

The Alignment Model of Indirect Communication

Asya Achimova, Michael Franke, Martin V. Butz

June 26, 2024

Abstract

Speakers often choose utterances under uncertainty about the potential opinions of the listener. In this case, utterances that do not signal the speaker’s opinion directly may allow the speaker to avoid possible conflict. Adopting the Rational Speech Act framework, we develop a model of indirect communication that is able to (1) rationalize the choice of indirect utterances when speakers’ opinions do not align; (2) capture complex reasoning about the interlocutor’s opinion based on their reaction to an initial statement. The model has several novel features: the social utility component in the speaker function is computed by simulating potential divergences of opinions between conversation partners. This simulation becomes possible due to the inclusion of higher-order beliefs in the model machinery. The model further considers multi-turn dialogues rather than isolated utterances when inferring listener’s opinion. We are therefore able to take a longer planning horizon into account when calculating the probability of utterances and use preceding discourse in social reasoning. The model, though complex, makes novel, non-trivial qualitative predictions, which are supported by data from three behavioral experiments reported here.

1 Introduction

Human linguistic communication is not a simple encoding-decoding process based on a fixed code, but relies heavily on pragmatic reasoning about the context and the conversation partners’ beliefs and intents (Sperber & Wilson, 1995). Building on the seminal work of Grice (1975) and fueled by a more recent “empirical turn” (Noveck & Sperber, 2004; Noveck, 2018), formal approaches to pragmatics, and particularly the Rational Speech Act framework (Frank & Goodman, 2012), have spawned probabilistic models of contextualized reasoning and expression choices, which account for many different phenomena, including the meaning of number words (Kao et al., 2014b), scalar implicature (Goodman & Stuhlmüller, 2013), metaphor (Kao et al., 2014a), reference (Frank & Goodman, 2012; Goodman & Frank, 2016), and vagueness (Égré et al., 2023; Schöller & Franke, 2017) to name but a few cases (see Degen, 2023, for a recent overview).

The prevalent focus of formal, probabilistic, and experimental pragmatics has long been on a particular mode of language use, namely the cooperative communication of relevant information about the world. A few exceptions include models of strategic communication where values of conversation partners do not align (Asher & Lascarides, 2013; Pinker et al., 2008) and models of use-conditional meaning (Qing & Cohn-Gordon, 2019). More recently, the field is undergoing a second, this time ‘social turn’, focusing increasingly on more

social aspects of meaning and communication, including models of signaling one’s persona (Burnett, 2017, 2019) and ideology (Henderson & McCready, 2019, 2021), learning about others (Achimova et al., 2022, 2023), and appearing polite (Carcassi & Franke, 2023; Yoon et al., 2020). Recent experimental studies have targeted the role of face management in determining the meaning of scalar expressions (Bonneton et al., 2009). Along related lines, the interactions of power, social distance, and gender as well as their influence on face management strategies in verbal communication were studied (Gotzner & Mazzarella, 2021).

This work seeks to contribute to this growing literature on probabilistic modeling and experimental investigation of social factors in language use. We introduce a novel extension of the Rational Speech Act family of models, which considers the gradual, dynamic, multi-turn sharing of opinions through the strategic use of indirect messages and pragmatic reasoning in cases where conversationalists may have reason not to reveal their beliefs, stances, and opinions directly.

1.1 Background

Sharing mental attitudes, such as beliefs, preferences, and assumptions—which we here collectively address as *opinions*—is a critical component of interpersonal relations, group formation, and bonding (Higgins, 2019; Rossignac-Milon et al., 2020). The motivation to share mental states with other people develops early in life. It manifests itself already in the apparent desire of infants to share significant experiences with their caretaker before their first birthday (Tomasello, 2019). Experimental evidence further suggests that preschoolers prefer puppet toys that are similar to themselves in physical appearance and food preferences (Fawcett & Markson, 2010). Mahajan & Wynn (2012) argue that the ‘like me/not like me’ dichotomy is important already to pre-linguistic infants, who prefer others who share similar traits with them. The authors maintain that similarity to self is an inherent preference exhibited by humans and further emphasize the importance of similarity for interpersonal attraction. Dissimilarity and conflicts in beliefs and attitudes, in turn, may damage the relationship between interacting partners. During a conversation, monitoring whether an utterance carries a risk to the relationship is one of the factors that determines the speaker’s utterance choices. For example, Brown & Levinson (1987) conceptualized such social considerations in the notion of *face* and argued that face preservation is a major motivational force that shapes human interactions. As a result, in social interaction contexts humans are confronted with the objectives to align with their conversation partners while staying true to themselves.

Aligning opinions requires care, restraint, and decency. Consider the case of two researchers in Cognitive Science, Alex and Bo, seeing a poster for a big conference on Machine Learning. If Alex and Bo do not know each other well, they might not know their respective attitudes towards ML. Maybe Alex fears that Bo strongly believes that the public attention on engineering-based ML is a risk to explanatory fields like Cognitive Science with its focus on natural intelligence. Yet Alex might herself believe that ML offers many interesting chances for comparing information-processing strategies in biological neural systems with suitable ML models. To start the conversation, a (non-sarcastic) statement like “The recent ML advances are truly amazing!” may not serve, so that, in order to see what Bo thinks about the matter, Alex might opt for an indirect utterance like in (1). From Bo’s subsequent

reaction, as exemplified in (2), Alex may be able to infer a great deal about Bo’s beliefs about the matter.

- (1) It will be interesting to see how Machine Learning will impact research in Cognitive Science.
- (2)
 - a. Yes, we will soon all be unemployed.
 - b. Yes, we will see exciting new work in computational cognitive modeling.

The model developed in this paper, the *Alignment Model of Indirect Communication* (AMIC), intends to capture exactly this kind of use of indirect language to explore whether or where opinions—used here as a catch-all term for various stances and attitudes—are shared, so as not to risk loss of face. AMIC builds on related earlier work on probabilistic models of politeness (e.g., Yoon et al., 2016, 2020; Carcassi & Franke, 2023) and other models of social meaning (e.g., Burnett, 2017, 2019; Henderson & McCready, 2019). Such models generalize more basic models of pragmatic language generation, which assume that the speaker’s preferences for selecting utterances pivot almost solely around considerations of truth and informativity (about the world). To explain social language use, these models assume that the speaker’s preferences also include a social payoff component, next to informativity. A key observation of previous modeling is that, if we define the speaker’s preferences for utterances as a composite utility function roughly like

$$\text{Utility}(u) = \omega \text{Informativity}(u) + (1 - \omega) \text{SocialValue}(u),$$

then we find that speakers tend to produce less informative utterances—here addressed as *indirect* (Asher & Lascarides, 2001; Pinker et al., 2008; Searle, 1975; Terkourafi, 2014)—to the extent that informativity and social value of an utterance diverge and the speaker prefers to emphasize the social dimension of language use over the informative, as captured by the weight parameter ω . For example, this explains why an utterance of a double negative as in (3), though literally compatible with a wide range of values, is a good evasive strategy if one wants to be polite (see Gotzner & Mazzarella, 2021, 2024, for an overview of the effect of negation on the meaning of adjectives).

- (3)
 - a. Alex: How did you like my cookies?
 - b. Bo: They were not bad.

Indirect language use is associated with ambiguity—a situation where an utterance features multiple possible meanings. More specifically, we will be concerned with pragmatic ambiguity—a property of utterances that emerges in discourse when the meaning of the utterance as a whole may change depending on the context (Winter-Froemel & Zirker, 2015), world knowledge, or beliefs of conversation partners. Evidence from computational modeling suggests that ambiguity may offer necessary flexibility to conversation partners to adjust word meanings to each other (Brochhagen, 2020).

1.2 Contribution

So far, most probabilistic models of social or polite language use have focused on situations where the social value function was essentially known to the speaker. This is a reasonable

assumption, for example, for the case of evaluating the impact of statements like (3) on the addressees self-esteem. But to explain the case of dynamically exploring potential (mis-)alignment of opinion like in (1)–(2), AMIC will include a novel social value function, which is (i) subject to uncertainty, and (ii) considers utility arising from alignment of opinions, with (higher-order) uncertainty about these opinions.

The second and arguably more critical contribution of this work is the inclusion of *multi-turn inferences*, which may capture the dynamic process of opinion alignment as the dialogue unfolds over several turns. Previous attempts to model multi-turn interactions have concerned question-answer pairs (Hawkins et al., 2015), dialogues that aim at establishing references to location (Vogel et al., 2013), and sequential language games where agents cooperate to jointly establish the identity of references given that each of the agents only has partial information about the target object (Khani et al., 2018). In this work, we turn to the inferences conversation partners draw about each other rather than objects in the world by reasoning about sequences of utterances that have been added to the conversation previously. This inference process allows conversation partners to establish the meaning of utterances in a dialogue.

While AMIC is generally able to model multi-turn inferences across many turns of dialogue, we here focus on two-turn dialogues for reasons of complexity. Just as established work in Conversation Analysis has shown that much can be learned from studying pairs of dialogical utterances, so-called adjacency pairs (Sacks et al., 1974; Schegloff, 1984), we argue that our focus on two adjacent turns in a conversation, rather than isolated utterances, on the one hand, or unstructured dialogue, on the other hand, enables systematic investigation of basic mechanisms of opinion inference during conversations.

Even with the restriction to two-turn conversations, AMIC showcases how (i) speakers may strategically choose to be indirect and (ii) how they learn about their partner’s opinion from the subsequent (linguistic) reaction to their (in)direct utterance. Indeed, we present novel empirical data that supports AMIC’s general predictions about the effects on speaker choices of utterances arising from beliefs about mutual opinion and speaker’s weighing of social value. More strikingly, based on simulation studies, we find that the model makes novel, subtle, but non-trivial predictions about interpretations in two-turn dialogues, which we test empirically as well.

The phenomenon of aligning opinions—widely construed—through multi-turned indirect communication is relevant beyond negotiating opinions about Machine Learning, as in the above example (1). While we will not pursue further potential applications of AMIC in this paper, there is an important parallel to indirect language used for *plausible deniability* (e.g. Pinker et al., 2008; Lee & Pinker, 2010). Consider the extended dialogue in (4), taken from Peet (2024).

- (4) Bassa: Is there a problem officer?
Officer: Do you realize you were going 50 in a 30 zone?
Bassa: I didn’t realize, I’m very sorry. Look, I have plenty of cash on me. Is there any way that we could settle this right now?
Officer: You know it is illegal to bribe a police officer?
Bassa: Of course! I wasn’t trying to bribe you! I was just wondering if I could pay the fine here and now!

While the previous literature on plausible deniability has largely focused on the conditions under which indirect utterances are plausibly deniable, and the implications of this question for linguistic theory and the philosophy of language, the dialogue in (4) demonstrates that the dynamics of incrementally reducing uncertainty about the interlocutor’s stance or opinion (e.g., on whether bribing is an option) also affect the indirect use of language to maintain plausible deniability until more is known about the opinion (beliefs, preferences, etc.) of the conversation partner. In sum, sidestepping conceptual questions after plausible deniability, AMIC offers a model of the general belief dynamics of aligning opinions with indirect language, which is likely relevant for plausible deniability and other interesting phenomena.

The paper is structured as follows. We first introduce the vanilla RSA architecture and discuss previous RSA-extensions that introduce complex utility functions. We then propose a novel model which implements social utility from opinion alignment. We showcase interesting predictions of this model, derived from simulation. Finally, we demonstrate that the qualitative patterns predicted by the model are confirmed with behavioral data.

2 The Alignment Model of Indirect Communication

The goal of the Alignment Model of Indirect Communication (AMIC) is to account for (i) the choice of indirect utterances when the speaker pursues multiple conversational objectives; (ii) the inferences the speaker draws about the opinion of the listener upon observing her reply to an indirect utterance. We develop the AMIC within the Rational Speech Act framework (Frank & Goodman, 2012; Franke & Jäger, 2016; Goodman & Frank, 2016), which models pragmatic communication in terms of speakers choosing utterances to maximize their conversational goals, and listeners using inverse reasoning about the speakers’ policy to infer the speakers’ intended messages from the observed utterance. Building on previous models in this tradition, we introduce a social utility function to model the speaker’s goal of avoiding apparent conflict in opinion and consider a larger interpretation horizon of multiple utterance. To do so, we introduce a novel formalization of (higher-order) belief about opinions, and explore several measures of opinion alignment in the context of the model. The following first introduces a vanilla formulation of the RSA model and previous extensions of it to cover social reasoning, before introducing the AMIC.

2.1 Vanilla RSA

The vanilla RSA model defines probabilistic choice policies for the speaker and the listener. The speaker selects an utterance u for a given state (a meaning to be communicated) with conditional probability $P_S(u | s)$, which is proportional to the utterance’s utility $U_{S_1}(u, s)$ for state s , using a parameterized soft-max function:

$$P_{S_1}(u | s) \propto \exp(\alpha \cdot U_{S_1}(u, s)) \quad (1)$$

For simple applications, e.g., for referring to an object from a list of potential referents or to describe a world state from a known set of alternatives, the utility function of the speaker can be defined in terms of the negative surprise of a literal interpreter L_0 :

$$U_{S_1}(u, s) = \log P_{L_0}(s | u) \quad (2)$$

This utility function can alternatively be motivated as the goal of minimizing the distance between the speaker’s belief (which is here assumed to be a degenerate probability distribution ruling out all but one world state s) and a literal interpreter’s belief after hearing utterance u (Goodman & Stuhlmüller, 2013; Scontras et al., 2021; Égré et al., 2023).

The literal listener can essentially be thought of as a construct to ground out pragmatic reasoning in a base layer of semantic meaning, and may simply be defined as a choice of a state proportional to how true it is, for some (Boolean or non-Boolean) semantic meaning function f that maps pairs of utterances and states onto numbers in the unit interval:

$$P_{L_0}(s | u) \propto \mathbf{f}(u, s) \tag{3}$$

Finally, pragmatic interpretation is formalized as a pragmatic listener who uses Bayes rule to solve the inverse problem of recovering the latent state s based on their prior beliefs and the speaker’s policy:

$$P_{L_2}(s | u) \propto P_{S_1}(u | s) \cdot P(s) \tag{4}$$

2.2 RSA models for social meaning

The speaker’s behavior in the vanilla RSA model is driven by the goal of signalling the intended meaning efficiently. As such, the vanilla RSA does not cover situations where speakers choose indirect utterances for social reasons, such as politeness. In order to account for such additional social considerations, the utility function of the speaker can be extended to also include additional goals beyond being informative. For example, work on politeness has suggested extending the speaker’s utility function to be a linear combination of the desire to be informative and to maximize the emotional well-being of the listener (e.g. Yoon et al., 2016, 2020; Carcassi & Franke, 2023).

$$U(u, s, \gamma) = \gamma U_{\text{informative}}(u, s) + (1 - \gamma) U_{\text{value}}(u, s) \tag{5}$$

Here, the utility component to be informative is as defined above, and the novel social utility component can be defined in terms of how much the literal interpretation of utterance u pleases the listener when the true state is s .

If the speaker optimizes solely the social utility ($\gamma = 0$), she would be expected to select only positive utterances. Setting priority to sending fully true information ($\gamma = 1$) would result in a preference for direct utterances. A combination of these goals ($0 < \gamma < 1$) leads polite utterances being chosen, which tend to be indirect. However, since social utility in previous work on politeness is defined via the emotional value that an utterance carries for the listener, these models do not generalize to other cases of indirectness, where the speaker’s goals may not be about producing positive feedback. For example, an utterance, such as (1), where the speaker wants to learn about listener’s opinion on machine learning, does not have a straightforward low or high politeness utility.

In the following section we introduce AMIC, which operationalizes the social utility of utterances by formalizing belief alignment objectives. We start with proposing a more general formalization of opinions, which are then exchanged in the alignment model.

2.3 Opinions and their degree of alignment

The general idea of the AMIC is that speakers choose more indirect expressions if they are not sure that their own opinion aligns well with that of the interlocutor(s). Spelling out this idea formally requires making assumptions about how to represent opinions and how to measure alignment between them. It is common in models of opinion dynamics (e.g. DeGroot, 1974; Hegselmann et al., 2002; Castellano et al., 2009) to focus on the simplest case of opinions, namely opinions about a binary issue (such as whether abortion should be legal, veganism is good, climate change is human-made, etc.), and to represent an agent’s opinion simply as a number $o \in [0; 1]$ on the unit interval. The opinion o is then a single number representing the agent’s *position*, i.e., how much the agent agrees with the binary issue. For our purposes, this representation of opinions is not fine-grained enough, because we would like to represent two relevant dimensions:

- (i) **position**: to what extent does the agent tend to agree with the issue?
- (ii) **opinionatedness**: how large or small is the range of positions on the issue that the agent would find acceptable?

We therefore represent an agent’s **opinion state** in terms of a probability distribution on the unit interval. As a choice of convenience, we consider the class of Beta distributions. A Beta distribution can be parameterized in terms of its mean $\mu \in [0; 1]$ and “sample size” $\nu > 0$.¹ The mean μ can be interpreted as the agent’s position or bias, and the sample size ν can be interpreted as the agent’s opinionatedness, where $\nu = 0$ corresponds to a uniform distribution over $[0; 1]$, that is, effectively no, or a fully ambivalent, opinion. As a result, ν reflects how much evidence the agent has accrued to back up her position. The set of all opinion states \mathcal{O} is then given by all Beta distributions (with $\mu \in [0; 1]$ and $\nu \geq 0$). We denote the listener’s and speaker’s opinion as O_L and O_S respectively. Figure 1 illustrates five opinions encoded as Beta distributions.

When representing an opinion by means of a density that is parameterized via two parameters, a measure of alignment between two agents’ opinions should be sensitive to both parameters. If we represent opinions as probability distributions, we can use information-theoretic measures of divergence or distance between probability densities, which are sensitive to both expected value and variance of the distributions they relate. Concretely, we want a measure of **opinion divergence** to be a function:²

$$\text{Div}: \Delta(\mathbb{R}) \times \Delta(\mathbb{R}) \rightarrow \mathbb{R},$$

that maps a pair of opinion states onto a real-valued measure of how much the opinion states diverge from each other. In the following, we use a symmetrized version of Kullback-Leibler divergence to measure alignment. If P and Q are probability distributions, we define opinion divergence as:

$$\text{Div}(P, Q) = D_{\text{KL}}(P||Q) + D_{\text{KL}}(Q||P),$$

¹A Beta distribution is usually defined with parameters α and β . We will use the symbols β_1 and β_2 , correspondingly to refer to these parameters to avoid confusion with the α parameter of the RSA models. Starting from $\beta_1, \beta_2 \geq 1$, as the usual parameters of the Beta distribution, this alternative parameterization is obtained via the one-to-one mapping: $\mu = \frac{\beta_1}{\beta_1 + \beta_2}$ and $\nu = \beta_1 + \beta_2 - 2$.

²As for notation, we write $\Delta(X)$ as the set or space of all probability distributions over X .

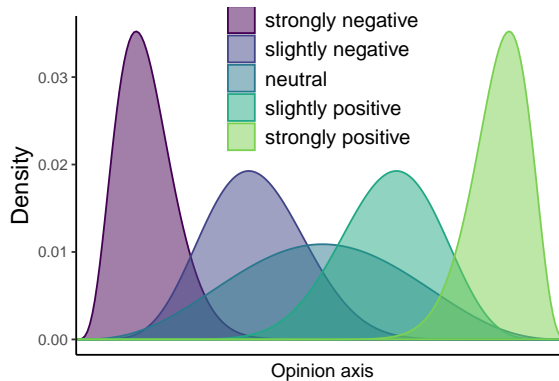


Figure 1: Examples of different opinion states as Beta distributions with different values for parameters ‘position’ μ and ‘opinionatedness’ ν . The five densities’ parameters are (in order of increasing ‘position’): $\mu_1 = .14$, $\nu_1 = 35$, $\mu_2 = .36$, $\nu_2 = 22$, $\mu_3 = .5$, $\nu_3 = 8$, $\mu_4 = .64$, $\nu_4 = 22$, $\mu_5 = .86$, $\nu_5 = 35$.

where D_{KL} is the KL-divergence.³

2.4 Higher-order beliefs about opinions

The AMIC assumes that the pragmatic choice of an utterance as well as the interpretation of an utterance are sensitive to opinion alignment. But there can be uncertainty about the interlocutor’s opinion (first-order uncertainty), uncertainty about the interlocutor’s first-order uncertainty (second-order uncertainty), and so on. While AMIC does not go beyond second-order uncertainty, it may be useful, nonetheless, to have a general notation for potentially even higher-order beliefs.

Let X be the listener L or the speaker S , and Y be the respective other agent. If O_Y is agent Y ’s opinion, then π_1^X is agent X ’s (first-order) belief about agent Y ’s opinion. Formally, $\pi_1^X \in \Delta(\mathcal{O})$ is a probability distribution over the space of all opinion states (here: the space of Beta distributions). For any $i > 1$, π_i^X is agent X ’s (i -th order) belief about agent Y ’s ($i - 1$)-th order belief. For example, a second-order belief of agent X is a probability distribution $\pi_2^X \in \Delta(\Delta(\mathcal{O}))$, i.e., a probability distribution over probability distributions over Beta distributions. In other words, the second-order belief of X , that is, π_2^X , denotes a distribution over potential first-order beliefs of Y , that is, π_1^Y , about the possible opinions of X , that is, \mathcal{O}_X .

As for notation, we interpret expressions π_i^X as random variables and write $P_{X_1}(O_Y | \pi_1^X)$ to represent the probability for a particular opinion O_Y . For example, we write $P_{S_1}(O_L | \pi_1^S)$ to represent a pragmatic speaker’s beliefs about the listener’s opinions.

³Other information-theoretic measures of divergence or distance are conceivable. Figure A.1 in the appendix shows divergences between the five opinion states from Figure 1, for symmetrized KL-divergence and some salient alternatives. Except for the very regular Earth Mover’s distance measure, simulations with the alternative divergence measures yield similar qualitative predictions when used in the final model.

2.5 Literal interpretation

As explained above, the vanilla RSA model grounds out pragmatic reasoning in a layer of literal interpretation, often formally represented as a literal listener. The formulation given above in Equation 3 assumes that there is a semantic meaning function $f : s, u \mapsto [0; 1]$ which maps pairs of states and utterances to truth (or “truthiness”) values. For our application, we are interested in what expressions like in (5) below reveal about the opinion of a speaker. In the present context, we sidestep the linguistic question of how such statements are related to information about opinions (in the wide sense that we endorse here). We will simply assume, for the time being, that there is a literal interpretation function $L_0(u) \in \mathcal{O}$, which assigns to each utterance u a distribution over positions (numbers in $[0; 1]$) usually associated with u based on its conventional meaning. For the simulations reported in this paper, we will provide an empirical measure of $L_0(u)$, as detailed below in Section 3.1.

2.6 Pragmatic speaker

The AMIC formalizes the situation in which a pragmatic speaker chooses between utterances in a multi-objective manner, attempting to (i) signal their own opinion and (ii) align with what they believe to be the listener’s opinions, given their first order belief π_1^S . The latter objective models the active avoidance of opinion conflicts, that is, strong opinion mismatches. For example, the AMIC will assign a higher probability to the utterance (5-b) than to (5-a) when $\pi_1^{S_1}$ is believed to oppose the speaker’s opinion O_{S_1} and the speaker has a strongly positive opinion about some matter. On the other hand, if the speaker has a positive opinion and believes that opinions are aligned, then the model will assign a higher probability to the utterance (5-a), since this utterance will have a high information utility and social utility.

- (5) [Speaker’s opinion in strongly positive.]
- a. The election outcome was amazing.
 - b. The election outcome was interesting.

Formally, the mental state of the pragmatic speaker is captured by their own opinion O_{S_1} and their beliefs about the opinion of the listener $\pi_1^{S_1}$. Given O_{S_1} , $\pi_1^{S_1}$ and the assumed literal listener’s interpretation of utterances, that is, $L_0(u)$, we can define the two goals of the speaker as:

1. **informative goal:** $L_0(u)$ should be as close as possible to the speaker’s own opinion O_{S_1} , and
2. **social goal:** $L_0(u)$ should be as close as possible to the believed listener’s opinion, that is, $\pi_1^{S_1}$.

These two goals translate into two utility functions, where the social utility corresponds to

an expected utility over potential opinions of the listener:⁴

$$\begin{aligned}
 U_{\text{inf}}(O_{S_1}, u) &= -\text{Div}(O_{S_1}, L_0(u)) \\
 U_{\text{soc}}(\pi_1^{S_1}, u) &= -\int P_{S_1}(O_L | \pi_1^{S_1}) \text{Div}(O_L, L_0(u)) \, dO_L
 \end{aligned}$$

We use the information-theoretic notion of Kullback-Leibler divergence, as specified above, in the calculation of both information and social utilities. This divergence measure takes into account both the location of the distribution on the negative-positive scale and how peaked the distributions are. Thus, we are able to consider both the polarity of an opinion and the speaker’s certainty. The *total utility* U_{total} is a linear combination of these two, with parameter γ weighing their relative importance:

$$U_{\text{total}}(O_{S_1}, \pi_1^{S_1}, u) = \gamma U_{\text{inf}}(O_{S_1}, u) + (1 - \gamma) U_{\text{soc}}(\pi_1^{S_1}, u) \quad (6)$$

The speaker’s *utterance choice probability*, given their own opinion and a belief about the literal listener’s opinion, is the usual soft-max of the total utility:

$$P_{S_1}(u | O_{S_1}, \pi_1^{S_1}) \propto \exp\left(\alpha U_{\text{total}}(O_{S_1}, \pi_1^{S_1}, u)\right) \quad (7)$$

Appendix A.3 shows examples for numerical utilities and resulting speaker probabilities.

As an example, Figure 2 shows the probabilities predicted by the AMIC for uttering one of the five opinions given particular opinions of speaker and listener—where the opinions are modeled by the beta densities shown in Figure 1—and given a relative weighting $\gamma = 0.8$ of the informative utility. The AMIC predicts that speakers are more likely to choose an indirect utterance when they expect the listener to have an opposing opinion. The more the opinions are believed to align, the more likely becomes the probability to choose the most direct opinionated statement.

2.7 Pragmatic listener

The pragmatic listener L_2 uses the utterance-generating model of the pragmatic speaker, in concert with Bayes rule, to infer which mental state of the speaker (consisting of an opinion and a belief about the literal listener) could plausibly have led to the observed utterance. Consequently, the pragmatic listener’s mental state is a triple $\langle O_{L_2}, \pi_1^{L_2}, \pi_2^{L_2} \rangle$ consisting of: (i) L_2 ’s own opinion $O_{L_2} \in \mathcal{O}$, (ii) L_2 ’s first-order beliefs $\pi_1^{L_2} \in \Delta(\mathcal{O})$ about the speaker’s opinion, and (iii) L_2 ’s second-order beliefs $\pi_2^{L_2} \in \Delta(\Delta(\mathcal{O}))$ about the speaker’s beliefs about the listener’s opinion. The posterior beliefs of the pragmatic listener about the speaker’s opinion are inferred by Bayes rule:

$$P_{L_2}(O_{S_1} | u, \pi_1^{L_2}, \pi_2^{L_2}) \propto \int P_{S_1}(u | O_{S_1}, \pi_1^{S_1}) P_{L_2}(O_{S_1} | \pi_1^{L_2}) P_{L_2}(\pi_1^{S_1} | \pi_2^{L_2}) \, d\pi_1^{S_1}, \quad (8)$$

⁴The model specification assumes that the speaker knows how the listener interprets utterances—it is part of their background beliefs. Further work may evaluate more complex scenarios where the listener’s interpretation is not fully transparent to the speaker. A potential solution lies in including lexical uncertainty into the model (Potts et al., 2015; Franke & Bergen, 2020; Bergen et al., 2016).

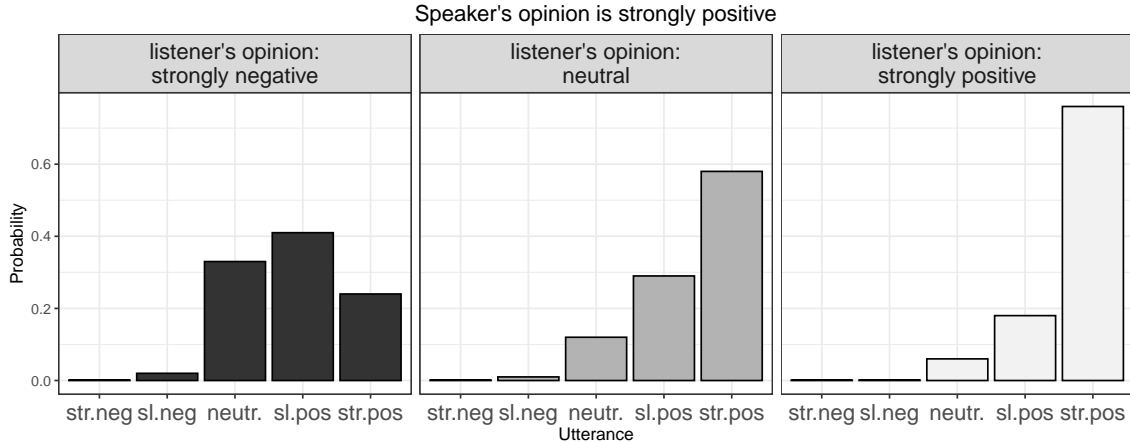


Figure 2: Utterance choice: model predictions. The speaker’s actual opinion is strongly positive. Her utterance choice depends on her opinion and the belief about the opinion of the listener, as well as the communicative goal. In this simulation, the informational and social goals are weighted at 0.8 and 0.2, respectively; the α parameter is set to 0.18. A higher value of α leads to more deterministic utterance choices that favor the utterance with highest utility. The left panel demonstrates that when the speaker believes that the listener has a conflicting opinion (*strongly negative*), she prefers a less direct utterance (*slightly positive* or even *neutral*) although her opinion is actually strongly positive.

which is the marginal distribution over the opinion O_{S_1} , marginalizing over the other component that the pragmatic listener is uncertain about, which is the speaker’s beliefs about the (literal) listener $\pi_1^{S_1}$. Notice that the AMIC only formalizes the inference of the mental state of the speaker that explains the observed utterance. It does not model how the listener may change her own opinion based on that inference—a challenge that we leave for future research.⁵

2.8 Learning about each other: a simulation

The pragmatic speaker protocol defined in Equation (7) describes a general way of choosing utterances in cases where the communication of opinions is important. Likewise, the pragmatic listener interpretation rule in Equation (8) describes a general format of inferring posterior beliefs about the speaker’s opinions after hearing an utterance, based on prior beliefs and the assumption that the speaker generates utterances following the protocol in Equation (7). Together, these production and interpretation rules provide a simple model of learning about each other’s opinion (see Figure 3). For example, after a first utterance u_A , which Alex chooses based on Equation (7), Bo may update her prior beliefs about Alex’s opinion using the rule in Equation (8). The posterior beliefs Bo obtains via Equation (8)

⁵How exactly listener’s update their own opinion based on what speakers say will require more elaboration, including factors like trust, status, competence, and the like, and possibly even considerations of utility (do I benefit, in the future, from adopting my neighbors’ beliefs?). A simple but compelling algorithm for opinion change is to adapt the parameters of the listener’s Beta distribution to be more aligned with the inferred speaker’s likely distribution.

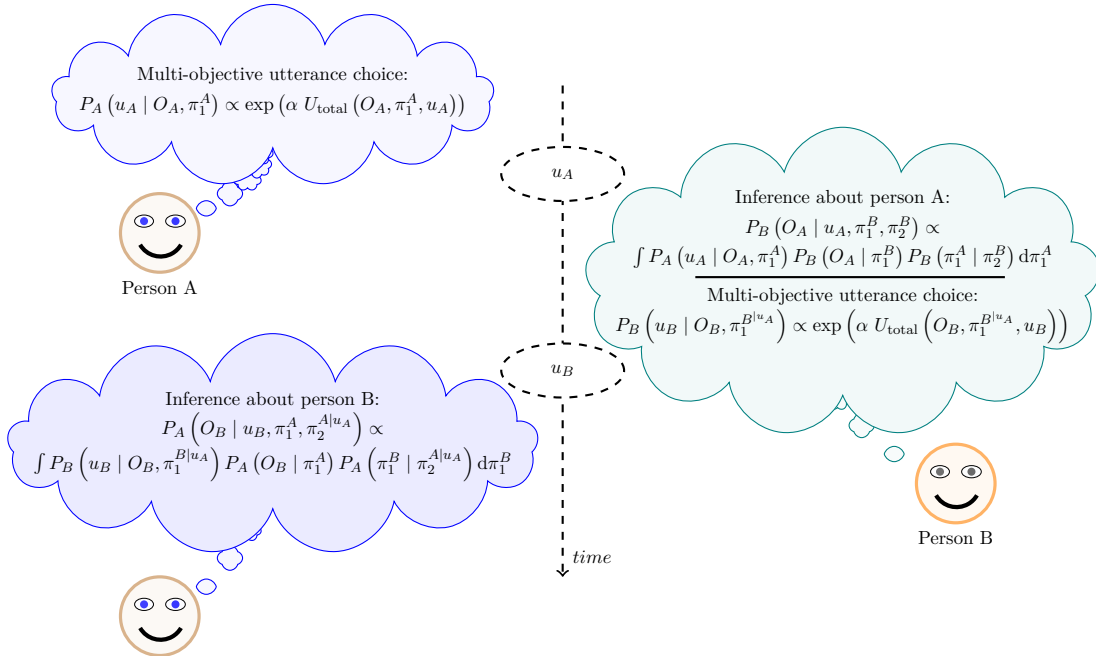


Figure 3: Multi-turn interaction in the alignment model. Each agent infers the other agent’s beliefs based on their prior beliefs about the interlocutor’s beliefs and the utterance probabilities that these prior beliefs about the interlocutor entail. In this plot, we diverge from talking about speakers and listeners and instead talk about persons A and B (e.g., Alex and Bo). This simplifies the notation so that, for example, π_2^B are person B ’s second order beliefs. We denote with $\pi_1^{B|u_A}$ person B ’s first-order beliefs (about person A ’s opinions) after interpreting utterance u_A and, similarly, $\pi_2^{A|u_A}$ person A ’s second-order beliefs (about person B ’s beliefs about person A ’s opinion) after having uttered u_A .

may then feed into her choice of a subsequent utterance u_B , again chosen via Equation (7). Finally, Alex may then interpret the utterance u_B via Equation (8) to learn from how Bo reacted (via u_B) to her utterance u_A . In this way, the model sketched here shows a path for agents to learn about each other’s beliefs.⁶

To assess the model’s predictions we simulate the interaction shown in Figure 3. In particular, we simulate the belief updates given person A chooses an utterance u_A to which person B provides a response u_B . The simulation assumes the informational weight $\gamma = 0.8$ —yielding a social weight of $1 - \gamma = 0.2$ —and a soft-max factor $\alpha = 0.18$ (Eq. 7).⁷ We represent opinions by means of the five beta distributions shown in Figure 1 and beliefs

⁶A particularly interesting possibility is that more sophisticated agents may use the sequential nature of this model to choose utterances strategically, based on their potential to reveal beliefs from anticipated follow-up utterances. For instance, Alex may choose a particular u_A also taking into account how much they will learn about Bo’s beliefs from the likely reactions u_B may trigger in Bo. Experimental evidence suggests that at least some speakers are capable of using ambiguity strategically to gain information about the interpreter’s prior preferences (Achimova et al., 2022, 2023), but we will not model this here.

⁷Similar values and similar densities yield similar results.

as probability masses over those five opinions. In particular, we commence with uniform prior beliefs over the five opinion densities as first and second order beliefs (i.e. π_1^A , π_1^B , π_2^A , π_2^B)—all of which we encode as probability masses over the five opinion densities. Moreover, person A and B are assumed to have one particular opinion O_A and O_B , respectively. The first utterance u_A leads to two updates. Person B updates her beliefs of the opinion of person A (i.e., $\pi_1^{B|u_A} \leftarrow \pi_1^B$). By the same process, person A updates what she believes that person B now believes about person A herself, seeing that she has revealed aspects about herself via her utterance u_A ($\pi_2^{A|u_A} \leftarrow \pi_2^A$). After the choice of the response u_B by person B, we finally compute the resulting belief that person A will have about person B’s opinion (i.e., $\pi_1^{A|u_B} \leftarrow \pi_1^A$).⁸

Table 1: Predicted probability distributions over inferred speaker B’s opinions (i.e., π_1^A) given a strongly positive, neutral, or negative statement of speaker A and either a slightly negative (*rather bad*) or a slightly positive (*decent*) response of speaker B.

	A’s posterior beliefs about B’s opinion				
	Strongly negative	Slightly negative	Neutral	Slightly positive	Strongly positive
<i>A: The election results are terrible</i> (strongly negative)					
B: <i>I find them rather bad</i>	0.16	0.35	0.28	0.19	0.02
B: <i>I find them decent</i>	0	0.08	0.18	0.32	0.42
<i>A: The election results are okay</i> (neutral)					
B: <i>I find them rather bad</i>	0.27	0.35	0.24	0.13	0.01
B: <i>I find them decent</i>	0.01	0.13	0.24	0.35	0.27
<i>A: The election results are amazing</i> (strongly positive)					
B: <i>I find them rather bad</i>	0.42	0.32	0.18	0.08	0
B: <i>I find them decent</i>	0.02	0.19	0.28	0.35	0.16

Table 1 shows selected results from these simulations.⁹ We see that when speaker A’s utterance u_A is strongly negative and speaker B chooses a slightly positive response, the model infers that speaker B’s actual opinion is most likely strongly positive (second row). A slightly negative response in this situation suggests that speaker B’s opinion might be slightly negative (35%) but also neutral (29%), or strongly negative (14%) (first row). If speaker A chooses a strongly positive utterance while speaker B responds with a slightly negative utterance, the model infers that the listener’s actual belief is most likely strongly

⁸As we implement π_2^A identical to π_1^A , that is, as a probability mass over five considered opinion densities, the integral effectively sums over all potential opinions that person A may have, when considering her utterance and starting from a uniform prior.

⁹A full simulation of all possible combinations of utterances and responses can be found in the Appendix (Figure 16).

negative (45% chance, previous to last row).

2.9 Summary: Modeling

In this section, we have presented the Alignment Model of Indirect Communication (AMIC), which formalizes the intuition that one reason for the use of indirect utterances is to avoid conflict between revealed opinions. It offers a mechanism based on formalizations of Bayesian inference that allows for the detection of opinion divergences without making them explicit. To represent the meaning of utterances and the opinions of conversation partners, we have used Beta distributions. We have then formalized beliefs about the beliefs of the conversation partner. While the formalization enables an infinite regress, our model uses the recursion up to the second order belief only, which is probably the typical cognitive limit in everyday conversations. We introduced a pragmatic speaker function S_1 that regulates the choice of utterances by balancing the informational and social goals. We have formalized the process of inferring the beliefs of a speaker following her utterance in the pragmatic listener function L_2 . The AMIC predicts that indirect utterances can become an optimal speaker’s choice when she is simultaneously pursuing informational and social goals. It further captures the fact that speaker’s opinion may differ from the literal meaning of her utterance. The AMIC also makes non-trivial predictions, as shown in Table 1, about opinion inferences in two-turn dialogues (of the kind shown in Figure 3). In the next section, we report on a set of empirical tests of the model and discuss to which extent AMIC reflects the qualitative patterns we witness in the data.

3 Behavioral data

In this section, we report the results of three experiments that were designed to obtain distributions that represent the meaning of predicates of personal taste (Experiment 1) and test the predictions generated by the alignment model. In particular, Experiment 2 targets the pragmatic speaker behavior, and Experiment 3 assesses how participants draw inferences about opinions, testing the predictions of the pragmatic listener layer of the model.

3.1 Experiment 1: Empirical baseline of utterance meanings

In Experiment 1, we obtain an empirical baseline of utterance meanings as it is represented within the literal listener layer of the model. The usual role of the literal listener in RSA models is to anchor pragmatic reasoning in literal interpretation. The AMIC requires a literal listener that captures how various utterances relate to opinion states. Concretely, we consider a literal listener as a function f that maps an utterance into opinion space, so that $L_0(u) \in \mathcal{O}$, where the precise distribution may be determined from empirical data as described next.

To represent the meaning of utterances in the form of a distribution, we conducted an online experiment via the Prolific crowd-sourcing platform ($n = 50$, data from 4 participants were excluded due to reported confusion of the participants, data from the remaining 46 participants were entered into the analysis). We have obtained written consent from all participants and reimbursed them for their participation.

We followed the data-elicitation paradigm proposed by Yoon et al. (2016), who asked participants to evaluate predicates of personal taste, such as ‘good’ and ‘not bad’, by mapping them to a Likert-scale: the participants assigned a different number of hearts depending on their perception of the description and the stated speaker’s goals. In our experiment, we asked the participants to evaluate similar statements within a carrier phrase (6) on a heart-scale from 1 ‘strongly negative’ to 5 ‘strongly positive’:

- (6) I find the election outcome...
- a. amazing.
 - b. decent.
 - c. interesting.
 - d. poor.
 - e. terrible.

A sample trial is shown in Figure 4.

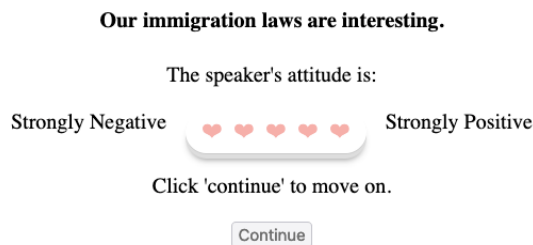


Figure 4: Experiment 1 (sample trial). Participants are asked to indicate the speaker’s attitude in hearts for ten predicates of personal taste.

We evaluated a total of ten different topics each featuring the same ten predicates of personal taste. Each participant assessed any given topic only once paired with one of the predicates selected from the list. Figure 5 displays the ratings assigned by the participants to each of these adjectives with all topics pooled together. It is these empirical distributions that we use as first approximations to the semantic meaning of utterances. The mean number of hearts assigned to each utterance also allows us to classify the utterances as strongly negative (rounded *mean* = 1 heart), slightly negative (2), neutral (3), slightly positive (4), and strongly positive (5). These distinctions are color-coded in Figure 5. Thus, for example, the utterances ‘terrible’ and ‘awful’ are strongly negative, while ‘amazing’ and ‘great’ are strongly positive.

The ratings we obtained do not directly indicate whether utterances are direct or indirect, since we define indirectness as a property of utterances that emerges in discourse rather than a characteristic of word semantics. Thus, an utterance, such as (7) may be judged as indirect if the speaker actually has a negative opinion about the election outcome (one or two hearts on our scale). The same utterance can be direct if the true belief state corresponds to four hearts.

- (7) I found the election outcome decent.

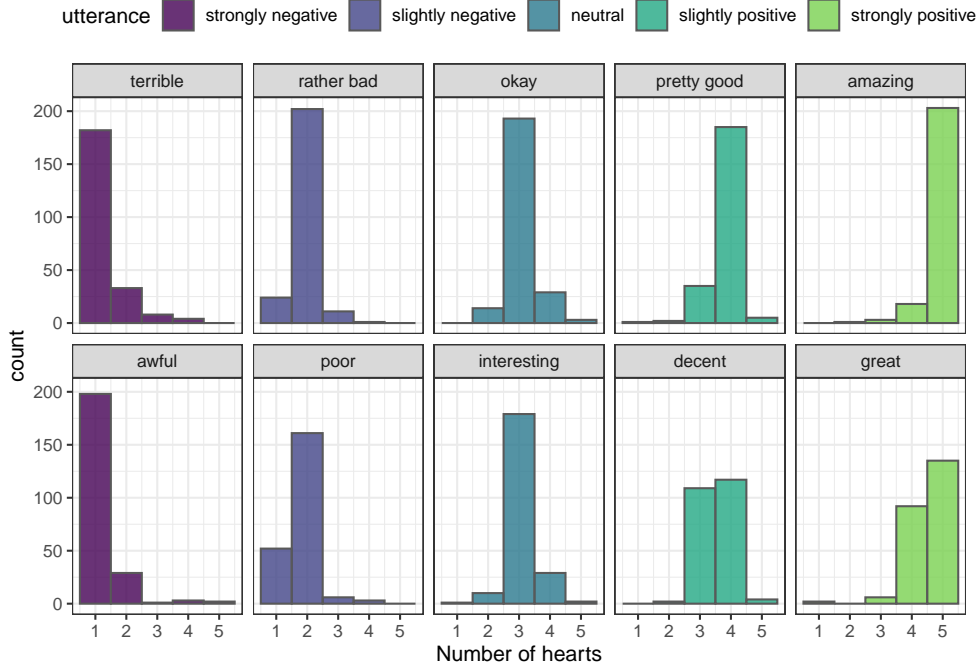


Figure 5: Utterance ratings for ten considered predicates of personal taste.

In sum, Experiment 1 provides a motivation for assigning the utterances to a scale from ‘strongly negative’ to ‘strongly positive’ and establishes a mapping between these categories and belief states represented in hearts.

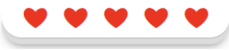
3.2 Experiment 2: Pragmatic speaker

Experiment 2 ($n = 97$, Prolific platform) was designed to assess how the communicative goal, the actual belief of the speaker, and an assumption about the listener’s belief affect utterance choices. Data from three participants was excluded since they reported they did not fully understand the instructions, data from one other participant was excluded due to a technical error. Thus, data from 93 participants was entered into the analysis. Figure 6 shows the experimental set up. The experiment was a $2 \times 3 \times 2$ within-subjects design. We manipulated the factor ‘match/mismatch’ of whether opinions of speaker and listener matched. A second factor for social goals has three levels: share opinion (informational), share opinion and avoid conflict (informational + social), or simply avoid conflict (social). Finally, we also varied whether the speaker’s opinion was positive or negative. Concretely, speaker’s and listener’s opinions were either strongly negative (one heart) or strongly positive (five hearts).

Based on the association of adjective meanings and the hearts scale established by Experiment 1, the alignment model predicts that speakers should select indirect utterances more often when they anticipate a mismatch in opinions and when they have social goals in addition to informational ones. Concretely, we were interested in two directional hypotheses about the proportions of choices of indirect utterance:

Adam wants to discuss the election results with Jeff.

Here is how Adam feels about the issue:

Strongly Negative  Strongly Positive

Adam thinks this is how Jeff feels about it, but he is not sure:

Strongly Negative  Strongly Positive

Adam wants to share his opinion and wants to be honest about it.

What would Adam say?

- The election results are awful.
- The election results are poor.
- The election results are interesting.
- The election results are decent.
- The election results are amazing.

Click 'continue' to move on.

[Continue](#)

Figure 6: Experiment 2 (sample trial). Participants are asked to select an utterance given the speaker's and listener's opinions and a communicative goal. In this case, the goal is purely informational.

H1: When social goals matter (the informational + social and the social conditions), we expect more indirectness in the mismatch condition than in the corresponding match condition.

- a. mismatch-informational + social > match-informational + social
- b. mismatch-social > match-social

H2: When opinions are mismatched, we expect more indirectness the more social goal matters:

- a. mismatch-social > mismatch-informational + social
- b. mismatch-informational+social > mismatch-informational

The distributions of participants' choices is shown in Figure 7. In this plot, we grouped utterances into three categories. Direct utterances correspond to the speaker's true opinion. Thus, strongly positive utterances were coded as direct if the true opinion of the speaker corresponded to five hearts. If the utterance matched the polarity of the opinion (e.g.

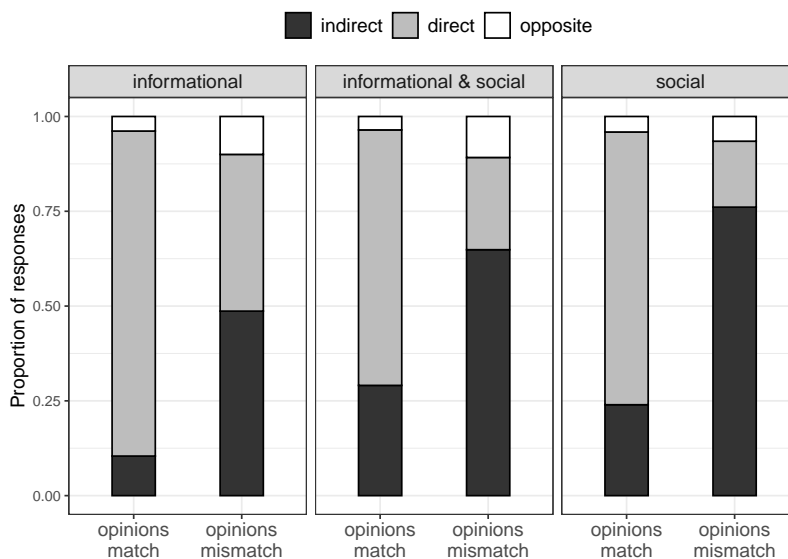


Figure 7: Experiment 2: Utterance choice. Utterances were categorized into three groups: indirect, direct, and opposite. Indirect utterances matched the true opinion in polarity but were less categorical. Direct utterances matched the opinion fully. Opposite utterances were those that did not match the true opinion polarity.

positive), but did not match the degree (slightly positive while the opinion was strongly positive), we counted such utterances as indirect. The ‘indirect’ category further included the neutral utterances. Finally, utterances that did not match the polarity of the true opinion were assigned to the ‘opposite’ category. The Figure 7 thus demonstrates the proportion of utterances in each of these categories broken by the type of communicative goals in the three facets. Figure 8 shows how the rates of different utterance types change depending on the communicative goal of the speaker and the polarity of the speaker’s opinion (strongly negative vs. strongly positive). Alternative figures with non-aggregated data are displayed in Appendix C.1 Figures 17 and 18.

To test our hypotheses, we ran a single Bayesian logistic regression model in which the dependent variable was binary (‘indirect utterance’ vs. ‘other’) using the `BRMS` package in R (Bürkner, 2018). The independent fixed effects were all main factors of the experimental design with all two- and three-way interactions. Additionally, we included the maximal by-participant random effects structure: by-subject random intercepts and by-subject random slopes for each fixed effect coefficient (including all interactions) (Barr et al., 2013). We retrieved samples for the posterior estimates of the predictors of central tendency for each design cell. Posterior samples were aggregated over the relevant subsets of cells and subtracted to test each one of the four hypothesized contrasts. We take the data and model to provide evidence in favor of a directed contrast if (the sampling-based approximation of) the posterior probability of the difference (aggregated central tendency of higher cell group minus that of the lower) is credibly bigger than zero, which we take to be the case if the 95% credible interval of the difference is entirely bigger than zero.

Based on this analysis protocol, we find that participants choose an indirect utterance

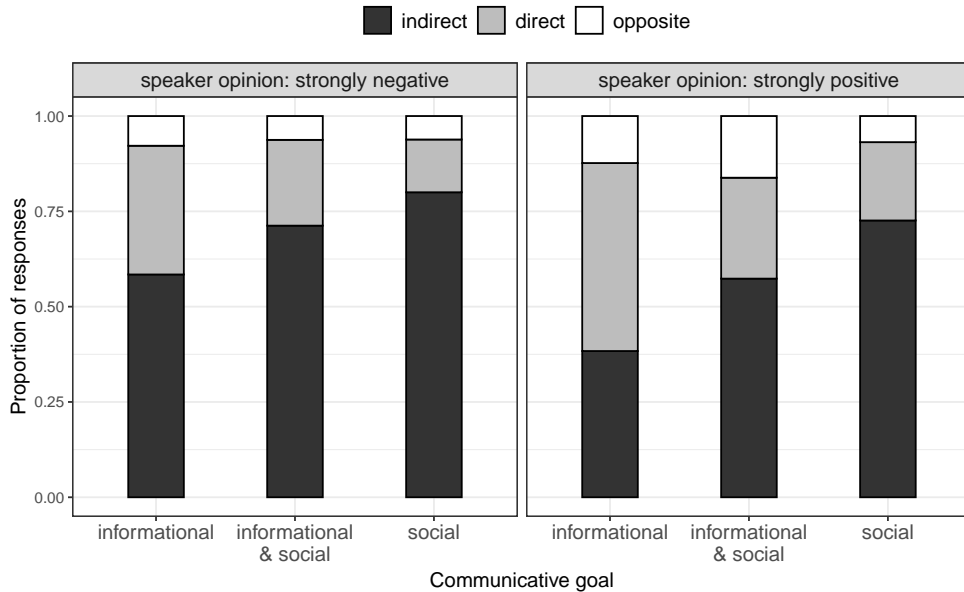


Figure 8: Experiment 2: Utterance choice when opinions mismatch broken by the polarity of speaker’s opinion.

reliably more often when the opinions of conversation partners did not match given that the speaker pursued a combination of informational and social goals (H_{1a} , posterior mean = 0.432, 95% credible interval = [0.339, 0.541]) or a social goal alone (H_{1b} , posterior mean = 0.545, 95% credible interval = [0.455, 0.634]). Contrary to our predictions, the rate of indirect utterances was not reliably larger for social goals compared to a combination of social and informational goals (H_{2a} , posterior mean = 0.0697, 95% credible interval = [-0.015, 0.0697]). This result suggests that speakers were not able to completely ignore the Gricean maxim of manner that prescribes speakers to prefer utterances that encode information most efficiently; the social goals condition required them to abandon this principle of cooperative communication. However the comparison of a combination of social and informational goals to purely informational ones conformed to our expectation: speakers were more likely to choose indirect utterances when social goals were on the table in addition to the informational ones (H_{2b} , posterior mean = 0.072, 95% credible interval = [0.171, 0.270]).

3.3 Experiment 3: Pragmatic listener

In order to evaluate the model’s opinion inference we designed an experiment where the conversation partners exchange opinion statements on a certain topic, and the task of the participants is to infer their actual opinion. The computational model of belief inference presented in Section 2.8 predicts that the same utterance of the second speaker can be interpreted differently depending on the first speaker’s statement and the communicative goals that the participants pursue in the conversation. To mimic the model setup, we informed the participants that the speakers want to exchange opinions but do not want to run into a conflict. We selected six adjectives (out of ten tested in Experiment 1) for the first

speaker’s utterance such that they reflect a full range of the scale from strongly negative to strongly positive with two adjectives representing the middle of the scale. The second speaker’s adjectives included six possible responses and excluded the most opinionated replies (strongly positive and strongly negative), since they were not compatible with the stated communicative goal. Figure 9 displays a sample trial for Experiment 2.

Mary and Rachel meet at a mutual friend's birthday party for the first time.
They would like to exchange opinions but don't want to run into a conflict.

Mary says: The election results are poor.

Rachel replies: I find them interesting.

How may Rachel actually feel about the issue?

Strongly Negative  Strongly Positive

Click 'continue' to move on.

Continue

Figure 9: Experiment 3 (sample trial). Participants are asked to read a dialogue and indicate how the second speaker actually feels about the issue.

We collected data from 286 participants on the Prolific crowd-sourcing platform. Each participant completed 6 trials, each featuring a separate topic. Data from 17 participants were excluded from the analysis since they reported that they did not fully understand the instructions, data from the remaining 269 participants was entered into the analysis.

We manipulated the first speaker’s statement (from strongly negative to strongly positive) and the second speaker’s response (from slightly negative to slightly positive). In critical trials, we then asked how the second speaker may have felt about the topic. In control trials (one trial out of six), we asked the participants to evaluate the statement of the first speaker. This manipulation served two purposes: first, it acted as a way to increase the participant’s engagement in the task. And second, the first speaker’s scores provided a baseline that allowed us to order the adjectives on the negative-positive scale and provide an additional confirmation of the scale we obtained in Experiment 1.

Figure 10 shows average scores from the experiment, alongside model predictions, for opinion inferences based on the first and the second speaker’s utterances. A plot of the non-averaged data can be found in the Appendix (Figure 19).

The key qualitative prediction of the model that we would like to assess is one of monotonicity, so to speak: the higher the rank (i.e., the position expressed by the first speaker), the lower the inferred opinion of the second speaker. Thus, for example, the model predicts that participants should assign a higher score to the adjective ‘pretty good’ if the first speaker statement was negative than when the first statement was strongly positive.

Based on visual inspection, this prediction seems to be supported, at least in tendency, by the data. To test this, we ran a Bayesian regression model, using a cumulative-logit link function to regress the Likert-scale rating data against monotonically ordered predictors

(Bürkner & Charpentier, 2020) of the ranks for the first and second speaker’s utterances, as well as their interaction, using the default priors of the R package `brms` (Bürkner, 2018). We find that the monotonicity coefficient associated with the first speaker’s utterance rank is indeed credibly negative (posterior mean: -0.194 ; 95% credible interval: $[-0.296; -0.0851]$). To further corroborate this result, we also compared this regression model, which has monotonically ordered predictors, against another regression model which allows all rank-levels to be estimated freely from the data (without constraints of monotonic ordering). We find that the model with monotonically ordered factors is substantially preferred under leave-one-out model comparison (difference in expected log-density: 11.2, estimated standard error of this difference: 4.7; see Vehtari et al. (2017)). Taken together, we interpret this as initial evidence in support of the AMIC’s predictions.

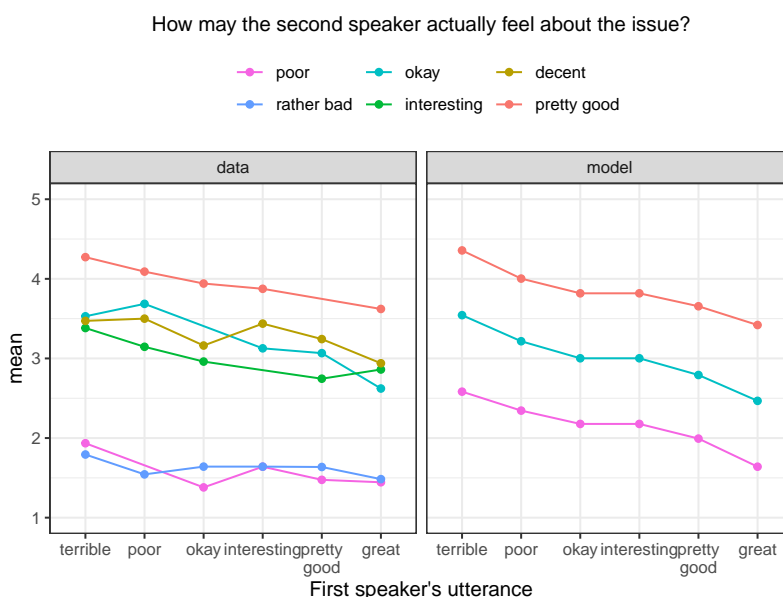


Figure 10: Opinion inference scores. The left panel displays the participants’ evaluation of the second speaker’s opinion. The right panel shows corresponding model predictions for slightly positive (*pretty good*), neutral (*okay*), and slightly negative (*poor*) utterances.

3.4 Summary: Behavioral data

Overall, we have reported the results of three experiments that were designed to provide an empirical assessment of different components of the introduced the AMIC. Experiment 2 targeted the behavior of speakers pursuing a range of communicative goals. The results confirm that social goals and a possible mismatch in the opinions of conversation partners favor the choice of indirect utterances, like the alignment model predicts. Finally, Experiment 3 demonstrates that participants were indeed able to interpret the speaker’s responses as indirect when they knew that conversation partners were pursuing social goals. The inferences about the actual speaker B’s opinion differed depending on the combination of speakers’ contributions. The direction of change corresponds to the one predicted by the

alignment model. The empirical findings thus qualitatively support our AMIC model.

4 Conclusion

One of the goals of theoretical pragmatics is to define how listeners arrive at the meaning of utterances beyond the literal meaning. Game-theoretic models, such as the Iterated Best Response theory (Franke, 2009) and the RSA framework (Frank & Goodman, 2012; Goodman & Frank, 2016) have answered this question by assuming that the listener reasons about the speaker, who is, in turn, reasoning about a lower level listener and maximizing the chance of the listener receiving the intended message. Thus, such models defined utterance utility solely by informational utility. Later models included social components, such as politeness (Yoon et al., 2020) and social meaning (Henderson & McCready, 2019), which additionally influence the speaker’s utterance choice. In this paper, we argued that the desire to avoid conflict of beliefs, which we defined in terms of opinion misalignment (or divergence), also affects the types of utterances speakers opt for. We have shown that indirect utterances allow the speaker to simultaneously satisfy informational and social goals. We have furthermore suggested that indirectness is a tool that allows the speaker to probe the state of the listener’s beliefs. Thus, conflict avoidance brings the additional benefit of implicitly checking if beliefs are shared.

On top of that, the alignment model proposed in this paper contains a novel inference mechanism that tracks reasoning about mutual opinions over extended stretches of dialogue. The mechanism infers the likely opinion of the listener upon registering her response. The principle behind the implemented inverse inference process is related to inverse reinforcement learning, which is able to infer the reward function that determines the behavior of observed other agents (Russell, 2020). It is furthermore related to computational models that infer hidden states of other agents, such as their knowledge and their preferences (Baker et al., 2009, 2017)—an ability that is indeed already observable in ten month old infants (Liu & Spelke, 2017). Our model embeds these mechanistic modeling principles into the realm of conversations, where utterance choices are modeled based on communicative and social objectives. We have used the inverse inference process to update beliefs about the covert opinions of conversation partners. Similar inference processes could be used to, for example, infer conflict avoidance utility weights (parameter γ in Eq.6).

Overall, we have brought together literature from social psychology, philosophy of language, psycholinguistics, and cognitive modeling to formalize the mechanisms that may underlie opinion alignment through the use of indirect utterances. We propose that indirectness can be viewed as a social means to foster the development of establishing shared opinions, which is possible as long as (i) prior opinions are not fully incompatible from the outset and (ii) the conversation partners are willing to adjust their individual opinions towards those of the conversation partner. From a computational standpoint, a certain degree of flexibility in the belief-encoding distribution ensures that conversation partners can adjust their belief systems to each other and reach consensus (Hegselmann et al., 2002).

From a sociological perspective, discovering whether opinions are shared serves two purposes: understanding the world through validating reality and belonging to a group (Andersen & Przybylinski, 2018; Higgins, 2019). The discovery of shared aspects signals to conversation partners that they may belong to the same social group. The discovery

of unexpected or rare alignment between two personal characteristics may lead to an even stronger bonding effect (Vélez et al., 2019). Thus, confirming that certain assumptions belong to the common ground may create the bonding and the ‘linguistic intimacy’ (Cohen, 1976) that emerges when an indirect utterance was apparently interpreted as intended.

The proposed model formalizes how production and interpretation of indirect utterances continuously provides conversation partners with signals of whether their belief systems align. While avoiding conflict, they monitor each other’s interpretation of indirect utterances and draw inferences about each other’s opinions. These inferences open space for dynamic belief alignment and contribute to establishing and maintaining social bonds while speakers navigate complex social environment.

References

- Achimova, Asya, Gregory Scontras, Ella Eisemann & Martin V. Butz. 2023. Active iterative social inference in multi-trial signaling games. *Open Mind* 7. 111–129.
- Achimova, Asya, Gregory Scontras, Christian Stegemann-Philipps, Johannes Lohmann & Martin V Butz. 2022. Learning about others: Modeling social inference through ambiguity resolution. *Cognition* 218. 104862.
- Andersen, Susan M & Elizabeth Przybylinski. 2018. Shared reality in interpersonal relationships. *Current opinion in psychology* 23. 42–46.
- Asher, Nicholas & Alex Lascarides. 2001. Indirect speech acts. *Synthese* 128(1-2). 183–228.
- Asher, Nicholas & Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics* 6. 2–1.
- Baker, Chris L., Julian Jara-Ettinger, Rebecca Saxe & Joshua B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1(4). 0064. doi:10.1038/s41562-017-0064.
- Baker, Chris L., Rebecca Saxe & Joshua B. Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113(3). 329–349. doi:10.1016/j.cognition.2009.07.005.
- Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3). 255–278.
- Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9.
- Bonnefon, Jean-François, Aidan Feeney & Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112(2). 249–258.
- Brochhagen, Thomas. 2020. Signalling under uncertainty: Interpretative alignment without a common prior. *British Journal for the Philosophy of Science* 71. 471–496.

- Brown, Penelope & Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*, vol. 4. Cambridge university press.
- Bürkner, Paul-Christian. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10(1). 395–411.
- Bürkner, Paul-Christian & Emmanuel Charpentier. 2020. Modelling monotonic effects of ordinal predictors in bayesian regression models. *British Journal of Mathematical and Statistical Psychology* 73(3). 420–451.
- Burnett, Heather. 2017. Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics* 21(2). 238–271.
- Burnett, Heather. 2019. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy* 42(5). 419–450.
- Carcassi, Fausto & Michael Franke. 2023. How to handle the truth: A model of politeness as strategic truth-stretching. In M. Goldwater, F. K. Anggoro, B. K. Hayes & D. C. Ong (eds.), *Proceedings of CogSci*, 222–228.
- Castellano, Claudio, Santo Fortunato & Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics* 81. 591–646.
- Cohen, Ted. 1976. Figurative speech and figurative acts. *The Journal of Philosophy* 72(19). 669–684.
- Degen, Judith. 2023. The rational speech act framework. *Annual Review of Linguistics* 9. 519–540.
- DeGroot, Morris H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69(345). 118–121.
- Égré, Paul, Benjamin Spector, Adèle Mortier & Steven Verheyen. 2023. On the optimality of vagueness: “around”, “between” and the gricean maxims. *Linguistics and Philosophy* 46(5). 1075–1130.
- Fawcett, Christine A & Lori Markson. 2010. Similarity predicts liking in 3-year-old children. *Journal of experimental child psychology* 105(4). 345–358.
- Frank, Michael C. & Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336. 998–998.
- Franke, Michael. 2009. *Signal to act: Game theory in pragmatics*, vol. DS-2009-11 ILLC dissertation series. Amsterdam: Institute for Logic, Language and Computation and Universiteit van Amsterdam.
- Franke, Michael & Leon Bergen. 2020. Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language* 96(2). e77–e96.

- Franke, Michael & Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft* 35(1). 3–44.
- Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences* 20(11). 818–829.
- Goodman, Noah D & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* 5(1). 173–184.
- Goodman, Noah D. & Andreas Stuhlmüller. 2013. Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science* 5. 173–184.
- Gotzner, Nicole & Diana Mazzarella. 2021. Face management and negative strengthening: The role of power relations, social distance, and gender. *Frontiers in psychology* 12.
- Gotzner, Nicole & Diana Mazzarella. 2024. Negative strengthening: The interplay of evaluative polarity and scale structure. *Journal of Semantics* .
- Grice, H Paul. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, 26–40. New York: Academic Press.
- Hawkins, Robert XD, Andreas Stuhlmüller, Judith Degen & Noah D Goodman. 2015. Why do you ask? good questions provoke informative answers. In David C. Noelle, Rick Dale, Anne Warlaumont, Jeff Yoshimi, Teenie Matlock, Carolyn Jennings & Paul P. Maglio (eds.), *Proceedings of the 37th annual meeting of the cognitive science society*, 878–883.
- Hegselmann, Rainer, Ulrich Krause et al. 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation* 5(3).
- Henderson, Robert & Elin McCready. 2019. Dogwhistles and the at-issue/non-at-issue distinction. In Daniel Gutzmann & Katharina Turgay (eds.), *Secondary content*, 222–245. Leiden, NL: Brill.
- Henderson, Robert & Elin McCready. 2019. Dogwhistles, trust and ideology. In *Proceedings of the 22nd amsterdam colloquium*, 152–160.
- Henderson, Robert & Elin McCready. 2021. Dogwhistles: Persona and ideology. In Nicole Dreier, Chloe Kwon, Thomas Darnell & John Starr (eds.), *Proceedings of the 31st semantics and linguistic theory conference*, 703–719.
- Higgins, E Tory. 2019. *Shared reality: What makes us strong and tears us apart*. Oxford University Press.
- Kao, Justine, Leon Bergen & Noah Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the cognitive science society*, vol. 36 36, .
- Kao, Justine T, Jean Y Wu, Leon Bergen & Noah D Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007.

- Khani, Fereshte, Noah D. Goodman & Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics* 6. 543–555. doi:10.1162/tacl_a_00037.
- Lee, James J. & Steven Pinker. 2010. Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review* 117(3). 785–807.
- Liu, Shari & Elizabeth S. Spelke. 2017. Six-month-old infants expect agents to minimize the cost of their actions. *Cognition* 160. 35 – 42. doi:10.1016/j.cognition.2016.12.007.
- Mahajan, Neha & Karen Wynn. 2012. Origins of “us” versus “them”: Prelinguistic infants prefer similar others. *Cognition* 124(2). 227–233.
- Noveck, Ira A. 2018. *Experimental pragmatics: The making of a cognitive science*. New York: Cambridge University Press.
- Noveck, Ira A. & Dan Sperber (eds.). 2004. *Experimental pragmatics*. Hampshire: Palgrave MacMillan.
- Peet, Andrew. 2024. The puzzle of plausible deniability. *Synthese* 203(5). 1–20.
- Pinker, Steven, Martin A Nowak & James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of sciences* 105(3). 833–838.
- Potts, Christopher, Daniel Lassiter, Roger Levy & Michael C Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics* 33(4). 755–802.
- Qing, Ciyang & Reuben Cohn-Gordon. 2019. Use-conditional meaning in rational speech act models. In M. Teresa Espinal, Elena Castroviejo, Manuel Leonetti, McNally Louise & Cristina Real-Puigdollers (eds.), *Proceedings of sinn und bedeutung 23*, vol. 23, 253–266. Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès).
- Rossignac-Milon, Maya, Niall Bolger, Katherine S Zee, Erica J Boothby & E Tory Higgins. 2020. Merged minds: Generalized shared reality in dyadic relationships. *Journal of Personality and Social Psychology* .
- Russell, Stuart. 2020. The purpose put into the machine. In John Brockman (ed.), *Possible minds: 25 ways of looking at ai*, chap. 3, 20–32. New York: Penguin Press.
- Sacks, Harvey, Emanuel A Schegloff & Gail Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. *Language* 50(4).
- Schegloff, Emanuel A. 1984. On some questions and ambiguities in conversation. In J. Maxwell Atkinson & John Heritage (eds.), *Structures of social action: Studies in conversation analysis*, chap. 3, 28–52. Cambridge: Cambridge University Press.
- Schöller, Anthea & Michael Franke. 2017. Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard* 3(1). 20160072.

- Scontras, Gregory, Michael Henry Tessler & Michael Franke. 2021. A practical introduction to the rational speech act modeling framework.
- Searle, John R. 1975. Indirect speech acts. In Peter Cole & Jerry L. Morgan (eds.), *Speech acts*, 59–72. Academic Press.
- Sperber, Dan & Deirdre Wilson. 1995. *Relevance: Communication and cognition (2nd ed.)*. Oxford: Blackwell.
- Terkourafi, Marina. 2014. The importance of being indirect: A new nomenclature for indirect speech. *Belgian Journal of Linguistics* 28(1). 45–70.
- Tomasello, Michael. 2019. *Becoming human: A theory of ontogeny*. Belknap Press.
- Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistical Computing* 27. 1413–1432.
- Vélez, Natalia, Sophie Bridgers & Hyowon Gweon. 2019. The rare preference effect: Statistical information influences social affiliation judgments. *Cognition* 192. 103994.
- Vogel, Adam, Max Bodoia, Christopher Potts & Dan Jurafsky. 2013. Emergence of gricean maxims from multi-agent decision theory. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, 1072–1081.
- Winter-Froemel, Esme & Angelika Zirker. 2015. Ambiguity in speaker-hearer-interaction: A parameter-based model of analysis. In Susanne Winkler (ed.), *Ambiguity*, 283–339. Berlin/Boston: de Gruyter.
- Yoon, Erica J., Michael H. Tessler, Noah D. Goodman & Michael C. Frank. 2020. Polite speech emerges from competing social goals. *Open Mind : Discoveries in Cognitive Science* 4. 71–87. doi:10.1162/opmi_a_00035.
- Yoon, Erica J., Michael Henry Tessler, Noah D. Goodman & Michael C. Frank. 2016. Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 2771–2776. Cognitive Science Society.

A Utterance choice simulations

A.1 Divergence measures

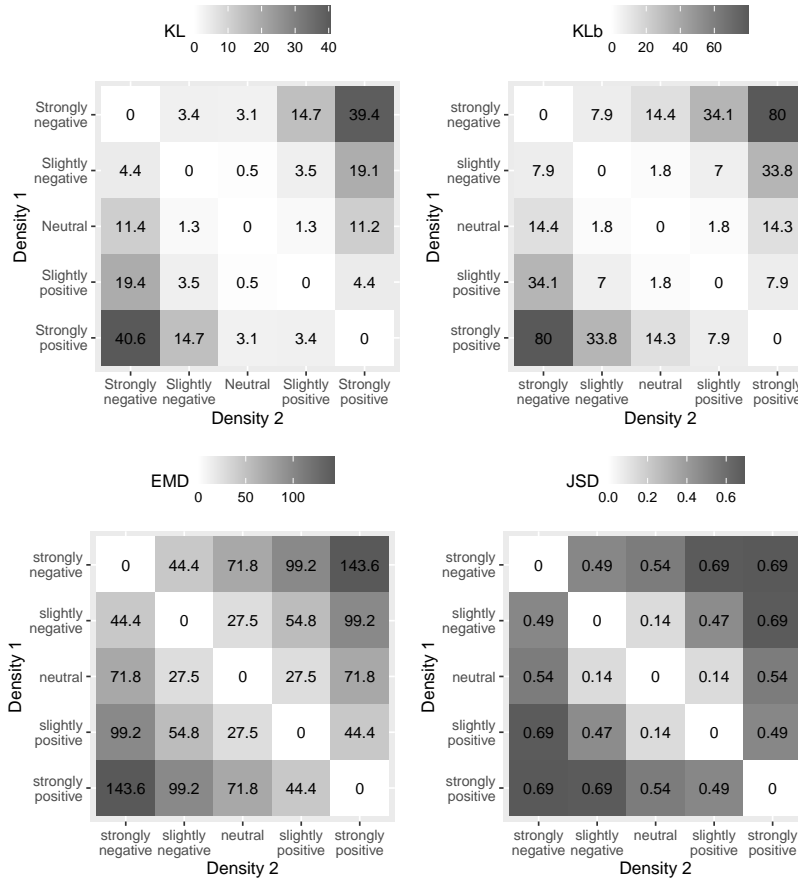


Figure 11: Measures of divergence between the opinion distributions may be interpreted as encoding the effort to change one belief into another one, or, in other words, belief compatibility. a) Kullback Leibler (KL) divergence; b) Bidirectional KL-divergence; c) Earth Mover's Distance; d) Jensen-Shannon Divergence.

A.2 Impact of divergence measure on the model predictions

Unidirectional KL-divergence produces a similar qualitative pattern compared to the bidirectional KL-divergence: the model infers a more positive opinion of speaker B given a more negative utterance of speaker A (Figure 12). However, the Jensen-Shannon Divergence (Figure 13) shows a weaker trend and the Earth Mover's Distance (Figure 14) even with modified parameters fails to capture the qualitative pattern observed in the data.

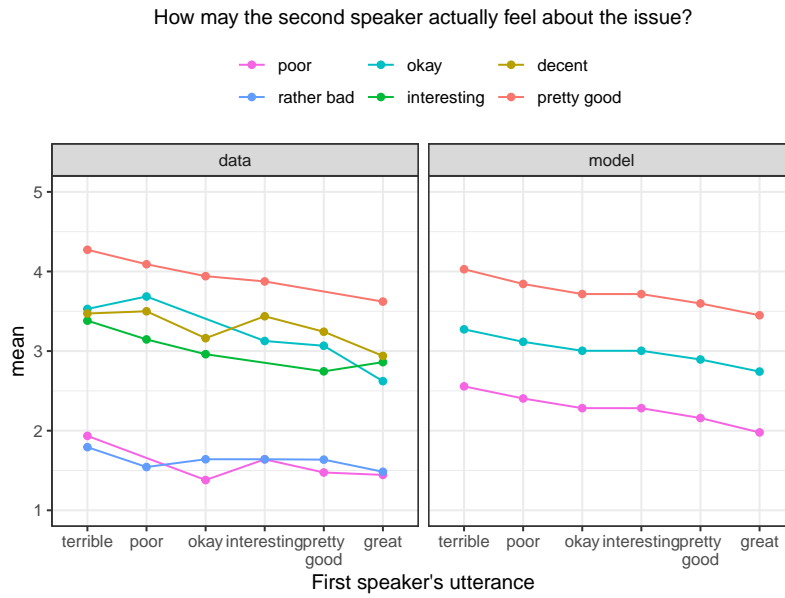


Figure 12: Opinion inference scores. The model relies on unidirectional KL as a divergence measure.

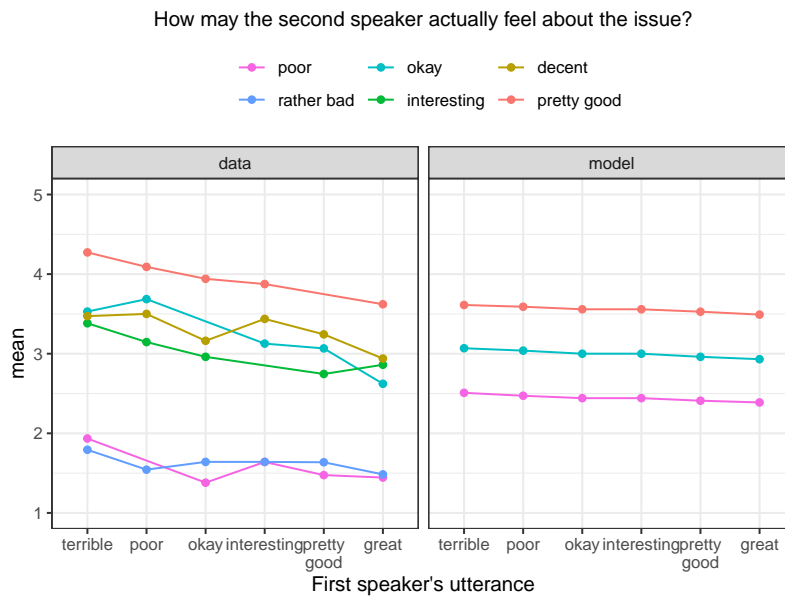


Figure 13: Opinion inference scores. The model relies on Jensen-Shannon Divergence measure.

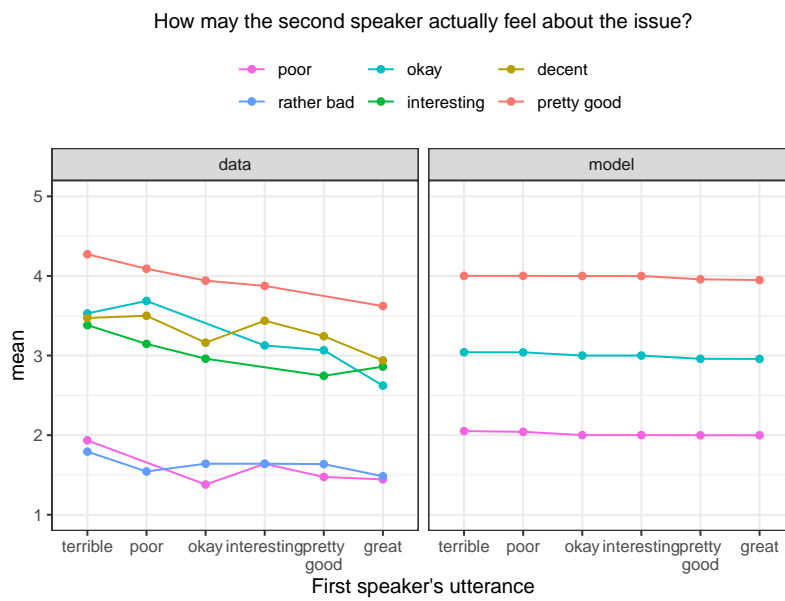


Figure 14: Opinion inference scores. The model relies on Earth Mover's Distance as a divergence measure.

A.3 Utterance utilities and probabilities

The left-hand side of Figure 15 shows utility values for utterances (rows) given different speaker beliefs about the listener’s opinion state $\pi_1^{S_1}$ (here assumed to be single-peaked distributions). The right-hand side of Figure 15 shows the corresponding utterance-choice probabilities computed via Equation 7 assuming $\omega_{inf} = 0.8$, $\omega_{soc} = 0.2$, and $\alpha = 0.18$. The values show that the model generates progressively smaller utilities for utterances that diverge from the speaker’s opinion (strongly positive in this case). Generally, utterances that offer the best compromise between the speaker’s opinion and the believed listener’s opinion are preferred.

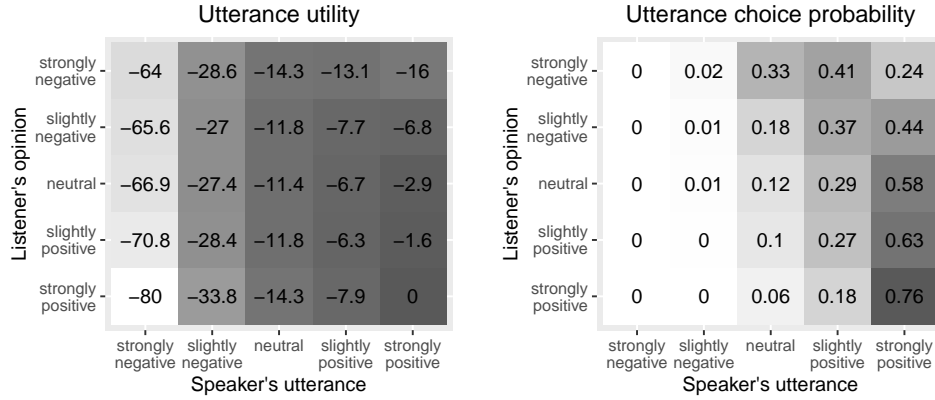


Figure 15: Utterance utility values (left side, to-be maximized), and corresponding utterance choice probabilities (right side). The calculations assume that the speaker’s opinion corresponds to a strongly positive ($\alpha = 30, \beta = 5$) opinion distribution.

B Opinion inference

B.1 Simulation

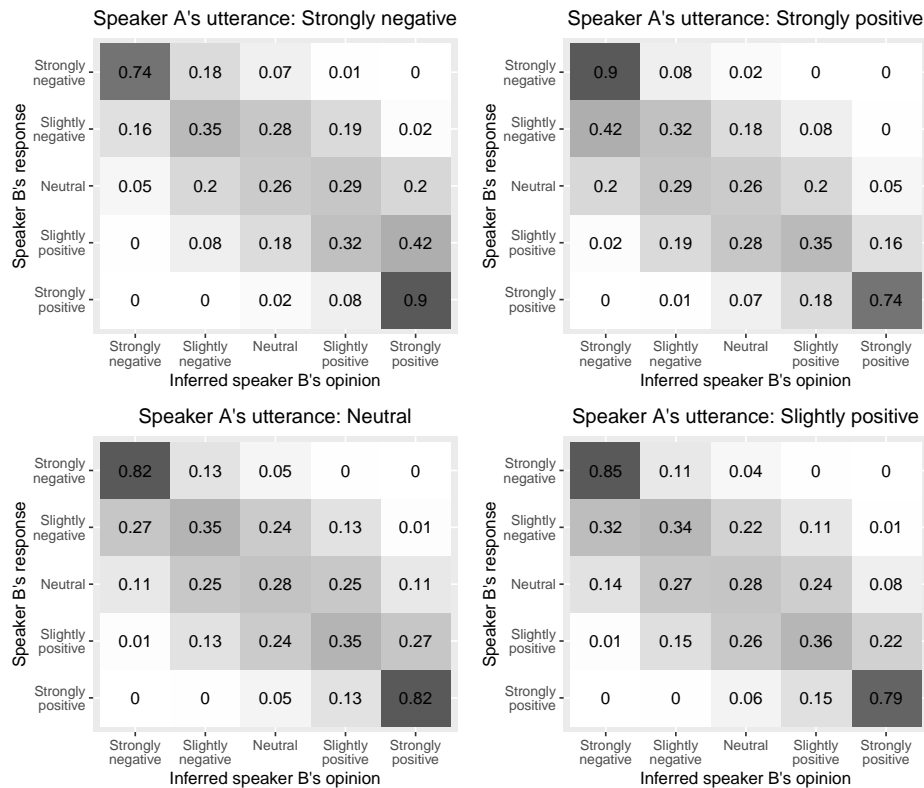


Figure 16: Model's posterior estimation of speaker B's opinion computed via Equation 8 given an initial utterance that is strongly negative (top left), strongly positive (top right), neutral (bottom left), or slightly positive (bottom right). Each row in each matrix encodes a particular posterior belief distribution π_1^A over speaker B's opinion given her response indicated in each row.

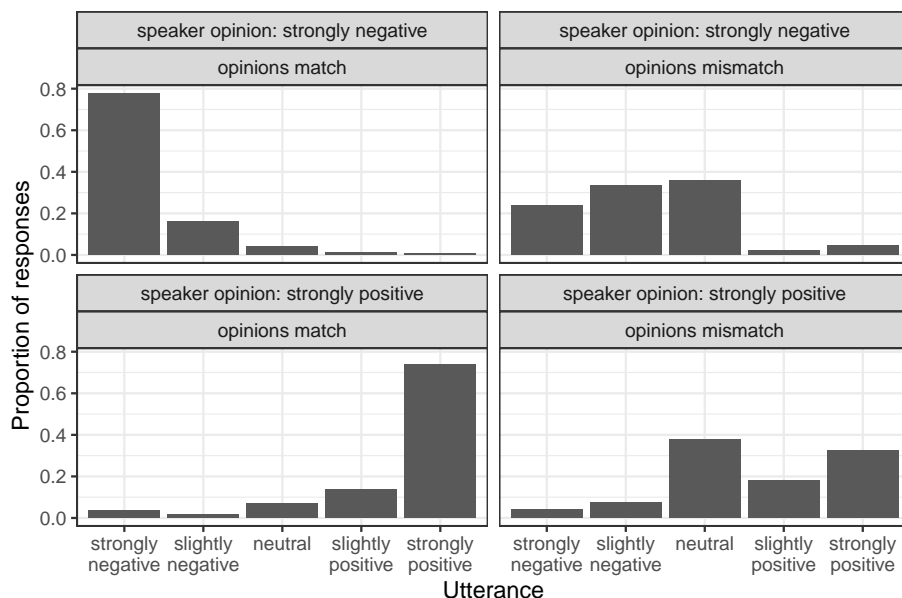


Figure 17: Experiment 2: Utterance choice (raw data). The left column shows the cases where the opinions of conversation partners match. Here speakers prefer utterances that correspond to their true opinion. The right column shows the cases of mismatch in opinion. Utterance choices shift towards the middle of the scale.

C Behavioral data

C.1 Experiment 2: Raw data

In Section 3.2, we reported aggregated data from the Pragmatic speaker experiment, where participants selected an utterances that would allow the speaker to communicate her opinion while pursuing one of the three announced communicative goals: informational, a combination of informational and social goals, or a social goal. For the analysis, we categorized the utterances as direct, indirect, or opposite. Here, we report raw data without the grouping. Figure 17 shows how often each of the utterance types (from strongly negative to strongly positive) was chosen depending on whether the opinions of conversation partners matched and the speaker’s opinion.

In figure 18, we report data from the mismatch cases broken by communicative goal and the speaker’s opinion.

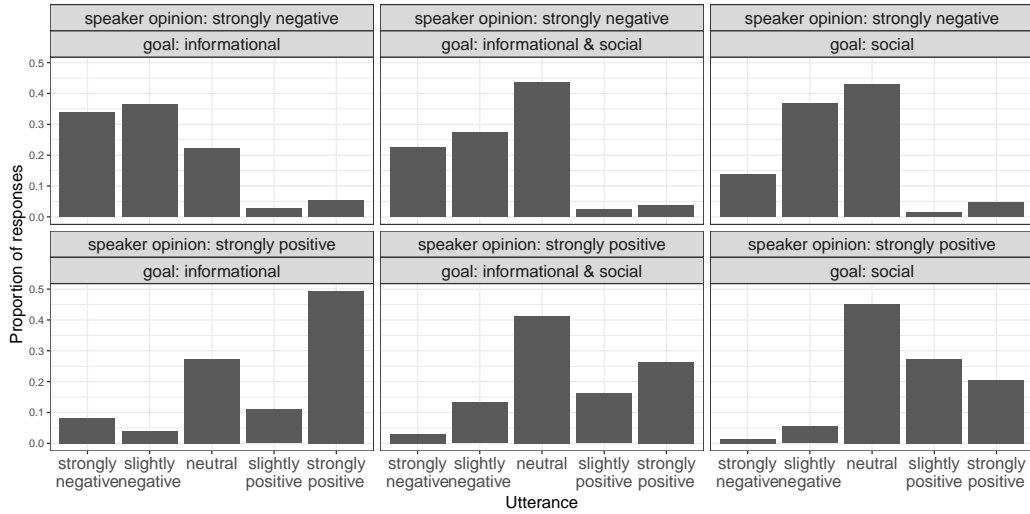


Figure 18: Experiment 2: Utterance choice (raw data). Cases where the opinions of conversation partners do not match. The top row corresponds to the strongly negative opinion of the speaker, the bottom row shows the strongly positive opinion of the speaker.

C.2 Experiment 3: Raw data

In Experiment 3, participants were asked to infer the opinion of the second speaker upon observing her response to the original speaker’s statement. In section 3.3, we reported aggregated data over inferred opinions for different utterances. Here, we report raw data: we are interested in the location of clusters of responses and their distribution on the vertical axis. The pattern we observe is qualitatively similar to the model predictions presented in Section 2.8.

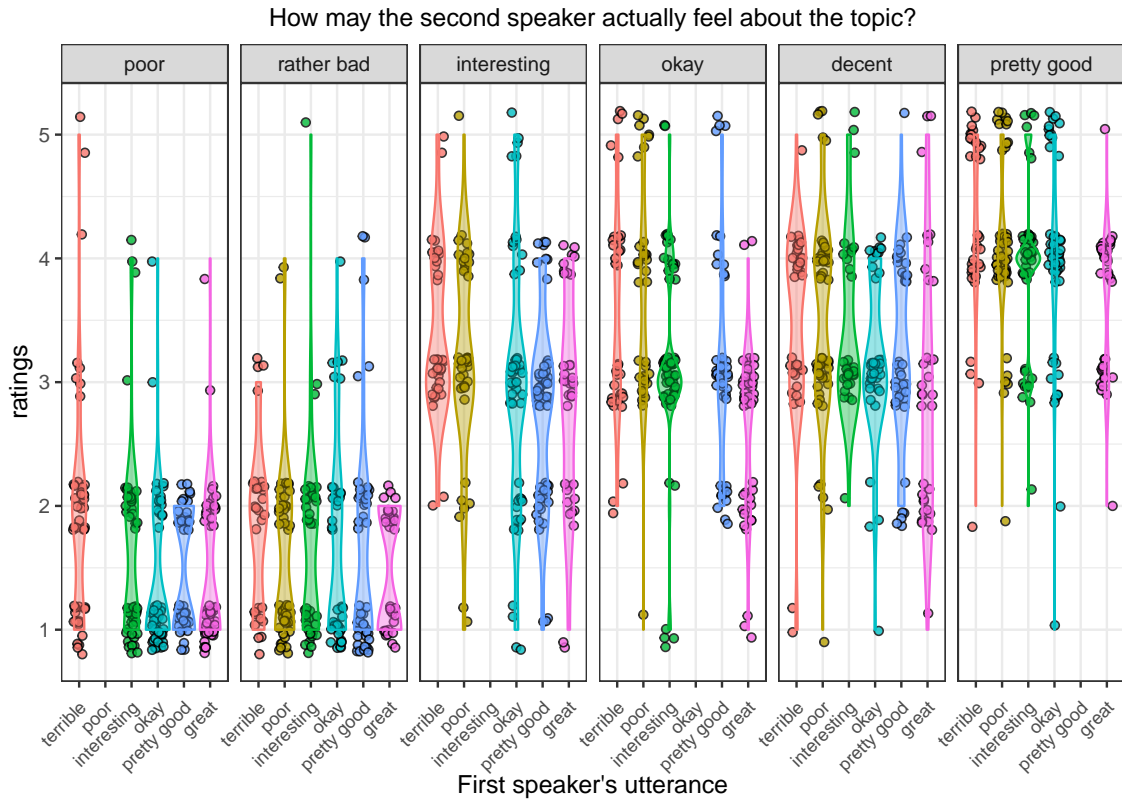


Figure 19: Utterance ratings for ten considered adjectives. Each data point represents a participant’s response. Jitter was added for visualization purposes. The location of the clusters along the vertical axis reflects how positive the inferred opinion is. Thus, the relevant contrasts lie within each facet between the adjectives at the opposite ends of the scale. The model predicts that upon hearing a predicate, such as ‘interesting’, participants should infer the opinion as more positive if the predicate follows a strongly negative statement compared to a strongly positive statement.