

Indirectness as a path to common ground management

Asya Achimova, Michael Franke, Martin V. Butz

Abstract Under the Gricean view of communication, cooperative speakers are expected to encode their messages efficiently. Probabilistic models of pragmatics formalize this requirement via a mechanism that assigns a higher probability to utterances that signal the intended meaning unambiguously. Indirect utterances seemingly stand at odds with this informational efficiency. In this paper, we take a social stance on indirectness: we maintain that the choice of indirect utterances is driven by a speaker pursuing both information-driven and social goals. We define the social goals in the model via the mechanism of avoiding a conflict of beliefs in common ground. We offer a Rational Speech Act formalism and introduce the Alignment Model of Indirectness (AMI). In this model, indirectness acts as a common ground management tool that allows the speaker to probe the state of common ground before committing to particular beliefs publicly. We furthermore formalize how the interpretation of an indirect utterance and a consequent listener's response reveal her background beliefs. Thus, indirectness serves as a tool for checking and adjusting the alignment of the speakers' beliefs and assumptions. **Wordcount:** 8607

1 Introduction

Sharing mental attitudes, such as beliefs, preferences, and assumptions, 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) conceptu-

alized such social considerations in the notion of *face* and argued that face preservation is a major motivational force that shapes human interactions.

In linguistic research, the relevance of shared beliefs and experiences for communication has been emphasized within the investigations of how common ground affects the interpretation of utterances (Clark 1996, 2015, Stalnaker 2002, 2014). The state of the common ground, and, as a consequence, whether beliefs are in fact shared, is never fully known to communication partners. In this paper, we look at indirectness as a way to probe the state of the common ground and check the alignment of shared beliefs. We maintain that indirect utterances allow speakers¹ to verify if their background beliefs are shared before offering to add new propositions to the common ground. Moreover, indirectness allows speakers to adjust their own beliefs before committing to them publicly. Thus, we view indirectness as a tool for common ground management.

The power of indirect utterances to act as a common ground management tool lies in their ambiguity. Indirect utterances, such as (1), are compatible with a range of diverse opinions that the speaker intends to express. The statement in (1) is compatible with a positive (1a), neutral (1b), or negative (1c) follow-up utterance.

- (1) [Candidate A won] The election outcome was interesting!
 - a. I'm so excited that candidate A won!
 - b. Both candidates had a lot to offer.
 - c. I fear what is to come now!

Upon hearing the indirect utterance in (1), the listener has to rely on her assumptions about what background beliefs are shared, or in other words, on the state of common ground to disambiguate the predicate “interesting”. Her interpretation, as evidenced by her reaction (1a - 1c), can thus reveal these assumptions.

The goal of this paper is to formulate a computational algorithm that predicts how speakers and listeners choose and interpret indirect utterances. We develop a probabilistic model of utterance choice and interpretation by extending and merging proposals developed within the Rational Speech Act framework (Frank & Goodman 2012, Goodman & Frank 2016). In particular, we formalize how the speaker's utterance choice may be driven by a combination of information-transfer and social goals, enhancing an RSA-based model that includes politeness considerations (Yoon et al. 2020). We furthermore model how the interpretation of indirect utterances can reveal the listener's background beliefs. Finally, we connect the linguistic analysis of indirectness to the sociological literature on sharing beliefs and discuss

¹ Throughout the paper we use the terms *speaker* and *listener* to refer to the person currently making an utterance and interpreting an utterance, respectively. Conversation partners alternate between taking these roles.

the implications of discovering and developing shared beliefs through ambiguity resolution for social bonding.

2 Choosing indirect utterances

Within a Gricean perspective, communication is viewed as a cooperative enterprise, where the speaker and listener share a goal of efficient information exchange. Game-theoretic (Frank 2009) and probabilistic pragmatic models (Frank & Goodman 2012, Goodman & Stuhlmüller 2013, Goodman & Frank 2016) have formalized this efficiency requirement as the probability of the listener choosing the right meaning upon hearing the utterance, and have used signaling games as a proxy of communication to model the utterance choice and utterance interpretation processes. In signaling games (Lewis 1969), the speaker intends to signal an object to the listener and chooses the utterance by estimating the chance of the listener arriving at the correct interpretation. Ambiguous utterances—those that apply to multiple objects—therefore appear inferior to unambiguous utterances because the probability of choosing the correct referent is smaller for the ambiguous ones. Indirect utterances appear problematic when viewed from the perspective of efficient information transfer since they are often ambiguous.

Psycholinguistic and modeling evidence suggests that ambiguity at the lexical level does not necessarily pose problems for the interpreters. Ambiguous words tend to occur frequently in informative contexts (Pimentel et al. 2020), so the context limits the range of relevant meanings of ambiguous elements. Brochhagen (2020) uses computational modeling of lexicon alignment to argue that the informativity of the context ensures the ecological validity of ambiguity: if the context expectations are shared, ambiguity can be correctly resolved. Furthermore, ambiguity may additionally offer necessary flexibility to conversation partners to adjust word meanings to each other (Brochhagen 2020). In this paper, we extend this argument to pragmatic ambiguity that stems from indirectness and show that the interpretation of indirect utterances can expose the interpreter’s assumptions about the state of common ground. Following this argument, we maintain that indirect utterances do not violate maxims of cooperative communication in the sense of Grice (1989) despite them being ambiguous. Rather, indirectness allows speakers to reach an intricate balance of information transfer goals and social goals and opens up the possibly to flexibly negotiate shared opinions.

We begin with the following informal definition of indirect utterances:

Definition 1:

- i. Indirect utterances are sub-optimal from the point of view of information transfer, that is, there exists an alternative utter-

ance that has a higher probability of signaling the intended message.

- ii. They are optimal if the speaker's goal considers multiple objectives, including, for example, information transfer and social ones.

Condition (ii) of the Definition allows separating indirect utterances from lies since those fail to meet information goals.

Our focus in this paper lies on indirectness as a way of managing the state of common ground between conversation partners. Not all indirect utterances can act as common ground management tools. To probe the state of common ground, an indirect utterance has to be ambiguous. The example in (2) fails this requirement.

(2) Could you please hand me the butter?

In this situation indirectness is conventionalized (Searle 1979): we expect all (adult) speakers of the speech community to recognize (2) as a request rather than an information seeking question about the physical abilities of the listener. We therefore formulate the final indirectness requirement:

Ambiguity requirement:

- iii. To act as a common ground management device, the indirect utterance must be pragmatically ambiguous.

Pragmatic ambiguity is 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.

We have claimed above that indirect ambiguous utterance choices may be explained by considering additional utterance choice objectives, besides the Gricean information transfer objective. Both theoretical and computational accounts have emphasized that social factors, such as politeness, face, as well as dominance and control play important parts in determining speakers' choice of utterances and affect how direct and explicit they are (Beaver & Stanley 2018, Brown & Levinson 1987, Carcassi & Franke to appear, Degen et al. 2015, Khani et al. 2018, Yoon et al. 2020). Accordingly, Yoon et al. (2020) demonstrated that speakers selected indirect utterances, such as "not terrible" instead of more straightforward "bad", when they were instructed to both send the listener truthful feedback and avoid hurting the listener's feelings (Yoon et al. 2020).

An alternative proposal attributes the choice of ambiguous utterances to the speaker's goal to learn about the listener's prior beliefs (Achimova et al. 2022a),

emphasizing furthermore that ambiguous utterances allow the speaker to learn more about the listener by observing how the listener resolves referential ambiguity. The current paper more generally investigates inferences of the listener's beliefs as a by-product of choosing indirect utterances. The choice of indirect utterances in our model is in turn motivated by a combination of informational (signal the intended world state) and social (avoid conflict of beliefs in common ground) goals. We thus highlight that ambiguity that stems from indirectness may not emerge directly from epistemic goals, but rather from other social objectives. Here, we consider conflict avoidance as one of the social goals that affects the choice of utterances and appeal to the notion of common ground to define this conflict in computational terms.

3 Common ground management

The term “common ground management” is often used in conjunction with work on the semantics of particular lexical items, which may directly signal the state of common ground (Krifka 2008, Döring 2018). Here, we extend the use of the term to also cover the cases when indirectness allows the speaker to probe the state of common ground. We propose that the social utility of indirect utterances is assessed by simulating the state of common ground for each potential utterance. The speaker chooses utterances in such a way that they do not lead to a conflict of beliefs in common ground.

The development of this account requires a particular view of common ground. A practically useful but eventually too simplistic picture treats common ground just as a set of worlds compatible with the information shared or accepted by the interlocutors. A more elaborate, but still standard view of common ground according to Stalnaker (2002) assumes that φ is common ground between agents A and B if both believe φ , both believe that they believe it and so on. This model both determines an objective common ground, i.e., what is actually common ground between A and B , but also allows an agent's beliefs about what is common ground to deviate from what common ground actually is. Essentially, it is possible that both A and B may have completely different beliefs about what is common ground. The idea of divergent perspectives on common ground has been developed in Conversation Analysis and computational models of dialogue. Ginzburg (2012), for example, argued for representing individually assumed common ground as Dialogue Game Boards (Ginzburg 1996).

Furthermore, speakers may be uncertain about the state of another person's beliefs as well as their own beliefs. Beaver (1997, 2001) emphasizes that the state of the common ground is never fully known even to the speaker or the listener themselves. In other words, both inevitably carry a level of uncertainty about its state.

Other recent models of common ground further emphasize that propositions in common ground may carry different levels of salience (Döring 2018). Moreover, memory-rich models (Brown-Schmidt & Duff 2016, Horton 2005, Horton & Gerrig 2016) presuppose that just like other types of information, propositions that constitute common ground are subject to different levels of availability and possibly decay over time. Therefore, verifying the state of common ground becomes a part of the communicative objectives. As a result, the speaker might want to check whether propositions still belong to common ground (Karagjosova 2004) or whether they are salient in common ground (Döring 2018). Salience is relevant for the interpretation of indirect utterances: if the listener arrives at the intended interpretation of such utterances, she signals that her salience map of beliefs aligns with the salience map of the speaker. Thus, the speaker receives a confirmation that the beliefs are shared. She can then publicly commit to those beliefs and make them part of common ground.

The gradient, uncertain nature of common ground is relevant to our current investigation. We propose and formalize that when speakers are not certain whether a particular belief belongs to common ground, they will be more likely to choose an indirect utterance. Accordingly, we propose the following common ground management protocol:

- i. The speaker chooses an indirect utterance due to multiobjective considerations, including her intention to both inform the listener about a particular opinion and avoid the overt exposition of potentially conflicting beliefs.
- ii. The listener interprets such indirect (ambiguous) utterances by relying on her background beliefs, inferring apparent speaker's beliefs. She then generates a reply, again typically pursuing multiple objectives.
- iii. Upon perceiving the listener's reply, the speaker can infer apparent listener's beliefs under the consideration of her own first utterance.

During such ongoing interactions, the conversation partners may adapt their beliefs towards each other, experience shared beliefs, or run into a conflict even if social considerations were at play.

4 Prior beliefs in utterance interpretation

Our analysis of indirect utterances relies on the assumption that their interpretation depends on the background beliefs of the interpreter and her assumptions about the common ground. The reliance on prior beliefs in the interpretation of utterances is not a process that is specific to indirect utterances. The analysis of discourse comprehension and its dependence on the particular beliefs of the reader has been

demonstrated, for example, for the case of news perception. Van Dijk (1982) shows that the interpretation of a news piece about particular events in Central America can be affected by the ideologies of the readers. Upon reading the same piece, readers carrying different ideologies may form different evaluative opinions of what actually happened (Van Dijk & Kintsch 1983).

4.1 Related empirical phenomena

The role of prior beliefs in utterance interpretation can also be traced at the level of individual utterances. Utterance comprehension depends not only on the ability to derive the meaning of an utterance compositionally, but also on reference assignment, ambiguity resolution, and interpreting underspecified expressions. Even simple sentences, such as (3), require the listener to rely on common ground to establish references and construct the truth conditions of a sentence.

(3) The Boston office called. Hobbs (2004: 733)

When common ground is viewed broadly as a common ground of a whole speech community, we can interpret it as world knowledge, or communal common ground (Clark 2015). For example, to infer the relation between the two parts of the compound “Boston office” in (3), the interpreter needs to know or infer that there is, or might be, an office located in Boston. Moreover, to be able to recognize the same expression as a metonymy, the interpreter needs to know that it is usually people who make calls. With the additional knowledge that people tend to work in offices, the whole event can be integrated into a scene (Butz 2017, Knott 2012, Kuperberg 2021): a company-representing person, who works in an office located in Boston, called.

The interpretation of utterances can further depend on argumentative reasoning patterns. Argumentative patterns—or *topoi*—play a particular part in the interpretation of enthymemes—syllogisms in which the conclusion does not necessarily follow from the premises. Unlike in full syllogisms, where the premises are spelled out, some of an enthymeme’s premises can be implicit². Enthymemes are in this sense similar to indirect utterances because they rely on the beliefs and knowledge of a conversation partner to be fully understood. Breitholtz (2021) maintains that understanding can take place when either the *topos* is shared between the conversation partners, or when the listener can accommodate the *topos*, similarly to how presuppositions are accommodated (Lewis 1979, Stalnaker 1974). However, as Breitholtz (2021) emphasizes, while both *topoi* and presuppositions can be accommodated, they are critically different: while a presupposition is triggered by specific

² The treatment of an enthymeme as a truncated syllogism developed in the XIX century philosophy and logic, including the works of Schopenhauer, Krug, Whatley, among others (Kraus 2013).

linguistic material in an utterance (such as, for example, a definite noun phrase), for topoi the range of possible relevant candidates is infinite. Thus, the accommodation relies on world model knowledge. To illustrate how understanding depends on shared world knowledge, Breitholtz (2021) considers the example in (4), where a speaker announces her intention to attend a birthday party despite the fact that she is also attending a wedding on the same evening:

- (4) Oh! I'm invited to a wedding that night. But the bride is pregnant so I might drop by in the wee hours. (Breitholtz 2021: 1)

To recognise the pregnancy of the bride as a reason why the speaker may still attend the party, the audience must share the topoi in (5):

- (5) a. If the bride is pregnant, she will be tired.
b. If she is tired the wedding night would not go for that long.

(Breitholtz 2021: 1)

Unless the topoi in 5 are shared or can be inferred by the listener, it is impossible to comprehend why a bride's pregnancy may be an argument to stop by later on.

The effect of prior beliefs has further been registered in the area of projection inferences. Degen & Tonhauser (2021) evaluated how speakers assess the speaker's commitment to the content of an utterances in sentences, such as (6), and demonstrated the listener's belief about the probability of the utterance content affects the likelihood of that content projecting: more plausible content (6b) was more likely to project than less plausible content (6a). The authors showed that the effect was robust for a range of clause-embedding predicates and types of prior beliefs.

- (6) Did Cole discover that Julian dances salsa?
a. Julian is German. Julian dances salsa.
b. Julian is Cuban. Julian dances salsa.

4.2 Listener beliefs in theories of meaning

The work on the effect of prior beliefs on projection strength has demonstrated the importance of including prior beliefs in meaning computation (Degen & Tonhauser 2021). Traditionally, formal theories of meaning have focused mainly on the speaker meaning, while beliefs of the listener/reader were left out of the formalism (Asher et al. 2021). An idealized theory of semantics presupposes that words have the same meaning for all speakers. This assumption excludes language practices and thus limits our understanding of what meaning actually is and how it is computed (Beaver & Stanley 2018).

Recent work at the intersection of literature and formal linguistics demonstrates that actual interpretation dynamics depend on the individual experiences and characteristics of the person constructing it. [Bauer et al. \(2021\)](#) use examples of dramatic irony—a situation when the listener in a fictional dialogue understands the speaker’s words differently than a knowledgeable audience—to show how a person’s knowledge and experience shape the constructed interpretation. [Bauer et al. \(2021\)](#) further introduce the notion of “meaning for a reader” and appeal to the notion of FictionalAssert ([Bauer & Beck 2014](#)) to formalize how the interpretation of an utterance depends on the personal history of the reader. The authors argue that ambiguity, among other aspects of a text, invites the reader to invest interpretative effort, and thus relate the text to her own experiences, giving volume to its meaning.

Within a game-theoretic framework, [Asher et al. \(2021\)](#) formalize a model of deriving meaning as dependent on individual characteristics of a person, which the authors call biases. Biases, in their framework, are deeper beliefs of conversation partners—beliefs that may be hard to access or be aware of. Focusing on the analysis of discourse relations, the authors demonstrate that these biases have a profound impact both on speaker’s choice of discourse moves and the listener’s interpretation of those moves. The theoretical model implements utterance choice and interpretation as strategic moves in a message exchange game.

The effect of prior listener’s beliefs has also been modeled in the Bayesian pragmatic reasoning framework ([Degen et al. 2015](#)). Such models build in priors over states (situations or objects that are described) and possible utterances. They offer a convenient formalism of combining the semantics of particular utterances and the individual preferences of the listener in the calculation of meaning ([Degen 2023](#)). If the meaning of the utterance is ambiguous and pragmatic reasoning is not sufficient for disambiguation, state priors determined by one’s preferences may affect the interpretation ([Achimova et al. 2022a](#)). Speakers appear to be sensitive to this effect of prior beliefs on the interpretation: when the anticipated listener’s prior beliefs are in conflict with the utterance content, speakers have been shown to select more phonologically overt forms to avoid misinterpretation ([Achimova et al. 2022b](#)).

Indirect utterances offer a particular example of utterances that depend on background beliefs for their interpretation due to their ambiguity. Ambiguous utterances can receive different interpretations depending on the background assumptions of the interpreter. If the interpreter arrives at the meaning that the speaker intended, it means that their prior beliefs align.

In the next section, we propose a model of indirectness that aims to capture both the choice and interpretation of indirect utterances. In particular, we formalize multiobjective utterance choice driven by the desire to signal own beliefs while avoiding conflict of beliefs in common ground. We further consider the process of utterance interpretation and formalize how overt responses to an utterance can

reveal the underlying assumptions about prior beliefs of the responding listener. The model reveals that indirect utterances are in fact optimal for encoding a message when the speaker jointly pursues informational and social goals. Moreover, the model shows how indirect utterances offer a means to come to align beliefs, thus making them a part of the common ground.

5 The Alignment Model of Indirectness (AMI)

In this section, we introduce the Alignment Model of Indirectness (AMI). The goal of the model is to predict 1) that indirect utterances are optimal in situations of uncertainty about the background beliefs of the listener; 2) what inferences the speaker draws about the actual beliefs of the listener upon observing her reply to an indirect utterance.

We develop the model formalism within the Rational Speech Act framework (Frank & Goodman 2012, Goodman & Frank 2016). In this paradigm, the speaker chooses utterances by reasoning about the listener. The speaker’s task is to choose utterances that maximize the chance of the listener at arriving at the intended meaning. The novelty of AMI is threefold. First, it formalizes a more complex utility function that modulates utterance choices. Intuitively, the function tends to generate indirect utterances particularly when two (or more) objectives are at odds with each other. Second, when assessing the utility of utterances, the speaker evaluates utterance compatibility with each possible belief state of the listener. AMI shows that ambiguous utterance choices can emerge simply by encoding opinions and beliefs as densities and optimizing utterance choice to match the considered densities. Finally, we formalize a Bayesian mechanism that models the recursive inference of posterior background beliefs of speaker and listener given their utterances and prior background beliefs. In the next subsections, we first introduce the basic RSA architecture and then show our extensions of the framework.

5.1 The RSA architecture

The RSA framework models communication as a combination of two processes: choosing utterances and inferring interpretations that rationalize the speaker’s behavior. The speaker’s goal here is to choose the most informative utterance from a set of alternatives, and the listener’s goal is to weigh the potential world states that could have triggered the utterance. The framework itself does not handle how the sets of alternatives are composed and leaves this task to phenomena-specific theories (Degen 2023).

5.1.1 Vanilla RSA

Utterance planning in the so-called vanilla RSA model starts with the speaker and a message that she wants to send. We will use the paradigm of reference games to illustrate the model concepts. In a reference game, the message is a reference to one of the objects that the listener should pick. The speaker chooses utterances by reasoning about the listener: for each utterance she evaluates the probability that the listener will choose the intended object. In turn, the listener evaluates whether each of the candidate objects qualifies for the reference based on the literal meaning of the utterance. This literal listener function may be formalized as:

$$(1) \quad P_{L_0}(s | u) \propto \mathfrak{f}(u, s),$$

where $\mathfrak{f}(u, s)$ denotes a truth function, which specifies which subset of possible belief states s is compatible with the utterance u . For example, in the scenario shown in Figure 1, if the listener heard the utterance “red” and was instructed to choose an object out of three available ones, the L_0 would assign the probability of 0 to object 1, 0.5 to object 2, and 0.5 to object 3.



Figure 1 Scenario: A green dotted cloud, a red striped cloud, and a red dotted circle. The L_0 hears the utterance “red” and assigns the probabilities of 0, 0.5, 0.5 to the three objects, respectively.

The speaker rates utterances based on their utility, which is determined by the probability of the listener choosing the intended object s given utterance u :

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

The speaker then assigns probabilities to utterances that are exponentially proportional to the utterance utility:

$$(3) \quad P_{S_1}(u | s) \propto \exp(\alpha \cdot U_{S_1}(u; s))$$

The top layer of the model is the pragmatic listener L_1 , which assigns probabilities to objects by reasoning about the speaker rather than relying on the literal meaning of utterances:

$$(4) \quad P_{L_1}(s | u) \propto P_{S_1}(u | s) \cdot P(s)$$

In sum, ambiguity of reference in the RSA model is resolved by applying a reasoning strategy: the pragmatic listener imagines a cooperative speaker and infers the intended meaning by reasoning about her communicative behavior. This type of reasoning has been subject to a recent debate, since often the ambiguity can be resolved through simpler strategies and heuristics (Sikos et al. 2019). In fact, in this paper, we will modify the literal listener function and enrich it with background beliefs. Thus, background beliefs will affect disambiguation by making some meanings more plausible than others. As a result, the model we propose in this paper offers a disambiguation path with fewer layers of recursion.

5.1.2 Previous RSA extensions

The RSA framework offers a versatile set of tools to include additional components into the calculation of utterance probabilities and the probabilities of choosing states. We will consider two extensions relevant to the current work: enriching the literal listener function L_0 with prior probabilities that make particular world states more likely than others and including social goals into the calculation of utterance utility on the speaker side.

By default, the literal listener function assigns equal probability to all qualifying objects. If we expect some objects to be chosen more often, we can add a prior probability over objects to the calculation. Thus, the choice of objects can be additionally determined by particular feature preferences f (Achimova et al. 2022a):

$$(5) \quad P_{L_0}(s \mid u, f) \propto \mathfrak{f}(u, s) \cdot P(s \mid f) \cdot P(f),$$

where $P(s \mid f)$ denotes prior world state interpretation preferences, such that object choices are biased towards preferred objects. In this paper, we will introduce a term that replaces feature preferences f with a term denoting the listener’s background beliefs. Following the formalization above, the interpretation of an utterance may thus be modeled by the product of two prior probability distributions that model the listener’s beliefs:

$$(6) \quad P_{L_0}(s \mid u, b) \propto P(u \mid s, b) \cdot P(s \mid b) \cdot P(b).$$

The addition of background beliefs into the model will allow us to show how the speaker’s choice of utterances depends on her assumptions about the listener’s background beliefs.

The speaker’s behavior in the vanilla RSA model is driven by the goal of signalling the intended meaning efficiently. However, the vanilla RSA fails to account for situations, where speakers choose indirect utterances. One such case involves the phenomenon of politeness. In order to account for polite utterances, Yoon et al.

(2020) added a social utility component into the overall utility calculation. They equated social utility to sending positive feedback to the listener. If the speaker optimized solely the social utility, she would be expected to select only positive utterances. Setting priority to sending fully true information would result in the preference for direct utterances. A combination of these goals led to polite utterances being chosen. Such utterances also turned out to be indirect. However, the politeness model does not scale to other cases of indirectness, where the speaker’s goal lies elsewhere than giving feedback to the listener.

In the following, we introduce the Alignment Model of Indirectness (AMI), which operationalizes the social utility of utterances by considering updates to the common ground. We start with proposing a more general notion of opinions, which are then exchanged in AMI.

5.2 Opinions and their degree of alignment

The general idea of AMI 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 Beta distribution, parameterized in terms of its mean $\mu \in [0; 1]$ and “sample size” $\nu > 0$.³ 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

³ A 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 $\nu = \beta_1 + \beta_2 - 2$.

distribution over $[0; 1]$. As a result, v 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 $v \geq 0$). We denote the listener’s and speaker’s opinion as O_L and O_S respectively. Figure 2 illustrates some opinions encoded as Beta distributions.

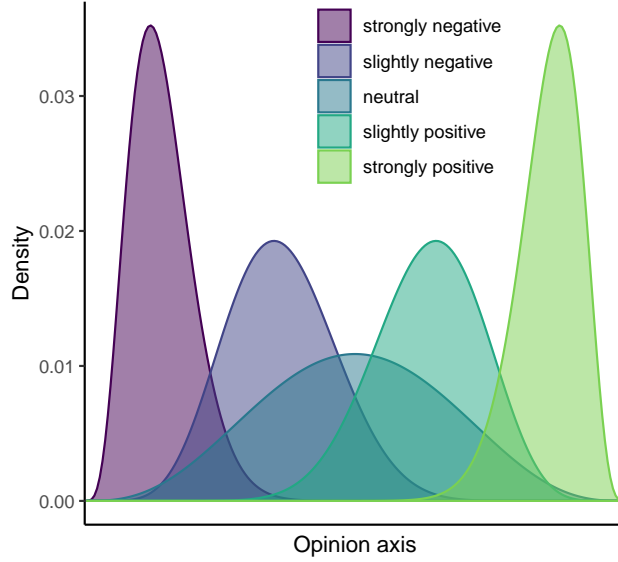


Figure 2 Examples of different opinion states as Beta distributions with different values for parameters ‘position’ μ and ‘opinionatedness’ v . The five densities’ parameters are (from order of increasing ‘position’): $\mu_1 = .14$, $v_1 = 35$, $\mu_2 = .36$, $v_2 = 22$, $\mu_3 = .5$, $v_3 = 8$, $\mu_4 = .64$, $v_4 = 22$, $\mu_5 = .86$, $v_5 = 35$.

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 function⁴

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

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

⁴ As for notation, we write $\Delta(X)$ as the set or space of all probability distributions over X .

of Kullback-Leibler divergence to measure alignment. If P and Q are probability distributions, we define opinion divergence as:

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

where D_{KL} is KL-divergence.⁵

5.3 Higher-order beliefs about opinions

AMI assumes that pragmatic choices of utterance are sensitive to opinion alignment, and that interpretation of an utterance is as well. 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. For present purposes, it is not necessary to go beyond second-order uncertainty, but it is useful nonetheless to have a general notation for any higher-order belief.

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.

5.4 Literal interpretation and the semantics of utterances

The usual role of the literal listener in RSA models is to anchor pragmatic reasoning in literal interpretation. For AMI, we require a literal listener that captures how various utterances relate to opinion states. While it may be possible, and ultimately desirable, to derive the way that utterances like “This was interesting!” change an interpreter’s opinion just in virtue of their denotational, truth-functional meaning, this exercise is difficult and orthogonal to our current purposes. But even without

⁵ Other information-theoretic measures of divergence or distance are conceivable. Figure A.1 in the appendix shows divergences between the five opinion states from Figure 2, for symmetrized KL-divergence and some salient alternatives. Simulations using several of these alternatives yield similar qualitative predictions.

a full-fledged theory of how opinion states change in light of literally interpreted utterances, we can test the genuine *pragmatic* implications of AMI, if we treat the semantics of utterances, for the time being, merely in terms of their “opinion-change potential” (OCP), a term coined in intentional analogy to the “context-change potential” of dynamic semantics (e.g. Heim 1983, Groenendijk & Stokhof 1991, Kamp & Reyle 1993). In this spirit, we empirically measure a plausible “opinion-change meaning” in an as neutral as possible context and consider this, for the time being, the starting point of pragmatic reasoning. 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.⁶

5.5 Experiment 1: Empirical baseline of utterance meanings

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 have collected judgements from naïve speakers, following the procedure also used by Yoon et al. (2016), who asked participants to evaluate expressions, 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 (7) on a heart-scale from 1 “strongly negative” to 5 “strongly positive”:

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

A sample trial is shown in Figure 3.

We evaluated a total of 10 different topics each featuring 10 adjectives. Figure 4 displays the ratings assigned by the participants to each of these adjectives with all

⁶ Future work might want to consider an additional complication, namely that a speaker may not know precisely how an utterance may be interpreted, similar to models that include the speaker’s uncertainty about lexical meaning (e.g. Bergen et al. 2016, Potts et al. 2015, Franke & Bergen 2020).



Figure 3 Sample trial of the utterance meaning assessment experiment

topics pooled together. It is these empirical distributions that we use as first approximations to the semantic meaning of utterances in terms of the de-contextualized “opinion-change potential”. 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 4. 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 (8) may be judged as indirect if the speaker actually has a negative opinion about the election outcome (1 or 2 hearts on our scale). The same utterance can be direct if the true belief state corresponds to 4 hearts.

(8) I found the election outcome decent.

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.

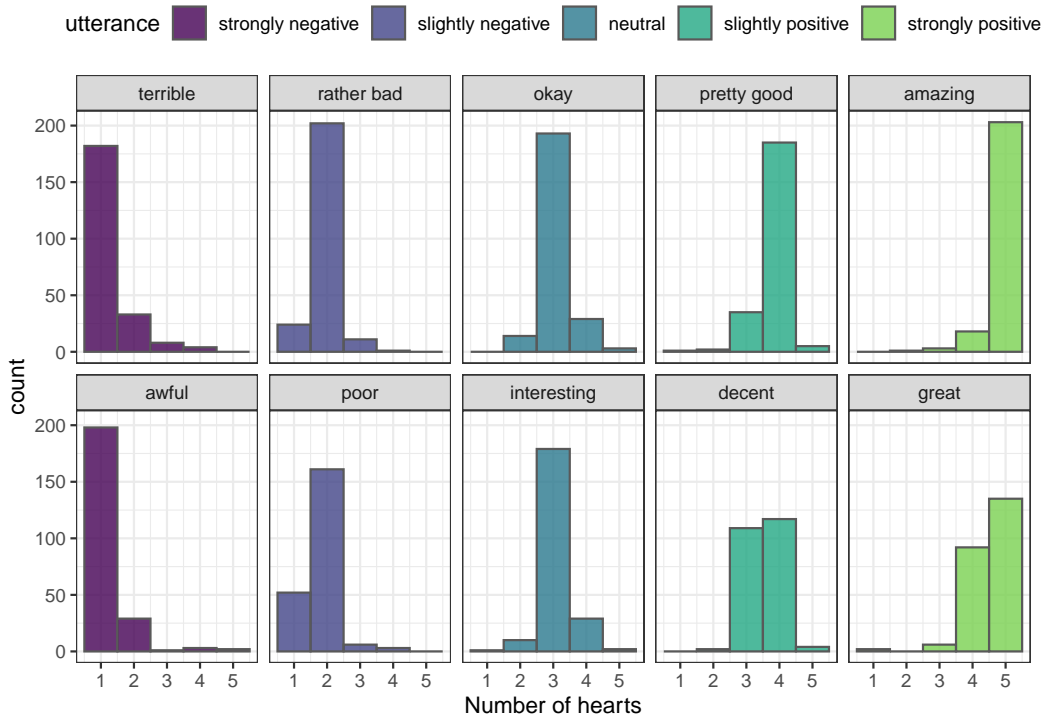


Figure 4 Utterance ratings for 10 considered adjectives

5.6 Pragmatic speaker

AMI proposes that a pragmatic speaker chooses between utterances in a multiobjective manner, attempting to (i) signal their own opinion and (ii) align with the listener’s opinions, that is, avoiding potential conflict with the listener. The mental state of the pragmatic speaker is sufficiently described by their own opinion O_{S_1} and their beliefs about the opinion of the listener $\pi_1^{S_1}$. With these two mental states, and the assumption about the literal listener’s interpretation of utterances, that is, $L_0(u)$, we can define the two goals:

- i. **informative goal:** $L_0(u)$ should be as close as possible to the speaker’s own opinion O_{S_1} , and
- ii. **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 listener⁷:

$$U_{\text{inf}}(O_{S_1}, u) = -\text{Div}(O_{S_1}, L_0(u))$$

$$U_{\text{soc}}(\pi_1^{S_1}, u) = - \int P_{S_1}(O_L | \pi_1^{S_1}) \text{Div}(O_L, L_0(u)) dO_L$$

The *total utility* U_{total} is a linear combination of these two, with parameter γ weighing their relative importance:

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

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

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

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

As an example, Figure 5 shows the model-determined probabilities of each of the five considered utterances being chosen given that the speaker’s actual opinion is strongly positive and depending on the assumption about the listener’s opinion. AMI 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 expected to align, the more likely becomes the probability to choose the most direct opinionated statement.

5.7 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

⁷ The model specification assumes that the speaker knows how the listener interprets utterances. 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, as, for example, in Bergen et al. (2016).

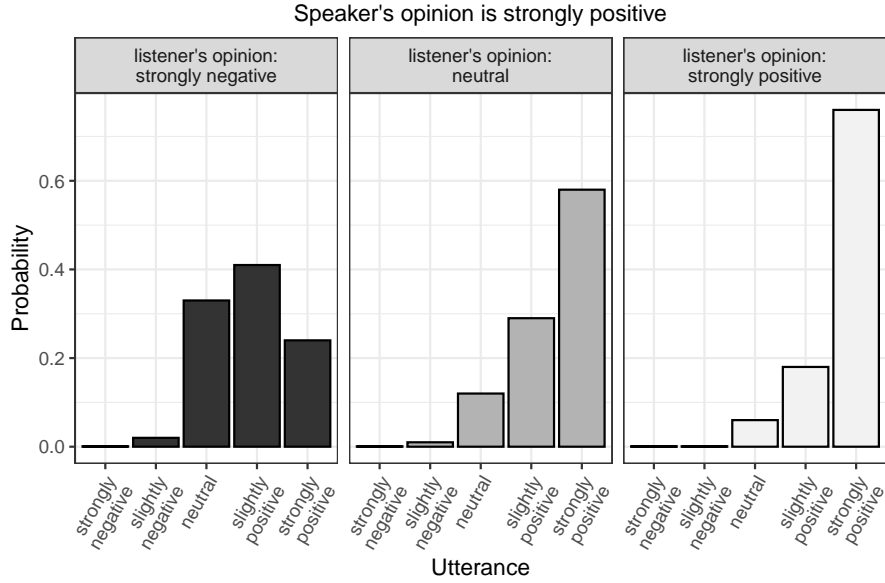


Figure 5 Utterance choice: model predictions. The speaker’s actual opinion is strongly positive. Her utterance choice depends on her opinion and the anticipated opinion of the listener, as well the communicative goal. In this simulation, the informational and social goals are weighted at 0.8 and 0.2, 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 anticipates a conflicting listener’s opinion (*strongly negative*), she prefers a less direct utterance (*slightly positive*) to signal her opinion that is actually strongly positive.

$\pi_2^{L2} \in \Delta(\Delta(\mathcal{O}))$ about the pragmatic speaker’s beliefs about the listener’s opinion. The posterior beliefs of the pragmatic listener are inferred by Bayes rule:

$$(8) \quad P_{L2} \left(O_{S1}, \pi_1^{S1} \mid u, \pi_1^{L2}, \pi_2^{L2} \right) \propto P_{S1} \left(u \mid O_{S1}, \pi_1^{S1} \right) P_{L2} \left(O_{S1}, \pi_1^{S1} \mid \pi_1^{L2}, \pi_2^{L2} \right)$$

Notice that AMI 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—a challenge that we leave for future research.⁸

⁸ How exactly listener’s update their own opinion based on what speaker’s say will require more elaboration, including factors like trust, status, competence and the like. 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.

Conversation partners take turns in producing utterances and interpreting them. Each partner therefore carries out both speaker and listener functions. AMI’s inference processes in a communication scenario are shown in Figure 6.

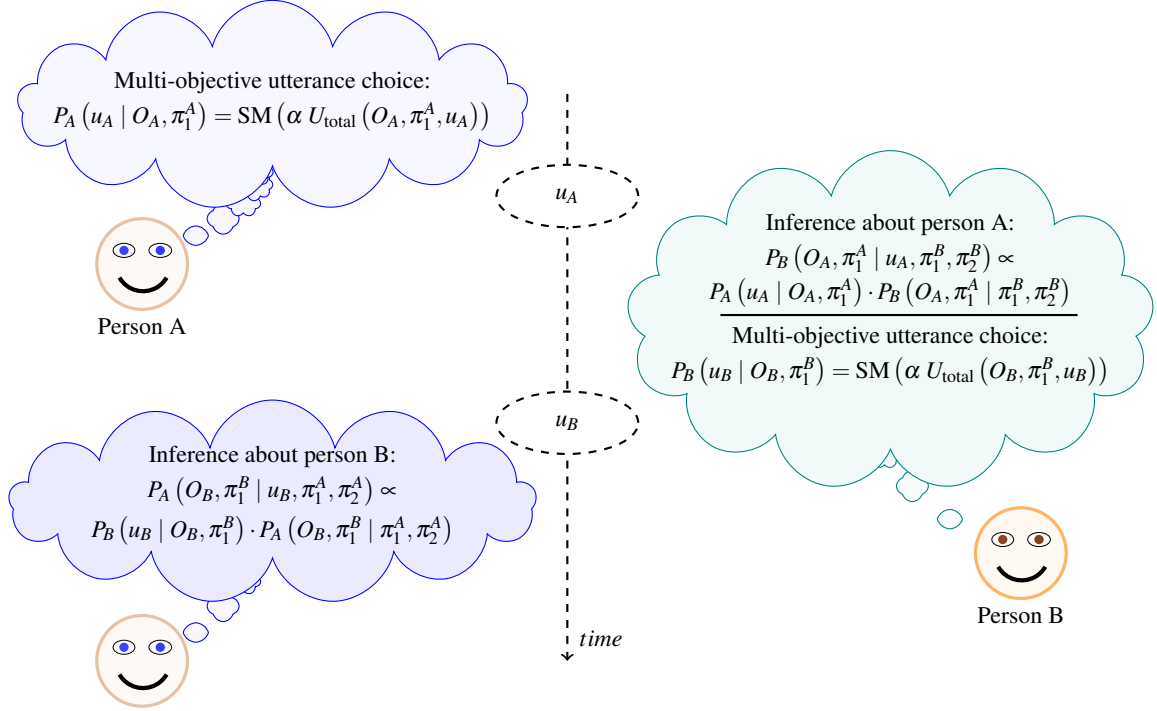


Figure 6 Multi-turn interaction in AMI. Each agent infers the other agent’s beliefs based on their prior beliefs about the interlocutor’s beliefs and the utterance probabilities that these prior beliefs about the interlocutor entail. In this plot, we diverge from talking about speakers and listeners and instead talk about persons A and B. This simplifies the notation so that, for example, π_2^B are person B’s second order beliefs.

5.8 Modeling learning about each other

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. For example, after a first utterance u_1 , 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) may then feed into their choice of subsequent utterance u_2 with Equation (7), which Alex in turn interprets via Equation (8) to learn from how Bo reacted (by u_2) to her utterance u_1 . In this way, the model sketched here shows a path for agents to learn about each other’s beliefs.

A 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 by anticipated follow-up utterances. For instance, Alex may choose a particular u_1 also taking into account how much they will learn about Bo’s beliefs from the likely reactions u_1 may trigger from 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. 2022a, 2023).

To assess the model’s predictions we simulate the interaction shown in Figure 6: person A chooses an utterance u_A to which person B then 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).⁹ We assume uniform priors of both persons about each other’s opinions and first order beliefs. Moreover, when inferring the person B’s opinion, person A assumes that person B’s belief about person A’s opinion on the matter corresponds to person A’s utterance, that is, we set π_2^A to a distribution that is single-peaked at u_A ¹⁰.

Table 1 shows selected results from these simulations.¹¹ We see that when the 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 the 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 as slightly negative utterance, the model infers that the listener’s actual belief is strongly negative (45% chance, previous to last row).

⁹ Similar values and similar densities yield similar results.

¹⁰ Alternatively, π_2^A may itself be inferred via Eq. 7 starting with a uniform prior over π_2^A and given utterance u_A . Our implementation yields nearly identical results for this more precise computation.

¹¹ A full simulation of all possible combinations of utterances and responses can be found in the Appendix (Figure 17).

		A’s posterior beliefs about of B’s opinion				
		Strongly negative	Slightly negative	Neutral	Slightly positive	Strongly positive
		<i>A: The election results are terrible (strongly negative)</i>				
B: <i>I find them rather bad</i>	0.14	0.35	0.29	0.2	0.02	
B: <i>I find them decent</i>	0	0.07	0.16	0.31	0.45	
		<i>The election results are okay (neutral)</i>				
B: <i>I find them rather bad</i>	0.26	0.36	0.24	0.13	0.01	
B: <i>I find them decent</i>	0.01	0.13	0.24	0.36	0.25	
		<i>The election results are amazing (strongly positive)</i>				
B: <i>I find them rather bad</i>	0.45	0.31	0.16	0.07	0	
B: <i>I find them decent</i>	0.02	0.2	0.29	0.35	0.14	

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


5.9 Summary: Modeling

In this section, we have presented AMI, the Alignment Model of Indirectness, which is intended to formalize the intuition that one reason for the use of indirect utterances is to avoid conflict in common ground, by allowing divergences in opinion to be revealed without direct conflict. The meaning calculation in the model is rooted in the Literal listener function L_0 , which returns distributions over potential meanings of an utterance. To represent the meaning of utterances and the opinions of conversation partners, we have appealed to Beta distributions. We have further proposed that they can serve as an approximation for empirically obtained distributions reported in Experiment 1. We then introduced a Pragmatic speaker function S_1 that regulates the choice of utterances by balancing the informational and social goals. Finally, we have formalized the process of inferring the beliefs of a speaker following her utterance in the Pragmatic listener function L_2 . This model architecture predicts that indirect utterances become an optimal speaker’s choice when she


Brianna wants to discuss the election results with Stephanie.

Here is how both really feel about the issue:

Brianna:

Strongly Negative  Strongly Positive

Stephanie:

Strongly Negative  Strongly Positive

Brianna wants to share her opinion but avoid possible conflicts.

What should Brianna say?

- The election results are awful.
- The election results are rather bad.
- The election results are interesting.
- The election results are decent.
- The election results are great.

Click 'continue' to move on.

Figure 7 Experiment 2: Sample trial

is simultaneously pursuing informational and social goals. It further captures the fact that speaker’s opinion may be different from the literal meaning of her utterance. AMI also makes non-trivial predictions, like shown in Table 1 about opinion inferences in two-turn dialogues (of the kind shown in Figure 6). In the next section, we carry out an empirical test of the model and discuss to which extent AMI reflects the qualitative patterns we witness in the data.

6 Behavioral data

In this section, we report the results of two experiments that were designed to test the predictions generated by AMI, and in particular the pragmatic speaker function and the pragmatic listener function.

6.1 Experiment 2: Pragmatic speaker

Experiment 2 (n = 98, 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 7 participants were excluded from the analysis, since these participants reported that they did not fully understand the instructions. Figure 7 shows the experiment set up. Participants had to choose an

utterance to satisfy one of the following communicative goals: share opinion (informational), share opinion and avoid conflict (informational + social), or simply avoid conflict (social). The speaker's and listener's opinions were either strongly negative (1 heart) or strongly positive (5 hearts). We therefore evaluated 4 situations: where the speaker's and listener's opinions aligned either on the positive or the negative side, and where there was a mismatch in either direction.

Based on the the association of adjective meanings and the hearts scale established by Experiment 1, AMI 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.

The distributions of participants' choices is shown in Figure 8. We analyzed the data by fitting a generalized linear mixed model with the type of utterance (direct or indirect) as the dependent variable and the communicative goal and the opinion constellation as the independent variables.¹² Strongly negative and strongly positive utterances were tallied as direct ones. The utterances that belonged to the middle of the scale (slightly negative, neutral, and slightly positive) were treated as indirect ones because there was a stronger alternative to represent the actual opinion that the speaker holds (either 1 or 5 hearts). The maximal random-effect structure (Barr et al. 2013) of the converging model included random intercepts for participants. The analysis revealed that the participants were more likely to choose an indirect utterance when the opinions of conversation partners did not match ($\beta = 3.328$, $SE = 0.226$, $z = 12.504$, $p < 0.001$) and less likely to select indirect utterances when the speaker pursued a purely informational goal ($\beta = -2.422$, $SE = 0.3$, $z = -8.087$, $p < 0.001$), which is in agreement with the predictions of AMI.

So far, we have explored the effect of communicative goals in the whole set of scenarios independent of whether the conversation partners' opinions matched or mismatched. We can now target the critical "mismatch" case and investigate whether the utterance choices depend on the communicative goals. The analysis revealed that the participants were more likely to choose an indirect utterance when the speaker was pursuing a social goal either in addition to the informational one ($\beta = 3.51$, $SE = 0.58$, $z = .227$, $p < 0.001$) or alone ($\beta = 2.871$, $SE = 0.5$, $z = 5.736$, $p < 0.001$), compared to purely informational goals (Figure 9).

¹² Here we report the model without the interaction between the independent variables. An evaluation of a model with the interaction showed that it did not result in a significant effect for either of the combinations of variable levels. Model comparison further revealed that the more complex interaction model did not improve the model fit ($\chi^2 = 1.67$, $df = 2$, $p = 0.434$).

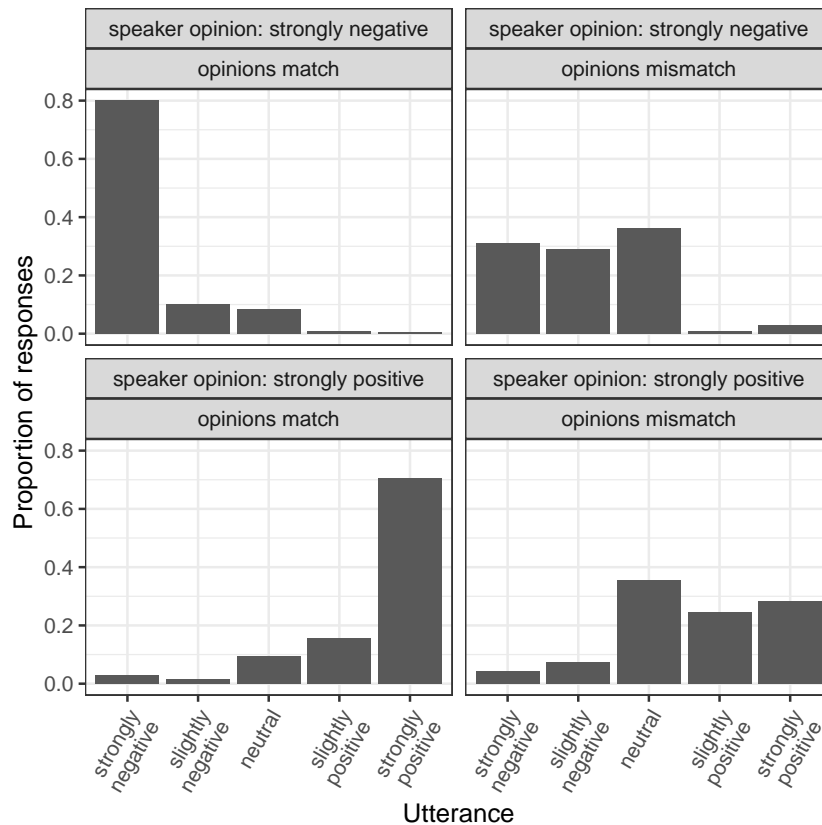


Figure 8 Experiment 2: Utterance choice. When opinions match (left column), participants choose strongly positive or strongly negative utterances. When opinions mismatch (right panels), they select indirect utterances more often.

6.2 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 5.8 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 6 adjectives (out of 10 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

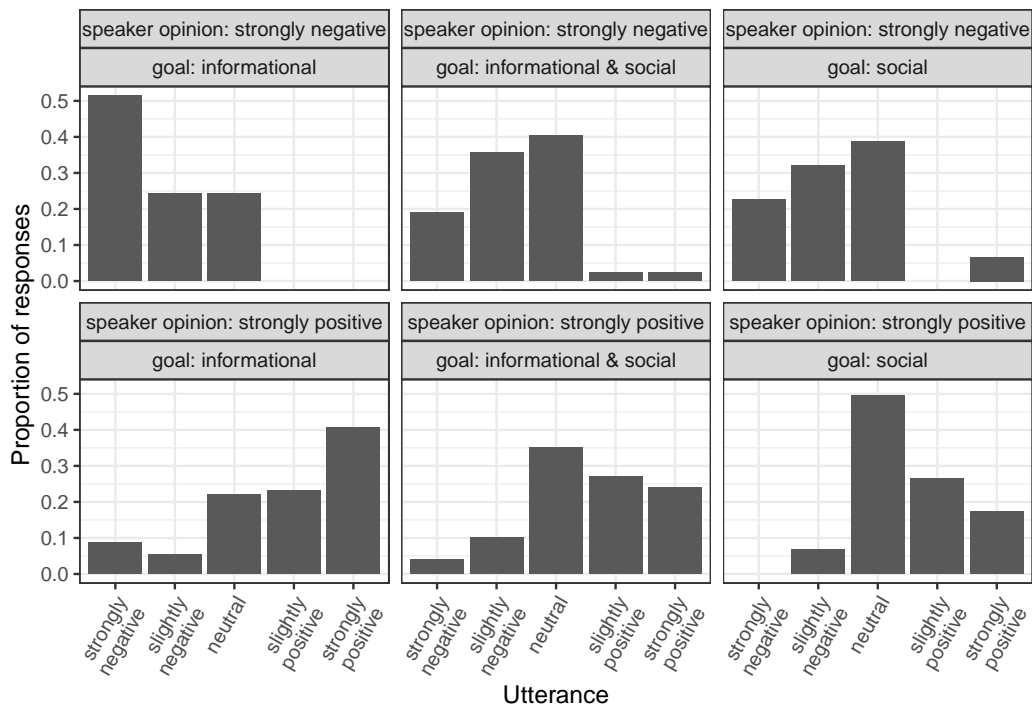


Figure 9 Experiment 2: Utterance choice when opinions mismatch. When speakers face an informational goal (left column), participants primarily chose utterances that directly match the speaker’s opinion (strongly positive, top row or strongly negative, bottom row). When participants also considered social goals (either alone or in combination with informational ones, middle and right column of both rows), they selected more indirect statements.

2 adjectives representing the middle of the scale. The second speaker’s adjectives included 6 possible responses and excluded the most opinionated replies (strongly positive and strongly negative), since they were not compatible with the stated communicative goal. Figure 10 displays a sample trial for Experiment 2.

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 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 pos-

Brandon and Matthew meet at a college event for the first time.
They would like to exchange opinions but don't want to run into a conflict.

Brandon says: The election results are okay.

Matthew replies: I find them pretty good.

How may Matthew actually feel about the issue?

Strongly Negative  Strongly Positive

Click 'continue' to move on.

Continue

Figure 10 Utterance ratings for 10 considered adjectives

itive). In critical trials, we then asked how the second speaker may have felt about the topic. In control trials (1 trial out of 6), 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 11 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 18).

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 AMI’s predictions.

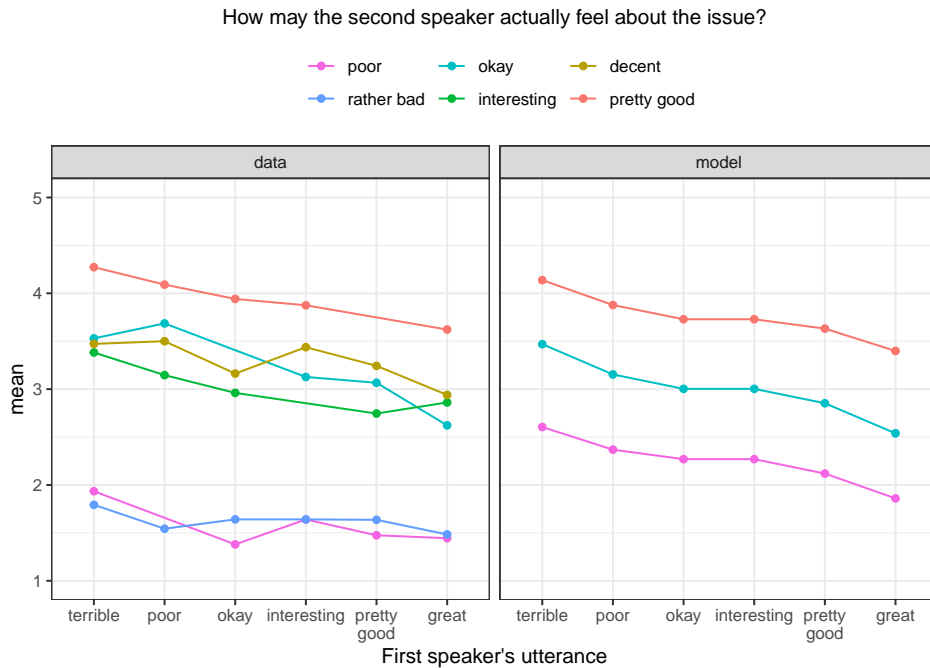


Figure 11 Opinion inference scores. The left panel displays the participants’ evaluation of the second speaker’s opinion. The right panel shows corresponding model predictions for slightly positive (*pretty good*), neutral (*okay*), and slightly negative (*poor*) utterances.

6.3 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 AMI model. The results of Experiment 1 support representing the meaning of utterances and speaker opinions in the form of distributions. 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 AMI 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 AMI. The empirical findings thus qualitatively support AMI. We leave a more detailed assessment of all model parameters and their interactions to future work.

7 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 (Frank 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 in common ground, which we defined in terms of incompatible opinions, also affects the types of utterances speakers opt for. We have treated indirectness as a common ground management tool that allows the speaker to simultaneously satisfy informational and social goals. We have furthermore suggested that indirectness allows the speaker to probe the state of common ground while leaving interpretation space to the listener. Thus, conflict avoidance brings the additional benefit of implicitly checking if beliefs are shared.

We have brought together literature from social psychology, philosophy of language, psycholinguistics, and cognitive modeling to describe the mechanisms that underlie the social implications of generating indirect utterances as well as interpreting such ambiguous utterances and generating according responses. We have argued that resolving ambiguities exposes the listener's opinions. Thus, monitoring the interpretation of ambiguous utterances reduces uncertainty over which opinions are shared and belong to the common ground. Moreover, it allows us to probe each other's opinions and to adjust our own stance before fully committing to a particular proposition publicly. Indirectness can thus 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 Alignment Model of Indirectness (AMI) formalizes how the interpretation of indirect utterances, and ambiguity resolution in particular, continuously provides conversation partners with signals of whether their belief systems align. Conversation partners monitor each other’s interpretation of indirect utterances and draw inferences about each other’s opinions. These inferences open space for belief alignment and contribute to establishing and maintaining social bonds.

Data availability

Experimental data and model simulations are available in this OSF repository: <https://tinyurl.com/indirectnessPaper>.

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. 2022a. Learning about others: Modeling social inference through ambiguity resolution. *Cognition* 218. 104862.
- Achimova, Asya, Christian Stegemann-Philipps, Susanne Winkler & Martin V. Butz. 2022b. Anaphoric reference in descriptions of surprising events. *Proceedings of Sinn und Bedeutung* 26 20–34. <https://doi.org/10.31234/osf.io/d5krz>.
- Andersen, Susan M & Elizabeth Przybylinski. 2018. Shared reality in interpersonal relationships. *Current opinion in psychology* 23. 42–46.
- Asher, Nicholas, Julie Hunter & Paul Soumya. 2021. Bias in semantic and discourse interpretation. *Linguistics and Philosophy* .

- 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.
- Bauer, Matthias & Sigrid Beck. 2014. On the meaning of fictional texts. In Daniel Gutzmann, Jan Köpping & Cécile Meier (eds.), *Approaches to meaning*, 250–275. Brill.
- Bauer, Matthias, Sigrid Beck, Julia Braun, Susanne Riecker & Angelika Zirker. 2021. Multiple contexts in drama: the example of Henry V. Ms., University of Tuebingen.
- Beaver, David & Jason Stanley. 2018. Toward a non-ideal philosophy of language. *Graduate Faculty Philosophy Journal* 39(2). 503–547.
- Beaver, David I. 2001. *Presupposition and assertion in dynamic semantics*, vol. 29. CSLI publications Stanford.
- Beaver, David Ian. 1997. Presupposition. In *Handbook of logic and language*, 939–1008. Elsevier.
- Bergen, Leon, Roger Levy & Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* 9.
- Breitholtz, Ellen. 2021. *Enthymemes and topoi in dialogue: The use of common sense reasoning in conversation*. Brill.
- 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.
- Brown-Schmidt, Sarah & Melissa C Duff. 2016. Memory and common ground processes in language use. *Topics in Cognitive Science* 8(4). 722–736.
- 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.
- Butz, Martin V. 2017. Which structures are out there? learning predictive compositional concepts based on social sensorimotor explorations. In Thomas K. Metzinger & Wanja Wiese (eds.), *Philosophy and predictive processing*, Frankfurt am Main: MIND Group. <https://doi.org/10.15502/9783958573093>.
- Carcassi, Fausto & Michael Franke. to appear. How to handle the truth: A model of politeness as strategic truth-stretching. In *Proceedings of the 45th annual meeting of the cognitive science society*, .
- Castellano, Claudio, Santo Fortunato & Vittorio Loreto. 2009. Statistical physics of social dynamics. *Reviews of Modern Physics* 81. 591–646.

- Clark, Eve V. 2015. *Common ground* 328–353. John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9781118346136.ch15>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118346136.ch15>.
- Clark, Herbert H. 1996. *Using language*. Cambridge, UK: Cambridge university press.
- 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.
- Degen, Judith, Michael Henry Tessler & Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings & P. P. Maglio (eds.), *Proceedings of the 37th annual meeting of the cognitive science society*, 548–553. Austin, TX: Cognitive Science Society.
- Degen, Judith & Judith Tonhauser. 2021. Prior Beliefs Modulate Projection. *Open Mind* 5. 59–70. https://doi.org/10.1162/opmi_a_00042. https://doi.org/10.1162/opmi_a_00042.
- DeGroot, Morris H. 1974. Reaching a consensus. *Journal of the American Statistical Association* 69(345). 118–121.
- Döring, Sophia. 2018. *Modal particles, discourse structure and common ground management*.: Humboldt-Universität zu Berlin, Sprach- und literaturwissenschaftliche Fakultät dissertation. <https://doi.org/http://dx.doi.org/10.18452/19449>.
- 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.
- Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation* 1. 221–237.
- Ginzburg, Jonathan. 2012. *The interactive stance: Meaning for conversation*. Oxford University Press.
- 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. 173–184.
- Grice, Paul. 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Groenendijk, Jeroen & Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and philosophy* 39–100.
- 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).
- Heim, Irene. 1983. File change semantics and the familiarity theory of definiteness. In Rainer Bäuerle, Christoph Schwarze & Arnim von Stechow (eds.), *Meaning, use, and interpretation of language*, 164–189. Berlin, Boston: De Gruyter. <https://doi.org/doi:10.1515/9783110852820.164>. <https://doi.org/10.1515/9783110852820.164>.
- Henderson, Robert & Elin McCready. 2019. Dogwhistles, trust and ideology. In *Proceedings of the 22nd amsterdam colloquium*, 152–160.
- Higgins, E Tory. 2019. *Shared reality: What makes us strong and tears us apart*. Oxford University Press.
- Hobbs, Jerry R. 2004. Abduction in natural language understanding. In Gregory Ward Laurence R. Horn (ed.), *Handbook of pragmatics*, 724–741. Blackwell Publishing.
- Horton, William S. 2005. Conversational common ground and memory processes in language production. *Discourse Processes* 40(1). 1–35.
- Horton, William S & Richard J Gerrig. 2016. Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science* 8(4). 780–795.
- Kamp, Hans & Uwe Reyle. 1993. *Tense and aspect* 483–689. Dordrecht: Springer Netherlands.
- Karagjosova, Elena. 2004. *The meaning and function of german modal particles*: Deutsches Forschungszentrum für Künstliche Intelligenz, DKFI; Saarland University, Dept. of Computational Linguistics and Phonetics. dissertation.
- 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. https://doi.org/10.1162/tacl_a_00037. https://doi.org/10.1162/tacl_a_00037.
- Knott, Alistair. 2012. *Sensorimotor cognition and natural language syntax*. Cambridge, MA: MIT Press.
- Kraus, Manfred. 2013. Enthymem. In Gert Ueding (ed.), *Historisches Wörterbuch der Rhetorik online: Bieeul*, De Gruyter. <https://www.degruyter.com/database/HWRO/entry/hwro.2.enthymem/html>.

- Krifka, Manfred. 2008. Basic notions of information structure. *Acta Linguistica Hungarica* 55(3-4). 243–276.
- Kuperberg, Gina R. 2021. Tea with milk? A hierarchical generative framework of sequential event comprehension. *Topics in Cognitive Science* 13. 256–298. <https://doi.org/10.1111/tops.12518>.
- Lewis, David. 1969. *Convention: a philosophical study*. Harvard, MA: Harvard University Press.
- Lewis, David. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8. 339–359.
- Mahajan, Neha & Karen Wynn. 2012. Origins of us versus them: Prelinguistic infants prefer similar others. *Cognition* 124(2). 227–233.
- Pimentel, Tiago, Rowan Hall Maudslay, Damián Blasi & Ryan Cotterell. 2020. Speakers fill lexical semantic gaps with context. *arXiv:2010.02172*.
- 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.
- 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*.
- Searle, John R. 1979. Indirect speech acts. In John R Searle (ed.), *Expression and meaning: Studies in the theory of speech acts*, 59–82. Cambridge University Press.
- Sikos, Les, Noortje Venhuizen, Heiner Drenhaus & Matthew Crocker. 2019. Reevaluating Pragmatic Reasoning in Web-based Language Games. <https://doi.org/10.13140/RG.2.2.30535.14249>.
- Stalnaker, Robert. 1974. Pragmatic presuppositions. In M.K. Munitz & P Unger (eds.), *Semantics and philosophy*, 197–214. New York: New York University Press.
- Stalnaker, Robert. 2002. Common ground. *Linguistics and philosophy* 25(5/6). 701–721.
- Stalnaker, Robert. 2014. *Context*. Oxford: Oxford University Press.
- Tomasello, Michael. 2019. *Becoming human: A theory of ontogeny*. Belknap Press.
- Van Dijk, Teun A. 1982. Opinions and attitudes in discourse comprehension. In *Advances in psychology*, vol. 9, 35–51. Elsevier.
- Van Dijk, Teun A. & Walter Kintsch. 1983. *Strategies of discourse comprehension*. New York: Academic 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.
- 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. https://doi.org/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

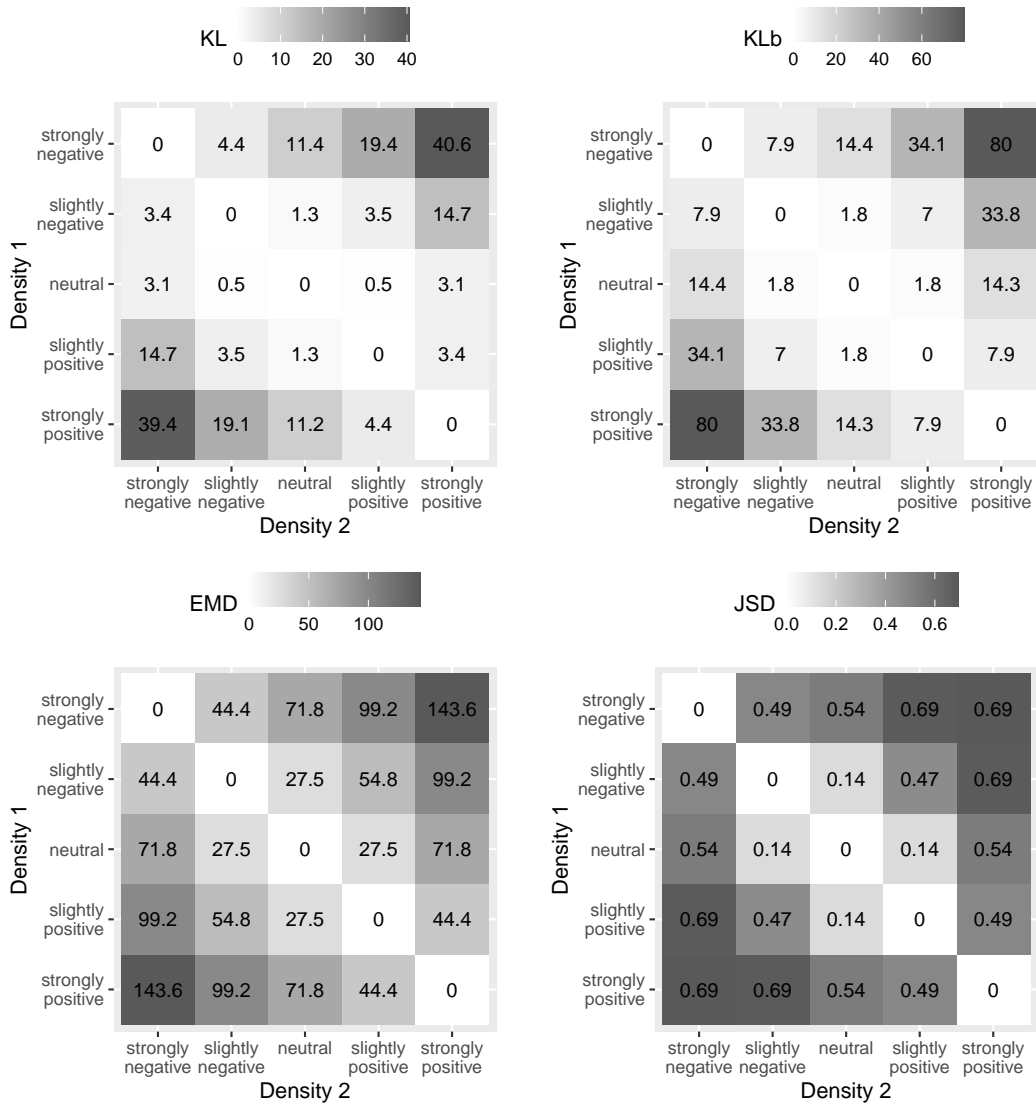


Figure 12 Measures of divergence between the opinion distributions may be interpreted as encoding the effort to change one belief into another one, or, in other words, belief compatibility. a) Kullback Leibler (KL) divergence; b) Bidirectional KL-divergence; c) Earth Mover's Distance; d) Jensen-Shannon 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 13). However, the Jensen-Shannon Divergence (Figure 14) shows a weaker trend and the Earth Mover's Distance (Figure 15) even with modified parameters fails to capture the qualitative pattern observed in the data.

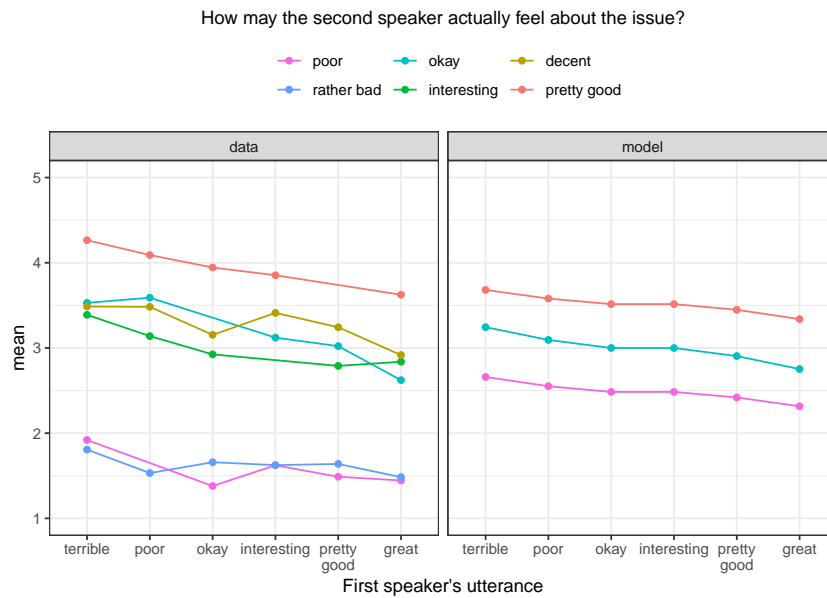


Figure 13 Opinion inference scores. The model relies on unidirectional KL as a divergence measure.

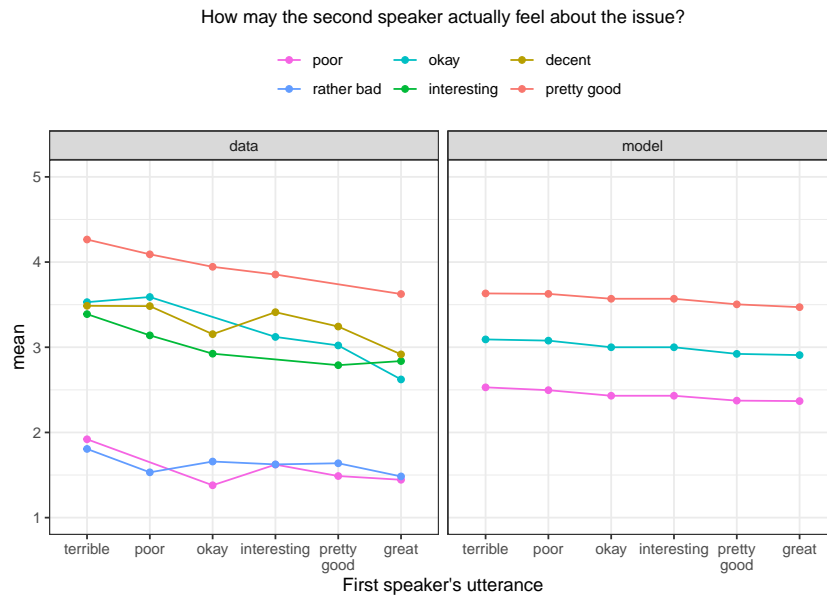


Figure 14 Opinion inference scores. The model relies on Jensen-Shannon Divergence measure.

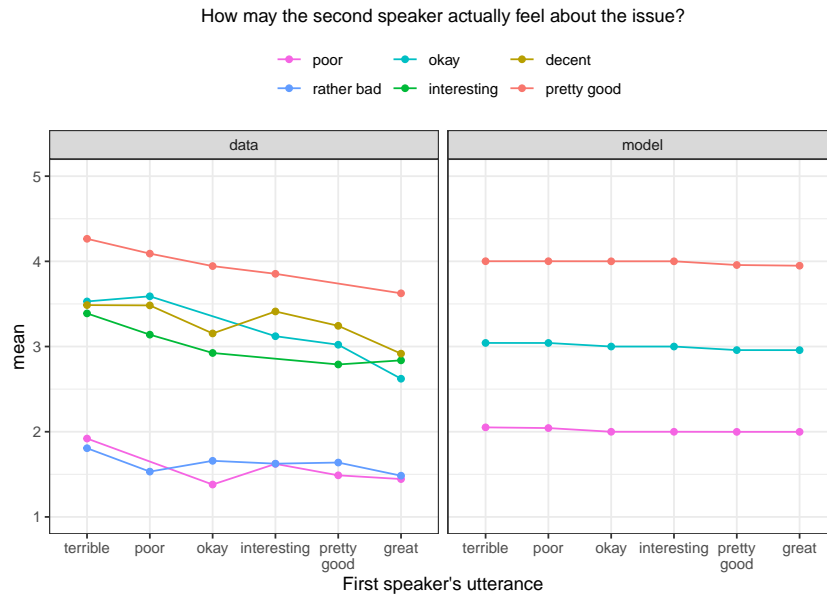


Figure 15 Opinion inference scores. The model relies on Earth Mover's Distance as a divergence measure.

A.3 Utterance utilities and probabilities

The left-hand side of Figure 16 shows utility values 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 16 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 considered listener’s opinion are preferred.

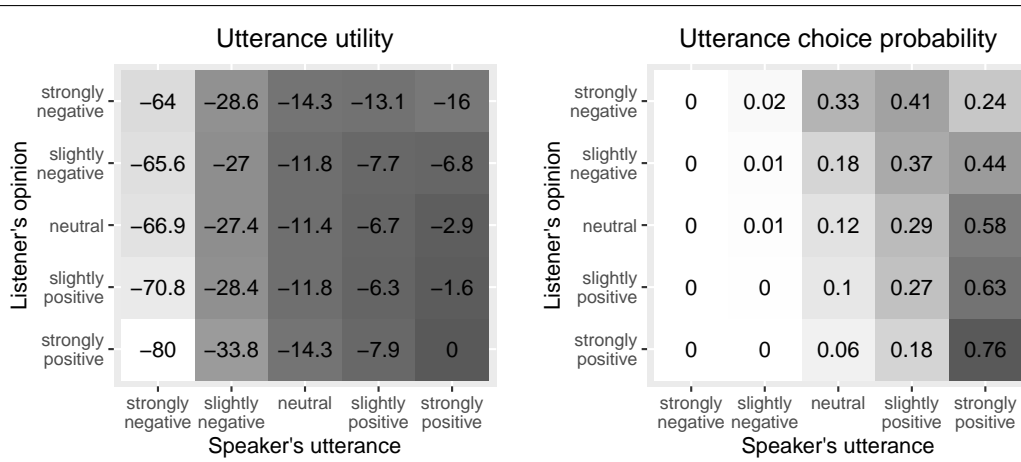


Figure 16 Utterance utility values (left side, to-be maximized), and corresponding utterance choice probabilities (right side). The calculations assume that the speaker’s opinion corresponds to a strongly positive ($\alpha = 30, \beta = 5$) opinion distribution.

B Opinion inference

B.1 Simulation

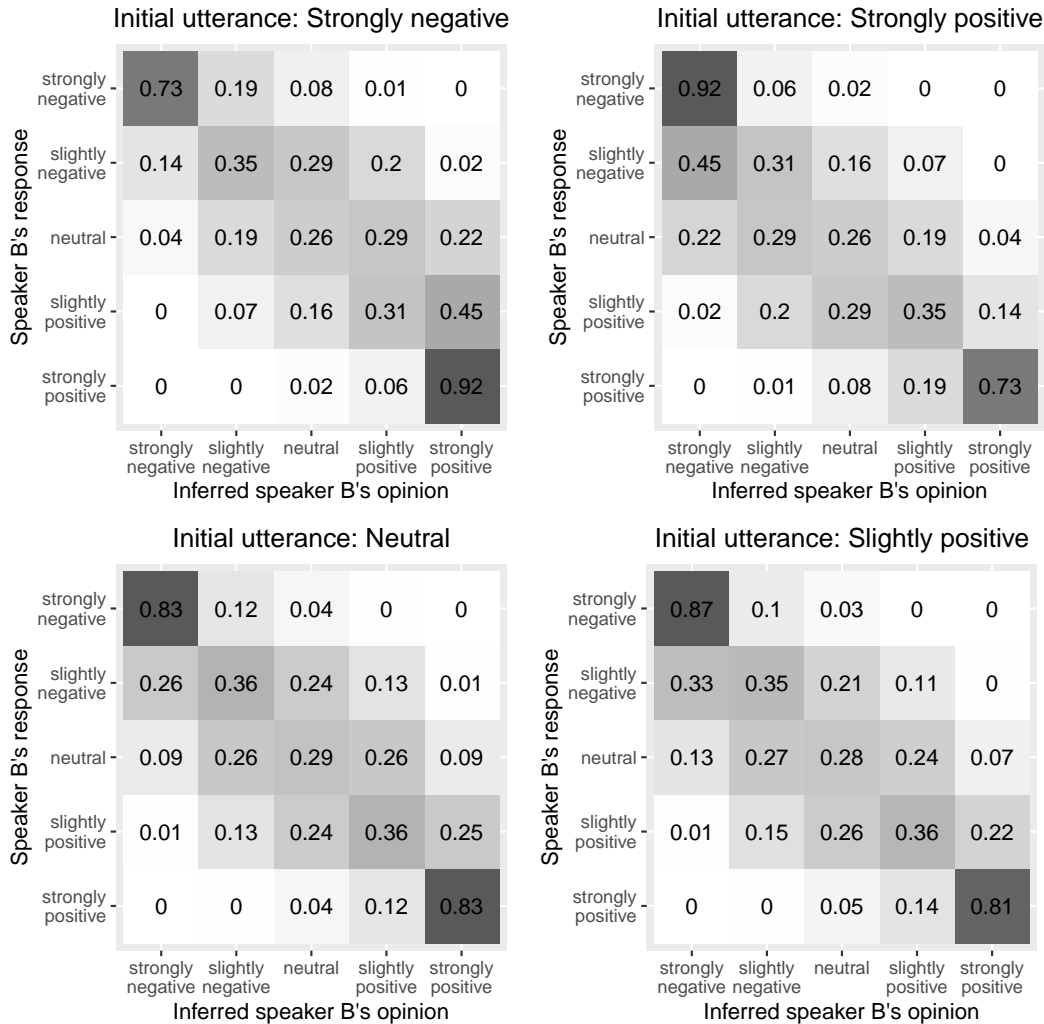


Figure 17 Model's posterior estimation of speaker B's opinion computed via Equation 8 given the initial strongly negative (top left), strongly positive (top right), neutral (bottom left), or slightly positive (top right) utterances. Each row in each matrix encodes a particular posterior probability distribution over speaker B's opinion given her response indicated in each row.

B.2 Behavioral data

Here, 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 5.8.

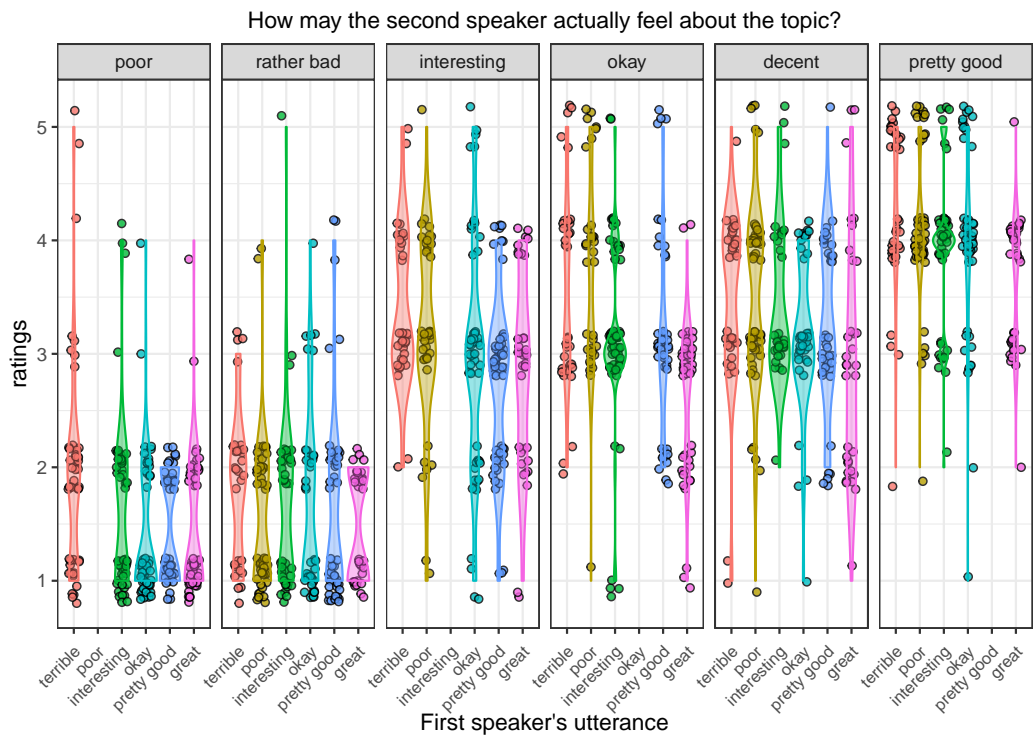


Figure 18 Utterance ratings for 10 considered adjectives