Conversational expectations account for apparent limits on theory of mind use

Robert X. D. Hawkins, Noah D. Goodman
{rxdh,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Theory of mind is a powerful cognitive ability: by the age of six, people are capable of accurately reasoning about others' beliefs and desires. An influential series of language understanding experiments by Keysar and colleagues, however, showed that adults systematically failed to take a speaker's beliefs into account, revealing limitations on theory of mind. In this paper we argue that these apparent failures are in fact successes. Through a minimal pair of replications comparing scripted vs. unscripted speakers, we show that critical utterances used by Keysar and colleagues are uncooperative: they are less informative than what a speaker would actually produce in that situation. When we allow participants to naturally interact, we find that listener expectations are justified and errors are reduced. This ironically shows that apparent failures of theory of mind are in fact attributable to sophisticated expectations about speaker behavior—that is, to theory of mind.

Keywords: Theory of mind; social cognition; pragmatics

Introduction

Humans can accurately and intelligently reason about the mental states of other humans. Among other things, this ability – called *theory of mind* (Premack & Woodruff, 1978) – allows us to infer the underlying beliefs and intentions that motivate others' actions, and to use these inferences to predict future actions (Baker, Saxe, & Tenenbaum, 2009). Children acquire this ability by at least age six (Wellman, Cross, & Watson, 2001) and it serves as an important landmark in the developmental trajectory of intuitive theory use (Gopnik & Wellman, 2012). While theory of mind use often appears to be automatic and effortless, Keysar and colleagues (Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; Lin, Keysar, & Epley, 2010) have argued that it is actually the opposite, even for adults: we are "mindblind" by default and only overcome our egocentric biases through an effortful process of perspective-taking. In other words, while adults are *capable* of applying theory of mind reasoning, we do not always apply it reliably.

In this paper we argue that the apparent failures used to support this view are in fact successes for sophisticated social reasoning. In particular, we argue that critical utterances used by Keysar and colleagues are *uncooperative*: they are less informative than what a speaker would actually produce in that situation; listeners who are sensitive to the pragmatics of the situation expect these more informative utterances and produce "errors" when their expectations are flouted.

The argument offered by Keysar and colleagues is based on an elegant experimental paradigm, where participants played a simple communication game with a confederate. The two players were placed on opposite sides of a 4×4 grid containing a set of everyday objects (see Fig. 1). The confederate played the role of 'director,' giving instructions about how to move objects around a grid, and the participant played the role of 'matcher,' attempting to follow these instructions. For example, the objects in one trial included a cassette tape. The director gave an instruction like 'move the tape up one square,' referring to the cassette. Critically, some objects were occluded such that only the matcher could see them, creating an asymmetry in the players' knowledge. To perform accurately on critical trials, the matcher would need to apply theory of mind to reason about which objects were shared and which were private. For example, imagine a roll of tape were placed in an occluded slot: if a participant failed to account for the director's (partial) knowledge, she might interpret 'tape' to mean the occluded roll of tape (which the director couldn't possibly know about). Indeed, Keysar et al. (2003) found that participants attempted to move the hidden item in 30% of cases: 71% of participants attempted to move this hidden item at least once (out of four critical cases) in the experimental condition, compared to 0% in a control condition where there was no ambiguity over the referent. Additionally, eye-tracking data showed that participants considered the hidden item more often and for longer in the experimental condition than the control condition.

While these results are compelling, the paradigm has been criticized from few different angles. Heller, Grodner, and Tanenhaus (2008) have pointed out that in many cases, the hidden object was a better fit for the referring expression than the one in common ground (e.g. the hidden roll of tape vs. the cassette tape for "the tape"), making the hidden object *a priori* more likely to be the referent; it would generally be fairer to compare two objects that fit the referring expression equally well. We validate this argument by empirically measuring relative fit of the expressions to the target and distractor items. Moreover, Hanna, Tanenhaus, and Trueswell (2003) argued that the viewpoint asymmetry paradigm is somewhat unnatural: common ground is typically built incrementally over the course of an interaction rather than presented all at once, and it is rare for a shared display to differ in perceptual accessibility.

In this paper, we offer an additional factor that helps account for Keysar's results. Theory of mind as applied within language understanding depends on an accurate model of what a speaker would say in different situations. Given an utterance, a listener can then reason backward to the most plausible situation (Grice, 1975; Clark, 1996; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). This suggests that we consider whether the utterances produced by the confederate in Keysar's critical conditions were actually what a speaker in that context would be expected to say. If not, then perhaps

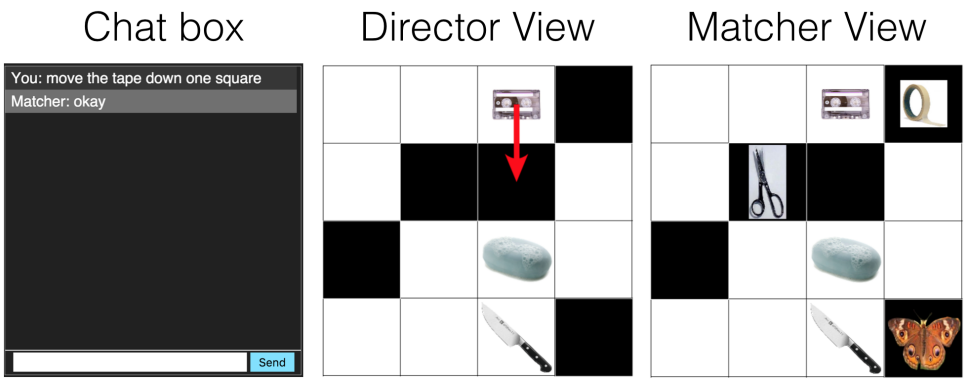

Figure 1: Interface used in the reported experiments. Objects behind the black squares were hidden from the director.

listeners are making choices that are in fact consistent with a correct pragmatic interpretation of the confederate's (uncooperative) utterance. More precisely, when both players know that objects are occluded in the display, the speaker may tend to add additional precision to references in order to avoid confusion. If the listener *expects* the speaker to do this, they will pragmatically pick the *a priori* more likely referent of the referring expression, which in critical trials will be the occluded object. In other words, it is precisely *because* the listener takes the speaker's mental state into consideration that they are tricked by an uncooperative confederate into choosing the wrong item.

We began by replicating Keysar et al. (2003) in a multiplayer web experiment. We recruited participants to be both director and matcher (instead of using a confederate), but instructions for critical items, as well as a random subset of filler items, remained scripted as in the original study. We replicated the original finding, but noted a tendency of directors to be overinformative in unscripted filler trials. We then ran the same experiment without using any scripted instructions, observing unconstrained director utterances. We found much greater precision in unconstrained director utterances, which match targets much better than distractors, and better performance of the matchers. This minimal pair of experiments demonstrates that listener mistakes are at least partially due to the pragmatics of the task, ironically showing that apparent failures of theory of mind are in fact attributable to sophisticated expectations about speaker behavior—that is, to theory of mind.

Expt. 1: Scripted Replication

Participants

We recruited 34 participants (17 pairs) from Amazon Mechanical Turk. All participants were from the U.S. Three pairs were excluded for making 2 or more errors on non-critical items.

Materials & Procedures

Participants interacted in a real-time, multi-player environment on the web (Hawkins, 2015). Pairs of participants—assigned randomly to 'director' and 'matcher' roles—interacted with one another through a web interface, shown in Fig. 1. On the left side of the screen, participants could freely type messages to one another; on the right side the screen, players could view a set of objects placed in a 4 x 4 grid. Five of the grid cells were occluded from only the director's perspective, and the remaining 11 were visible to both matcher and director. Six or seven objects were displayed in the grid at a given time. One of these objects was a 'target', such as a cassette tape, placed in an unoccluded cell such that both participants could see it. Another object, such as a roll of tape, was placed in an occluded slot such that it was only visible to the matcher. The rest of the objects were unrelated 'fillers' placed in random locations. We used the same set of targets and occluded alternatives as Keysar et al. (2003), but we were unable to obtain the filler objects from the original experiment and created our own. Before entering the game environment, every participant independently passed a short quiz about the task's instructions, ensuring that they understood the interface. Among other items on the quiz, we verified that both participants understood that items behind black cells were only visible to the matcher.

The experiment was composed of eight items, with each item using a different set of objects. Each item included one 'critical pair' of objects, one of them the target and the other hidden, such as the cassette tape and the roll of tape. For each item, we gave the director a series of four instructions to move objects around, which were displayed as a series of arrows pointing from some object to an unoccupied cell. To collect clean mouse-tracking data, we began every instruction by asking the matcher to click a small circle in the center of the grid. After this small circle was clicked, the director was allowed to communicate the next instruction and we started recording from the matcher's mouse. One of the instructions was a 'critical instruction,' which referred to the target object. For half the instructions, directors were free to communicate however they wished. For the other half, including all the critical instructions, their messages to the matcher were prescripted using the precise wording from Keysar et al. (2003). For example, when giving instructions on how to move the cassette tape, the director would be forced to use the ambiguous utterance "Move the tape down one square." (That is, in

	% attempted at least once			% attempted at least twice			% of total cases		
	Orig.	Expt. 1	Expt. 2	Orig.	Expt. 1	Expt. 2	Orig.	Expt. 1	Expt. 2
Experimental	71	93	61	46	57	32	30	43	24
Baseline	0	7	0	0	0	0	0	2	0

Table 1: Side-by-side comparison of error rates in Keysar et al. (2003) and our two replications. The first two sections show the percentage of *participants* attempting to move the occluded distractor at least once, or twice, of the four possible cases. The third section shows the percentage of *all experimental trials* that the participant actually tried to move the occluded object.

these conditions, the scripted message would automatically appear in the director's chat window, and they would have to click 'send' for the experiment to continue.)

We collected baseline performance for each condition by replacing the hidden alternative (e.g. a roll of tape) with an object that did not fit the critical instruction (e.g. a battery); we used the same unambiguous replacements as Keysar et al. (2003). Each participant received half the items in the experimental condition and half in the baseline condition. The assignment of items to conditions was randomized across participants, and the order of conditions was randomized under the constraint that the same condition would not be used on more than two consecutive items. All object sets, object placements, and corresponding instruction sets were the same for all participants.

This paradigm differs from those used by Keysar et al. (2003) in three primary ways. First, participants were not seated across from each other at a table: they each saw a view of the 4 x 4 grid on their screen and communicated via a text box. Second, we did not use a trained confederate. We randomly assigned one of the players to the role of the instruction giver, and maintained the original wording by scripting a subset of their instructions. Finally, the hidden object was not placed in a bag, in which respect our design more closely resembles Keysar et al. (2000).

Results and discussion

In Table 1 we show the error rates on critical items, and compare to the data from Keysar et al. (2003). We find that 93% of participants (all but one) attempted to move the hidden distractor at least once in the Experimental condition, out of four possible items, compared to only 7% (only one) in the baseline condition. This is similar to the effect observed by the authors in the original study, which found 71% and 0%. Our errors were larger across the board, perhaps due to the interface or the population, but the gap between the two conditions is roughly the same size. In Table 2, we break down the pattern of errors by item. We note that several items have much higher error rates than others – for example, 75% of participants in the experimental condition of item 6 made an error (the "whiteboard eraser" vs. the "pencil eraser") while only 17% of participants in item 8 made an error (the "computer mouse" vs. the "toy mouse"). Informally, it seems as though the more difficult items are the ones where the utterance fits the distractor better than the target. We empirically substantiate this observation in our results for Expt. 2 below.

This item-wise variability suggests that the dependent variable highlighted in the original study (i.e. "percentage of participants who moved the critical item at least once") is somewhat problematic: it could look like 100% of participants made errors even if they all made those errors on one particularly difficult item. Indeed, if we exclude the three 'hard' items where over 60% of participants in the experimental condition made errors, this dependent variable drops from 93% to only 43% of participants.

As a proxy for the eye-tracking analyses reported by Keysar et al. (2003), we conducted a mouse-tracking analysis. We define the decision window as the span of time between the point when the matcher received their instruction message and when they started moving an object. If it took them multiple attempts to move the correct object, we restricted our analysis to the first attempt. Within the decision window, we computed the total amount of time spent hovering over the cell containing the target and divided by the total length of the decision window to get a measure of the relative time spent considering the target. We had to exclude an additional 3 participants for this analysis, because the timestamps for director and matcher did not align and we could not establish the decision window properly. A paired-samples t-test found that people tended to spend less time hovering over the target cell on critical experimental trials than on baseline trials, $t(10) = -2.65, p = 0.02$ (see Fig. 2), indicating that the presence of a hidden distractor interfered with participants ability to directly choose the target.[1] .

By running this replication as a multi-player web experiment we have available an additional source of data beyond the original experiments: half of the instructions were unscripted, providing observations of natural production of referential descriptions for filler items. Informally, we noted a tendency toward additional, possibly unnecessary, precision in descriptions. Instead of "move the stuffed animal down", participants said "move the stuffed panda bear down." Or, instead of saying "move the plane to the right" when there is only one plane, participants said "move the red airplane to the right." Perhaps directors were taking the time to make more precise descriptions because they believed it was contextually relevant: both parties know that there are hidden objects in the environment increasing the chance of miscommunication from imprecise descriptions. If the matcher expected the di-

[1] This analysis includes participants who actually made errors, since the data is too sparse to exclude them.

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
	instruction	"glasses"	"bottom block"	"tape"	"large measuring cup"	"brush"	"eraser"	"small candle"	"mouse"
	target	sunglasses	block (3rd row)	cassette	medium cup	round hairbrush	board eraser	medium candle	computer mouse
	hidden distractor	glasses case	block (4th row)	scotch-tape	large cup	flat hairbrush	pencil eraser	small candle	toy mouse
Expt. 1	# incorrect	0	5	1	3	2	6	6	1
	# correct	6	3	1	8	2	2	4	6
Expt. 2	# incorrect	0	1	1	3	9	7	1	5
	# correct	12	14	11	13	7	8	12	8

Table 2: Item-wise error rates for critical trials

rector to be precise, then they would be justified in picking the first or best object that meets the description (rather than worrying excessively about occluded cells). That is, the scripted instructions used by the director for critical trials may have been *uncooperative* for this situation, and thus led matchers astray. We tested this prediction in Expt. 2, where we removed the scripted instructions and allowed speakers to refer to items however they wished. By the reasoning above we expected to see more precise descriptions by unscripted directors and fewer errors by matchers in the critical trials.

Expt. 2: Unscripted Replication

Participants

We recruited 64 participants (32 pairs) from Amazon Mechanical Turk, roughly doubling the sample size from Expt. 1. All participants were from the U.S. Three participants were excluded for making 2 or more errors on non-critical items, and one additional participant was excluded because they were not a native English speaker.

Materials & Procedures

Everything was the same as Expt. 1, except we did not use scripted messages for critical instructions.

Results

Error rates are reported in Table 1, alongside the results from Keysar et al. (2003) and our scripted replication in Expt. 1.

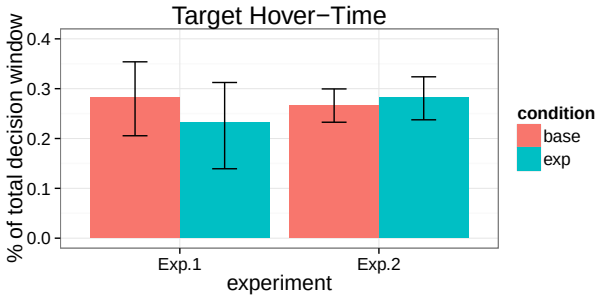

Figure 2: Mean percentage of time spent hovering over target cell in the two conditions for Expt. 1 (left) and Expt. 2 (right). Error bars are boostrapped 95% confidence intervals.

Participants never moved the hidden object in the baseline condition, and total error rates for experimental trials are significantly lower than the rates found in Expt. 1, $\chi^2(1) = 5.35, p = 0.02$.

We find that patterns of errors in Expt. 2 diverge significantly from a uniform distribution across items, $\chi^2(7) = 24.8, p < 0.001$. Looking more closely at these patterns, we see that the only items where errors are consistently made are those where the more precise utterances used remain ambiguous. For example, in item 5, both the target and distractor are hair brushes: one is round and one is flat. Many participants produced the sub-class label "hair brush," which was more precise than the scripted basic-level "brush" from Expt. 1, but the two objects were still confusable at the sub-class level. If we remove the two most difficult items (the "hair brushes" and the "erasers"), the total percentage of errors on experimental trials drops from 24% to 10% and the percentage of participants making at least one error drops from 61% to 32%.

When we conducted a mouse-tracking analysis identical to the one reported for Expt. 1, we found no significant decrease in target hover time between experimental and baseline trials $t(27) = 0.89, p = 0.38$. To directly test for differences in hover time patterns across the two experiments, we used a mixed-effects model with a random intercept for game ID and an interaction between condition and experiment on target hover time. We found a marginally significant interaction, $b = 0.09, t = 1.89, p = 0.066$ (see Figure 2), providing some evidence that the presence of a hidden distractor no longer interfered target selection in Expt 2.

Next, we test whether these improvements in performance are in fact due to more informative speaker behavior. We recruited twenty judges on Amazon Mechanical Turk, who provided ratings for how well the 71 unique labels used by speakers across both experiments (including scripted labels) fit the target and hidden distractor objects. Their responses were given on a slider with endpoints labeled "not at all" and "perfectly." Inter-rater reliability was relatively high, with intra-class correlation coefficient of 0.6 (95% CI = $[0.54, 0.66]$). In a mixed model including random intercepts for raters and items, we found a significant crossover interaction of ex-

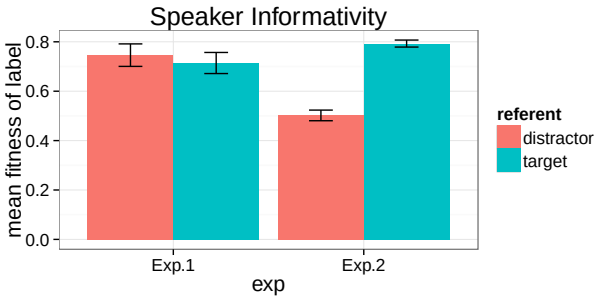

Figure 3: Mean fitness ratings provided by judges on Amazon Mechanical Turk. Error bars are bootstrapped 95% confidence intervals.

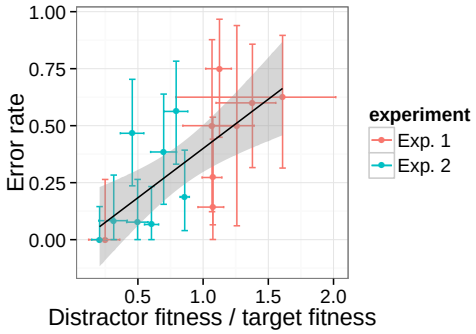

Figure 4: Item-wise error rates compared with label fitness ratios, across both experiments. Error bars are bootstrapped 95% confidence intervals for fitness and 95% highest posterior density intervals for error rates.

periment and referent on mean fitness rating (see Fig. 3), $b = 0.32, t = 9, p < 0.001$. In Expt. 1, the scripted label fit the hidden distractor just as well or better than the target, but in Expt. 2, the unconstrained labels fit the target much better and the hidden distractor much worse. In other words, the scripted labels used in Keysar's studies were less informative than speakers normally produce in this scenario.

Does differential informativity of scripted utterances account for the variability across items that we noted in Expt. 1? In Figure 4, we compare the item-wise error % and ratio of target fit to distractor fit. We find that across both experiments, participants have significantly higher error % on items where the speaker's label fits the distractor better than the target, $b = 0.43, t(14) = 3.97, p = 0.001$, capturing a significant portion of variance, $R^2 = .5, F(1,14) = 0.001$. This suggests that item-wise variability across the two experiments is primarily driven by relative informativity of speaker utterances.

After separately establishing that matchers in Expt. 2 make fewer mistakes, that directors in Expt. 2 produce more informative utterances, and that informativity captures item-wise variability in error rates in both experiments, we tested the link between these effects in our aggregated data: does label informativity generally predict errors on critical trials? We used a mixed effects logistic regression model to estimate the effect of target and distractor fit on the probability of making a critical error in both experiments, including a random intercept for game ID. We found that participants are less likely to make errors when the target fit is higher, $b = -0.8, z = -4.1, p < 0.001$ and more likely to make errors when the distractor fit is higher, $b = 1.7, z = 3.7, p < 0.001$. Furthermore, a model including target fit and distractor fit in addition to item-level fixed effects is significantly better than a model including item alone, $\chi^2(2) = 36.2, p < 0.001$, implying that speaker informativity captures residual variance beyond the item-wise effects reported above.

General Discussion

Pragmatic language understanding requires sophisticated social reasoning. To interpret an utterance, a listener must consider how a speaker is likely to behave in context. The rational use of theory of mind for a listener thus depends on her expectations about the speaker: If, in a particular context, she expects a speaker to provide sufficiently informative utterances, she may be justified in neglecting his epistemic state when resolving reference.

In our replication (Expt. 1), we found evidence that listeners neglect the speaker's epistemic state, as Keysar and colleagues claim. We also found that when the speaker's utterance fit the hidden distractor better than the target, matchers were more likely to make errors. Indeed, we found that the extremely heterogeneous pattern of errors across items was well explained by the relative fit of the utterance to target and distractor. This suggests that reference disambiguation was driven primarily by *a priori* label fitness rather than consideration of occluded vs. mutual visible items. However, this does not necessarily imply limitations on theory of mind—if speakers could be expected to naturally provide expressions which apply better to the target, then this behavior would be appropriate. In Expt. 2, we found that speakers did naturally produce more precise, informative utterances than required and these unconstrained utterances fit the target significantly better than they fit the hidden distractor. For example, no speaker in Expt. 2 produced "the bottom block," which was used as a critical instruction in Expt. 1. Instead, they said "the bluish block with a B" or "block with the blue writing," which relied on less confusable perceptual features and thereby decreased error rates. Thus, the errors observed in Expt. 1 can be explained as a result of an uncooperative confederate (speaker)—the experiment set up certain pragmatic expectations through the task context, then deliberately flouted them in critical trials.

The Gricean maxim of quality dictates that cooperative speakers should make their utterance as informative as is required for current purposes, but no more so. If the referring expressions used in Expt. 1 uniquely pick out the target from the speaker's perspective, then why would unscripted speakers in Expt. 2 be more informative than this? One possibility is that awareness of the complex mismatch in epistemic state with the listener leads the speaker to provide extra information in an attempt to avoid miscommunication. Another possibility is that more general dynamics of language are in play. Studies of over-informativity in referring expressions

have uncovered a general tendency of speakers to provide redundant information. This tendency can depend on a number of situational factors, such as the tendency to mention perceptually salient features to speed up the identification process (Koolen, Gatt, Goudbeek, & Krahmer, 2011). The exact origin is a subject that must be followed up by future research (e.g. Gann & Barr, 2014), but it is clear that the tendency of speakers to produce highly informative referring expressions is useful to, and relied on by, listeners.

This finding is consistent with a recent proposal by Heller, Parisien, and Stevenson (2016) that referring expressions are interpreted by probabilistically integrating multiple sources of information: when conversational expectations lead participants to expect over-informative utterances, they (rationally) place relatively less weight on features of the environment determining common ground, such as shared perceptual access. Further work in a wider variety of tasks is necessary to pin down the various factors determining the relative weighting of these different sources of information, but we have argued that pragmatics play a crucial role.

It is worth noting several significant differences between our study and Keysar et al. (2003). The primary difference between our study and the original, of course, is that it was run on the web with participants connected through a virtual environment, instead of face-to-face in a room. We believe we addressed the major concern about exploring theory of mind in web experiments – that participants do not truly believe they are interacting with another human – by allowing instantaneous, responsive, real-time interaction. On the other hand, it is known that textual communication, as in our chat box, can differ from face-to-face verbal communication. Additionally, aspects of the interface such as the graphical representation of occluded cells may be less intuitive, or require more training, on the web than in the lab.

A related difference is our decision not to use a confederate. While confederates are useful for reducing variation across instances of the experiment and delivering carefully targeted manipulations, their use may have unexpected consequences. Beyond the difficulties of conducting large-scale experiments with confederates, it is difficult to exactly replicate all the subtleties of the confederate's behavior that might influence results. The pair of experiments we report is a reminder that manipulations administered by a trained confederate can interact in unexpected ways with a participant's social and communicative expectations. Regardless of the experimental context, it is illuminating to see how real participants naturally interact.

Acknowledgements

We're grateful to Boaz Keysar for providing select materials for our replication. Expt. 1 was originally conducted under the supervision of Michael Frank, with early input from Desmond Ong. This work was supported by ONR grants N00014-13-1-0788 and N00014-13-1-0287, and a James S. McDonnell Foundation Scholar Award to NDG. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

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

Clark, H. H. (1996). *Using language*. Cambridge university press Cambridge.

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

Gann, T. M., & Barr, D. J. (2014). Speaking from experience: Audience design as expert performance. *Language, Cognition and Neuroscience*, *29*(6), 744–760.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085–1108.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 43–58). New York: Academic Press.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

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

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

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

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

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.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(04), 515–526.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 655–684.