Joint inferences of speakers' beliefs and referents based on how they speak

Paula Rubio-Fernández (paula.rubio-fernandez@ifikk.uio.no)
Department of Philosophy, University of Oslo, Blindernveien 31, 0315 Oslo

Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)
Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06520-8205

Abstract

For almost two decades, the poor performance observed with the so-called Director task has been interpreted as evidence of limited use of Theory of Mind in communication. Here we propose a probabilistic model of common ground in referential communication that derives three inferences from an utterance: what the speaker is talking about in a visual context, what she knows about the context, and what referential expressions she prefers. We tested our model by comparing its inferences with those made by human participants and found that it closely mirrors their judgments, whereas an alternative model compromising the hearer's expectations of cooperativeness and efficiency reveals a worse fit to the human data. Rather than assuming that common ground is fixed in a given exchange and may or may not constrain reference resolution, we show how common ground can be inferred as part of the process of reference assignment.

Keywords: common ground; computational modeling; reference resolution; Theory of Mind

Introduction

Imagine you are on a plane and the passenger next to you is reading the news and comments: 'Trump has done it again'. You would probably interpret 'Trump' to mean Donald Trump, but what if your best friend in college also went by the name 'Trump': would you even consider that your fellow passenger could be talking about your friend?

An old debate in theoretical and experimental pragmatics addressed precisely this question: whether names (or definite descriptions, more generally) are interpreted relative to the interlocutors' mutually shared knowledge, or *common ground*. Clark and Marshall (1981) argued that indeed, considerations of common ground should constrain demonstrative reference. However, Keysar (1997) responded that a real test of this view should separate the speaker's and listener's perspectives (as in the example above), otherwise the listener may simply rely on their own private knowledge and assume common ground with the speaker.

Keysar and colleagues designed the so-called 'Director task' to test whether listeners use common ground to constraint reference interpretation. In this task, a participant follows the instructions of a confederate to move around various objects in a vertical grid of squares. The confederate sits on the other side of the grid and cannot see all of the objects, because some of the cells are occluded on her side. Crucially, the confederate is supposed to be ignorant of the contents of those cells, and when she asks the participant to 'move the small candle,' for example, the smallest of three candles is visible only to the participant. Over a long series of studies, participants have shown a tendency to consider, and sometimes even reach for, the smallest candle in their privileged view before picking up the medium-sized candle in open view (e.g., Keysar et al., 2003; Lin et al., 2010).

Keysar et al. interpreted this pattern of results as evidence of an 'egocentric bias' in communication, according to which listeners initially comprehend language egocentrically and only use common ground as a correction mechanism. This view renewed the old debate on reference and common ground when other studies using the Director task showed that listeners can use common ground information from the earliest stages of interpretation (e.g., Nadig & Sedivy, 2002; Hanna & Tanenhaus, 2004). However, the negative results observed with the Director task have also been interpreted in social cognition research as evidence that we make limited use of Theory of Mind in communication (e.g., Apperly & Butterfill, 2009; Apperly et al., 2010).

We have recently argued that the Director task is not a reliable test of Theory of Mind use in communication since optimal performance in the task (according to the usual metrics of interference) is possible by using a selective-attention strategy, without necessarily deriving any epistemic inferences about the speaker (Rubio-Fernández, 2017).

Inferring common ground

While allowing to separate the speaker's and hearer's perspectives, the Director task makes some unnatural assumptions that rarely apply in everyday communication. The first is that participants must assume that the confederate only knows about the objects that she can see in the grid and will not refer to any other object. In reality, however, speakers often refer to entities outside their visual field. Given the high selective attention demands of this paradigm, participants' fixations on the hidden objects in the grid need not be a form of egocentric behavior.

A second unnatural assumption in the Director task is how common ground is fixed at the start of the game, rather than being inferred during the exchange. A more reliable test of Theory of Mind use in communication would be to see whether participants are able to infer common ground given the Director's instructions. For example, if the confederate asked the participant for 'the blue cup' and there was a red cup in an occluded cell, would participants infer that the confederate knows about the red cup and used color contrastively? The results of Rubio-Fernández (2017) show precisely this, suggesting that when participants keep track of the contents of the occluded cells in the grid, they may still be making sophisticated epistemic inferences, rather than failing to use their Theory of Mind.

Heller et al. (2016) have recently proposed a probabilistic model of reference resolution based on the results of the Director task. Rather than assuming that participants interpret the instructions either from their own egocentric perspective, or according to their common ground with the Director, this model integrates both perspectives by giving each a probabilistic weight. Heller et al.'s model accounts for some discrepancies in the results of previous studies but assumes that common ground is determined by shared visual context, and does not allow for the possibility that (1) the speaker may be aware of objects that she cannot currently see, or (2) that the listener can infer and reconsider what the speaker knows.

In this study we present and test a probabilistic model of referential communication that assigns reference to an expression in a given visual context by jointly deriving epistemic inferences based on the speaker's choice of referential expression and adjusting their expectations about the speaker's linguistic preferences. For example, if a rational and cooperative speaker produced an under-specific description (e.g., 'the cup' when there are two cups from the listener's perspective), the listener would assume that the speaker only knows about one of the objects. Likewise, if the same speaker produced a modified description (e.g., 'the blue cup'), the listener could assume that the speaker was either preempting an ambiguity (between the two cups) or using the adjective redundantly (rather than contrastively). Our model therefore tries to account for three pragmatic phenomena given a referential expression: what the speaker is talking about in the visual context (referent), what she knows about the context (beliefs) and how she talks (efficiency).

Computational framework

Our model (http://github.com/julianje/CommonGround) consists of two components: a generative model of how speakers choose their utterances given a target referent, and a Bayesian model of how listeners infer speakers' referents and beliefs given their utterances. Our framework builds upon the strengths of reference resolution models in language (Frank & Goodman, 2012; Franke & Degen, 2013; Kehller & Rohde, 2013; Shafto, Goodman, & Griffiths, 2014; Stevens, 2017) and mental-state inference models (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Jara-Ettinger, Schulz, & Tenenbaum, *under review*). We begin by describing the generative model of a speaker, and we then explain how our model of a listener uses this speaker model to infer the speaker's beliefs and referents given their utterances.

Speaker model

In our generative model, the speaker has a set of beliefs (which, in our task, corresponds to what the speaker can see) and a goal (which, in our task, is to communicate a referent) that together determine the speaker's utterance. To generate the utterance, the speaker has an intuitive model of a simple listener, which she uses to reason which potential utterances are sufficiently informative.

The simple listener model takes a set of beliefs and an utterance and returns a uniform probability distribution over all potential referents that match the utterance. For instance, the utterance 'the triangle', combined with a belief that there is only one triangle among all the objects, returns a probability of 1 for the triangle and a probability of 0 for all other potential referents in the space of beliefs. Through this model, the speaker would determine that the utterance 'the triangle' is sufficiently informative. By contrast, if the simple listener's beliefs contained two triangles, then it would return a probability of 1/2 for each of these triangles, and a probability of 0 for all other potential referents. The speaker would therefore conclude that the utterance is not sufficiently informative. Using this model of a simple listener, the generative model of a speaker finds an utterance which is sufficiently informative to identify the intended referent (i.e. where the referent has a probability of 1 based on the simple listener model).

Intuitively, speakers can accidentally be under- or over-specific. Thus, we include a small probability that the speaker will produce an utterance that is insufficiently informative (the *Uncooperativeness* parameter[1]), and a small probability that the speaker will produce redundant modifiers (the *Redundancy* parameter). We estimate both parameters through participant judgments (see Parameter estimation study). Formally, the Uncooperativeness parameter is the probability that the speaker will believe that a proposed utterance is sufficiently informative, independently of the output from the simple listener model. Similarly, the Redundancy parameter is the probability that the speaker will consider using a modified expression without evaluating if a simpler one would have been sufficiently informative.

Listener model

Our model of participants as listeners consists of a Bayesian inference mechanism for inferring a speaker's beliefs and intended referent through the generative model of the speaker.

We treat the probability of under-specification (the Uncooperativeness parameter) as observable and constant across all speakers. That is, we assume that listeners do not question that speakers are generally cooperative, but they nonetheless understand that they can accidentally fail to specify the referent.

By contrast, we treat the probability of over-specification (the Redundancy parameter) as unobservable and variable across speakers. That is, we assume that listeners believe that different speakers may be more or less likely to use adjectives redundantly and that each speaker's individual tendency to use redundant adjectives must be inferred. Nonetheless, we assume that participants have prior beliefs about how often people speak redundantly.

[1] Naturally, speakers can be under-informative for many reasons, including distraction, accidents, and maliciousness. Here, we call the under-specification parameter the 'Uncooperativeness parameter' for simplicity, but it is intended to capture the general expectation that speakers may be under-informative, regardless of the underlying reason.

Given an utterance, our listener model performs a joint inference over the speaker's beliefs, intended referents and degree of redundancy using Bayes' rule:

$$p(b,t,r|u) \propto p(u|b,t,r)p(b,t,r) \quad (1)$$

where b is the speaker's belief, t is the target (i.e. the speaker's intended referent), r is the speaker's level of redundancy, and u is the utterance the speaker produced. The prior distribution, $p(b,t,r)$, is given by

$$p(b,t,r) = p(t|b)p(b)p(r) \quad (2)$$

where the prior beliefs about the speaker's level of redundancy ($p(r)$) and the speaker's beliefs ($p(b)$) are independent, and the probability of a target referent depends on the speaker's beliefs ($p(t|b)$), such that only objects that the speaker knows about have positive probability of being the target. In our task (see Experiment), we use a prior distribution over beliefs, a beta distribution (fit to participants' priors in the Parameter estimation task) over redundancy, and a uniform distribution over the referents, conditioned on the speaker being aware of these potential referents. Finally, the likelihood function, $p(u|b,t,r)$, is computed through the generative model described above.

Parameter estimation study

Methods

Participants 50 participants from the US (as determined by their IP addresses) were recruited using Amazon's Mechanical Turk Framework.

Stimuli 24 displays of shapes of different colors were generated. 20 of these displays consisted of a single shape (circle, rectangle, square, star and triangle) in 4 colors (blue, green, red and yellow) surrounded by a black border. The remaining 4 displays consisted of two shapes of the same type in different colors with one of these shapes (the target) surrounded by a black border (target side counterbalanced). The single shapes were used to measure over-specification (and estimate expectations about redundancy) and the double shapes to measure under-specification (and estimate expectations about cooperativeness).

Procedure Participants were told they would see a set of images with a target surrounded by a black border and that their task would be to select which of two utterances an average speaker would use to refer to it given the visual display. The two utterances were always an unmodified description of the target (e.g., 'The triangle') and a modified description of the target (e.g. 'The blue triangle'). Thus, selecting the modified description in the single-shape trials (e.g., preferring 'The blue triangle' when there is only one triangle) reveals expectations about over-specification, while selecting the unmodified description in the dual-shape trials (e.g., preferring 'The triangle' when there are two triangles) reveals expectations about under-specification.

Results

Our model's Uncooperativeness parameter (see Computational Framework) was set to the proportion of times that participants chose an under-specific description in the dual-shape trials: 5.5% of trials. By contrast, because our model infers each speaker's degree of redundancy, we used participants' choices in the single-shape trials to build a prior distribution (see prior over Redundancy parameter in Computational Framework). To do so, we fit a beta distribution to participants' choices using maximal likelihood. The resulting prior distribution was a Beta distribution with parameters α=0.39 and β=0.32.

Experiment

Methods

Participants 60 participants (mean age (SD) = 35.22 years (10.66 years), range = 18-73 years) from the US (as determined by their IP address) were recruited using Amazon's Mechanical Turk Framework.

Stimuli

Each trial included two displays of 4 geometrical shapes (circles, squares, stars and triangles) in 4 different colors (blue, green, red and yellow), each with a referential expression for the target (see Figure 1 for examples). The description of the target appeared above each display, and could be either modified (e.g., 'The blue triangle') or unmodified (e.g., 'The triangle'). The combination of shapes and instructions yielded four conditions for each individual display: Unique (single shape/ no color adjective), Contrastive (two shapes/ color adjective), Redundant (single shape/ color adjective), Ambiguous (two shapes/ no color adjective). The possible overlap between the positions of the target and the contrast shape (when present) in the two displays yielded six types of position overlap: No Overlap, Target-Target, Contrast-Contrast, Target-Contrast, Double-Same (2 Targets and 2 Contrasts), Double-Crossed (2 Target-Contrast). A total of 28 combinations were included in 2 lists of 14 trials with a balanced number of condition combinations. We only excluded 3 combinations because one did not allow any common ground inference (Ambiguous-Ambiguous/No Overlap) or rendered two impossible combinations where the target or the contrast in one display would correspond with the blind spot in the other (Ambiguous-Contrastive/Contrast-Contrast and Ambiguous-Contrastive/Double-Crossed).

Procedure

Participants played a coordination game with a virtual speaker and followed her instructions to select a shape in a display. The virtual speaker giving the instructions could only see 3 shapes in each display, whereas participants could see 4. The virtual speaker did not know that she had a blind spot, but always tried to be helpful. Each trial contained two displays and the speaker's blind spot was the same quadrant in both displays, although it varied across trials. The speaker's choice of referential expression to single out the

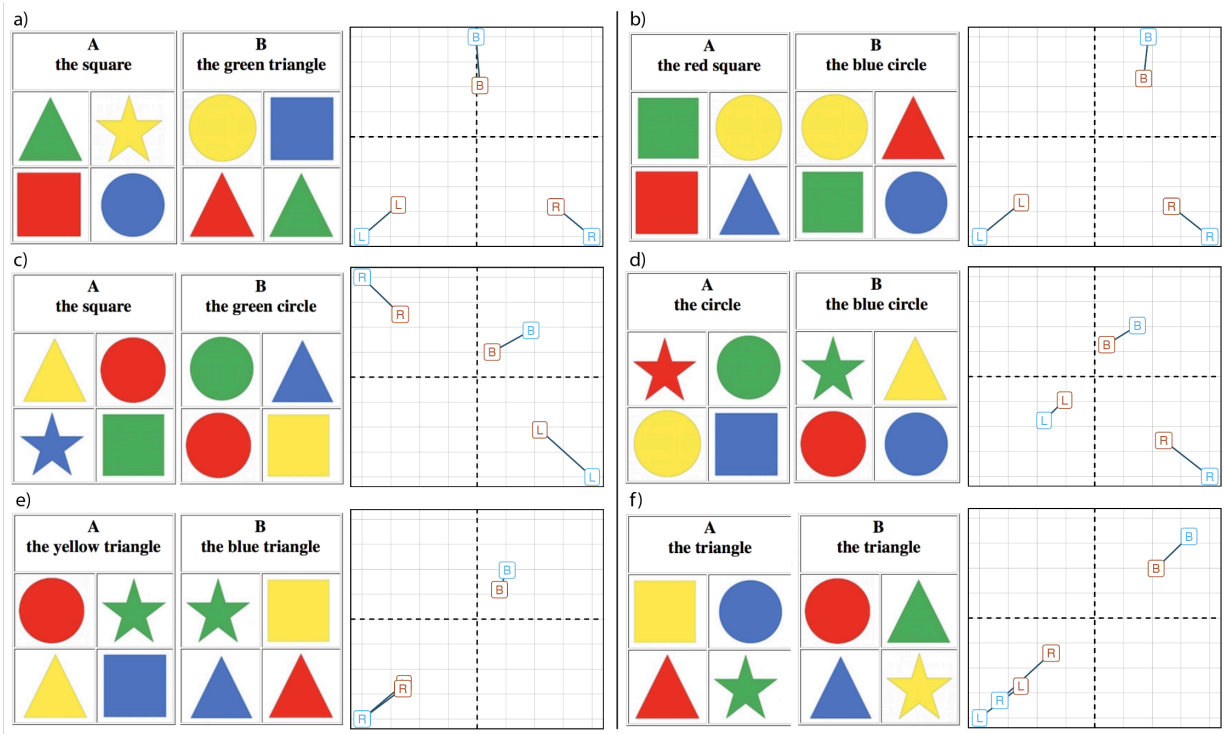

Figure 1. Six trials from the Experiment along with model predictions. Each trial consisted of two displays of four shapes and an instruction for each display. Using separate 2-dimensional trackpads, participants had to infer (1) which cell the speaker was referring to on the left-hand side display, (2) which cell the speaker was referring to on the right-hand side display, and (3) which cell was the speaker's blind spot in both displays. In each panel, the region of the right shows average participant judgments on the overlaid trackpads, along with model predictions. Speaker judgments are shown in red and model predictions are shown in blue. Each relation from a model prediction to a participant judgment is connected by a black line. L refers to the inferred referent on the left-hand side display, R refers to the inferred referent on the right-hand side display, and B refers to the inferred blind spot.

target was written above each display. See Figure 1 for examples. The pairs of displays were randomly ordered and rotated in each trial.

Participants had to answer three questions in each trial: which shape the virtual speaker was referring to in each display and which quadrant was the blind spot in both displays. Participants used three separate 2-dimensional trackpads shown on the screen to enter their responses while indicating their certainty (i.e. the closer they moved the button towards a corner, the greater their certainty that that was the referent or the blind spot; see Figure 1). Participants were given two examples of how to use the 2D trackpads and two examples of complete trials to show them how to reason about the blind spot by considering both displays.

Results

Participant judgments on the trackpad were interpreted as marginal probabilities that the referents or blind spots were on the left or right side (x value) and on the top or bottom (y value). Model predictions were transformed to points in the 2D trackpad. The top row of Figure 2 shows our model predictions (x-axis) plotted against average participant judgments (y-axis). Our model showed a correlation of 0.95 for belief inferences (95% CI: 0.92-0.97) and a correlation of 0.99 (95% CI: 0.989-0.997) for referent inferences.

Figure 1 shows the six trials and the corresponding graphs showing participant inferred referents in red (L for the referent in the left-hand side display and R for the referent in the right-hand side display) and inferred blind spot (B) along with model predictions in blue (connected by a black line).

Figure 1a (Unique-Contrastive) shows how our model and participants infer common ground based on the inferred referents. The target in the display on the left overlaps with the contrast shape in the display on the right, making the probability that the blind spot is in each of the two top cells 1/2. Figure 1b (Contrastive-Contrastive) shows how contrastive adjectives affect our model and participant inferences. Again, the speaker in Figure 1b refers to each of the two bottom cells, but because the two contrast shapes are in the top left cell, participants beliefs about the blind spot shift towards the top right cell.

Figure 1c (Unique-Contrastive) shows how our model and participants infer common ground using contrast. The two instructions unambiguously identify targets in opposite quadrants, but people and our model infer that the contrast shape in the right display is also in common ground. Figure 1d (Ambiguous-Contrastive) shows how our model and participants can combine under-specification with contrast to jointly infer common ground and resolve referential ambiguity. The left display suggests that the speaker is either referring to the bottom left cell or to the top right cell, and that she can only see one of them. Although the right display makes no direct reference to either of these cells, the contrast shape suggests that the speaker can see the bottom left cell. Having inferred common ground, participants and our model infer that the speaker was referring to the bottom left cell in

the left display and that she cannot see the top right cell. Note that our model does not show full confidence in this joint inference (because it is also possible that the speaker was uncooperative) and neither do participants.

Figure 1e (Redundant-Contrastive) shows the effects of redundancy in our model predictions. Here, because the speaker is redundant in the left display, speakers and our model do not treat the contrast on the right display as informative when inferring common ground. Finally, Figure 1f (Unique-Contrastive) shows how our model and participants inferences are sensitive to the possibility that the speaker is being uncooperative. The speaker unambiguously refers to the triangle in the left display, revealing that she can see the bottom left cell. The speaker then ambiguously refers to either of the two triangles on the right display. Under perfect rationality, the speaker must be referring to the bottom left cell in both displays and her blind spot would be the right top cell. However, our model's confidence about the inferred referent decreases in the right display because of the speaker's possible uncooperativeness, accurately predicting this fine-grained difference in participant judgments.

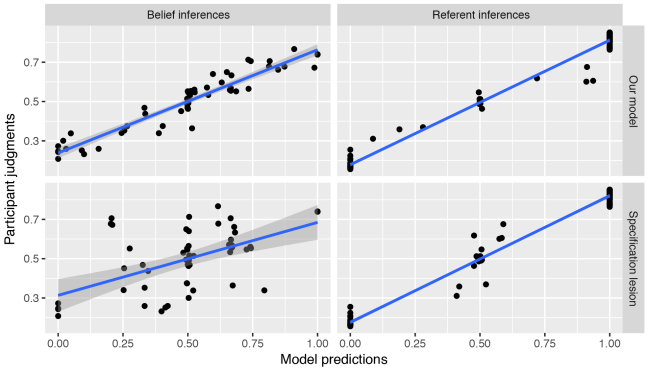

Figure 2. Model predictions against participant judgments. The top row shows our model and the bottom row shows the model after specification lesion (where the model no longer draws any inferences through the presence or absence of modification). Each point corresponds to a participant judgment. Blue lines show best linear fit.

Model lesion predictions

Having found that our model predicted participant judgments with high quantitative accuracy, we next evaluated the role of under-specification (Uncooperativeness parameter) and over-specification (Redundancy parameter) by lesioning the model. In the lesioned model we set the Uncooperativeness parameter to 0.99 (i.e. an expectation that speakers rarely recognize when they are being under-specific, making the absence of adjectives uninformative) and we set the prior distribution over Redundancy to a Beta distribution with parameters α=10 and β=1 (i.e. an expectation that speakers are often redundant, making the presence of adjectives uninformative). Thus, our lesioned model continues to expect that the speaker will correctly identify the referents, but now assumes that the use or absence of adjectives is uninformative.

The bottom row of Figure 2 shows the inferences from the lesioned model. This model showed a correlation of 0.55 (95% CI: 0.28-0.75) on belief inferences and a correlation of 0.99 (95% CI: 0.989-0.996) on referent inferences. Our main model was reliably better than the lesioned model on belief inferences (correlation difference = 0.4; 95% CI on difference: 0.22-0.67) but not on referent inferences (correlation difference = 0.0006; 95% CI on difference: -0.0039 – 0.0051).

Although the lesioned model was generally able to infer referents (largely because the target is unambiguously identifiable in all cases, except when the speaker is under-specific), Figure 2 suggests that the lesioned model was less sensitive to features of the trials relative to participants. To investigate this, we did a post-hoc analysis of trials where the lesioned model failed to identify the referents. Two of these corresponded to the trials shown in Figures 1d and 1f. Figure 3 shows the lesioned model's inferences along with participant judgments in these trials. In the displays in Figure 1d, the lesioned model incorrectly infers that the blind spot is in the top left cell and fails to make any inferences about which circle the speaker is talking about in the left-hand side display (see left display in Figure 3). This shows how loss of sensitivity to contrast impairs the model's ability to infer the referents and the blind spot. In Figure 1f, participants make stronger inferences about the speaker's blind spot and the inferred referent in the right display. Our lesioned model fails to derive these inferences because it does not rely on the under-specification to infer the blind spot and consequently uncover the referent (see right display in Figure 3).

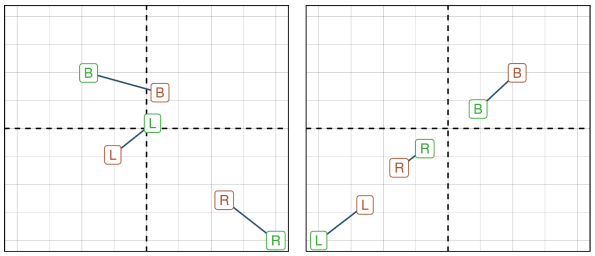

Figure 3. Model lesion against participant judgments. Predictions correspond to the trials shown in Figure 1d (left) and Figure 1f (right). Consistent with Figure 1, average participant judgments are shown in red. Model lesion predictions are shown in green.

Discussion

We presented a formal model of definite reference interpretation and common ground that captures three fundamental pragmatic inferences in referential communication: what the speaker is referring to, what she knows about the context, and what preferences she has when choosing referential expressions. Our model inferences closely mirrored participant judgments, while an alternative model compromising the hearer's expectations of cooperativeness and redundancy was less successful.

Our model shows that common ground can be computed as part of the process of reference assignment, rather than being established a priori, as assumed in the Director task (e.g., Keysar et al., 2003) and related computational models (Heller et al., 2016). Our results are consistent with work showing that participants in a modified version of the Director task can derive sophisticated epistemic inferences given a speaker's choice of referential expression (Rubio-Fernández; 2017). Critically, participants in that study derived pragmatic inferences spontaneously, suggesting that interlocutors can derive epistemic inferences in referential communication without being instructed to do so.

Although our model performs three inferences from each utterance (see Eq. 1), here we only evaluated people's inferences about speaker's intended referents and their beliefs, but we did not ask participants to explicitly infer the speaker's level of redundancy. Existing work already suggests that people can infer speaker's redundancy and adjust their inferences accordingly (Grodner & Sedivy, 2011). In future work, we will evaluate this capacity quantitatively.

Similarly, our model framework and implementation can handle an arbitrary number of useful adjectives, favoring more informative adjectives over less informative ones, and combining them when necessary. Here we focused on simple situations where the potential referents could only be disambiguated by their shape or their color. In future work, we will explore situations where speakers have several ways of drawing contrast to evaluate how listeners adjust their inferences based on their priors for redundancy (e.g., listeners tend to expect color to be used redundantly more often than size) and the efficiency of these contrasts.

Finally, our results suggest that testing people's ability to derive epistemic inferences in referential communication is a more reliable test of Theory of Mind use in communication than the standard Director task, which imposes highly unnatural demands on participants' selective attention. Although our model fits do not imply that participants were actively mentalizing when doing our task, they do show that, if people are not mentalizing, whatever mechanisms they use to circumvent mentalistic reasoning must be sufficiently complex to accurately approximate Theory of Mind inferences.

Acknowledgments

This research was supported by a *Young Research Talent Grant* from the Research Council of Norway (Ref. 230718) awarded to PRF and a *Google Faculty Research Award* to JJE.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., & Moffitt, G. (2010). Why are there limits on Theory of Mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology*, *63*(6), 1201-1217.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 0064.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, and I. Sag (Eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

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

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual- vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter and E. Gibson (Eds.), *The processing and acquisition of reference*. MIT Press: Cambridge, MA

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105-115.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104-120.

Jara-Ettinger, J., Schulz, E., & Tenenbaum, J.B., (*under review*). The naïve utility calculus as a foundation for action understanding.

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1-2), 1-37.

Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, *24*(2-3), 253-270.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on Theory of Mind use in adults. *Cognition*, *89*(1), 25-41.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using Theory of Mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551-556.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329-336.

Rubio-Fernández, P. (2017). The Director task: A test of Theory-of-Mind use or selective attention? *Psychonomic Bulletin & Review*, *24*(4), 1121-1128.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55-89.

Stevens, J. (2017). Biases and labeling in iterative pragmatic reasoning. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.