from an ambiguous utterance?

We developed a model in the Rational Speech Act (RSA) tradition (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) to capture the social and epistemic inferences elicited by words with social meaning, specifically potentially offensive descriptors. Vanilla RSA models predict pragmatic inferences listeners make for literally ambiguous statements by considering the alternative statements the speaker could have said. Recent work has modeled inferences about speakers' *social* goals, specifically the desire to be kind to the listener (Polite RSA; Yoon, Tessler, Goodman, & Frank, 2016, 2017). The polite RSA model defines the social utility of an utterance as the quality of the world it makes the listener believe they are in. We extend this work by having potentially offensive utterances incur a *social cost* to the speaker. A listener who is aware of these social costs can resolve otherwise ambiguous utterances to infer a speaker's intended referent.

In our experiments, participants were introduced to a world where the words "blue" or "green" were potentially offensive. With their new social understanding, they played reference games in which they were asked to interpret a speaker's utterance (e.g., "person with the hat") in terms of which character in a scene the speaker was trying to refer to (see Figure 1).

We hypothesize that listeners reason about the social cost of producing potentially offensive speech a) to contextually understand ambiguous utterances, and b) to evaluate speakers. Experiment 1 tests participants' inferences about who an ambiguous utterance refers to. Experiment 2 measured participants' inferences about the speaker's goals. Across these two experiments, we find that our model accounts for the fine-grained inferences listeners draw when reasoning about potentially offensive speech.

Computational Model

We built a rational model of communication within the Rational Speech Act framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Our model belongs to the class of "uncertain RSA" models, which involve reasoning about aspects of the speaker beyond just their intended meaning (Goodman & Frank, 2016). We used this framework to understand the phenomenon wherein a speaker is underinformative so as to not use potentially offensive speech, but listeners are nevertheless able to infer who speakers are referring to. In other words, when listeners are aware of a speaker's alterna-

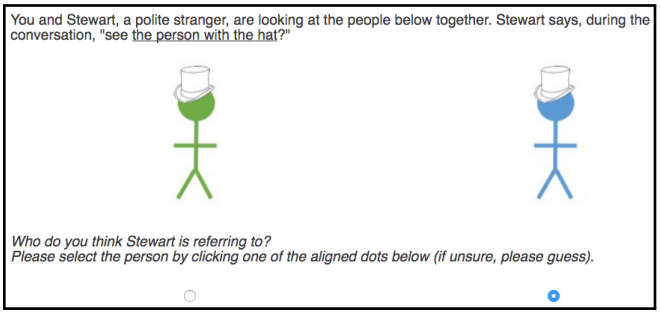

Figure 1: Example context from Experiment 1.

tive utterances and the associated social costs, they can reason backwards to infer the speaker's intended referent.

Specifically, this work builds on an RSA model for polite language use (Polite RSA; Yoon et al., 2017). The listener in Polite RSA reasons about whether the speaker was trying to be epistemically informative (à la Vanilla RSA) or considerate to the listener's feelings (a social goal). The Polite RSA model operationalizes the social utility of an utterance u in terms of the subjective value of the world state that the listener would believe themselves to be in upon hearing u. For example, positive social utility is incurred by making the listener believe they are in a good state (e.g., that the cookies they baked were delicious). The model predicts that speakers who try to balance being informative and kind will choose to produce more *indirect speech* (e.g., saying "it wasn't amazing" as opposed to "it was terrible"), and this prediction was borne out empirically (Yoon et al., 2017).

We took inspiration from the Polite RSA model, but parametrized the reasoning slightly differently. We modeled a listener who reasons about a potential social cost to an utterance. That is, words could be costlier to produce by the speaker by virtue of their social stigma of use. We assumed, for example, that a socially-minded speaker would incur a cost by referring to an overweight person as "fat". Rather be on the word form itself, this kind of cost can likely be derived out of a more basic mechanism analogous to that used by Yoon et al. (2017), a point we return to in the Discussion.

Model details

The RSA framework models utterances and inferences as deriving from recursive social reasoning: a speaker S_1 produces an utterance u reasoning about how a literal listener L_0 would interpret it. A pragmatic listener L_1 interprets the utterance u reasoning about what speaker S_1 would say.

We start with the literal listener L_0, who literally interprets the meaning of any utterance u to determine the intended referent r within the context C:

$$P_{L_0}(r \mid u, C) \propto [\![f(u)]\!](r) \cdot P(r) \tag{1}$$

$[\![u]\!](r)$ is u's literal meaning, mapping to 1 if u matches referent r and 0 otherwise given context C. $f(u)$ expresses the noisy semantics model: with probability γ the listener doesn't condition on the utterance heard and instead samples a referent from the prior (Degen, Hawkins, Graf, Kreiss, & Goodman, in prep; Graf, Degen, Hawkins, & Goodman, 2016). Mathematically, $P\big(f(u_{w-1})|f(u)\big) = 1-\gamma,\ \forall\, w \in u$, where each w represents a word in the utterance u. $P(r)$ is a uniform distribution over possible referents given the context C.

Speaker S_1 produces an utterance based on a utility function U, which has two parts. The first part represents an epistemic utility which we define as the literal listener L_0 uncertainty about the referent r after hearing the utterance u: $\ln\big(P_{L_0}(r \mid u, C)\big)$. This uncertainty is weighted by an utterance prior $P(u)$ that assigns more probability to utterances with fewer words (uttering words is effortful). If $\sum_w$ is the utterance's word count, W is the maximum number of words possible in an utterance, and ξ parameterizes, then $P(u) = \frac{\exp(-\xi \cdot \sum_w)}{\sum_{w=0:W} \exp(-\xi \cdot \sum_w)}$. We introduce a weighting parameter β_{epi} which captures how much the speaker cares about reducing the listener's uncertainty about the true referent.

The second part of speaker S_1's utility function represents a social utility. In our experiments and model, color terms are potentially offensive. The speaker is aware of a specific color word which is considered potentially offensive and designated as *badWord*. The speaker's social utility is $V(u) = 0$ if $badWord \in u$, and 1 otherwise. We introduce another weighting parameter β_{soc} which captures how much the speaker cares about avoiding potentially offensive language. By combining both epistemic and social utility, we get S_1's utility function as follows:

$$U(u, r, C, \hat{\beta}) = \beta_{epi} \cdot \ln\big(P_{L_0}(r \mid u, C) \cdot P(u)\big) + \beta_{soc} \cdot V(u)$$

Overall, the speaker chooses an utterance softmax-optimally, where λ_1 represents S_1's optimality:

$$P_{S_1}(u \mid r, C, \hat{\beta}) \propto \exp\big(\lambda_1 \cdot U(u, r, C, \hat{\beta})\big) \tag{2}$$

The pragmatic listener L_1 then reasons about the speaker S_1, jointly inferring the referent r and how much weight the speaker S_1 places on the epistemic β_{epi} and social β_{soc} utility (Goodman & Lassiter, 2015). $P(r)$ is uniform over possible referents given context C, and $P(\hat{\beta})$ is a uniform distribution across the set $\{.1, .3, .5, .7, .9\}$.

$$P_{L_1}(r, \hat{\beta} \mid u, C) \propto P_{S_1}(u \mid r, C, \hat{\beta}) \cdot P(r) \cdot P(\hat{\beta}) \tag{3}$$

We implemented the model in WebPPL, a probabilistic programming language (Goodman & Stuhlmüller, 2014). The model has three free parameters: a parameter for the noisy semantics (i.e., the overall extent to which utterances are not truth-functional) γ, a cost to producing more words ξ, and the speaker optimality parameter λ_1. In parameter fitting, γ was fixed at .1, and the other parameters were fit to the data, but restricted to the following ranges (consistent with models of the same model class): ξ fell between 0-1 and λ_1 fell between 1-20 (cf., Yoon et al., 2016, 2017). The best-fitting parameter settings were: ξ = .5, and λ_1 = 20, determined through minimizing the least-squared error between model predictions and behavioral results.

In our experiments, utterances u could be any combination of the following: n/a (in the experiment, we added "person" to all utterances, so participants saw "the person" instead), one color term ("blue", "green", or "orange"), "scarf", and "hat". So, for example, an utterance could be "the person" or "the orange person with the scarf". The intended referent r could be any of the two or three possible referents that appeared within a context C. The potentially offensive color term *badWord* was either "blue" or "green", counterbalanced across participants.

We tested our model against human behavior in two experiments. In Expt. 1, listeners inferred the intended referent r given an utterance u and context C. In Expt. 2, listeners inferred $\hat{\beta}$ given a referent r, utterance u, and context C.

General Experiment Methods

Participants

We recruited participants from Amazon Mechanical Turk, with U.S. IP addresses and no reported color blindness. In Experiment 1, 45 participants were recruited, and three were removed (two for later reporting colorblindness, and one for failing a catch-trial). In Experiment 2, 46 participants were recruited, and one removed for later reporting colorblindness.

Stimuli and Procedure

Training Participants began by viewing training scenes. Training scenes were designed to inform participants that using a particular color (*badWord*: either "blue" or "green") was potentially offensive. Participants first read and were tested on an explicit description of the manipulation: "In a parallel world, some people are different colors. In this world, calling someone a '[color] person' is potentially offensive," where [color] was *badWord*. Participants then viewed several counterbalanced scenes, in which characters were selectively scolded by other characters for saying *badWord*.

Main Experiment Following the training scenes, participants viewed reference game contexts in the main experiment. Within each context, two or three people were aligned left to right, were colored blue, green, or orange, and possibly wore hats and scarves. In the accompanying text, participants were observing the possible referents with a speaker named [Name]. [Name]s were selected by random selection with replacement from a list of 172 names for each context. The order of the possible referents in the context was randomly sampled at the beginning of the experiment and was fixed for all participants. The order of trials was randomized.

Context selection Contexts were selected to test the inference that if a speaker did not explicitly refer to a person by their color, then perhaps that color was a *badWord* and potentially offensive. We sought examples that produced a range of model predictions. Contexts were selected to be roughly consistent across the experiments, so that the different methods of probing potential offensiveness could be compared. Finally, contexts were chosen to have built-in controls, such that if an image was presented where the referent color was *badWord*, the same image type was presented in a different context where the colors were switched so that the referent was now not *badWord*. In the rest of this paper, we describe the analysis with respect to the *badWord* being "blue".

Experiment 1: Inferring the referent

Experiment-Specific Methods

This experiment contained 35 contexts. In each context, a speaker presented an utterance and the participant was asked to select which of the 2-3 referents the speaker was likely referring to (simple multiple-choice task, see Figure 1).

Results and Discussion

In calculating statistics, because probabilities for the last referent of each context were entirely determined by probabilities assigned to the other referent(s), values from a randomly chosen referent were removed from further statistical analyses in order to meet assumptions of independence.

Participants' social inferences closely mirrored the inferences predicted by the model. Specifically, if the speaker's statement was ambiguous, participants selected the person with the potentially offensive color as being the referent, as predicted by the model (e.g. see contexts 1A, 1G in Table 1). When no referents of potentially-offensive colors were available, participants and the model were approximately ambivalent between the referents (e.g. see context 1B). In "positive control" contexts, in which the referent was unambiguously indicated by an utterance describing the intended referent's color ("the blue person"), participants selected the designated referent, as predicted by the model (e.g. see context 1F).

The left plot in Figure 2 shows a scatterplot with model predictions and participants' inferences across all contexts. Our model explained participants' inferences to a high degree of quantitative accuracy with bootstrapped 95% confidence intervals for adjusted R^2 of [.86,.96] and for Spearman's ρ of [.88,.97]. The model's incorporation of social utility was critical to fit participants' inferences: when social utility was removed in a lesioned version of the model, bootstrapped confidence intervals for adjusted R^2 dropped to [.27,.60], and for Spearman's ρ to [.62,.89] (Figure 2, right). Moreover, the moderately high correlations from the lesioned model were mostly driven by the presence of the positive control contexts in Expt. 1, which did not require social knowledge. When the eight positive control contexts were removed, the lesioned model's bootstrapped confidence intervals for adjusted R^2 dropped to [0.06,0.43], and for Spearman's ρ to [.37,.83].) 10000 samples were drawn in all cases.

While the model generally captured participants' inferences well, there was a subset of contexts for which the model's predictions did not match participants' inferences. In these contexts, the model was reluctant to make the inference that the speaker was referring to the person with the potentially-offensive color when that person wore an item which was not specified in the utterance. Specifically, first consider the normal case: in context 1D, the only way to pick out the blue person would be to refer to their color. Given this fact, upon hearing "the person" instead, the model correctly predicted that people would choose the blue person as the intended referent. However, in context 1C, the blue person could also be unambiguously identified by referring to their scarf. Upon hearing the utterance "the person" in this context, the model was unsure who the intended referent was, whereas people considered the blue person with the scarf to be most likely. A similar phenomenon occurred in context 1E. One possible explanation for the deviation between model predictions and people's judgments here is that participants may have learned to associate the utterance "the person" with a

Table 1: Example Expt. 1 contexts. For each context, the 2-3 referents are separated by "/" and can be blue ("Bl"), green ("Gr"), or orange ("Or"). Results were collapsed across conditions so that "blue" was the potentially offensive word in all contexts. In the "Utterance" column, "n/a" stands in for "the person", and "blue hat scarf" for "the blue person with the hat and the scarf". The behavioral, model, and lesioned model (without social inference) proportions allocated to each referent are shown.

	Referents (*Bl=potentially offensive)	Utterance	Behavioral Mean *(Std.)* Props.	Model Props.	Lesioned Props.
1A	Or / Bl	"n/a"	.14 *(.05)* / .86 *(.05)*	.12 / .88	.5 / .5
1B	Gr-hat / Or-scarf	"n/a"	.50 *(.08)* / .50 *(.08)*	.5 / .5	.5 / .5
1C	Or / Bl-scarf / Gr	"n/a"	.17 *(.06)* / .67 *(.07)* / .17 *(.06)*	.35 / .30 / .35	.41 / .17 / .41
1D	Or-scarf / Bl / Gr-hat	"n/a"	.10 *(.05)* / .83 *(.06)* / .07 *(.04)*	.02 / .95 / .02	.22 / .56 / .22
1E	Bl-hat / Or / Gr-scarf	"n/a"	.57 *(.08)* / .40 *(.08)* / .02 *(.02)*	.38 / .44 /.18	.22 / .56 / .22
1F	Bl-scarf-hat / Gr / Or-scarf-hat	"blue hat scarf"	.93 *(.04)* / 0 / .07 *(.04)*	1 / 0 / 0	1 / 0 / 0
1G	Bl-scarf-hat / Or-scarf-hat / Gr-scarf-hat	"hat scarf"	.81 *(.06)* / .12 *(.05)* / .07 *(.04)*	.87 / .07 /.07	.33 / .33 / .33

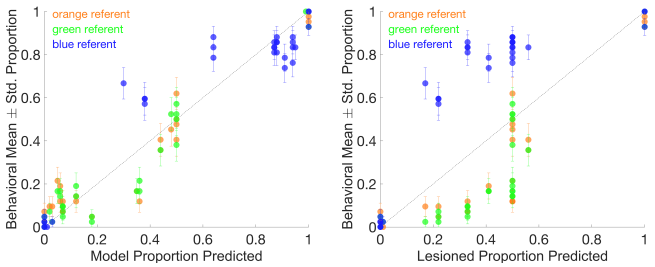

Figure 2: Behavioral and model comparison for Expt. 1. Participants saw an utterance and inferred which of the 2-3 referents the speaker was referring to for 35 contexts. Referents were orange, green, or blue. Results were collapsed across conditions so that "blue" was the potentially offensive word in all contexts. Behavioral results show the proportion of participants selecting each referent; model predictions show the proportions that the model allocated to each referent. **Left**: Full model. **Right**: Lesioned model (social utility set to 0).

blue person based on inferences drawn in previous contexts.

Experiment 2: Inferring speaker goals

We placed participants in a world where certain words were potentially offensive in Expt. 1. Given this knowledge, we found that listeners could infer a speaker's intended referent even if the speaker was ambiguous, as predicted by the model. In Expt. 2, we tested whether listeners could infer a speaker's goals (informational or social) based on how the speaker referred to someone.

Experiment-Specific Methods

After viewing the same training scenes that participants had seen Experiment 1, participants saw additional training scenes that clarified that the dimension of "offensiveness" corresponded to the use of *badWord*, and that the dimension of "ambiguity" referred to how much the utterance specifically identified the intended referent. Participants answered a comprehension check question, and then saw 40 different contexts in the test phase. Figure 3 shows a screenshot of the test phase. In each context, an intended referent (out of two or three possible referents) was circled, and two possible utterances the speaker could say were shown on the left and right sides of the screen. Participants moved two separate sliders ranging from 0 to 100 to indicate which of the two utterances they considered to be more offensive, and which to be more ambiguous. The sliders were initially set at 50, which represented ambivalence.

Results and Discussion

Similarities between model predictions and judgments Overall, the model again provided an accurate account of participants' inferences. If one utterance better distinguished the referent, participants rated that utterance as less ambiguous, as predicted by the model. This rating of lower ambiguity appeared over relatively subtle distinctions, like when the utterance reduced the number of valid possible intended referents from 3 to 2 (see for example context 2G in Table 2) or from 2 to 1 (e.g. contexts 2D, 2E). If utterances were both equally informative, participants roughly rated them as equally informative (e.g. context 2B) though behavioral exceptions exist.

With respect to offensiveness, if a single utterance contained the word "blue", then that utterance was rated as more offensive (e.g. context 2F). If neither utterance contained

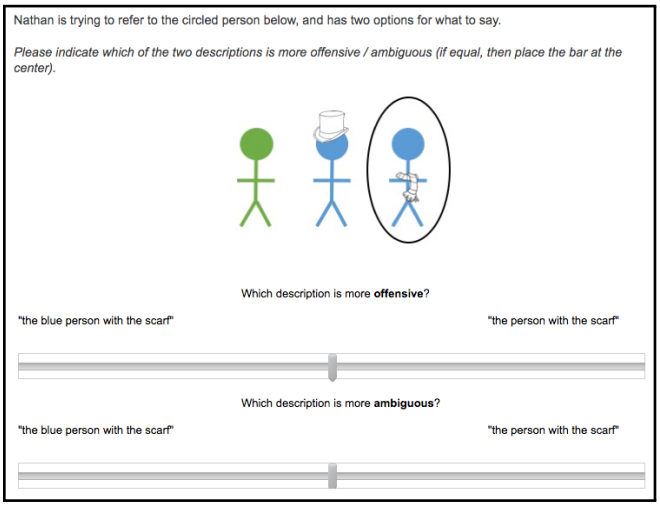

Figure 3: Example context from Experiment 2.

Table 2: Example Experiment 2 contexts. For each context, referents are separated by slashes (the intended referent is in bold) and could be blue ("Bl"), green ("Gr"), or orange ("Or"). Results were collapsed across conditions so that "blue" was the potentially offensive word in all contexts. Each context had two utterances: "Utt. 1" was positioned on the left of the screen at score 0, and "Utt. 2" was positioned on the right at score 100. Thus, lower scores indicate that Utt. 1 was rated higher (more ambiguous / offensive) than Utt. 2, and higher scores indicate Utt. 2 was rated higher (more ambiguous / offensive) than Utt. 1. In the experiment, these utterances were longer than the abbreviations shown here: "the person" was shown rather than "n/a", and "the blue person with the scarf" rather than "blue scarf". In the results columns, "Amb" indicates ambiguity ratings: behavioral mean and italicized standard errors are shown ("Amb"), as are model predictions ("AmbM") and lesioned model predictions without social inference ("AmbL"). "Off" indicates offensiveness ratings.

	Referents (*Bl=potentially offensive)	Utt. 1	Utt. 2	Amb	AmbM	AmbL	Off	OffM	OffL
2A	Gr-scarf / **Bl-scarf**	"blue scarf"	"scarf"	92 *(2)*	74	90	9 *(2)*	11	50
2B	Bl-scarf / **Gr-hat**	"green"	"hat"	52 *(3)*	50	50	43 *(3)*	50	50
2C	**Gr-hat** / Bl-scarf	"hat"	"green hat"	33 *(3)*	37	37	55 *(3)*	50	50
2D	**Bl-scarf** / Bl-hat / Gr	"blue"	"blue scarf"	13 *(4)*	9	9	40 *(3)*	50	50
2E	**Gr-scarf** / Gr-hat / Bl	"scarf"	"green"	83 *(4)*	89	89	55 *(3)*	50	50
2F	Gr-hat / **Bl-hat** / Bl-scarf-hat	"n/a"	"blue"	7 *(1)*	30	14	91 *(3)*	92	50
2G	**Gr-hat** / Gr-scarf-hat / Bl-hat	"hat"	"green hat"	15 *(3)*	14	14	52 *(3)*	50	50
2H	Bl-hat / Gr-scarf-hat / **Gr-hat**	"hat"	"n/a"	74 *(5)*	50	50	54 *(2)*	50	50

"blue", those utterances were rated as equally (un)offensive (e.g. context 2G). If both utterances contained the word "blue", then those utterances were roughly rated as equally offensive (e.g. context 2D), but see minor trends below.

Overall, model predictions and participants' judgments were highly correlated (Figure 4). Bootstrapped confidence intervals (alpha = .025, adjusted for multiple comparisons, 10^4 samples) for adjusted R^2 were [.72,.90] for ambiguity and [.90,.98] for offensiveness; for Spearman's ρ intervals were [.85,.96] for ambiguity and [.66,.90] for offensiveness.

Differences between model predictions and judgments The behavioral responses did, however, differ from the model in a few systematic ways. An important trend that occurred in behavior was that people found utterances to be much less informative if redundant traits were not listed (e.g. context 2H). While the model predicted that saying "the person with the hat" and "the person" would be equally informative if all possible referents were wearing hats, participants found "the person" to be much more ambiguous. While this desire for "redundant overinformativity" is not captured in our model, it is often observed in referent games (Degen et al., in prep).

However, some preference for information redundancy was indeed captured by the model through the noisy semantics assumption. In context 2C, the model predicted that an utterance with two informative words is less ambiguous than an utterance with one informative word— because a listener with "noisy hearing" might miss one.

Another systematic divergence between model predictions and participants' judgments was that when asked about ambiguity, the model engaged in social inference more than participants did (e.g. contexts 2A, 2F). For example, if an utterance was "the person" when one possible referent was blue and the other green, the model made the social inference that the speaker was trying to refer to the blue person and predicted the utterance "the person" to be less ambiguous than it would have been without the social inference. However, in this setup, participants rarely appeared to make this inference. Instead, participants seemed to treat the ambiguity question as separate from the knowledge they were demonstrating in the offensiveness question (in which they were indicating that the term "blue" was potentially offensive.)

The results comparing the full model to the lesioned model (social utility set to 0) support the above hypothesis. When the social considerations were removed, the model predictions for ambiguity became *closer* to the behavioral results (e.g. context 2A). Numerically, for ambiguity ratings, bootstrapped confidence intervals (alpha = .025, 10^4 samples) for adjusted R^2 were [.78,.95] for the lesioned model (compared to [.72,.90] for the full model) and for Spearman's ρ were [.91,.98] for the lesioned model (compared to [.85,.96] for the full model). (The equivalent comparison with the lesioned model for offensiveness ratings was trivial by design, as the lesioned model was always ambivalent over utterances.)

The finding that participants did not engage social reasoning when asked about ambiguity may be due to question framing. "Offensiveness" and "ambiguity" ratings were clearly delineated in Experiment 2, and the focus on answering each separately (in addition to the extra training scenes that differentiated them) may have discouraged social reasoning to crossover into inferences about ambiguity.

On the offensiveness question, the differences between model predictions and participants' judgments were relatively small. Interestingly, participants considered any mention of color as slightly more offensive than model predictions, even if that color was non-offensive (e.g. contexts 2B, 2C). Participants also considered it slightly more offensive to say a color term if no other features were mentioned (e.g. contexts 2D, 2E), or to say "the person" alone (e.g. context 2H). These results are intuitive: if a feature like "blue" is offensive, it sug-

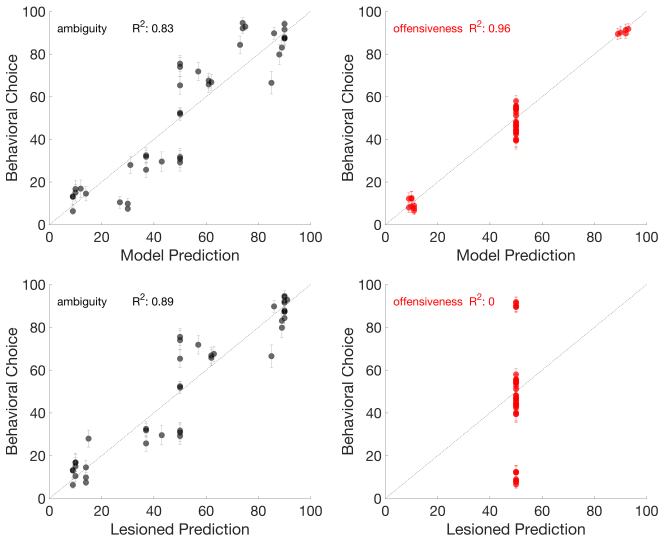

Figure 4: Behavioral and model comparison for Experiment 2. Participants rated which of two utterances describing a scene was more ambiguous (**left**), and which was more offensive (**right**) in 40 contexts. Behavioral results are the mean and standard error of participants' ratings of utterances, ranging from 0 (the utterance to the left of the screen was rated most ambiguous/offensive) to 100 (the utterance to the right was rated most ambiguous/offensive). Thus, lower scores indicate that the left utterance was rated more highly (more ambiguous / offensive) than the right utterance, and higher scores indicate that the right utterance was rated more highly than the left utterance. Model responses are the rescaled difference between β_{epis} / β_{soc} for the left and right utterances. Adjusted R^2 values are reported. **Top**: Full model. **Bottom**: Lesioned model (social utility set to 0).

gests that the general category of color might be be avoided; and it feels rude to not say anything when referencing someone. Future work will probe how to add these intuitions into a richer, hierarchical model that draws generalizations ("don't refer to color") from specific instances ("don't say blue").

Conclusion

Some words are potentially offensive. This means that in some situations, the most efficient way of referring to someone may incur a social cost, creating a tension between efficiency and social adeptness of speech. We hypothesized that when listeners and speakers have shared knowledge of this tension, speakers can avoid using offensive speech and listeners can resolve otherwise ambiguous utterances to correctly infer the speaker's intended referent.

To make these ideas precise, we built on an existing model of polite language understanding by introducing a *social cost* that a speaker incurs for producing potentially offensive language. The model captures the inference that people make in determining a speaker's intended referent given an utterance that is ambiguous but constrained by social cost (Experiment 1), and also captures the explicit access that participants have to a speaker's epistemic and social goals given their utterance and context (Experiment 2). This work shows how the general mechanism of reasoning about the social function of language employed by the speaker (Yoon et al., 2016, 2017) can begin to explain how listeners reason from the absence of potentially offensive language to resolve reference in context. While the model overall provides a very good fit to participants' inferences and judgments in both experiments, there were also some discrepancies which motivate future extensions of the model.

In our model, we directly mark potentially offensive words with a social utterance cost, but the same word might be offensive in one context and not another, or if said by one speaker but not another. One possibility is that it is a derivative property of subjective values associated with world states, in the style of Yoon et al. (2016), perhaps by speakers putting themselves in the listener's shoes and imagining themselves being referred to in a particular way. Another possibility is that these costs arise from social signaling: the speaker does not want the listener to infer that they are the type of person that calls people "blue". In future work we hope to investigate how the social cost of potentially offensive speech is grounded in the complex social inferences that listeners and speakers draw about each other.

Acknowledgments This work was supported by NSF STC award CCF-1231216 to the Center for Brains, Minds & Machines, the Marine Biological Lab. in Woods Hole, the Berkeley AI Travel Stipend & NIH Berkeley Neuro. Grant to MAG, the NDSEG & the UCSF Discovery Fellowship to TV, a Philippe Foundation grant to LB, & NSF Graduate Research Fellowship DGE-114747 to MHT.

References

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

Degen, J., Hawkins, R. X., Graf, C., Kreiss, E., & Goodman, N. D. (in prep). Over 'overinformativeness': rational redundant referring expressions.

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

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. *The handbook of contemporary semantic theory, 2nd ed.. Wiley-Blackwell*.

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.

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proc. of 38th conf. of cog. sci. society.*

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

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 *Proc. of 38th conf. of cog. sci. society.*

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proc. of 39th conf. of cog. sci. society.*