Cognition 193 (2019) 104011

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

How do you know that? Automatic belief inferences in passing conversation

Paula Rubio-Fernández[a,b,*], Francis Mollica[c], Michelle Oraa Ali[a], Edward Gibson[a]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States
[b] Department of Philosophy, University of Oslo, Norway
[c] Department of Brain and Cognitive Sciences, University of Rochester, United States

ARTICLE INFO

Keywords:
Theory of Mind
False-belief tasks
Automatic and controlled processes
Pragmatics
Belief inferences

ABSTRACT

There is an ongoing debate, both in philosophy and psychology, as to whether people are able to automatically infer what others may know, or whether they can only derive belief inferences by deploying cognitive resources. Evidence from laboratory tasks, often involving false beliefs or visual-perspective taking, has suggested that belief inferences are cognitively costly, controlled processes. Here we suggest that in everyday conversation, belief reasoning is pervasive and therefore potentially automatic in some cases. To test this hypothesis, we conducted two pre-registered self-paced reading experiments (N1 = 91, N2 = 89). The results of these experiments showed that participants slowed down when a stranger commented 'That greasy food is bad for your ulcer' relative to conditions where a stranger commented on their own ulcer or a friend made either comment – none of which violated participants' common-ground expectations. We conclude that Theory of Mind models need to account for belief reasoning in conversation as it is at the center of everyday social interaction.

1. Introduction

A simple communicative act such as deciding who to ask a question requires estimating other people's knowledge. For example, if you got lost in a new city, you may ask a passerby for directions, but if you could not remember your mother's birthday, you would not ask a random passerby on the street. Estimating another person's knowledge (also known as *belief reasoning*) is a key component of human Theory of Mind: our capacity to interpret and predict other people's behavior by reference to their mental states. The large majority of Theory of Mind studies in the last 30 years have investigated the development of belief reasoning in laboratory tasks where a protagonist holds a false belief (e.g., about the location of an object) and the child has to predict the protagonist's course of action, without defaulting to their own knowledge of the situation (Scott & Baillargeon, 2017; Wimmer & Perner, 1983). Rather than using a false-belief task, the present study tried to recreate belief reasoning in conversation.

Experimental pragmatics studies have long acknowledged the role of belief reasoning in communication (e.g., Brennan, Galati, & Kuhlen, 2010; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986; Heller, Gorman, & Tanenhaus, 2012). Researchers examining narrative texts have also investigated the degree to which readers keep track of story characters' beliefs during their dialogues (e.g., Gerrig, Brennan, & Ohaeri, 2001; Graesser, Bowers, Bayen, & Hu, 2000; Lea, Mason, Albrecht, Birch, & Myers, 1998; Weingartner & Klin, 2005). Social cognition research, on the other hand, has not investigated belief reasoning in conversation. We will frame the present study from a Theory of Mind perspective in order to show how the study of belief reasoning in dialogue can advance theoretical debates in social cognition research.

By relying on false-belief tasks to investigate humans' capacity for belief reasoning, theoretical models of Theory of Mind fall short of explaining the data observed in everyday communication. Thus, current accounts aim to explain infants' and children's performance in false-belief tasks (e.g., Butterfill & Apperly, 2013; Helming, Strickland, & Jacob, 2014; Heyes & Frith, 2014; Ruffman, 2014; Scott & Baillargeon, 2017) but do not explain the Theory of Mind development reported in the pragmatics literature, which does not always parallel false-belief performance: communication studies often suggest that toddlers have an immature understanding of other people's knowledge (e.g., Dunham, Dunham, & O'Keefe, 2000; Matthews, Lieven, Theakston, & Tomasello, 2006; Moll, Carpenter, & Tomasello, 2011). The reliance on false-belief tasks is also problematic for theoretical models of adult social cognition that fail to account for belief reasoning in conversation (e.g., Apperly & Butterfill, 2009; Heyes, 2014) and as a result leave out of their scope the bulk of the data observed in everyday social interaction. Here we argue that belief reasoning is pervasive in communication and therefore needs to be investigated and characterized in conversational settings, and not

* Corresponding author.
E-mail address: prubio@mit.edu (P. Rubio-Fernández).

https://doi.org/10.1016/j.cognition.2019.104011
Received 22 September 2018; Received in revised form 18 June 2019; Accepted 19 June 2019
Available online 27 June 2019

only in false-belief tasks.

2. Does communication involve belief reasoning?

A popular theoretical view, both in philosophy and psychology, is the hypothesis that attributing mental states to others (not only beliefs, but also intentions and desires) is too cognitively demanding to be the basis for real-time social interaction (e.g., Bermúdez, 2003; Gallagher, 2001; Gauker, 2003; Heyes, 2014; Millikan, 2005; Pickering & Garrod, 2004; cf. Borg, 2018). Geurts and Rubio-Fernández (2015) have argued that this theoretical view is often based on introspection, which is not reliable when estimating the potential complexity of a mental process. Perceiving and recognizing a chair, for example, may seem intuitively simple, yet research on visual cognition shows that such inferences are highly complex. More importantly, no theoretical account has yet offered an explanation as to how adult conversation works the way it does if it does not rely on belief reasoning (cf. Grice, 1989; Sperber & Wilson, 1995). For example, no such account explains how we make informative contributions to ongoing conversations, remind each other of upcoming events or pre-empt a misunderstanding (all of which require some form of mindreading).

In line with the general view that Theory of Mind inferences are cognitively costly, Apperly and Butterfill (2009) and Butterfill and Apperly (2013) characterize belief reasoning as a "System 2" type of reasoning: by definition, reasoning that is slow, controlled, flexible, resource-demanding and effortful. They give the example of anticipating what a group of students might know in preparation for a lecture or working out afterwards how one had misjudged their expertise, both of which require deliberative belief-reasoning (Apperly & Butterfill, 2009:966). To challenge the view that belief reasoning may be the basis for real-time communication, Apperly and Butterfill (2009) and Apperly (2018) discuss Keysar, Lin, and Barr (2003) and Epley, Keysar, Van Boven, and Gilovich (2004) studies using the *director task*, in which a participant follows the instructions of a confederate to move around various objects in a vertical grid of squares. The confederate sits on the other side of the grid and cannot see all of the objects because some of the cells are occluded on her side. Crucially, the confederate is supposed to be ignorant of the contents of those cells, and when she asks the participant to 'move the small candle,' for example, the smallest of three candles is visible only to the participant. Over a long series of studies, participants have shown a tendency to consider, and sometimes even reach for, the smallest candle in their privileged view before picking up the medium-sized candle in open view (i.e. the one intended by the confederate).

Keysar et al. (2003) and Epley et al. (2004) interpreted these results as evidence for an egocentric bias in communication, whereby people suffer interference from their own perspective when deriving Theory of Mind inferences about their interlocutors. However, other studies using the same paradigm have challenged the view that language comprehension is initially insensitive to perspective taking (e.g., Hanna & Tanenhaus, 2004; Heller, Grodner, & Tanenhaus, 2008). Furthermore, Rubio-Fernández (2017) provided evidence for the view that the director task is an unnatural task that requires selective attention, and not necessarily Theory of Mind. Thus, whereas in everyday life people normally know about and refer to entities that they cannot see, participants in the director task must assume that the speaker only knows about the objects in her visual field, which poses artificial constraints when interpreting her instructions.

Rather than establishing a priori which objects the director knows about, Rubio-Fernández and Jara-Ettinger (2018) have tried to create more naturalistic versions of this task where the participant is supposed to infer which objects are in common ground depending on how the director refers to those objects. In addition, they proposed a computational model that jointly infers which object the director is referring to and which objects she can and cannot see in a display given her instructions. Model predictions closely mirrored human data, suggesting that belief inferences can be derived as part of the pragmatic process of reference assignment. In other words: people appreciate what others know as part of the process of understanding what they mean, which suggests that belief reasoning need not be an optional, controlled process.

In summary, theoretical discussions on the complexity of mental state reasoning have often been based on introspection, whereas the conclusions from empirical studies have sometimes disregarded the specific task demands of paradigms investigating common ground. Lin, Keysar, and Epley (2010), for example, showed that performance on the standard version of the director task relied on participants' attentional resources. However, rather than concluding that participants' poor performance in the director task might be related to their executive control (rather than to their Theory of Mind, as is generally assumed), Lin et al. interpreted their results as evidence that using Theory of Mind in communication is cognitively costly. While we agree that the setup of the director task is likely to tax executive control, we disagree that that particular task be representative of common ground use in normal conversation (Rubio-Fernández, 2017). Therefore, concluding from the results of the director task that people have difficulties using Theory of Mind in everyday communication seems unwarranted.

3. Can belief inferences be derived automatically in conversation?

Apperly (2011:95) argues that belief inferences may be derived spontaneously, but not automatically: "spontaneous belief inferences require some motivation. In an experiment this might be the frequency of judgements about belief. In real life, I am sure that people are frequently motivated to infer what others are thinking. But in the absence of such motivation there is no evidence at all that beliefs are inferred". Apperly (2011, 2018) also claims that experimental paradigms tapping the derivation of pragmatic inferences have in-built incentives to motivate participants to engage in pragmatic reasoning. For example, McKoon and Ratcliff (1986) asked participants to read sentences such as 'The woman, desperate to get away, ran to the car and jumped in' and then later decide whether the word 'driving' had appeared in the sentence. Participants often responded positively, suggesting they had inferred that the woman drove away in her car. However, according to Apperly (2011:92), these inferences were triggered by the participants' awareness of the ensuing memory test: without such motivation, these kinds of inferences are not automatically derived.

While some forms of belief reasoning may be deliberative (e.g., assessing an audience's expertise when preparing a talk) and others may be spontaneous (e.g., as when poker players try to guess what the others are thinking), here we want to challenge the view that belief inferences cannot be derived automatically and propose that everyday conversation is the natural arena for testing such a hypothesis. Consider the following example (adapted from Geurts & Rubio-Fernández, 2015; Rubio-Fernández, 2017): imagine that you are eating at a restaurant when a customer at another table tells you 'That greasy food is terrible for your ulcer.' If this person were a stranger, his comment would immediately strike you as creepy. The reason why you would react with unpleasant surprise is that you would have automatically inferred that this man knew about your health, which is unexpected. That also explains why you would have reacted differently if your best friend had made the same passing remark.

In pragmatics terminology, the stranger's comment would have violated your *common-ground expectations*: we normally assume a certain amount of shared knowledge with our interlocutors (which may range from today's weather to very personal information, depending on how well we know each other), and the stranger's comment would immediately suggest that your common ground was much more extensive than you had first assumed. Our surprise in this situation suggests that we monitor common ground by default and a violation of our expectations automatically triggers a belief update. We investigated this hypothesis using a self-paced reading task, but we assume that if a

stranger commented on our personal life in a real-life situation, one would also react with surprise.

4. How and when do we use common ground in conversation?

The violations of common ground tested in this study are particularly striking because they involve strangers having knowledge of one's personal life, which is a rare experience. However, equally salient violations of common ground could be observed when communicating with people we know, so long as they reveal clearly unexpected knowledge (e.g., if your boss commented on the nightmares you had last night). Admittedly, though, establishing whether something is in common ground or not is not always straightforward: we do not always remember whether we have shared a certain piece of information with our interlocutors (e.g., 'Did I tell you that I got a promotion?'), and likewise, we sometimes assume our interlocutors know more, or less, than we thought (e.g., 'Oh, sorry, I assumed Jake would have told you!' vs 'Oh, I didn't know they had already informed you'). It is therefore not possible to generalize from the clear-cut cases and assume that common ground has set boundaries in every interaction. However, what the clear-cut cases (such as conversations with strangers) should allow us is to test whether violations of common ground can be detected automatically, which has theoretical implications for models of Theory of Mind that claim it is not possible.

It must be noted that while the results of this study should have implications for models of Theory of Mind, it is beyond the scope of this paper to elucidate how common ground is used in communication. The only view we are defending here is that common ground is monitored *by default* in conversation, and so when a violation is detected, that automatically triggers a belief update. Our basis for adopting this position is that speakers must take into account who they are talking to in order to make themselves understood (e.g., whether their interlocutor knows the person they are talking about, or they need to first introduce this person in the conversation), and likewise, listeners must interpret language in relation to their common ground with the speaker if they want to make sense of a message (e.g., by interpreting a person's name as referring to someone mutually known). Thus, monitoring common ground may be more or less demanding of cognitive resources and violations may be more or less accessible depending on the interlocutor and the context of the conversation, but once a common ground violation is detected, a belief inference should be automatically triggered.

The intended contribution of this study is therefore to provide empirical evidence calling for an analysis of common ground that goes beyond the poor performance often observed in the director task. Given the specific task demands of that paradigm, the results of those studies should not suffice to conclude that belief reasoning is cognitively costly and optional (cf. Apperly & Butterfill, 2009; Keysar et al., 2003; Lin et al., 2010). As we pointed out in the introduction, the simplest speech act requires keeping track of who knows what in conversation, a requirement that should make Theory of Mind experts keenly interested in dialogue, rather than leaving that research field to the linguists and the pragmatists (Rubio-Fernández, 2019).

5. Experiment 1

We used self-paced reading to evaluate the automatic vs. spontaneous views of belief reasoning in a simulated dialogue. In an early study, Lea et al. (1998) observed that participants slowed down their reading when a story character used a pronoun that was not anchored in common ground. Similarly, in our study, the automatic view predicts that when strangers refer to the participants' personal life (e.g., 'That greasy food is terrible for your ulcer'), participants' reading times would be slow relative to conditions where strangers referred to their own lives (e.g., 'for my ulcer'), or where friends referred to either themselves or the participant. We interpret these longer reading times as surprise, or more generally, as an index of cognitive effort revealing an infelicity. In contrast, the spontaneous view predicts not such difference in the absence of specific motivation to figure out what the speaker is thinking.

The critical argument behind the automatic-inference hypothesis is that belief reasoning can be triggered in conversation without needing to induce an inquisitive mood in the interlocutors, or experimentally motivate them to figure out what another person is thinking (cf. Apperly, 2011). By contrast, the spontaneous-inference hypothesis would predict slower reading times when strangers refer to the participant's personal life only if participants were asked to compute the speaker's beliefs (see, e.g., Apperly, Back, Samson, & France, 2008). Since we did not ask participants to figure out what the speaker knew in each scenario, the spontaneous-inference hypothesis would predict comparable reading times in all conditions.

5.1. Methods

5.1.1. Participants

Ninety-two participants were recruited through Amazon Mechanical Turk with the goal of retaining 80 participants. One of the recruited participants did not complete the task and their data were not used. Simulated power analyses based on a pilot study (see Supplementary Materials) indicated this sample size would have greater than 80% power to detect our effect, while maintaining a false positive rate of less than 5%. The task took approximately 30 min and participants were paid $3. Recruitment was limited to participants located in the US territory (according to their IP address) and who had a 95% reliability rate from previous performance on MTurk tasks.

5.1.2. Materials

Materials consisted of short vignettes made of two sentences: (i) a context sentence, which described a scenario in a public space where 'you' (the participant reading) were co-present with another person (the speaker); and (ii) a comment sentence addressed to you by the speaker. There were 24 items in this format in a 2 × 2 design crossing (1) the participant's relation to the speaker in the vignette (*stranger* vs. *friend*) and the pronoun reference in the target sentence (*my* vs. *your*). Comments were written in the 1st and 2nd person in order to put participants in a position more akin to that of a conversational partner.

Table 1
Sample item in the four conditions of the experiment.

Relation	Pronoun reference	Context	Comment
Friend	Their-life	You are having dinner with your dad at a restaurant when he says:	This greasy food/is terrible/for my ulcer/but it's an/old favorite and/those are hard/to give up.
Stranger	Their-life	You are having dinner at a restaurant when a customer at another table says:	
Friend	Your-life	You are having dinner with your dad at a restaurant when he says:	This greasy food/is terrible/for your ulcer/but it's an/old favorite and/those are hard/to give up.
Stranger	Your-life	You are having dinner at a restaurant when a customer at another table says:	

Note. Regions in the comment are separated by/marks.

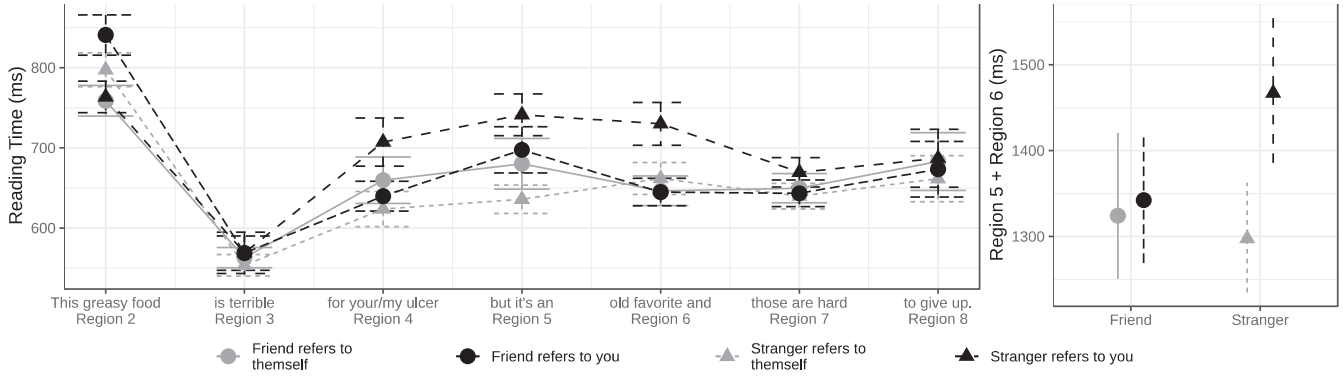

Fig. 1. Left plot: Mean Reading Times for critical conditions by region. Region 1 corresponds with the context sentence and is not included in the graph because it was presented at once and reading times were considerably longer. Error bars reflect SEM. Right plot: Summed reading times for Regions 5 and 6 combined. Line ranges reflect 95% bootstrapped confidence interval.

Table 1 shows a sample item in the four conditions.

The seventh word of the comment sentence was always the critical pronoun (*my*, *your*). Target sentences were divided into multi-word regions for presentation as follows: Region 1 was the context sentence (*You are having dinner with your dad at a restaurant when he says:* in Table 1) and was presented all at once. Region 2 was the first three words of the comment sentence (*This greasy food*). Region 3 was the next two words (*is terrible*). Region 4 was the next three words (*for my/your ulcer*), whose second word was always the critical pronoun. This was the first region where we might have been able to detect a critical difference in our conditions. The following regions (5–8 in Table 1 and up to as many as 9 for an item) were all three words each and continued to refer to either the speaker or the participant's life (*but it's an/old favorite and/those are hard/to give up*). Regions sometimes made up syntactic constituents (as in Region 2 in Table 1, *This greasy food*) but sometimes did not (as in Region 6, *old favorite and*), but importantly, this was always the same across conditions.

Critical items were divided in four separate lists using a Latin square design so that each participant would see only one version of each item and six critical items from each condition. The two critical variables (speaker relation and pronoun reference) were therefore manipulated within participants.

We also created 24 filler items, which were used in all four lists, for a total of 48 items per list. Matching the targets, there was also a speaker in the filler trials, half of whom were strangers and half friends. To ensure that participants were paying attention to their reading, they were asked a Yes/No comprehension question at the end of each trial. For half of the critical items, the comprehension question (whether the speaker was a stranger (Yes/No) or someone they knew (Yes/No)), also served as a manipulation check. Yes/No responses were counterbalanced across items and conditions. All materials, code, data and our pre-registration can be found at OSF (https://osf.io/zq3dg/).

5.1.3. Procedure

On each trial, participants were presented with sequences of hyphens (e.g., '–') marking the word-regions of the vignette. Participants were instructed to press the space bar to read through the vignette region by region following a moving window procedure. At the end of each vignette, participants were presented with a Yes/No comprehension question. They made their response using two keys on their keyboard. Stimulus presentation and response collection were controlled by Ibex Farm, a Web-based experimental platform for self-paced reading (Drummond, 2013).

Participants were randomly assigned to a stimulus list, and the order of presentation of the critical and filler items was randomized individually. Participants completed a total of 48 trials. An extra three practice items were included at the start of the task so that participants could accustom to reading on Ibex Farm. Practice and filler trials were not analyzed.

5.2. Results and discussion

The automatic-inference hypothesis predicted that the stranger/your-life condition would violate participants' expectations about what other people know about them. If participants monitor their interlocutor's knowledge states by default, they should incur processing costs for updating their common ground in the stranger/your-life condition, resulting in an interaction between Pronoun (my/your) and Relation (Friend/Stranger) such that items in the stranger/your-life condition are read slower than items in the other three conditions. If participants do not keep track of others' knowledge by default, then participants should demonstrate no difference in reading times between conditions (i.e., no interaction).

According to the results of our pilot study, reading time differences should be observed after the critical pronoun in Regions 5 and/or 6. We interpret this delay as due to (i) spillover-effects, which are commonly observed in self-paced reading tasks where slower responses are often observed a few words after the critical region (e.g., Smith & Levy, 2013); and (ii) as a form of pragmatic reasoning, belief inferences may take a second or two to be derived, rather than being computed instantaneously (e.g., Noveck, Bianco, & Castry, 2001; Noveck & Posada, 2003).

As per our pre-registration, we removed from further analyses all participants who self-reported as non-native English speaker (n = 3) and whose accuracy in the comprehension questions was less than 80% on average (n = 2). The remaining 86 participants still provide adequate power to detect our effect. Overall, participants were 94% accurate on the task. Reading times ± 2 SD away from each participant's mean reading time were removed from further analyses to ensure that our conclusions were not driven by outliers. As a result, 4.4% of the data was not subject to further analyses. We report our analyses following our pre-registration. However, the trends hold even when all data are retained (see Supplementary Materials).

Fig. 1 displays the average reading times for each region and the combined reading times for our regions of interest. As can be seen in the rightmost panel, participants were slower to read our regions of interest in the stranger/your-life condition. Following our pre-registration, we conducted three linear mixed effect regression models predicting Reading Time (ms) for (1) Region 5, (2) Region 6 and (3) the sum of Regions 5 and 6, with fixed effects for Relation and Pronoun, and their interaction. The maximal random effect structure was utilized — i.e. random intercepts and slopes for Item and Participant. Relation and Pronoun were sum coded with the friend/their-life condition as reference level. The model was fit using the lme4 package (Bates, Machler, Bolker, & Walker, 2015) in R (R Core Team, 2018). The parameter estimates can be found in Table 2.

Table 2
Coefficients and t-values from linear mixed effect models.

	Estimate	Std. Error	*t* Statistic
5th Region			
Intercept	688.93*	39.94	17.25
Pronoun	61.89*	25.66	2.41
Relation	0.09	27.22	0.00
Pronoun × Relation	88.77	52.66	1.69
6th Region			
Intercept	670.84*	39.56	16.96
Pronoun	34.01	19.85	1.71
Relation	50.79*	21.79	2.33
Pronoun × Relation	70.14*	35.08	2.00
5th + 6th Region			
Intercept	1357.92*	76.22	17.82
Pronoun	94.21*	36.40	2.59
Relation	49.07	36.20	1.36
Pronoun × Relation	151.40*	63.01	2.40

Asterisks denote significance level.

As predicted, we found a significant interaction such that reading times for the stranger/your-life condition were significantly longer than reading times to the other three conditions for both Region 6 and Regions 5 and 6 combined, consistent with a default preference to monitor our interlocutor's knowledge states. We did not interpret the main effects as they contradict across regions and are most likely driven by the predicted interaction.

A reviewer suggested that we carried out parallel analyses over Region 4 and Regions 4 and 5 combined in order to include the critical pronoun in the analysis window. As in the pre-registered analyses, we found significant Pronoun × Relation interactions in both Region 4 ($\beta = 103$, $t = 1.93$) and Regions 4 and 5 combined ($\beta = 189$, $t = 2.82$), driven by the longer reading times observed in the stranger/your-life condition. Therefore, the analyses of the pronoun region further support the automaticity hypothesis (for the regression coefficients of these analyses, see Supplementary Materials).

Participants in Experiment 1 slowed down their reading when a stranger in the vignette referred to their personal life, compared to other conditions where the same stranger referred to their own life, or the speaker was a friend. These results support the hypothesis that people can derive automatic belief inferences in conversation, without requiring specific motivation to reason about what their interlocutors know.

6. Experiment 2

The results of our first experiment suggest that people can derive belief inferences automatically in conversation, and not only spontaneously. However, in half of the experimental trials (12 out of a grand total of 48), participants in Experiment 1 were asked whether they knew the speaker or not. This manipulation was intended to ensure that participants kept track of who the speaker was in each vignette, but it is also possible that by drawing participants' attention to their relationship with the speaker, considerations of common ground may have been made more salient, resulting in the derivation of spontaneous (rather than automatic) inferences. Thus, in Experiment 2, we tried to replicate the results of Experiment 1 but asking participants general comprehension questions throughout the task, without drawing their attention to the speaker in any trial.

As in Experiment 1, we conducted the pre-registered analyses over Region 5, Region 6 and Regions 5 and 6 combined, plus the earlier analyses suggested by a reviewer over Region 4 (including the critical pronoun) and Regions 4 and 5 combined.

6.1. Methods

6.1.1. Participants

Eighty-nine participants were recruited through Amazon Mechanical Turk with the goal of retaining 80 participants. As in Experiment 1, recruitment was limited to participants located in the US territory and who had a 95% reliability rate from previous performance on MTurk tasks.

6.1.2. Materials and procedure

The same materials and procedure that had been employed in Experiment 1 were used again in Experiment 2, with the exception of the 12 comprehension questions that had probed participants about their relation to the speaker. Those questions were replaced by general comprehension questions, similar to the other ones used in the task.

6.2. Results and discussion

As per our pre-registration, we removed from further analyses all participants who self-reported as non-native English speakers (n = 2) and whose accuracy in the comprehension questions was less than 80% on average (n = 7). The remaining 80 participants still provide adequate power to detect our effect. Overall, participants were 95% accurate on the task. Reading times ± 2 SD away from each participant's mean reading time were removed from further analyses to ensure that our conclusions were not driven by outliers. As a result, 4.2% of the data was not subject to further analyses. We report our analyses following our pre-registration. However, the trends hold even when all data are retained (see Supplementary Materials).

Fig. 2 displays the average reading times for each region and the combined reading times for our regions of interest. As can be seen in the rightmost panel, participants were slower to read our regions of interest

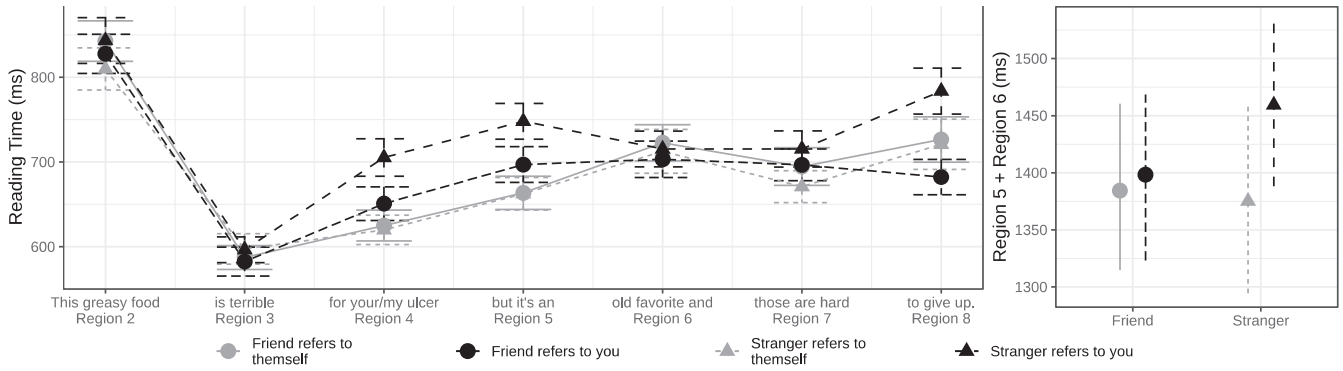

Fig. 2. Left plot: Mean Reading Times for critical conditions by region. Region 1 corresponds with the context sentence and is not included in the graph because it was presented at once and reading times were considerably longer. Error bars reflect SEM. Right plot: Summed reading times for Regions 5 and 6 combined. Line ranges reflect 95% bootstrapped confidence interval.

Table 3
Coefficients and t-values from linear mixed effect models. Asterisk denotes statistical significance.

	Estimate	Std. Error	*t* Statistic
5th Region			
Intercept	695.61*	38.86	17.90
Pronoun	59.29*	17.82	3.33
Relation	17.92	17.81	1.01
Pronoun × Relation	66.55*	32.09	2.07
6th Region			
Intercept	718.08*	44.02	16.31
Pronoun	−11.25	18.71	−0.60
Relation	9.91	18.14	0.55
Pronoun × Relation	20.83	42.99	0.48
5th + 6th Region			
Intercept	1410.12*	80.41	17.54
Pronoun	48.98	24.77	1.98
Relation	29.52	27.36	1.08
Pronoun × Relation	84.99	52.49	1.619

in the stranger/your-life condition. Following our pre-registration, we conducted the same analyses as in Experiment 1 (see Table 3 for the parameter estimates).

As predicted, we found a significant interaction such that reading times for the stranger/your-life condition were significantly longer than reading times to the other three conditions for Region 5, consistent with a default preference to monitor our interlocutor's knowledge states. Again, we did not interpret the main effects as they contradict across regions and are most likely driven by the predicted interaction.

Parallel analyses on Region 4 and Regions 4 and 5 combined (i.e. shifting our analysis window one region earlier to include the critical pronoun) revealed similar results in the pronoun region ($\beta = 62$, $t = 1.88$) and in Regions 4 and 5 combined ($\beta = 129$, $t = 2.65$), further supporting the automaticity hypothesis. The regression coefficients for these analyses are provided in the Supplementary Materials.

As in Experiment 1, participants slowed down their reading when a stranger in the vignette commented on their personal life, compared to other conditions where the same stranger commented on their own life, or the participant knew the speaker. Importantly, this pattern of results was replicated without probing participants to keep track of whom the speaker was, offering stronger support to the automaticity hypothesis. Interestingly, this effect emerged again earlier than we had predicted, revealing longer reading times in the region including the critical pronoun and not only in the spillover regions. This suggests that deriving belief inferences does not require slow controlled reasoning. In sum, the results of Experiment 2 offer further support to the hypothesis that people can derive automatic belief inferences in conversation, without requiring specific motivation to reason about what their interlocutors know.

7. General discussion

The results of our studies suggest that people can derive belief inferences automatically when reading a dialogue. If someone makes a passing remark about some private matter in a conversation, we understand that they know about our personal life. This kind of belief inference makes up our common ground with our closest interlocutors, and normally goes unnoticed in conversations with friends and family. However, when our participants noticed that a stranger had remarked on their personal lives, the same inference was unexpected, which slowed down their reading times, as early as the critical word region. The results of our second experiment suggest that these findings are not an artifact of the comprehension questions used in the first experiment.

It may be argued that using a self-paced reading task defeats the purpose of a study aiming to promote the investigation of belief reasoning in conversation (see also Verga & Kotz, 2019). While we agree that the ideal test case of our proposal would be a study of belief reasoning in naturalistic interaction, testing the difference between the automatic and spontaneous views of belief reasoning requires a highly controlled experimental setup. We therefore see the use of a self-paced reading task mimicking naturalistic dialogue as a methodological compromise that allowed us to test a very specific research question, while generally supporting the view that Theory of Mind use should be investigated in conversation. In addition, there is an extensive literature investigating belief reasoning in narrative comprehension, including dialogue, which we take to support our methodology (e.g., Gerrig et al., 2001; Graesser et al., 2000; Lea et al., 1998; Weingartner & Klin, 2005).

In any case, we acknowledge the difficulty of studying naturalistic conversation and even designing laboratory tasks that mirror everyday communication. However, when experimental pragmatics studies do not attain such ambitious goals, researchers should at least acknowledge the specific demands of their paradigms relative to everyday conversation, before drawing conclusions about the limits of human communication from laboratory tasks (cf. Keysar et al., 2003; Lin et al., 2010). In our study, participants had to read recreated dialogues in a self-paced reading task, which is markedly different from naturalistic conversation. However, our point still holds that if the dialogues with strangers took place in real life, people would also react with surprise. Therefore, models of Theory of Mind defending the view that belief reasoning is cognitively costly may need to test their claims in carefully controlled laboratory tasks, but they must also account for belief reasoning in everyday conversation – be that experimentally, or just theoretically.

Like other defendants of the view that belief reasoning is cognitively costly, Apperly (2011) argues that people must be specifically motivated to infer what others are thinking in order to be able to spontaneously infer beliefs. However, our task did not specifically incentivize participants to figure out what the speakers knew or did not know. More generally, our results are not limited to a laboratory setting, or even to conversations with strangers: one would also react with surprise if a close friend made a remark that suggested they knew about a secret (e.g., 'How many people are coming to my surprise birthday party?'). One way to accommodate our results with Apperly's view would be to argue that whenever we engage in conversation, we are intrinsically motivated to figure out what the speaker is thinking (or trying to communicate), in line with mindreading accounts of communication (Grice, 1989; Sperber & Wilson, 1995). Another way, more in line with Apperly's own proposal, would be to try to determine which aspects of communication may become automatized during communicative development. For example, it may be possible to detect common-ground violations automatically through associative memory representations (see Horton & Gerrig, 2005, 2016), whereas other common ground calculations may require more effort.

While uncommitted to any specific view of common ground, we would like to point out that the debate on whether common ground requires the use of Theory of Mind or relies on 'ordinary processes' is a false dichotomy that should be reconsidered. There is no principled reason why associative memory processes may not be recruited in the deployment of our Theory of Mind abilities, especially in highly frequent situations that could lead to the automatization of underlying processes. Thus, building and using common ground may often rely on purely associative processes that allow communicating fast and efficiently, yet that does not mean that the output of those processes are not belief inferences (the hallmark of Theory of Mind use, by all accounts). If your boss unexpectedly commented on the nightmares you had last night, your immediate response would probably be to ask her 'How do you know I had nightmares last night?', yet the speed of your response (or the relative 'unintelligence' of the underlying processes) need not be evidence that you did not use your Theory of Mind. After all, those processes led you to infer that your boss *knew* what you dreamt last night. Therefore, associative memory processes may be at the heart of belief reasoning, rather than being an alternative to using

our Theory of Mind (cf. Apperly & Butterfill, 2009).

While the present results do not settle the debate as to whether Theory of Mind is systematically deployed in human interaction (e.g., Borg, 2018; Heyes, 2014; Millikan, 2005), they suggest that belief inferences may be automatically derived in certain conversational contexts. The passing conversations with strangers recreated in our study made violations of common ground particularly salient compared to conversations with people we know, which probably made belief inferences particularly easy to derive. Future theoretical and experimental work in Theory of Mind should therefore investigate how belief inferences may be derived in different conversational contexts as a way to understand which aspects of belief reasoning may be automatic. Some speech acts, for instance, may lend themselves to automatic belief inferences: when a speaker makes a statement (e.g., 'It's raining'), we normally understand that the speaker believes its contents, which explains why it is pragmatically odd to say 'It's raining but I don't believe it's raining' (what is known as *Moore's paradox*; Moore, 1993). Limitations in memory and attention also suggest that low-level associative processes must play an important role in building and using common ground, especially given the speed and flexibility with which we communicate (see Brown-Schmidt, Yoon, & Ryskin, 2015).

Overall, our results highlight how pragmatic measures of belief reasoning in communication are relevant to models of Theory of Mind. Currently, these models try to account for the results of false-belief tasks and other experimental paradigms, but if they are to explain the bulk of the data from everyday social interaction, they must also account for belief reasoning in conversation.

Acknowledgements

This research was supported by a *Young Research Talent Grant* from the Research Council of Norway awarded to PRF (230718).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.104011.

References

Apperly, I. A. (2011). *Mindreaders: The cognitive basis of theory of mind.* Hove, UK: Psychology Press.
Apperly, I. A. (2018). Mindreading and psycholinguistic approaches to perspective taking: Establishing common ground. *Topics in Cognitive Science, 10*, 133–139.
Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential Theory of Mind task. *Cognition, 106*, 1093–1108.
Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*, 953–970.
Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.
Bermúdez, J. L. (2003). The domain of folk psychology. *Royal Institute of Philosophy Supplement, 53*, 25–48.
Borg, E. (2018). On deflationary accounts of human action understanding. *Review of Philosophy and Psychology, 9*, 503–522.
Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialogue: Coordinating speaking and understanding. *Psychology of learning and motivation: Vol. 53*, (pp. 301–344). Academic Press.
Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. *Psychology of learning and motivation: Vol. 62*, (pp. 59–99). Academic Press.
Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal Theory of Mind. *Mind & Language, 28*, 606–637.
Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, & I. A. Sag (Eds.). *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.
Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1–39.
Dunham, P., Dunham, F., & O'Keefe, C. (2000). Two-year-olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues? *British Journal of Developmental Psychology, 18*, 519–532.
Drummond, A. (2013). Ibex farm. Online server: < http://spellout.net/ibexfarm > .
Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*, 327–339.
Gallagher, S. (2001). The practice of mind. Theory, simulation or primary interaction? *Journal of Consciousness Studies, 8*, 83–108.
Gauker, C. (2003). *Words without meaning.* Cambridge, Massachusetts: MIT Press.
Gerrig, R. J., Brennan, S. E., & Ohaeri, J. O. (2001). What characters know: Projected knowledge and projected co-presence. *Journal of Memory and Language, 44*, 81–95.
Geurts, B., & Rubio-Fernández, P. (2015). Pragmatics and processing. *Ratio, 28*, 446–469.
Graesser, A. C., Bowers, C., Bayen, U. J., & Hu, X. (2000). Who said what? Who knows what? Tracking speakers and knowledge in narrative. In W. van Peer & S. Chatman (Eds.), *Narrative perspective: Cognition and emotion* (pp. 255–272).
Grice, P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.
Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science, 28*, 105–115.
Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science, 4*, 290–305.
Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition, 108*, 831–836.
Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*, 167–170.
Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9*, 131–143.
Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*, 1243091-1–6.
Horton, W. S., & Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. *Discourse Processes, 40*, 1–35.
Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science, 8*, 780–795.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on Theory of Mind use in adults. *Cognition, 89*, 25–41.
Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551–556.
Lea, R. B., Mason, R. A., Albrecht, J. E., Birch, S. L., & Myers, J. L. (1998). Who knows what about whom: What role does common ground play in accessing distant information? *Journal of Memory and Language, 39*, 70–84.
Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics, 27*, 403–422.
McKoon, G., & Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*, 82–91.
Millikan, R. G. (2005). *Language: A biological model.* Oxford University Press.
Moll, H., Carpenter, M., & Tomasello, M. (2011). Social engagement leads 2-year-olds to overestimate others' knowledge. *Infancy, 16*, 248–265.
Moore, G. E. (1993). Moore's paradox. In Thomas Baldwin (Ed.). *G. E. Moore: Selected writings* (pp. 207–212). London: Routledge.
Noveck, I. A., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol, 16*, 109–121.
Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language, 85*, 203–210.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169–190.
R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Rubio-Fernández, P. (2019). Theory of mind. In C. Cummins, & N. Katsos (Eds.). *The handbook of experimental semantics and pragmatics* (pp. 524–536). OUP.
Rubio-Fernández, P. (2017). The director task: A test of Theory-of-Mind use or selective attention? *Psychonomic Bulletin & Review, 24*, 1121–1128.
Rubio-Fernández, P., & Jara-Ettinger, J. (2018). Joint inferences of speakers' beliefs and referents based on how they speak. In Proceedings of the 2018 meeting of the cognitive science society, Madison, Wisconsin, USA.
Ruffman, T. (2014). To belief or not belief: Children's Theory of Mind. *Developmental Review, 34*, 265–293.
Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*, 237–249.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, United Kingdom: Blackwell.
Verga, L., & Kotz, S. A. (2019). Putting language back into ecological communication contexts. *Language, Cognition and Neuroscience, 34*, 536–544.
Weingartner, K. M., & Klin, C. M. (2005). Perspective taking during reading: An on-line investigation of the illusory transparency of intention. *Memory & Cognition, 33*, 48–58.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.