HARVARD | BUSINESS | SCHOOL

Topic Preference Detection: A Novel Approach to Understand Perspective Taking in Conversation

Michael Yeomans
Alison Wood Brooks

Working Paper 20-077

Topic Preference Detection: A Novel Approach to Understand Perspective Taking in Conversation

Michael Yeomans
Imperial College London

Alison Wood Brooks
Harvard Business School

Working Paper 20-077

Funding for this research was provided in part by Harvard Business School.

RUNNING HEAD: Topic Preference Detection

Topic preference detection:

A novel approach to understand perspective taking in conversation

Michael Yeomans, Imperial College London

Alison Wood Brooks, Harvard Business School

Abstract

Although most humans engage in conversations constantly throughout their lives, conversational mistakes are commonplace— interacting with others is difficult, and conversation requires quick, relentless perspective-taking and decision making. For example: during every turn of every conversation, people must decide: Should we stay on this topic or switch to another one? In this paper, we introduce topic preference detection—accurately predicting a conversation partner’s topic preferences (i.e., what they want to talk about)—as a novel approach to evaluate perspective taking. Across synchronous and asynchronous conversations between people in close relationships and among strangers, we find that people want to accommodate their partner's topic preferences in conversation. However, they routinely fall short in learning what other people prefer, compared to benchmark models that rely on natural language processing algorithms to detect topic preferences from people's conversational behavior. Even after a ten-minute face-to-face conversation, people were still missing important cues to their partner's topic interest. Compared to earlier perspective-taking paradigms, our work suggests that topic preference detection captures many naturalistic elements of perspective taking, and demonstrates that conversation can fail to produce mutual understanding, even when people want to accommodate each other’s preferences.

Keywords: Conversation; Perspective-Taking; Natural Language Processing; Decision-Making

Disclosure: The authors have no conflicts of interest to disclose. All of our exact data, code, preregistrations, and experimental stimuli are available in an anonymized repository on the Open Science Framework at https://goo.gl/yXijxr.

Introduction

Conversation—turn-by-turn natural language communication—is one of the most common and important tasks humans do together. People reveal their preference to converse over and over again in life (i.e., by talking to others frequently), and conversations can be a deep source of happiness and enjoyment (Diener & Seligman, 2002; Dunbar, Marriott & Duncan, 1997; Mehl, Vazire, Holleran, & Clark, 2010; Epley & Schroeder, 2014; Kumar & Gilovich, 2015). The choices people make outside of their conversations influence, and are influenced by, what happens during their conversations (Crawford & Sobel, 1982; Lerner & Tetlock, 1999; Berger, 2014).

People pursue a wide array of diverse goals in conversation, often simultaneously, such as impression management, enjoyment, conflict management, persuasion, information exchange, social connection, divergent and convergent thinking, helping others achieve their goals, and on and on. Though conversations, and the motives underlying them, are a dynamic and uncertain decision-making environment, one of the primary functions of conversation is to develop and maintain a shared understanding with others (Hardin & Higgins, 1996; Pickering & Garrod, 2004; Schegloff, 2007). As people take turns speaking back and forth, they must constantly monitor what other people have just said and what they intended, and decide what to say next. While people are generally able to understand the meaning of what they say to each other, we suspect they have a poorer understanding of the motives and preferences that lie beneath each other’s words. Instead, cognitive limitations, such as low perspective-taking ability, are likely to affect how well people learn about one another during conversation (Epley, 2008; Eyal, Steffel, & Epley, 2018). Specifically, we predict that one reason people fail to achieve their conversational goals is because they misunderstand each other’s goals and preferences during the conversation itself.

The current research focuses on one of the most frequent (and fraught) decisions people make in conversation: topic selection. Every time someone takes a turn speaking in conversation, s/he decides whether to stay on the current topic or switch to a different one. The sequence of all the topic choices people make over many turns is integral to the structure of a conversation. Prior work suggests that people often manage topic boundaries explicitly, using verbal, nonverbal, and prosodic (i.e., paralinguistic) cues to transition away from the current topic and to start a new one (Hearst, 1997; Drew & Holt, 1998; Galley, McKeown, Fowler-Lussier & Jing, 2003). While research has shown that people manage the flow of topics (Passonneau & Litman, 1997; Nguyen et al., 2014), we know little about how they decide, in the moment, whether to stay on topic or switch to a new topic—and which topic to try next. Here, we explore topic selection as a perspective-taking task. We aim to understand how people infer and weigh their partners’ topic preferences (if at all) in their topic choices, and consider how people might manage topic search, selection, and shift more effectively.

Perspective Taking

There is a rich literature documenting the profound limits of people’s ability to understand the minds of others (Epley, 2008). To navigate the social world successfully, people must actively reason about what others see, know, believe, and desire. This capacity to consider others’ mental states, commonly referred to as “theory of mind,” is essential for communication and social coordination. Without direct access into others’ minds, however, people frequently use intuitive strategies to guide their inferences about others’ mental states. One such strategy entails consulting the contents of one’s own mind (Goldman, 2006).

Although one's egocentric perspective can be a good proxy for making social predictions (Dawes, 1989; Hoch, 1987), people often rely too heavily on accessible self-knowledge during mental-state reasoning (e.g., Birch & Bloom, 2007; Keysar, Lin, & Barr, 2003; Sommerville, Bernstein, & Meltzoff, 2013). By failing to adjust for ways in which others' perspectives might differ from their own (Epley, Keysar, Van Boven, & Gilovich, 2004; Tamir & Mitchell, 2013), they set the stage for potential misunderstanding and conflict (Ross & Ward, 1995). In fact, failures of perspective taking and egocentric perceptions are two of the main psychological barriers to resolving conflict (e.g., Friend & Malhotra, 2019; Bruneau et al., 2017; Greenberg, 1983; Nasie et al., 2014; Scharer et al., 2018, Bruneau et al., 2015; Hameiri & Nadler, 2017; Hopthrow et al., 2017).

Many factors can affect the extent of egocentrism during mental-state reasoning, including characteristics of both targets and perceivers. For example, egocentric projection tends to be greater with close others (e.g., friends and romantic partners) and those perceived as similar to oneself (e.g., ingroup members) than with strangers (Krienen, Tu, & Buckner, 2010; Savitsky, Keysar, Epley, Carter, & Swanson, 2011) or dissimilar others (Ames, 2004; Todd, Hanko, Galinsky, & Mussweiler, 2011). People also tend to be more egocentric when they are distracted by a concurrent task (Lin, Keysar, & Epley, 2010; Schneider, Lam, Bayliss, & Dux, 2012), under pressure to respond quickly (Epley et al., 2004), are members of individualistic cultures (Wu, Barr, Gann, & Keysar, 2013; Wu & Keysar, 2007), or occupy high-power roles (Galinsky, Magee, Inesi, & Gruenfeld, 2006; Overbeck & Droutman, 2013).

Earlier Perspective-Taking Paradigms

Previous perspective-taking research has relied on a set of dependent measures that are simplified approximations of everyday perspective-taking. Most earlier perspective-taking tasks require participants to consider another person’s spatial or situational perspective compared to their own (Schober, 1993; Keysar et al., 2000). For example, Galinsky and colleagues (2006) captured visio-spatial perspective-taking by measuring whether people wrote the letter “E” from their own visual orientation or from their partner’s orientation (from across a table). Similarly, Yip & Schweitzer (2019) captured perspective taking by measuring whether people wrote their own time zone or a partner’s different time zone (or both) in an email work meeting request. These perceptual tasks capture whether individuals are thinking about another person’s orientation at all, but do not account for individual or contextual variation—the many ways that minds differ across people and situations.

Similarly, Baron-Cohen and colleagues (2001) designed a measure of empathic understanding that asks people to read others’ emotions by looking at a curated set of photos of people's eyes only. The photos are decontextualized: perspective takers did not know anything about the person in the photo, or what the person in the photo was looking at when the photo was taken—information that is often critical for understanding a person's emotional state (Barrett, Mesquita & Gendron, 2011; Zhou, Majka & Epley, 2017). These experimental paradigms rely on stimuli that were chosen to match the canonical definitions of each emotion, which may or may not represent the mixed, ambivalent, suppressed, or misrepresented emotions people experience, express, and perceive in everyday life (e.g., Brooks, 2013; Filipowicz et al., 2011; Larsen et al., 2001; Rees, L., Rothman, N.B., Lehavy, R., Sachez-Burks, J., 2013).

Likewise, many interpersonal judgment tasks ask people to evaluate others’ preferences or personality traits, using the target's self-report as the correct answer (e.g., Salovey & Mayer, 1990; Ickes, 1997; Vazire, 2010). This requires personal knowledge, at least, but is often untethered to specific contexts or behaviors for which the judgment might be relevant. That is, the correlation between self-reported attitudes and behaviors is not always perfect. Furthermore, a person's own self-report presents an unrealistic performance standard—perfect accuracy is impossible because the target has information (their internal mental states) that is unknown—or, at best, partially known—to the judge (and to the experimenter). The judgments themselves can be uncertain, private, and/or contextual. In these cases, researchers often fail to calibrate their results against a reasonable benchmark of what could have been known from the information available to others (see Youyou, Kosinski & Stilwell, 2018; Yeomans et al., 2019).

Taken together, though previous operationalizations of perspective taking help us begin to understand isolated aspects of human theory of mind, they fail to capture the complex and continuous ways that humans develop understanding of other people's minds. Further, the primary tool humans have to maintain their understanding of others—conversation—has been largely forbidden or neglected in prior research. Though results from prior tasks have been used to infer more general properties about perspective-taking in the wild, perspective-taking failures that emerge from constrained stimuli may not be good models for more common perspective-taking failures, when information is abundant, individual differences are critical, conversation is pervasive, and inaccuracy may reflect deeper cognitive limitations. Simple paradigms may exaggerate the benefits of simple solutions to resolve difficult misunderstandings, such as merely prompting people to consider other people's perspectives (see Eyal, Steffel & Epley, 2018; McAuliffe et al., 2019).

Perspective Taking in Conversation

In this paper, we suggest that conversation provides a uniquely compelling approach to capture how people naturally take (and mistake) each other's perspectives. Recent work suggests that while perspective-taking is quite difficult, perspective-*getting*—asking people to describe their thoughts in conversation—may improve perspective-taking accuracy (Eyal, Steffel & Epley, 2018). Indeed, there are many features of conversation that make it an ideal setting to study perspective taking.

First, conversation is an ideal setting to study perspective taking because it is ubiquitous. People reveal their preference to converse over and over again in life—*we talk to each other all the time* (Coupland, 2014 ; Dunbar, Marriott & Duncan, 1997; Mehl et al., 2007), at times too frequently (e.g., Kushlev et al., 2018 ; Quoidbach et al., 2019). Every human relationship can be understood simply as a sequence of conversations over time.

Second, conversation is inherently interpersonal. Conversation is structured to allow people to understand each other, with one person speaking at a time over a series of turns. This structure allows people to listen to what the speaker is saying, before they decide what to say next (Duncan, 1972; Sacks, Schegloff & Jefferson, 1978; Stivers et al., 2009). This is a rich and personal “perspective-getting” task, as a partner's comments on the current topic could be a very timely source of information about their state of mind, which could inform the decision to stay on topic or switch to a different topic.

Third, conversation is one of the most important skills humans develop. Large parts of the human brain are devoted to the efficient computation of listening to others, understanding what they are saying, and generating timely responses (Pickering & Garrod, 2004). Effective conversation requires a multitude of cognitive tasks, and humans are quite good at *some* of them. For example, people mostly comprehend the meaning of what has been said, which helps them coordinate within a shared reality of mutual understanding (Clark & Schaefer, 1989, Hardin & Higgins, 1996; Pickering & Garrod, 2004; Schegloff, 2007). In computational linguistics, "topic models" approximate this human capacity for detecting the topics covered by a swath of text (Hearst, 1997; Blei, Ng & Jordan, 2003).

On the other hand, people also make many systematic mistakes during their conversations—errors of omission and commission. As a decision-making environment, conversation is uncertain, fast-paced (relentless), and high-dimensional—conversing is difficult. Though many isolated aspects of language generation can be understood in terms of efficient information transmission (Zajonc, 1960; Grice, 1967; Misyak et al., 2014; Goodman & Frank, 2016), the *compounded* decisions necessary for conversation are quite difficult, and recent work describes how and why conversations are fraught (e.g., Gilovich et al., working). Even the simplest conversations exert a deluge of cognitive demands, with every turn prompting a judgment about what was just said, and a decision about what to say next.

Emerging research across large swathes of naturalistic conversation data has begun to uncover common, systematic errors in conversational behavior—as well as some ways people can avoid them. For example, people may be particularly vulnerable to egocentric bias in conversation (Epley, Keysar, Van Boven & Gilovich, 2004; Goel, Mason & Watts, 2010), while many forms of conversational success are linked to verbal behaviors that directly reflect perspective-taking. For example, people can make better impressions on others and learn more information by asking more follow-up questions (Huang et al., 2017) and offering reciprocal self-disclosure (Sprecher et al., 2013), behaviors that explicitly show that the speaker was listening to what their partner just said. Similarly, people who disagree can build interpersonal trust by using more receptive language to express their acknowledgment of other people's views and to convey their own views in a way that is more likely to be heard (Yeomans et al., 2020).

Here, we suggest that perceiving and accommodating other people's topic preferences is another way for conversationalists to directly demonstrate perspective-taking—and for researchers to measure it.

Topic Preference Detection: A New Approach

In this paper, we suggest a new test of perspective-taking by focusing on a pervasive and understudied conversational phenomenon: *topic preference detection*, the extent to which conversationalists understand others' topic preferences. In addition to the abundant cognitive demands of conversation broadly, there are several reasons why people may struggle with topic preference detection specifically—explanations that originate from both the speaker and the listener. First, speakers may obscure their topic preferences to be polite (Brown & Levinson, 1987), to be strategic (Crawford & Sobel, 1982), because they overestimate their transparency to others (Gilovich, Savitsky & Medvec, 1998), or because their preferences are genuinely uncertain to themselves. It is likely that people's preferences across all possible conversation topics may be obscured (to themselves and their partners) by salient contextual cues, such as whatever is happening in their immediate environment (Berger & Schwartz, 2011; Shiller, 1995), or by the cognitive demands of maintaining the conversation itself.

We suggest that topic preference detection is a convenient and rigorous approach to measure perspective-taking, and that it is an important psychological phenomenon in its own right. People make judgments continuously during their interactions with others. At every turn of every conversation, every speaker must decide: should we stay on this topic or switch to a different one? Topic decisions affect the course of a conversation, and—over the long term—the course of a relationship. Though the extent to which conversationalists actually rely on others' preference to inform their topic choices is yet unknown, logically, people should rely on their detection of others’ topic preferences to achieve any interpersonal goals. In conversation, individuals have rich, timely, and relevant information about their partner’s preferences—from the verbal (words), nonverbal, and prosodic (paralinguistic) cues their partners express. But are people capable of using that information to learn about their partner’s true preferences? Do people use the cues available to them to understand the minds of others?

Overview of Current Work

Across three studies with hundreds of synchronous and asynchronous conversations among close others and strangers, we explore whether people want to accommodate their partners’ topic preferences in conversation, and how well they can do so. Each study provides evidence for these hypotheses using a slightly different paradigm and sample. Most notably, participants exchange conversational turns asynchronously in Studies 1 and 2, while in Study 3 they chat synchronously face-to-face; and Studies 1 and 3 pair strangers, while Study 2 observes pairs of participants who know each other well (e.g., friends, family, significant others).

Although most conversations in life are synchronous (as in Study 3), our asynchronous conversational paradigm in Studies 1 and 2 has important practical advantages. For example, using asynchronous turn taking allowed us to collect topic preference predictions immediately after someone's turn, rather than after a whole conversation. This measurement approach closely resembles the conversational topic selection process, during which beliefs about topic preferences could theoretically be updated after every turn (to inform the next turn). This design also relieved some cognitive constraints for the participants (e.g. time pressure), and allowed us to share a single writer's responses with multiple readers (to approximate the wisdom of crowds).

Finally, we focus on cooperative conversations in which people have a clear, explicit goal to maximize enjoyment. We did this for two reasons. First, enjoyment is a common—and commonly desired—outcome of both cooperative and competitive conversations (Epley & Schroeder, 2014; Kumar & Gilovich, 2015; Sun, Harris & Vazire, 2019). Indeed, we suspect that conversational enjoyment is a core input into subjective well-being in life (Diener & Seligman, 2002; Quoidbach et al., 2019; Mehl, Vazire, Holleran, & Clark, 2010). Furthermore, the consequences of perspective-taking in strategic or competitive interactions are more complex (Crawford & Sobel, 1982; Epley, Caruso & Bazerman, 2006; Galinsky et al., 2008)—focusing on cooperative interactions makes our results easier to interpret. Third, if people don't understand each other's minds during cooperative conversations, we suspect that accurate perspective-taking during strategic encounters is even more difficult and uncommon. In this way, studying cooperative interactions is a conservative test of our hypothesis that people often fail to understand others' topic preferences.

Developing the Topic Preference Detection Task

Previous research has measured perspective-taking accuracy with tasks that substantially simplify natural social interactions. In particular, these simplifications typically constrain participants to a one-dimensional judgment where the truth is objective (and, at times, obvious). In contrast, our topic preference detection task incorporates several naturalistic complexities that have been previously overlooked. First, the space of possible conversation topics is vast, and far more complex than a single choice. Second, topic preferences are subjective and personal—they differ from person to person. Third, topic preferences are subtle—that is, people's preferences may be truly undetectable, which means that perfect accuracy is an unrealistic benchmark.

We developed our topic preference detection task to capture all three of these complexities—to more closely capture naturalistic perspective taking. Rather than simplify the task itself, we developed an innovative experimental protocol to objectively evaluate naturalistic conversations: First, we created a "topic space," a list of twelve topics of conversation, so all participants could evaluate a uniform set of stimuli. Second, we evaluated perspective-taking accuracy: how accurately can people detect others' self-reported topic preferences among the twelve topics based on their words in conversation? Third, we used machine learning algorithms to generate a normative benchmark to compare against human performance: how accurately can machines detect people's self-reported topic preferences based on the same words available to human conversation partners? Though other aspects of the studies differ, all three studies employ the topic preference detection task. We describe the three steps of this shared method below.

Defining the Topic Space

In a pilot study (n = 300; see Appendix A), we pretested 50 conversation topics drawn from previous research (Aron et al., 1997; Huang et al., 2017). From the pilot data, we selected twelve topics as stimuli for the other studies in the paper. We selected topics that people rated as average in terms of "interestingness," with high variance. This ensured that people's preferences for the 12 topics would not be obvious (e.g. "what do you like to do with your free time" was universally liked, and "when was the last time you cried" was universally disliked), and that a judge would need to learn something personal about the target to predict their topic preferences. This meant we could measure topic preference detection accuracy over a common set of stimuli for all judges, while still allowing each target to have a different set of preferences over the stimuli. The final list of twelve topics is shown in Table 1.

Table 1: The twelve conversation topics provided to participants in Studies 1-3.

What do you do for work? What do you like about it?
Why do you do these kinds of studies?
Are you a religious person? Why?
Do you have any fruit trees, plants, or a garden?
What’s the strangest thing about where you grew up?
What is the cutest thing you’ve seen a baby or child do?
Would you like to be famous? In what way?
When did you last sing to yourself? To someone else?
If you were able to live to the age of 90 and retain either the mind or body of a 30-year- old for the last 60 years of your life, which would you want?
If you could change anything about the way you were raised, what would it be?
What do you value most in a friendship?
Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?

Evaluating Accuracy

For each of the 12 topics, we asked people to respond to the topic question and then rate their preference for the topic: to what extent would you like to stay on this topic (+10) or switch to a different one (-10)? Next, we asked different people to predict the target's topic preferences based on their words. Each target produced 12 ratings, and each judge made multiple predictions (6 or 12) for each target they saw. This meant that, unlike in previous perspective-taking tasks, we could evaluate within-person preference detection accuracy: for any pair of topics and a single target person, could a judge determine which of those topics a person preferred? This design has several advantages. First and foremost, within-person judgments mimic real topic choices made in conversation, where one's partner remains constant but the topic(s) under discussion can change. This design also allows for much greater statistical power, because we can control for person-level differences, and compare across a person's preferences between topics.

We used a non-parametric measure of accuracy (Kendall's tau), which compared the relative ordering of a judges' predictions to the ordering of the target's ratings, rather than the absolute levels. This is essential for measuring accuracy in subjective domains, because humans often have idiosyncratic interpretations of scale ranges. A parametric comparison (e.g. Mean Absolute Deviation or Root Mean Squared Error) would unfairly make human judges look much less accurate, simply because they interpreted the scales differently. In this way, Kendall's tau is more generous to human judges, compared to parametric measures of accuracy.

Machine Learning Benchmark

Ground truth in perspective-taking tasks has often been objective and deterministic (e.g., the letter “E” is oriented egocentrically or not). In our topic preference detection task, though, ground truth is subjective: targets may be uncertain about their own topic preferences, or they may not be explicit about them in conversation. In either case, it would be impossible for any judge—even a machine—to detect targets' preferences exactly. Even simple reading tasks like sentiment analysis of product reviews cannot reliably achieve perfection (Pang & Lee, 2008). Instead, we suggest an alternative benchmark to empirically determine *what could have been known* from the information at hand.

To estimate an upper bound of preference detection accuracy in this domain, we compared human prediction accuracy to machine learning prediction accuracy. Specifically, we developed a natural language processing model to predict someone's preference for a topic, using only their words (i.e., the text of their response to the topic questions). We processed the text using a combination of four approaches: topic modeling (Roberts et al., 2014); word embeddings (Mikolov et al., 2017); politeness markers (Yeomans, Kantor & Tingley, 2018); and some polynomial combinations of word count, sentiment, and topic fixed effects (full details in Appendix B). The algorithms we used to make predictions from these feature sets are relatively simple (compared to many other machine learning algorithms). Namely, we rely on the LASSO regression for supervised learning (Friedman, Hastie & Tibshirani, 2010), which only estimates linear terms. We believe that even simple LASSO predictions can provide a conservative normative benchmark against which to measure human performance.

For all ensemble models, we created a higher-level set of features, which used the human and NLP predictions themselves as features, along with topic fixed effects. We also created features representing the quadratic terms of this set (including interactions). In part, this allowed for topic-level variation in the contribution of each feature. We generated predictions using a leave-one-person-out cross-validation.

Study 1

The data for Study 1 were collected in two parts. In Study 1A, working adults were told to imagine a conversation with someone, and read all 12 topic questions. After writing a response to each topic question (e.g., “What’s the cutest thing you’ve seen a baby or child do?”), they rated their interest in staying on this topic versus switching to a new topic. In Study 1B, a separate group of judges from the same participant pool saw the writers' responses and predicted the writers’ topic preferences (our main measure of topic preference detection accuracy).

Study 1 Methods

Study 1A Sample. Writers were recruited from Mechanical Turk and completed attention checks before starting or being counted in our sample. We intended to recruit 400 writers, which was chosen to double the sample from the pilot, to accommodate a between-subjects condition. Writers were initially counted in the sample if they passed the attention checks, but we later excluded one for writing nonsense in the text boxes, and seven for giving near-identical answers to the 12 topic preference questions ($\sigma<1$), leaving a final sample of 392 writers.

Study 1A Protocol. Participants were told to imagine they were in a conversation with someone. They were randomized into one of two between-subjects conditions: half were told to imagine they were conversing with someone they speak to often (“close target”), while the other half were told to imagine they were conversing with someone they had never met (“distant target”). They were presented with the twelve topics, one at a time, and were prompted to respond to each topic with the question “How would you respond to this question in conversation?” (see Appendix C for the exact prompts). After writing their response, they rated their interest in staying on the topic (+10) versus switching to a different topic (-10), using a slider tool on a single scale (which was initially positioned at zero, and required a response before continuing). Finally, they completed demographic questions (age, gender).

Study 1B Sample. Judges were recruited from Mechanical Turk and completed attention checks before starting or being counted in our sample. We intended to recruit enough judges so that each written text would receive at least three judges. 693 judges passed the attention checks, but 38 did not complete the survey, and one person made the same guess for every person ($\sigma<1$), leaving 654 judges in the final sample.

Study 1B Protocol. Each judge was randomly assigned to read responses to only six of the twelve topics, written by four writers from the writer study (Study 1A), for a total of 24 topic preference judgments made by each judge. The writers were assigned so each judge only saw writers from one experimental condition of Study 1A (in which the writers imagined conversing with a close or distant other), but there were no new randomizations in Study 1B. At the beginning of the study, they were shown exactly what the writers saw (i.e. matching the writers' instructions from Study 1A), as well as a histogram of all the writers' ratings, to get a sense of how people used the stay/switch preference scale in aggregate. Judges were also incentivized for performance (i.e., preference detection accuracy), with the ten most accurate participants receiving a bonus of $2.

For each of the 24 topic responses, the judge saw the topic question, the writers' response in text, and made three judgments (see Appendix D for exact prompts). First, they predicted the writer's preference for the topic, using the same -10 to +10 slider that the writers used (*"prediction"*: this was our primary measure, and was used to determine the incentives). Then, they reported their own preference for staying on that topic in conversation with the writer herself (*"partner-specific preference"),* also on the -10 to +10 scale. Third, they reported whether they would intend to follow up on the topic with the writer, or switch to a new topic (“*intent to follow up*") on a 1 to 7 likert scale. Judges were shown all six responses for each writer in a single block, before moving to the next writer, and at the end of each block, judges evaluated their overall impression of that writer. After all of the writer judgments, they gave their own preference for all twelve topics without any partner in mind (*"generic preference"*). Finally, they completed demographic questions (age, gender).

Study 1 Results

Topic Preferences. The judges seemed to care about their conversation partners' topic preferences. Judges said they were more likely to follow up on the topic with a writer if they thought that writer preferred the topic (τ = .584, CI_{95} = [.565, .603]). And their own partner-specific preferences were highly correlated with what they believed the writer wanted to talk about (τ = .531, CI_{95} = [.511, .550]).

The correlation was caused by a combination of two important mechanisms. First, the judges' own preferences affected their predictions. That is, there was some egocentric bias in the judges' predictions. Judges' predictions were also correlated with their generic preference for each topic (τ = .128, CI_{95} = [.111, .145]). However, neither sets of one's own preferences were reliable guides to the writers' actual preferences (generic: τ = .034, CI_{95} = [.020, .049]; partner-specific: τ = .097, CI_{95} = [.083, .112]). That is, judges were putting too much weight on their own preferences as a guide to predict the writer's preferences.

Additionally, the judges' predictions affected their own preferences. That is, judges wanted to accommodate the writers' topic preferences, and adjusted their own preferences to reflect what they learned about the writer. In fact, judges' preferences were correlated with what *other judges* thought the writers wanted to talk about (τ = .189, CI_{95} = [.173, .205]), which cannot be attributable to egocentric projection. Instead, it shows that when judges picked up on signals that writers were interested in a topic, the judges themselves became more interested in discussing the topic with them.

Judge Accuracy. In Figure 1, we compare the accuracy of various methods of topic preference detection. The judges' predictions were closer to the writers' true preferences than to their own preferences, (τ = .142, CI_{95} = [.127, .158]). This demonstrates that judges were somewhat able to adjust from their own preference and take their partners' perspectives. Judge predictions were also more accurate as aggregated groups of 3-6 judges per text (τ = .185, CI_{95} = [.161, .210]). Statistical aggregation (the wisdom of crowds) mitigates egocentrism but it has limited practical value in real conversations, when individuals typically need to detect their partners' topic preferences on their own. Also, aggregating judgments does not account for biases and blindspots that are shared across people.

NLP Accuracy. For each of the twelve topics, we trained a separate prediction model, which based its judgments on the same text data that the human readers used. Within each topic, we employed 20-fold nested cross-validation to estimate out-of-sample prediction accuracy (Stone, 1974). To smooth out error due to cross-validation, we repeated this procedure five times and averaged the results from each run to make a final prediction for each text. The predictions that resulted from this algorithm significantly outperformed individual humans (τ = .174, CI_{95} = [.151, .196]; paired t(653)=3.7, p<.001).

When we combined groups of predictions across multiple human judges, the unweighted average of those judges performed similarly to the NLP model (paired t(391)=0.9, p=.393). However, we found that the most accurate predictions we found were from an ensemble that combined the aggregated human predictions and the NLP predictions, (τ = .218, CI_{95} = [.193, .241]; paired t(391)=3.9, p<.001). This shows that the algorithm detected topic preference cues in the writers’ responses that humans did not. This benchmark is conservative, given our limited training set and simple algorithm, but still finds that humans had room for improvement in detecting others’ topic preferences from the text.

Although our model trained on each topic separately, there were some linguistic differences that were consistent across topics. We demonstrate some examples in Figure 2, using the politeness R package (Yeomans, Kantor & Tingley, 2018). We identify politeness features that were more common in the top third most-preferred topics, compared to the bottom third least-preferred topics, across the pool of all twelve topics. For example, preferred topics tended to mention other people ("you", "we", "ours") and use positive words. In contrast, disliked topics tended to mention the self ("I", "my"), and used negations ("not") and truth intensifiers ("actually", "in fact").

Writer Transparency. Study 1A randomized writers into two conditions. Half were told to imagine conversing with a close other, and the other half were told to imagine someone they had never met. Preference detection accuracy for human readers was lower when writers imagined close others (τ = .164, CI_{95} = [.131, .197]) than strangers (τ = .208, CI_{95} = [.171, .243], unpaired t(390)=1.7, p=.079). Interestingly, responses written for strangers were also more transparent to the the NLP algorithm (τ = .145, CI_{95} = [.111, .177]) than responses written for close others (τ = .205, CI_{95} = [.174, .234]; unpaired t(390)=2.6, p=.008). This finding suggests that people can modify the transparency of their topic preferences. In this case, they increased their transparency when they imagined conversing with an unfamiliar partner.

Fig. 1. Topic preference detection accuracy in Study 1. Each bar shows a group mean and standard error. The accuracy of the algorithm and the ensemble was estimated out-of-sample, using a nested-cross-validation procedure.

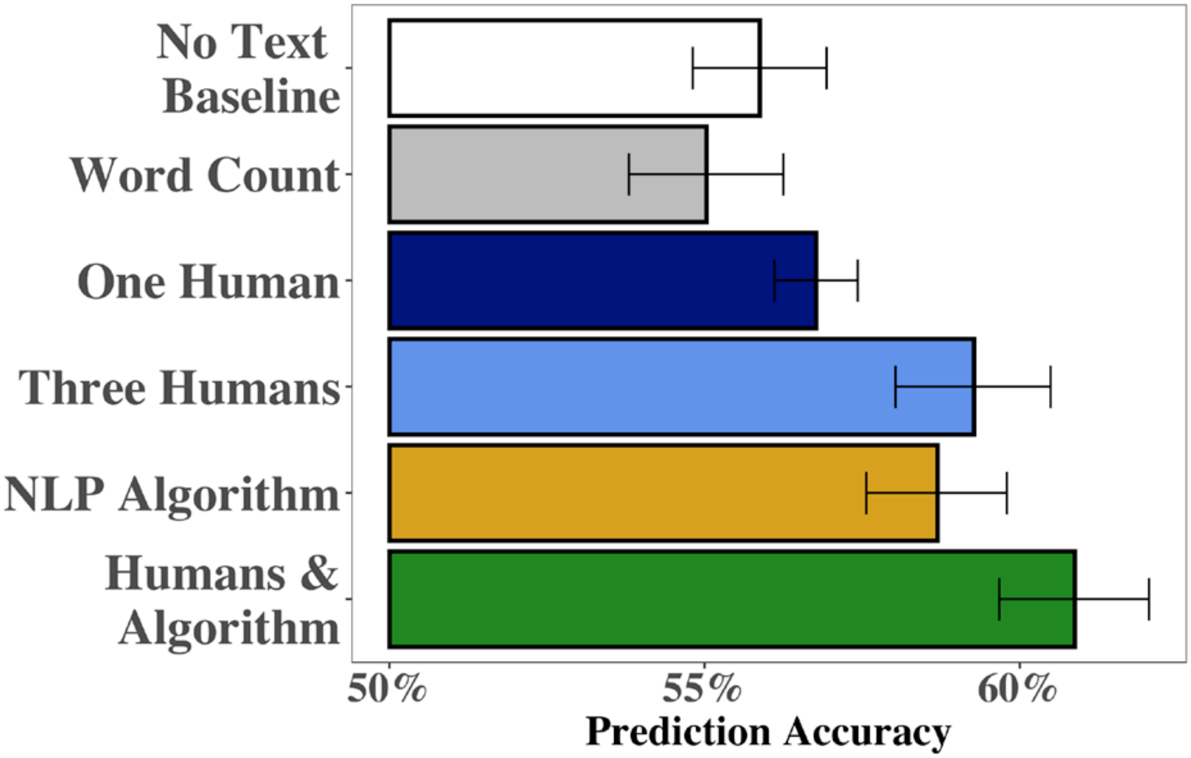
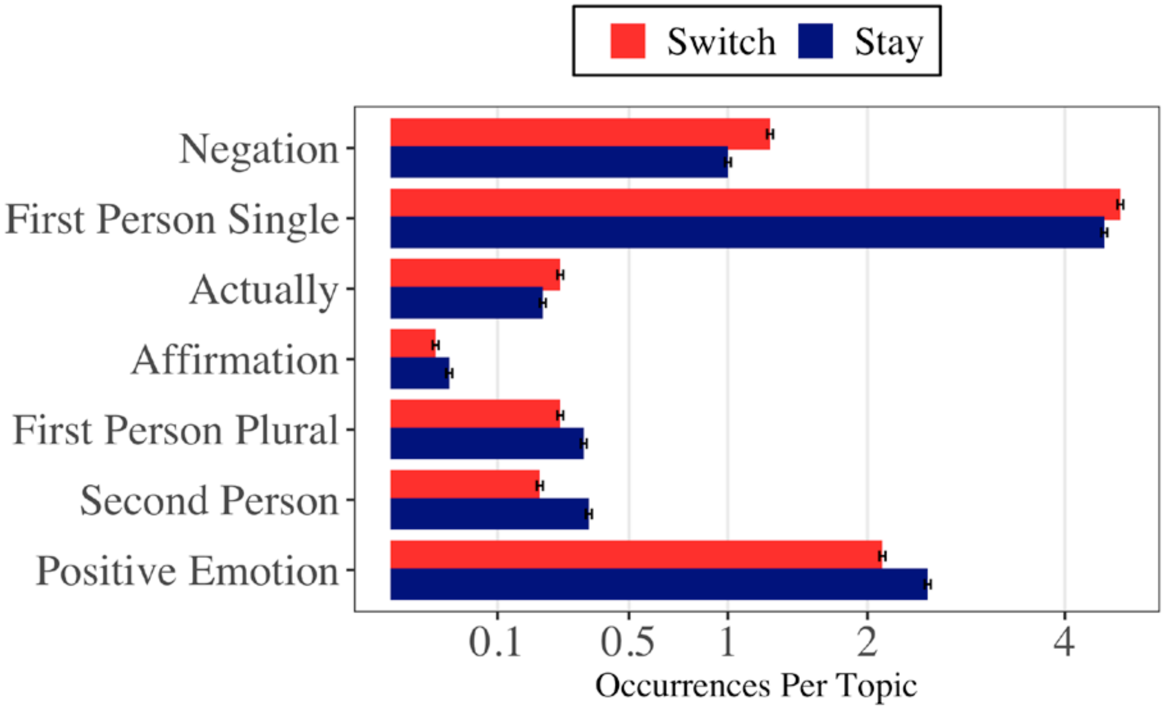


Fig. 2. Politeness features that predict topic preferences in Study 1. Responses are split into upper and lower terciles, with the middle tercile removed. Features are included if they pass a chi-squared test across the two groups (with a threshold of p < .05, un-adjusted) and occur in at least 2% of the written topic responses. Each bar shows a group mean and standard error.



Study 1 Discussion

Taken together, the results of Study 1A-B demonstrate that topic preference detection captures many basic properties of perspective taking. First, topic preference detection is important. People want to make topic selection decisions that accommodate their partner's topic preferences, as well as their own. This reassures us that our domain is a cooperative one (Crawford & Sobel, 1982), and that people often want to learn their partner's preferences in conversation. However, we also found that even after getting someone's perspective on a topic, people still had blind spots in their understanding of what that person wanted to talk about.

People were quite egocentric and over-relied on their own preferences to predict other people's preferences. Egocentric projection was not a useful way to predict other people's preferences in this study, unlike other work on preference prediction (Hoch, 1987; Krueger & Clement, 1994; Gilbert et al., 2009; Zhou, Majka & Epley, 2017), perhaps because our task focused on judgments that were truly personal, and could not be solved by simply imagining situational factors. Instead, people were able to accurately adjust their predictions by incorporating information about the target from their text response.

We also found that topic preference detection demonstrates the limitations of perspective getting (Eyal, Steffel & Epley, 2018). Even when targets were allowed to express their thoughts on a topic in open-ended text, their preferences were not very clear, to humans or to algorithms. We also found that there is room for improvement in people's preference detection accuracy, compared to an objective, and relatively unsophisticated, empirical benchmark (that did not include interaction detection, sentence embeddings, and used less that 400 training examples per predictions). Still, the algorithmic benchmark outperformed humans who have had conversations their whole lives.

We also found that people have volition over how transparently they express their conversation preferences. When writers believed their audience was less familiar, they expressed their preferences more directly. This could indicate that people communicate more clearly with strangers because they do not assume strangers understand them as well as close others. Alternatively, close others may communicate in idiosyncratic ways that are undetectable by strangers or by algorithms trained on other peoples' responses. We explore these explanations in Study 2, among participants who knew each other.

Study 2

In Study 2, we modified the topic preference detection task to accommodate pairs of participants who knew one another. Each person in the pair was both a writer and a reader for their partner. First, each person wrote their response and rated their preference for each of the twelve topic questions at separate computers. Then, they switched seats with their partner, and tried to guess their partners' preferences.

Study 2 Methods

Sample. To collect data from people who knew each other, we recruited writers and readers simultaneously as pairs to a behavioral lab (e.g., spouses, partners, friends, coworkers, etc.). We intended to collect 200 participants (i.e. 100 pairs) to match the size of a single condition in Study 1, and based on our expectation of participant availability. and in our preregistration we decided to exclude pairs based on a research assistant's judgment that one or both of the participants were not following instructions, did not finish the survey, or had given the same rating to every topic (sd<.5). We recruited 202 people: 14 people were excluded by the research assistant, and another 12 for giving the same ratings to every topic, leaving a final sample of 172 participants.

Protocol. First, every person wrote a response and rated their preference for staying on each topic. They were specifically told to consider these responses in the context of a conversation with their partner, at that moment. Afterwards, both people predicted one another's preferences for all twelve topics, in two blocks of six items per block (block order and topic assignment were randomly shuffled). In one block, they read the topic question and their partner's response, and predicted their partner’s preference once (just as in Study 1). In the other block, they made two predictions for each topic: first without seeing their partner's text (i.e. based on what they already knew about their partner); then, they were allowed to adjust that initial prediction after seeing the response text.

We also measured readers' confidence in their predictions. After every prediction, they reported the probability that their guess was within three scale points (above or below) of their partner’s true rating. At the end of the study, we asked them to make a summary judgment about how many of their twelve with-text predictions were within three scale points of their partner’s true rating. Finally, we collected demographics (age, gender), as well as information about how well the partners knew each other.

In addition to pre-registering the sample size, we also pre-registered our prediction algorithm. That is, we posted the code to extract the natural language processing feature sets above. The entire dataset from Study 1 was used for training data, and we estimated a separate model for each topic using the same algorithm, which combined all of the NLP features together. We then produced held-out predictions for the Study 2 data one topic at a time, and generated ensemble models using the same procedure as above.

Study 2 Results

The accuracy of different predictions are plotted in Figure 3. participants in Study 2 had a similar prediction accuracy (τ = .155, CI_{95} = [.121, .189]) to the NLP algorithm (τ = .128, CI_{95} = [.095, .159]; paired t(171)=1.3, p=.190). This suggests these close others were somewhat more accurate than the strangers in Study 1. However, we again found that an ensemble of the NLP and human predictions was better than humans alone (τ = .188, CI_{95} = [.156, .219]; paired t(171)=2.3, p=.023). This suggests the humans judges were still missing out on information in the text. Furthermore, people were just as accurate before they read their partners' response (τ = .162, CI_{95} = [.106, .217]) as they were afterwards (τ = .163, CI_{95} = [.111, .214]; unpaired t(171) = 0.0, p = .965). They did not seem to incorporate anything useful from the text. Instead, their improved accuracy came from their prior knowledge of their partner’s topic preferences. Furthermore, humans were again overly egocentric. Their predictions of their partners’ preferences were more correlated with their own preferences (τ = .164, CI_{95} = [.127, .200]) than with their partners' actual preferences (τ = .056, CI_{95} = [.021, .091]; paired t(171)=4.5, p < .001).

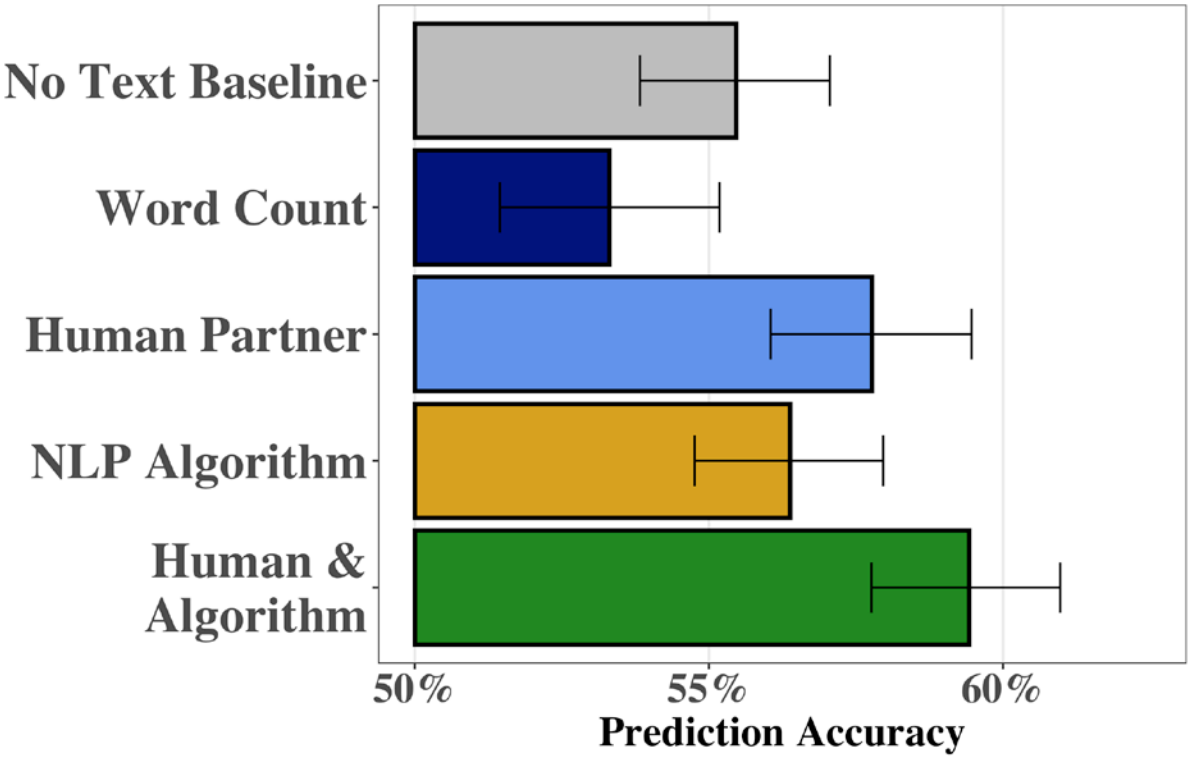
People also underestimated the difficulty of the task. On average, they expected their predictions would be within +/- 3 of the correct answer 70.8% of the time (CI_{95} = [70.0%, 71.7%]). And at the end of the study, they predicted that 55.9% of their guesses would have been within +/- 3 of the correct answer (52.9%, 58.9%). However, their true hit rate by this definition was only 39.3% (CI_{95} = [37.2%, 41.4%]), much lower than they had expected.

We collected 98 friends, 18 spouses, 22 significant others, 20 work colleagues, 6 family members, and 8 others. This included a range of personal history—70 people had known their partner for less than 1 year; and 24 people had known their partner for more than 10 years. Using an ordered factor regression, we found evidence that a longer personal history predicted increased confidence (linear term: β=10.18, SE=3.11; t(166)=3.3, p=.001), but this did not translate into increased accuracy (β=.027, SE=.045; t(166)=0.6, p=.552).

Study 2 Discussion

We collected participants who knew one another, and some results from Study 1 were replicated. Participants were once again egocentric in their predictions of other people's preferences. And the human predictions again missed out on information in the text that was captured by the NLP models. Participants were able to match the accuracy of the NLP algorithms on their own, however this was not because they listened to what their partner was saying. Instead, they were equally accurate when they did not read any text on the topic. We also found that participants were quite overconfident, and they underestimated how difficult the task was. Furthermore, this overconfidence increased as they gained more information about their partner (either a longer relationship, or reading the text). This suggests a breakdown in how well people process new information, as well as a potential mechanism for why topic preference detection failures persist. However, Study 2 (like Study 1) was limited to asynchronous conversations, so we extend the preference detection task to synchronous conversations in Study 3.

Fig. 3. Topic preference detection accuracy in Study 2. Each bar shows a group mean and standard error. The accuracy of the algorithm and the ensemble was estimated out-of-sample, using a nested-cross-validation procedure.



Study 3

In Study 3 we created a topic preference detection paradigm using synchronous, face-to-face conversations. This context introduced new sources of information, and made our paradigm more externally valid in many respects. Speakers engaged in natural turn-taking with one another, choosing how much to talk about each topic and when to ask follow-up questions or change topics. They could also listen to their partner's prosody, and see their nonverbal cues. This information could be helpful in all sorts of ways for predicting topic preferences.

The synchronous paradigm also changed how people formed their topic preferences. Speakers were asked about their preference for each topic after the conversation, and with their specific partner in mind. Speakers could draw directly from their recent experience to forecast their preference for the topic. In particular, they might consider what they learned about their partner's preferences when they state their own preferences.

We conducted this study in two parts. In Study 3A, we recruited participants ("speakers") to sit down and converse about the topic list, and then predict one another's topic preferences afterwards. These conversations were videotaped. In Study 3B, we recruited a separate group of participants ("observers") to watch the videos and predict the preferences of the speakers. We also transcribed and annotated the conversations, and used the text to develop another NLP algorithm as an accuracy benchmark.

Study 3A Methods

Sample. Individuals were recruited separately to a behavioral lab. We intended to collect 200 participants (i.e. 100 pairs), to match the size of one condition in Study 1 and based on expectations of participant availability. We planned to have the research assistant conducting the study identify any dyads that had a technical malfunction, failed to follow instructions, or otherwise did not complete the study. In our data, 210 people started the study, and the exclusions removed 14 people (7 dyads), leaving 196 speakers (98 dyads) in the final sample.

Protocol. First, speakers separately read the twelve topic questions in a random order and gave their generic preference for each. Then they met another participant and spent ten minutes conversing about the 12 topics, with a clear, shared goal: enjoy the conversation (see Appendix E for the exact instructions). Although we do not discuss it in detail here, participants were randomized into one of two conditions - half were told to switch topics frequently, while the other half were not told how often to switch topics. After the conversation, each speaker rated their preference for each topic with their partner, and also predicted their partner's same preference rating for each topic (see Appendix F for the exact prompts). All of the conversations were recorded on video and transcribed. We then had research assistants annotate each turn in every conversation with one of one of the twelve topic labels (see Appendix G for annotation protocol).

Study 3B Methods

Sample. Individuals were recruited separately to a behavioral lab. We intended to collect 300 participants, so that each video would be watched by at least three people. We planned to have the research assistant conducting the study identify any dyads that had a technical malfunction, failed to follow instructions, or otherwise did not complete the study. In our data, 335 people started the study, and the exclusions removed 39 people, leaving 296 observers in the final sample.

Protocol. First, observers separately read the twelve topic questions in a random order and gave their generic preference for each. Then they read the instructions that the previous participants had seen, and watched a randomly-selected video. After the video was over, they predicted the topic preferences of each person in the video, and answered some other questions about them (see Appendix H for the exact prompts). Participants were incentivized to be accurate, with a $20 prize going to the participant with the most accurate predictions.

Study 3 Results

Like Study 2, all judges made predictions for all twelve topics. We again calculated the accuracy of each set of twelve predictions as the Kendall's tau correlation with the true values. In Figure 4, we plot this accuracy for several of the topic preference detection strategies.

Topic Choices. In synchronous conversation, people can choose how much to say on each topic - i.e. their "revealed preference" during conversation. We constructed a standardized multiple regression model, to predict the total time on topic for each person (clustering standard errors at the person and dyad level). And we found that the amount of time a speaker spends on a topic is strongly determined by two variables: the speaker's own pre-conversation preference for the topic (standardized β = .117, SE = .023, t(2349) = 5.1, p < .001), and their partner's pre-conversation preference for the topic (standardized β = .072, SE = .023, t(2349) = 3.2, p < .001).

We conducted the same analysis with speakers' post-conversation preferences, and found again that participants preferred to accommodate their partner's preferences. Post-conversation preferences were dependent on both the speaker's own pre-conversation preference for the topic (standardized β = .561, SE = .027, t(2349) = 21, p < .001), and their partner's actual pre-conversation preference for the topic (standardized β = .061, SE = .025, t(2349) = 2.4, p = .017). Speakers' post-conversation preferences were even more strongly determined by their predictions of their partner's preferences (standardized β = .557, SE = .031, t(2348) = 18, p < .001), even after controlling for the actual preferences of both the speaker and their partner. This evidence again shows that topic preferences affect how people conduct conversation, and that our participants wanted to accommodate their partner's topic preferences (if they had known what they were).

Human Predictions. The human predictions were somewhat more accurate than in Study 1, on average. The speakers were more accurate for their partners (τ = .203, CI_{95} = [.166, .240]) than the individual observers (τ = .150, CI_{95} = [.128, .171], paired t(298) = 3.8, p < .001). However, each dyad was seen by 3-4 observers, and when the observers were pooled into an unweighted average, their accuracy was similar to the speakers themselves (τ = .220, CI_{95} = [.186, .253]; paired t(198) = 0.8, p = .434). Like Study 1B, this pooling improved accuracy, in part, by mitigating the egocentrism within each independent observer's predictions (τ = .147, CI_{95} = [.122, .172]).

The speakers' own predictions of their partner were also quite egocentric: a multiple regression, controlling for their partner's actual pre-conversation and post-conversation preferences, suggests that speakers still put too much weight on their own pre-conversation preferences when forming their predictions of their partners (standardized β = .300, SE = .027, t(2348) = 11, p < .001) . We could not aggregate multiple speakers, as they each predicted for a different person. However when we combined the speaker's predictions with the observer's predictions to create an all-human ensemble, accuracy improved once again (τ = .249, CI_{95} = [.215, .282]).

NLP Predictions. To create a more systematic benchmark for accuracy, we pooled all of each person's speaking turns for each topic as a single "document" to predict their topic preferences. The total word count on each topic was substantially more correlated with their post-conversation preferences than in Studies 1 & 2 (τ = .247, CI_{95} = [.206, .288]). This result already suggests that the benchmark for preference detection in synchronous conversation should be higher than in asynchronous conversation.

We next tried to apply the asynchronous NLP algorithm developed in Study 1. We followed the approach from Study 2, and only used the Study 1 data to train the model. This approach had limited prediction accuracy (τ = .105, 95% CI = [.077, .134]). But this model may not be the right benchmark for data from synchronous conversation. It was trained on data from the asynchronous conversations, so this model does not account for any of the social cues in the context of the conversation itself.

Synchronous conversation includes many kinds of dialogue acts that can communicate topic preferences during the conversation. For example, the transcripts identified interruptions, and laughter automatically (we counted laughter separately, depending on whether someone was laughing at something their partner had said, or something they had said themselves). The topic annotations also identified which partner mentioned a topic first, which partner ended a topic, and whether that end was a segue or an explicit switch. We also labeled backchannels (e.g., “yea,” “uh huh”) where the transcript timestamps revealed short turns amidst two much longer turns. We also parsed the types of questions people asked each other into one of three types: "switch," "follow-up," and "mirror" questions, using an existing NLP question type classifier (Huang et al., 2017).

In Figure 5, we plot some of the most distinctive dialogue acts. When someone mentions a topic first, they are likely to be interested in it, whereas when their partner mentions it first, they are less likely to be interested in it. In addition, people who laughed at the end of their own statements tended to want to stay on topic, although other kinds of laughter did not signal topic enjoyment for either person. Additionally, topic segues were good signs that a topic was over, although when someone explicitly ended a topic, that indicated that their partner probably still wanted to talk more about the topic.

One perhaps surprising feature is interruptions: some recent papers have found interruptions as evidence of competitive status-seeking behavior (Mendelberg, Karpowitz & Oliphant, 2014; Jacobi & Schweers, 2017). We find here that in co-operative conversation, interrupting one another is a sign of a dynamic, bubbling discourse, in which the participants are interested and listening attentively (Fay, Garrod & Carletta, 2000). This is also consistent with the finding that backchannels - brief, active indications of listening - were the single most consistent indicator of topic interest.

We estimated a new topic preference detection model that was topic-generic, treating every combination of topic and participant as a separate observation. We generated out-of-sample predictions for all dyads this way, using either the dialogue act features described above, or else the set of 36 social cues (e.g., gratitude, hedges, acknowledgment), from the politeness R package (Yeomans, Kantor & Tingley, 2018). These two algorithmic predictions (along with the predictions from the asynchronous model) were then included in an ensemble model similar to that in Studies 1 and 2, which included all first- and second-order terms from the algorithms' predictions, the speakers' word counts, and topic fixed effects. Although we did not seem to have enough data to have estimate a separate model for each topic, the interaction terms in this ensemble allowed the relative strength of each NLP component to vary for each topic.

At each level of prediction, we conducted a nested cross-validation, estimating a LASSO regression in each inner fold and partitioning the outer folds by dyad, so that for any given loop, all the held-out documents were from the same dyad. The final set of predictions from the model outperformed all of our other benchmarks ($\tau = .338$, $CI_{95} = [.305, .369]$), including both groups of humans (partners: paired $t(195) = 7.3$, $p<.001$; observers: paired $t(195) = 5.9$, $p<.001$). We built an identical ensemble model that also included the predictions from the two groups of humans (partners and observers) as well as the NLP predictions, speaker word count, and topic fixed effects. The ensemble model that included the human predictions ($\tau = .338$, 95% CI = [.304, .369]) was no better than the ensemble model that only drew on the NLP predictions (paired $t(195)=0.1$, $p = .992$).

Fig. 4. Topic preference detection accuracy in Study 3. Each bar shows a group mean and standard error. The accuracy of the algorithm and the ensemble was estimated out-of-sample, using a nested-cross-validation procedure.

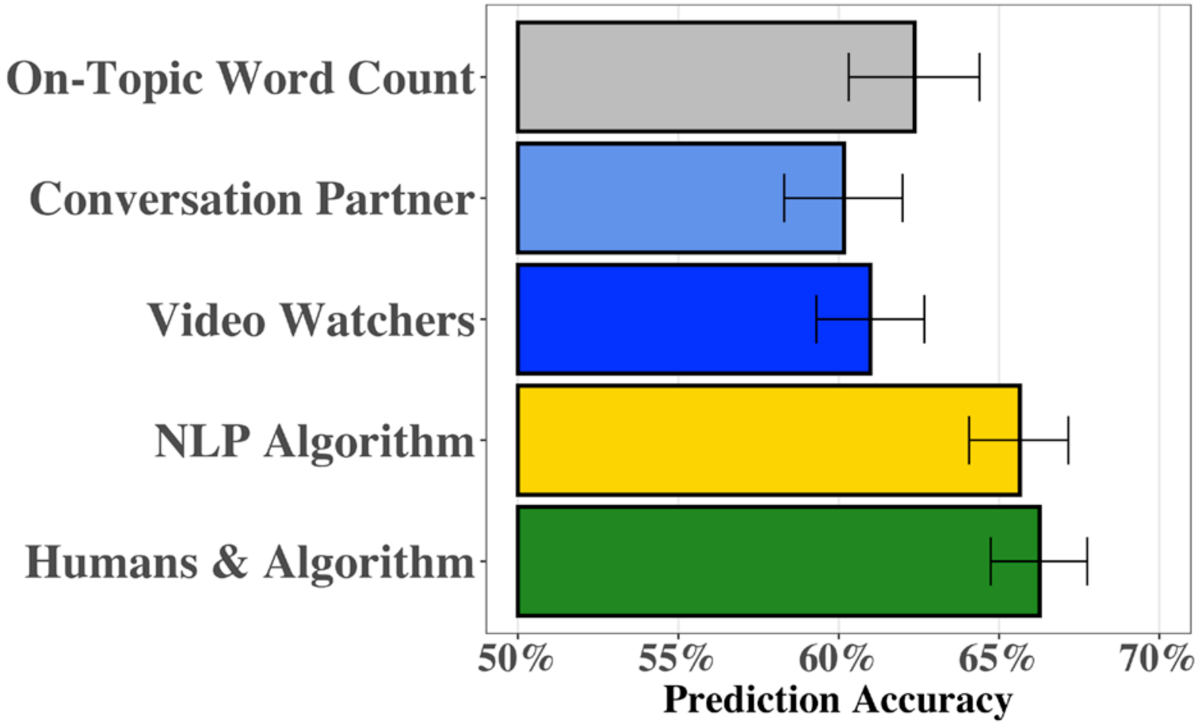
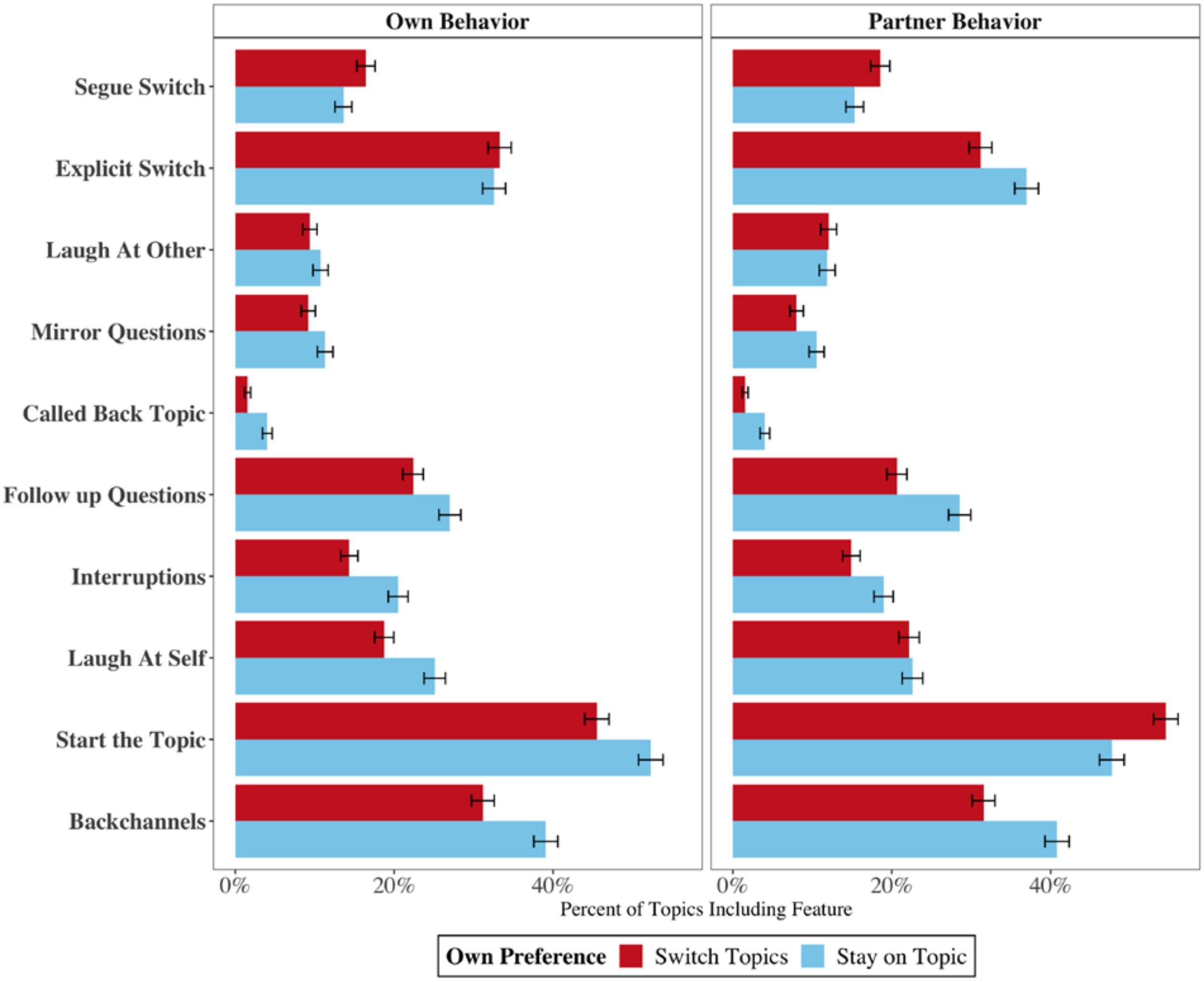


Fig. 5. Dialogue act features that predicted topic preferences in Study 3. Bars compare means (and standard errors) from two groups - the top third of topics that were most preferred, and the bottom third of topics that were least preferred. Each bar represents the average rate that each conversational feature was used at least once, within each group of topics. The data are split in two panels to separate the conversational features spoken by the person who rated the topic (left panel), and those spoken by their partner (right panel).



Study 3 Discussion

In Study 3, we allowed participants to have full, open-ended conversations with one another. They could listen to what their partner had to say about any of the topics and ask follow-up questions, add their own perspective, or switch topics at leisure. They also had access to prosody, gesture, facial expressions, and other non-verbal information from their partners. Still, neither the speakers nor the observers detected topic preferences as well as our NLP algorithm, which only had access to the transcript. However, this does not preclude the possibility that an algorithm could make better use of non-verbal information, if we were to quantify those features of conversation and include them in the model.

The synchronous conversations also showed some consequences of topic preferences in conversation. First, each person's speaking time on a topic was strongly predicted by two factors - their own pre-conversation topic preferences, and their partner's pre-conversation preferences. This meant that stated topics preferences were a good predictor of revealed preferences during conversation. Furthermore speakers' preferences converged with their predictions about their partners by the end of the conversation. This meant that speakers wanted to accommodate their partners' preferences, to the extent that they knew what those preferences were. However, our research suggests that detecting their topic preferences can be a surprisingly difficult task.

General Discussion

These results suggest that topic selection is a difficult and important perspective-taking task, even in cooperative conversations. Across thousands of dyads, people were egocentrically biased and inaccurate in their judgments of others' topic preferences. Although people filled their conversational speech with information about their topic preferences that were accessible to algorithms, their human partners failed to pick up on many of those cues (or ignored them), and they were slow to act on them. Taken together, our results suggests that there is ample room for improvement in conversational perspective-taking.

Although the current research focused on enjoyable, co-operative conversations, we suspect topic selection is integral to many other conversational goal pursuits as well, such as persuasion, impression management, information exchange, productivity, social connection, and conflict management. Many of these goals can be entangled with conversational enjoyment. But we still think it is likely that the context and goals of a conversation affect people's preferences for different topics. For example, optimal topic choices will differ based on how many people are talking together, and their relationships with one another (Cooney et al., 2019).

Enjoyment is still a useful model for conversational decision-making because it is a cooperative, positive-sum goal, with (relatively) few strategic or competitive barriers. Instead, our work identifies more tractable barriers effective topic selection, like cognitive limits and coordination failures, and we suspect that the same cognitive barriers may affect other goals.

In general, our work suggests three implications for improving conversational perspective-taking. First, our results stress the importance of listening - our machine learning results show that although there is ample information in what people say about different topics, our participants failed to pick up on those cues. Our results also suggest that people can modulate their transparency to others. While there may be good reasons to be indirect (e.g. politeness; Brown & Levinson, 1987), prior work also suggests that people overrate their own transparency to others (Gilovich, Savitsky & Medvec, 1998; Gilovich & Savitsky, 1999). Finally, our work suggests that people could to offload some of the cognitive effort in advance. Conversation already poses an immense cognitive load (Pickering & Garrod, 2004; Goodman & Frank, 2016). Just as we provided generic topic lists to participants in our studies, people could prepare their own list of possible topics before their conversations, written down or in mind, to assist in generating smooth transitions to new topics.

Conversational decisions have long been examined from the outside in, using networks, vignettes, and surveys to begin to understand human interaction. But recent advances in natural language processing (Hirschberg & Manning, 2015; Jurafsky & Martin, 2017) have made it possible for researchers to examine conversations from the inside out—to better understand what people actually say to each other. Our findings suggest that topic selection may also matter for algorithmic conversationalists. Automated dialogue agent technology is slowly advancing beyond constrained applications (e.g. question answering, customer service) to open-domain dialogue, where topic preference detection may be critical for successful conversations (Fang et al., 2018; Dinan et al., 2019).

Our results contribute to a growing body of evidence that conversational behavior is filled with a variety of heuristics and biases that suggest ample room for improvement. Linguistic models often build from examples where communication is successful (Grice, 1967; Misyak, Mekonyan, Zeitoun & Chater, 2014; Goodman & Frank, 2016). But, in practice, communication is often unsuccessful, especially when measured against higher-level goals such as enjoyment, or impression management. Examples of unsuccessful communication may better reveal the social mind, and motivate conversational interventions that help people communicate more effectively together.

References

Ames, D. R. (2004). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of personality and social psychology*, *87*(5), 573.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. Personality and Social Psychology Bulletin, 23(4), 363-377.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. Current Directions in Psychological Science, 20(5), 286-290.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. Journal of Consumer Psychology, 24(4), 586-607.

Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? Journal of Marketing Research, 48(5), 869-880.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382-386.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Brooks, A.W. (2013). Get excited: Reappraising pre-performance anxiety as excitement. *Journal of Experimental Psychology: General, 143(3),* 1144-1158.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage.

Bruneau, E.G. & Saxe, R. (2012). The power of being heard : The benefits of 'perspective-giving' in the context of intergroup conflict. *Journal of Experimental Social Psychology, 48(4) :* 855-866.

Bruneau, E.G., Cikara, M., & Saze, R. (2015). Minding the gap : Narrative descriptions about mental states attenuate parochial empathy. *PLoS ONE, 10(10).*

Clark, H.H., & Schaefer, E.F. (1989). Contributing to discourse. *Cognitive science, 13(2),* 259-294.

Cooney, G., Abi-Esber, N., Mastroianni, A., & Brooks, A.W. (2019). The many minds problem: Disclosure in dyadic vs. group conversation. *Current Opinion in Psychology*.

Coupland, J. (2014). *Small talk. Routledge.*

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. Econometrica: Journal of the Econometric Society, 1431-1451.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1-17.

Diener, E., & Seligman, M. E. (2002). Very happy people. *Psychological science*, *13*(1), 81-84.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., ... & Prabhumoye, S. (2019). *The Second Conversational Intelligence Challenge (ConvAI2). arXiv preprint arXiv:1902.00098.*

Drew, P., & Holt, E. (1998). Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in society, 27(4),* 495-522.

Dunbar, R. I., Marriott, A., & Duncan, N. D. (1997). Human conversational behavior. *Human nature, 8(3),* 231-246.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, *23*(2), 283.

Epley, N. (2008). Solving the (real) other minds problem. *Social and personality psychology compass*, *2*(3), 1455-1474.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. Journal of personality and social psychology, 87(3), 327.

Epley, N., Caruso, E. M., & Bazerman, M. H. (2006). When perspective taking increases taking: reactive egoism in social interaction. *Journal of personality and social psychology, 91*(5), 872.

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. Journal of Experimental Psychology: General, 143(5), 1980.

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. Journal of personality and social psychology, 114(4), 547.

Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. & Ostendorf, M. (2018). Sounding Board: A User-Centric and Content-Driven Social Chatbot. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations(pp. 96-100).

Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. Psychological science, *11(6),* 481-486.

Filipowicz, A., Barsade, S., & Melwani, S.(2011). Understanding emotional transitions : The interpersonal consequences of changing emotions in negotiations. *Journal of Personality and Social Psychology, 101(3),* 541.

Friend, W. & Malhotra, D. (2019). Psychological barriers to resolving intergroup conflict : An extensive review and consolidation of the literature. *Negotiation Journal,* 407-442.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33(1),* 1-22.

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological science, 17*(12), 1068-1074.

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological science, 19*(4), 378-384.

Galley, M., McKeown, K., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 562-569).

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. *Science, 323(5921),* 1617-1619.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology, 75(2),* 332.

Gilovich, T., & Savitsky, K. (1999). The spotlight effect and the illusion of transparency: Egocentric assessments of how we are seen by others. *Current Directions in Psychological Science, 8(6),* 165-168.

Goel, S., Mason, W., & Watts, D. J. (2010). Real and perceived attitude agreement in social networks. Journal of Personality and Social Psychology, 99(4), 611.

Goldman, A. I. (2006). Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press.

Goodman, N. D., & Frank, M.C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science, 20(11),* 818-829.

Greenberg, J. (1983). Overcoming egocentric bias in perceived fairness through self-awareness. *Social Psychology Quarterly, 46(2),* 152-156.

Hameiri, B. & Nadler, A. (2017). Looking backward to move forward. *Personality and Social Psychology Bulletin, 43(4),* 555-569.

Hardin, C. D., & Higgins, E. T. (1996). Shared reality: How social verification makes the subjective objective. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition, Vol. 3*. Guilford Press.

Hearst, M. A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics, 23(1),* 33-64.

Hirschberg, J., & Manning, C.D. (2015) Advances in natural language processing. *Science, 349(6245),* 261-266.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53(2),* 221.

Hopthrow, T., Hooper, N., Hooper, L., Meier, P., & Weger, U. (2017). Mindfulness reduces the correspondence bias. *Quarterly Journal of Experimental Psychology, 70(3),* 351-360.

Grice, H. P. (1967). *Logic and conversation.*

Huang, K., Yeomans, M., Brooks, A.W., Minson, J., & Gino, F. (2017). It doesn’t hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology, 113(3),* 430-452.

Ickes, W. J. (Ed.). (1997). *Empathic accuracy*. Guilford Press.

Jacobi, T., & Schweers, D. (2017). Justice, interrupted: The effect of gender, ideology, and seniority at Supreme Court oral arguments. *Virginia Law Review*, 1379-1485.

Jurafsky, D., & Martin, J. H. (2017). Speech and language processing (Vol. 4). London:: Pearson.

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.

Krienen, F. M., Tu, P. C., & Buckner, R. L. (2010). Clan mentality: evidence that the medial prefrontal cortex responds to close others. *Journal of Neuroscience, 30*(41), 13906-13915.

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: an ineradicable and ego-centric bias in social perception. *Journal of Personality and Social Psychology, 67(4),* 596.

Kumar, A., & Gilovich, T. (2015). Some “thing” to talk about? Differential story utility from experiential and material purchases. *Personality and Social Psychology Bulletin, 41(10),* 1320-1331.

Kushlev, K., Heintzelman, S.J., Oishi, S., & Diener, E. (2018). The devlining marginal utility of social time for subjective well-being. *Journal of Research in Personality. 74,* 124-140.

Larsen, J.T., McGRaw, A.P., Cacioppo, J.T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and social psychology, 81(4),* 684.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological bulletin, 125(2),* 255.

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.

McAuliffe, W., Carter, E.C., Berhane, J., Snihur, A.C., & McCullough, M.E. (2019). Is empathy the default response to suffering? A meta-analytic evaluation of perspective taking’s effect on empathic concern. *Personality and Social Psychology Review.*

Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological science*, *21*(4), 539-541.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*(5834), 82-82.

Mendelberg, T., Karpowitz, C. F., & Oliphant, J. B. (2014). Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics*, *12*(1), 18-44.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in Pre-Training Distributed Word Representations. *arXiv preprint arXiv:1712.09405.*

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. Trends in cognitive sciences, 18(10), 512-519.

Nasie, M., Bar-Tal D., Pliskin, R., Nahhas, E., & Halperin, E. (204). Overcoming the barrier of narrative adherence in conflicts through awareness of the psychologyical bias of naive realism. *Personality and Social Psychology Bulletin, 40(11), 1543-1556.*

Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning, 95(3),* 381-421.

Overbeck, J. R., & Droutman, V. (2013). One for all: Social power increases self-anchoring of traits, attitudes, and emotions. *Psychological Science*, *24*(8), 1466-1476.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics, 23(1),* 103-139.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences, 27(2),* 169-190.

Quoidbach, J., Taquet, M., Motjoye, Y., & Gross, J.J. (2019). Happiness and social behavior. *Psychological Science.*

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science, 58(4),* 1064-1082.

Ross, L., & Ward, A. (1995). Psychological barriers to dispute resolution. In *Advances in experimental social psychology* (Vol. 27, pp. 255-304). Academic Press.

Rees, L., Rothman, N.B., Lehavy, R., & Sanchez-Burks, J. (2013). The ambivalent mind can be a wise mind: Emotional ambivalence increases judgment accuracy. *Journal of Eperimental Social Psychology, 49(3),* 360-367.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7-55). Academic Press.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*(1), 269-273.

Schaerer, M., Kern, M., Berger, G., Medvec, V., & Swaab, R.I. (2018). The illusion of transparency in performance appraisals : When and why accuracy motivation explains unintentional feedback inflation. *Organizational Behavior and Human Decision Processes, 144,* 171-186.

Schegloff, E. A. (2007). *Sequence organization in interaction: Volume 1: A primer in conversation analysis (Vol. 1).* Cambridge University Press.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, *23*(8), 842-847.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*(1), 1-24.

Shiller, R. J. (1995). Conversation, information, and herd behavior. *The American Economic Review, 85(2),* 181-185.

Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring beliefs in centimeters: Private knowledge biases preschoolers' and adults' representation of others' beliefs. *Child Development, 84*(6), 1846-1854.

Sprecher, S., Treger, S., Wondra, J. D., Hilaire, N., & Wallpe, K. (2013). Taking turns: Reciprocal self-disclosure promotes liking in initial interactions. *Journal of Experimental Social Psychology, 49(5),* 860-866.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B:Methodology, 36,* 111–147.

Sun, J., Harris, K., & Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions? *Journal of personality and social psychology, in press.*

Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General, 142*(1), 151.

Todd, A. R., Hanko, K., Galinsky, A. D., & Mussweiler, T. (2011). When focusing on differences leads to similar perspectives. *Psychological Science*, *22*(1), 134-141.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of personality and social psychology*, *98*(2), 281.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological science, 18*(7), 600-606.

Wu, S., Barr, D. J., Gann, T. M., & Keysar, B. (2013). How culture influences perspective taking: differences in correction, not integration. *Frontiers in human neuroscience, 7,* 822.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036-1040.

Yeomans, M., Kantor, A., & Tingley, D. (2018). The politeness Package: Detecting Politeness in Natural Language. *R Journal*, *10*(2).

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making, 32(4),* 403-414.

Yeomans, M., Minson, J., Collins, H., Chen, F. & Gino, F. (2019). Conversational Receptiveness: Improving Engagement with Opposing Views. *Organizational Behavior and Human Decision Processes, in press.*

Yip, J. A., & Schweitzer, M. E. (2019). Losing your temper and your perspective: Anger reduces perspective-taking. *Organizational Behavior and Human Decision Processes, 150*, 28-45.

Zajonc, R. B. (1960). The process of cognitive tuning in communication. *The Journal of Abnormal and Social Psychology, 61(2),* 159.

Zhou, H., Majka, E. A., & Epley, N. (2017). Inferring perspective versus getting perspective: Underestimating the value of being in another person’s shoes. *Psychological science*, *28*(4), 482-493.