The neural bases of argumentative reasoning

Jérôme Prado, Jessica Léone, Justine Epinat-Duclos, Emmanuel Trouche, Hugo Mercier

The neural bases of argumentative reasoning

Jérôme Prado[1, 2], Jessica Léone[1, 2], Justine Epinat-Duclos[1, 2], Emmanuel Trouche[2, 3] & Hugo Mercier[2, 4]

[1] Lyon Neuroscience Research Center (CRNL), Experiential Neuroscience and Mental Training Team (EDUWELL), INSERM U1028 - CNRS UMR5292, University of Lyon, Lyon, France.

[2] Marc Jeannerod Institute of Cognitive Science, CNRS UMR 5304, University of Lyon, Lyon, France

[3]University Mohammed 6 Polytechnic, Faculty of Governance, Economic and Social Sciences, Ben Guerir, Morocco

[4]Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, Paris, France

Correspondence should be addressed to H.M. (hugo.mercier@gmail.com) or J.P. (jerome.prado@univ-lyon1.fr), Lyon Neuroscience Research Center, CH Le Vinatier, 95 bd Pinel, 69675 Bron Cedex, France. Phone: +33 (0)4 72 13 89 16

Declarations of interest: none

Word count (including figure legends and references): 8,533

Abstract

Most reasoning tasks used in behavioral and neuroimaging studies are abstract, triggering slow, effortful processes. By contrast, most of everyday life reasoning is fast and effortless, as when we exchange arguments in conversation. Recent behavioral studies have shown that reasoning tasks with the same underlying logic can be solved much more easily if they are embedded in an argumentative context. In the present article, we study the neural bases of this type of everyday, argumentative reasoning. Such reasoning is both a social and a metarepresentational process, suggesting it should share some mechanisms, and thus some neural bases, with other social, metarepresentational process such as pragmatics, metacognition, or theory of mind. To isolate the neural bases of argumentative reasoning, we measured fMRI activity of participants who read the same statement presented either as the conclusion of an argument, or as an assertion. We found that conclusions of arguments were associated with greater activity than assertions in a region of the medial prefrontal cortex that was identified in quantitative meta-analyses of studies on theory of mind. This study shows that it is possible to use more ecologically valid tasks to study the neural bases of reasoning, and that using such tasks might point to different neural bases than those observed with the more abstract and artificial tasks typically used in the neuroscience of reasoning. Specifically, we speculate that reasoning in an argumentative context might rely on mechanisms supporting metarepresentational processes in the medial prefrontal cortex.

Keywords: Reasoning; argumentation; metarepresentation; theory of mind; fMRI.

Introduction

Since Aristotle, many scholars have posited a deep relationship between reasoning and argumentation—as Piaget argued, "logical reasoning is an argument which we have with ourselves, and which reproduces internally the features of a real argument" (Piaget, 1928, p. 204). More recently, two strands of research have drawn attention to the importance of argumentation in the study of reasoning. First, work on Bayesian modeling has shown that participants appropriately evaluate ecologically valid arguments—the type of arguments we encounter in everyday life—instead of falling prey to fallacies of argumentation (for review, see Hahn & Oaksford, 2007). Second, the interactionist theory of reasoning (Mercier, 2016; Mercier & Sperber, 2011, 2017) suggests that argumentation is one of the main functions of human reason, highlighting the contrast between how poorly participants perform on many reasoning tasks when facing them on their own, and the same participants' superior performance when they exchange arguments with each other (e.g. Trouche, Sander, & Mercier, 2014; for reviews, see Laughlin, 2011; Mercier, 2016).

Both the Bayesian and the interactionist theories concur in pointing out that, in everyday life, people rarely face the kind of tricky, abstract reasoning problems—the Wason selection task (Wason, 1966), abstract syllogisms (e.g. Khemlani & Johnson-Laird, 2012), the Cognitive Reflection Test (Frederick, 2005)—that have been the most studied by psychologists of reasoning. By contrast, every day we are exposed to countless arguments that we must evaluate: when we engage in discussion, read a newspaper, or watch TV. In line with this observation, a recent series of experiments has highlighted the contrast between how people evaluate abstract arguments and more ecologically valid arguments. Politzer and his colleagues constructed ecologically valid versions of standard categorical syllogisms (Politzer, Bosc-Miné, & Sander, 2017; see also Politzer, 2010). For example, the syllogism in (1):

(1) All A are B

All C are A

Therefore all C are B

was turned into the story in (2):

(2) On his way to school in the morning, Pierre always walks across the park where one can observe shrubs, trees, and many kinds of flowers: roses, asters, and tulips. On arriving at school, he told his schoolmate Marie:

In the park, all the flowers are frozen.

And Marie replied:

Therefore in the park, all the roses are frozen (Politzer et al., 2017, p. 1040)

The three main differences between the abstract problem in (1) and the ecologically valid problem in (2) are: (i) the content of the premises, (ii) the inclusion in a context that renders the syllogism somewhat relevant and, (iii) the omission of one of the premises (here, all roses are flowers), which is taken as part of the participant's background knowledge (syllogisms with such an implicit premise are called enthymematic). These ecologically valid syllogisms significantly improved participants' performance, allowing 11-year-olds to perform as well as adults, and significantly better than same-age children on less ecologically valid versions of the same syllogisms (Politzer et al., 2017).

The idea that we constantly and effortlessly evaluate simple, relevant arguments presented in a social setting suggests that human reasoning is more similar than often thought to other cognitive mechanisms with a social function, such as theory of mind (ToM) (i.e., how one interprets other agents' intentions; Mitchell, 2009), pragmatics (i.e., how language is used in context; Noveck, 2018), and metacognition (see Shea et al., 2014 for the social function of metacognition). This raises the possibility that argumentative reasoning and other social mechanisms may share some neural substrates. For example, ToM—arguably the social ability whose neural bases have been the most studied—is known to be supported by several brain regions, including the medial prefrontal cortex

(mPFC), the precuneus (PC) as well as the left and right temporo-parietal junction (lTPJ and rTPJ) (Molenberghs, Johnson, Henry, & Mattingley, 2016; Van Overwalle & Baetens, 2009). Studies suggest that at least some of these regions also contribute to pragmatic processing (Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2013; Paunov, Blank, & Fedorenko, 2019; Spotorno, Koun, Prado, Van Der Henst, & Noveck, 2012; Van Ackeren, Casasanto, Bekkering, Hagoort, & Rueschemeyer, 2012) and metacognition (Vaccaro & Fleming, 2018). Thus, it is possible that at least some regions involved in ToM may also support argumentative reasoning in an ecologically valid context such as discourse processing.

Consistent with this idea, neuroimaging research indicates that processing language at the discourse-level generally involves brain regions that go well beyond the left-lateralized perisylvian areas known to support sentence-level processing (Fedorenko & Thompson-Schill, 2014; Ferstl, Neumann, Bogler, & von Cramon, 2008; Mar, 2011). Specifically, brain regions that contribute to discourse comprehension overlap with some of the ToM areas (Mar, 2011). This appears to be particularly the case of the mPFC. For example, Ferstl & von Cramon (2002) found enhanced activity in the mPFC when participants judged sentence pairs to be coherent (e.g., "Sometimes a truck drives by the house. That's when the dishes start to rattle.") as compared to pairs that were judged to be incoherent (e.g., "The lights have been on since last night. That's when the dishes start to rattle."). In a recent study, Jacoby & Fedorenko (2018) found that regions of the mPFC that support ToM are also activated when participants read coherent narratives, as compared to incoherent narratives (see also Ferstl, Rinck, & von Cramon, 2005; Ferstl & von Cramon, 2001; Fletcher et al., 1995; Lin et al., 2018).

Two main hypotheses have been proposed to account for the recruitment of the mPFC during discourse processing. A first possibility is that this region is involved because understanding narratives often requires participants to access the mental states of the characters or of the narrator (Fletcher et al., 1995; Gallagher et al., 2000). This mentalizing hypothesis, however, is difficult to reconcile with studies showing that activity in that region can be observed even when such mental state content is minimal (Ferstl & von Cramon, 2002; Jacoby & Fedorenko, 2018; Lin et al., 2018). Another

possibility is that activity in the mPFC during discourse processing reflects domain-general mechanisms involved in both ToM and discourse comprehension. For instance, it has been argued that activity in the mPFC during both types of tasks may indicate that this region supports domain-general mechanisms supporting inference-making, such as "the initiation and maintenance of nonautomatic cognitive processes" (Ferstl & von Cramon, 2002, p. 1611; see also Ferstl et al., 2008; Friese, Rutschmann, Raabe, & Schmalhofer, 2008; Kuperberg, Lakshmanan, Caplan, & Holcomb, 2006). This would naturally suggest a role for this region in argumentative reasoning.

It has also been suggested that the mPFC may support a domain-general capacity for metarepresentation (i.e., representing another representation, or the relationship between representations; Sperber, 2000; Stone & Gerrans, 2006), which is involved in ToM (which deals with representations of others' states of mind), pragmatics (which deals with representations of speakers' intents), or metacognition (which, at least in part—see Proust, 2007—represents features of our own mental states). Critically, arguments are representations of logical or evidential relationships between premises and conclusions. Since reasoning produces and evaluates arguments, it is also a metarepresentational mechanism, one that examines the qualities of representations (here, arguments), rather than things in the world (faces, foods, etc.). This is the most theoretically relevant commonality between reasoning and ToM, pragmatics, and metacognition: they may share metarepresentational mechanisms, and their neurobiological substrates. Therefore, there are several reasons to believe that argumentative reasoning may rely on neural mechanisms that are also involved in ToM as well as in other social, metarepresentational abilities.

Yet, studies that specifically investigated the neural bases of so-called logical reasoning—i.e. reasoning tasks that are commonly modelled by classical logic (such as the syllogism in (1))—have largely failed to identify reasoning-related activity in ToM regions, and more particularly the mPFC. These studies, which used tasks that appear closely related to the control condition of Politzer *et al.* (2017), have found instead activity in a lateral frontoparietal network that includes the posterior parietal cortex, the inferior and middle frontal gyrus, and the rostrolateral prefrontal cortex (Prado,

Chadha, & Booth, 2011). Thus, this literature provides relatively little support for the idea that argumentative reasoning involves the mPFC or other regions involved in ToM.

The lack of involvement of brain regions supporting social cognition in the studies mentioned above, however, may be more apparent than real. First, even though the tasks used in the previous literature on the neural bases of logical reasoning, and many everyday life arguments (such as those of Politzer et al. 2017), can be modeled using classical logic, it is not clear they are actually processed as logical tasks by participants. For example, if the classical syllogisms used both by Politzer et al. (2017), and by most neuroimaging studies (Prado et al., 2011), are often treated as fitting with logical reasoning, there is no agreement in the psychological literature regarding the type of reasoning triggered: some psychologists claim that participants use logical reasoning (Braine & O'Brien, 1998), but others have suggested instead that participants use mental models (Johnson-Laird & Byrne, 1996), probabilistic reasoning (Chater & Oaksford, 1999), or semantics (Geurts, 2003).

Moreover, even if two tasks have the same underlying logic—such as abstract syllogisms and their concrete, socially embedded variant developed by Politzer et al. (2017)—this does not mean that they recruit the same cognitive processes. Past studies have largely relied on the abstract reasoning tasks typically used by psychologists of reasoning, which are difficult and may trigger slow and effortful processing. By contrast, nearly all everyday arguments are processed quickly and effortlessly—think of the back-and-forth of an animated discussion (see Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993). It is thus possible that (at least some of) the frontoparietal activity captured by standard, abstract tasks may have more to do with enhanced demands in executive control and working memory than with reasoning (logical or otherwise) *per se*. In other words, previous studies investigating logical reasoning tasks similar to the syllogism in (1) have failed to examine what is, arguably, one of the most important facets of reasoning: how we evaluate simple, relevant arguments presented in a social setting. Investigating such argumentative reasoning might reveal the implication of at least some ToM regions, in keeping with findings from the discourse comprehension literature (Mar, 2011).

The present study investigates the brain regions supporting reasoning in argumentative context. In doing so, we developed a novel paradigm that is more ecological than those used in most reasoning experiments. Specifically, we presented participants with ecologically valid arguments (inspired by those of Politzer et al., 2017), and asked them to evaluate their conclusion. However, in everyday life, we are often not asked to give explicit feedback on the logic of the arguments we encounter. Therefore, we primarily analyzed activity associated with the conclusion of the argument when it was simply read by participants (i.e., before they had to answer the question). In order to isolate activity associated with argumentative reasoning, we compared activity associated with conclusions of arguments to a condition in which the same conclusions were presented as assertions. We anticipated differences both in reading times and in fMRI responses when participants read the conclusion of an argument rather than an assertion, in particular when the evaluation of the conclusion was not trivial (i.e. when the conclusion was neither trivially true nor false).

Materials and Methods

Participants

Fifty-seven French-speaking volunteers were recruited in the Lyon area. All participants were presented with the exact same task in which they evaluated a series of arguments and assertions (see below). Twenty-seven of these participants performed the task outside of the MRI scanner, while 30 participants performed the task in the scanner while their brain activity was measured. Three participants were excluded from the behavioral analyses because of missing data (n=1) or diagnosed neurological disorders (n=2). Therefore, the final sample for the behavioral analyses consisted of 54 participants (21 males) aged from 18 to 31 years (mean age = 23 years). Three additional participants were excluded from the fMRI study because of MRI contraindications (n=1) and excessive movement in the scanner (n=2). This resulted in a sample of 25 participants (14 males) for the fMRI analyses, aged between 20 to 30 (mean age = 24 years). All subjects in the fMRI experiment were right-handed. All participants provided written informed consent to participate in the study, which was approved by

a local ethics committee (CPP Sud-Est III, Lyon). Subjects were paid 10€ for their participation in the experiment outside of the MRI scanner and 50€ for their participation in the fMRI experiment.

Task

We used enthymematic categorical syllogisms such as the following (translated from French):

In the yard, some neighbors are discussing the neighborhood's greengrocer.
They are wondering whether some of the fruits in that shop are organic.
At that point, Julian arrives and says:
"None of the apples are organic in that shop."
Then, another neighbor speaks again, saying:
"So you see that some fruits are organic in that shop."

In the assertion condition, the last two sentences were modified, to read:

Then, another neighbor joins them, saying:
"I've seen that some fruits are organic in that shop."

As can be seen from this example, the content of the last statement is identical in both conditions, the only difference being whether it is introduced as a conclusion following from the previously introduced premise (i.e. "None of the apples are organic in that shop."), or as an assertion offered independently of that previous statement. To further emphasize the similarity between the conditions, the same question was asked for both: "Do you agree with what has just been said? Yes / No."

We used six figures (i.e. logical forms) for the syllogisms. In four figures (i.e., AA3, AE3/EA3, IA1/AI4, OA1/AO4), the conclusion was neither necessary nor impossible given the

premise; however, some conclusions were more plausible than others, as shown by the fact that participants are more likely to accept as logically valid the plausible rather than the implausible conclusions (Chater & Oaksford, 1999; Evans, Handley, Harper, & Johnson-Laird, 1999, a reaction that could be normative within a Bayesian framework, see Chater & Oaksford, 1999). The example above presents an *implausible* argument, while a *plausible* argument may read: "Some fruits are organic in that shop." / "So you see that some apples are organic in that shop." Additionally, trivial stories in which the conclusion was either necessary nor impossible were also presented and used as fillers. An example of a figure (i.e., AA4/AA1) in which the conclusion was *necessary* (i.e. logically valid) is the following: "All the fruits are organic in that shop." / "So you see that all the apples are organic in that shop." An example of a figure (i.e., OA3/AO3) in which the conclusion was *impossible* (i.e. in direct contradiction with the premise) is the following: "None of the fruits are organic in that shop." / "So you see that some of the apples are organic in that shop." Twenty-four different contents were created, resulting in a total of 288 different stimuli (24 contents * 6 figures * 2 conditions).

Procedure

Stimuli were presented with Presentation software (Neurobehavioral Systems, www.neurobs.com). Participants performed the task in 3 runs of 24 stories each. Participant read the stories line by line in a self-paced manner (i.e., each sentence, conclusion, and question remained on the screen until the participant pressed a key) (see **Fig. 1**). All lines were presented in white on a black background. The interval between the disappearance of a line and the presentation of the next one was 500 ms. After the conclusion (line 6) disappeared, a white fixation cross appeared for a random interval ranging from 3.000 ms to 5.000 ms to introduce jittering. The question was then presented and the participants pressed one of two buttons on a keypad (yes/no response). Another variable period of visual fixation (between 3.000 and 5.000 ms) was added after the disappearance of the question. Each run contained 16 stories with a conclusion that was either plausible or implausible (4 plausible arguments, 4 plausible assertions, 4 implausible arguments, and 4 implausible assertions), as well as 8 filler stories with a conclusion that was either necessary or impossible (i.e., 2 necessary arguments, 2

necessary assertions, 2 impossible arguments, and 2 impossible assertions). Each story in a run had a unique content. Although content was repeated across runs, a content was presented in a given condition only once for each participant. The presentation order of the stories was pseudo-randomized, such that each participant was presented with the stories in a different sequence to balance out order effects.

Each line was displayed in a left-justified manner at the center of the screen. Participants were instructed to read at a normal rate and to respond as accurately as possible to the questions. Three practice trials were presented at the beginning of the behavioral and the fMRI experiments.

Behavioral data analysis

Analyses of the behavioral data from both the behavioral and the fMRI experiments were conducted using the lme4 package implemented in R. Responses to the question were analyzed using logistic mixed models, while reading times of the conclusion (i.e., line 6) were analyzed using a linear mixed model. All full models included in their fixed effects an intercept, a main effect of Condition (i.e., Argument versus Assertion), a main effect of Plausibility (i.e., Plausible versus Implausible), a main effect of Environment (i.e., Outside the scanner versus Inside the scanner) and the interactions between these factors. The factor Condition was deviation coded as -0.5 for Arguments and 0.5 for Assertions. The factor Plausibility was deviation coded as -0.5 for Implausible and 0.5 for Plausible. The factor Environment was deviation coded as -0.5 for Inside the scanner and 0.5 for Outside. All models had maximal random effects structure, with by-subject random intercepts and slopes for Condition, Plausibility, and their interaction, as well as by-story random intercept and slope for Environment. Main effects and interactions were tested using likelihood ratio tests between mixed effect models differing only in the presence or absence of fixed effects of interest.

fMRI data acquisition

Images were collected with a Siemens Prisma 3T MRI scanner (Siemens Healthcare, Erlangen, Germany) at the CERMEP Imagerie du vivant in Lyon, France. The BOLD signal was measured with a susceptibility weighted single-shot EPI sequence. Imaging parameters were as follows: TR = 2000 ms, TE = 24 ms, flip angle = 80°, matrix size = 128 × 120, field of view = 220 × 206 mm, slice thickness = 3 mm (0.48 mm gap), number of slices = 32. A high-resolution T1-weighted whole-brain anatomical volume was also collected for each participant. Parameters were as follows: TR = 3500 ms, TE = 2.24 ms, flip angle = 8°, matrix size = 256 × 256, field of view = 224 × 224 mm, slice thickness = 0.9 mm, number of slices = 192.

fMRI data analyses

FMRI data analysis was performed using the Statistical Parametric Mapping software (SPM12; Functional Imaging Laboratory, UCL, London, UK, http://www.fil.ion.ucl.ac.uk/spm). Each fMRI run started with six dummy scans to allow for magnetization equilibration effects. The functional images were corrected for slice acquisition delays and spatially realigned to the first image of the first run to correct for head-movements. The realigned functional images and the anatomical scans for each subject were then normalized into the standard Montreal Neurological Institute (MNI) space. This was done in two steps. First, after co-registration with the functional data, the structural image was segmented into gray matter, white matter and cerebrospinal fluid by using a unified segmentation algorithm (Ashburner & Friston, 2005). Second, the functional data were normalized to the MNI space by using the normalization parameters estimated during unified segmentation (normalized voxel size, $2 \times 2 \times 4$ mm^3). Finally, the functional images were spatially smoothed with a Gaussian filter equal to twice the voxel size ($4 \times 4 \times 8$ mm^3 full width at half-maximum).

Statistical analysis of fMRI data was performed according to the GLM. Activity associated with the conclusion (i.e., line 6) was modeled using epochs that started with the appearance of the line and ended with its disappearance (i.e., epoch length corresponded to reading time). Other sentences as well as the question were not explicitly modeled (i.e., they were part of background noise). Because

the task was self-paced, different regressors were constructed for each participant based on their own timings. All epochs were convolved with a canonical hemodynamic response function (HRF). The time series data were high-pass filtered (1/128 Hz), and serial correlations were corrected using an autoregressive AR(1) model.

For each subject, we calculated the contrasts corresponding to (i) the difference between Arguments and Assertions, (ii) the difference between Implausible and Plausible stories, and (iii) the interaction between these 2 factors. Individual contrasts were then submitted to second-level one-sample t-tests and thresholded using a FWE corrected cluster-level threshold of $p < .05$ (uncorrected voxel height threshold: $p < .001$).

In additional analyses, we extracted brain activity from regions of interest (ROIs) that were identified in a manual coordinate-based meta-analysis of the ToM network (Van Overwalle & Baetens, 2009). These ROIs, which were also used in previous studies from our group (Schwartz, Epinat-Duclos, Noveck, & Prado, 2018; Spotorno et al., 2012), included all voxels within a 6-mm radius of the following coordinates: x=0 y= 50 z= 20 (mPFC), x= 0 y=−60 z=40 (PC), x=−50 y=−55 z=25 (lTPJ), and x= 50 y=−55 z=25 (rTPJ). For each participant, we calculated the average parameter estimate for each condition within an ROI by averaging the fMRI signal across all voxels within that ROI.

Data availability

Datasets and R scripts for the behavioral analyses, as well as the parameter estimates for each ROI and each participant in the fMRI study, are available on Figshare: https://figshare.com/s/4a94548695934551c201. The un-thresholded t-map for the contrast of Argument versus Assertion is available in NeuroVault: https://identifiers.org/neurovault.collection:5994.

Results

Behavior

Using behavioral data from both the behavioral and the fMRI experiment (n=54), we first analyzed responses to questions in a logistic mixed model (see **Fig. 2A**). First, patterns of responses did not differ between participants who performed the task Inside versus Outside the scanner, as revealed by a lack of main effect of Environment (β = -.310, *SE* =.481, $\chi^2(1)$ = .417, *p* =.519, *partial* η^2 =.00), and a lack of interaction between Environment and other factors (all $\chi^2(1)$s < .749, all *ps* > .387). Second, participants were more in agreement with conclusions when those were Plausible than Implausible, as indicated by a significant main effect of Plausibility (β = 1.056, *SE* =.374, $\chi^2(1)$ = 7.417, *p* =.006, *partial* η^2 =.13). Third, participants were also more in agreement with conclusions in Assertions than in Arguments, as indicated by a significant main effect of Condition (β = 3.262, *SE* = .456, $\chi^2(1)$ = 40.72, *p* < .0001, *partial* η^2 =.61). Fourth, this effect was similar across Plausible and Implausible stories, as shown by a lack of interaction between Condition and Plausibility (β = -.001, *SE* =.548, $\chi^2(1)$ = .00, *p* =.999, *partial* η^2 =.01).

We then analyzed reading times of the conclusion (i.e., line 6) in a linear mixed model (see **Fig. 2B**). First, reading times were similar Inside versus Outside of the scanner, as shown by a lack of main effect of Environment (β = .119, *SE* =.084, $\chi^2(1)$ = 1.997, *p* =.158, *partial* η^2 =.02), and a lack of interaction between Environment and other factors (all $\chi^2(1)$s < 2.168, all *ps* > .141). Second, participants took longer to read conclusions of Implausible stories than Plausible stories, as suggested by a main effect of plausibility that tended to be significant (β = -.062, *SE* = .032, $\chi^2(1)$ = 3.425, *p* = .064, *partial* η^2 =.13). Third, the difference in reading time between Arguments and Assertions was not significant (β = -.050, *SE* = .032, $\chi^2(1)$ = 2.284, *p* = .131, *partial* η^2 =.11). However, this difference was significantly larger in Implausible than in Plausible stories, as demonstrated by a significant interaction between Plausibility and Condition (β = .165, *SE* = .055, $\chi^2(1)$ = 8.186, *p* = .004, *partial* η^2 =.11). Specifically, conclusions of Arguments were read slower than conclusions of Assertions in

Implausible stories (β = -.189, *SE* = .061, $\chi^2(1) = 8.945$, $p = .003$, *partial* $\eta^2 = .15$), but not in Plausible stories (β = -.009, *SE* = .036, $\chi^2(1) = .062$, $p = .803$, *partial* $\eta^2 = .00$).

fMRI (whole-brain analyses)

Activity associated with the conclusion of stories was then compared between Arguments and Assertions. As shown in **Fig. 3** and **Table 1**, there was more activity for conclusions of Arguments than Assertions in the medial prefrontal cortex (mPFC). In contrast, we found more activity for conclusions of Assertions than Arguments in the left intraparietal sulcus (IPS). There was no difference between Plausible and Implausible stories. The difference in activity between Arguments and Assertions also did not differ between Plausible and Implausible stories (i.e., there was no interaction between Condition and Plausibility).

To evaluate whether the mPFC cluster in which activity was greater for conclusions of Arguments than Assertions overlapped with brain regions involved in ToM, we performed a large-scale automated meta-analysis of studies investigating the neural bases of ToM using the software "Neurosynth" (http://neurosynth.org; Yarkoni et al., 2011). The search terms "theory mind" resulted in 181 studies (see Neurosynth website for a complete list of studies). Brain regions that were preferentially related to the terms "theory mind" (i.e., voxels that are reported more often in articles that included the terms "theory mind" in their abstracts than articles that did not) are displayed in light green in **Fig. 3** ($p < .01$ FDR corrected). This automated meta-analysis revealed a large brain network that included the bilateral TPJ, the PC and the mPFC. Critically, the cluster of the mPFC in which activity was greater for conclusions of Arguments than Assertions overlapped with the region of the mPFC that was identified in the meta-analysis (see **Fig. 3**). More specifically, all 3 peaks in that cluster (see **Table 1**) were located in the brain region of the mPFC identified by the automated meta-analysis.

fMRI (ROI analyses)

To gather additional evidence that argumentative reasoning involves at least some brain regions supporting ToM (and estimate the size of our effects in non-circular analyses; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009), we also extracted brain activity associated with conclusions of Arguments and Assertions from 4 ROIs that were previously identified in a manual coordinate-based meta-analyses of ToM: mPFC, PC, lTPJ, and rTPJ (see **Fig. 4**). Brain activity was entered in a 4x2x2 ANOVA with the within-subject factors ROI (mPFC, PC, lTPJ, rTPJ), Plausibility (Implausible, Plausible), and Condition (Argument, Assertion). Overall levels of activity differed between ROIs, as indicated by a main effect of ROI ($F(3,72) = 44.25$, $p < .001$, *partial* $\eta^2 = .65$). More importantly, the difference in activity between Arguments and Assertions also differed between ROIs, as shown by an interaction between ROI and Condition ($F(3,72) = 14.29$, $p < .001$, *partial* $\eta^2 = .37$). To explore this interaction, activity in each ROI was analyzed in separate 2x2 ANOVAs with the factors Plausibility and Condition. We found more activity for conclusions of Arguments than conclusions of Assertions in the mPFC ($F(1,24) = 8.96$, $p = .006$, *partial* $\eta^2 = .27$), but not in any other ROIs (all $Fs < 3.06$, all $ps > 0.093$).

Exploratory analyses

The results above suggest that conclusions of arguments as associated with greater activity than assertions in the mPFC, but not in other regions involved in ToM. It is possible, however, that activity in these other regions may be detected when participants have to explicitly evaluate whether they agree or not with the conclusions (i.e., at the time of the question). To test this hypothesis, we modeled activity associated with the question in a set of exploratory analyses. Whole brain analyses (conducted with a FWE corrected cluster-level threshold of $p < .05$, see **Methods**) indicated no difference of activity between questions that followed Arguments and questions that followed Assertions. There was also no difference between Plausible and Implausible conclusions, and not interaction between Plausibility and Condition.

Brain activity associated with the question was then extracted from the 4 ROIs identified in the coordinate-based meta-analysis from Van Overwalle & Baetens (2009). Activity was analyzed in a 4x2x2 ANOVA with the within-subject factors ROI (mPFC, PC, lTPJ, rTPJ), Plausibility (Implausible, Plausible), and Condition (Argument, Assertion) (see **Fig. 5**). Overall levels of activity differed between ROIs, as indicated by a main effect of ROI ($F(3,72) = 24.44$, $p < .001$, *partial* $\eta^2 = .50$). The difference in activity between Arguments and Assertions also differed between ROIs, as shown by an interaction between ROI and Condition ($F(3,72) = 4.37$, $p = .007$, *partial* $\eta^2 = .15$). To explore this interaction, activity in each ROI was analyzed in separate 2x2 ANOVAs with the factors Plausibility and Condition. There was more activity for conclusions of Arguments than conclusions of Assertions in the mPFC ($F(1,24) = 5.28$, $p = .031$, *partial* $\eta^2 = .18$) as well as in the PC ($F(1,24) = 6.11$, $p = .021$, *partial* $\eta^2 = .20$), but not in the lTPJ or rTPJ (all $Fs < 0.03$, all $ps > 0.902$).

Discussion

In the present experiment, participants were asked to evaluate a statement that was either the conclusion of an argument (i.e. following from a previously introduced premise), or an assertion (i.e. offered independently of the premise). When the conclusion was neither trivially true nor trivially false, we observed that participants took longer to read it in the context of an argument than in the context of an assertion. Moreover, conclusions of arguments were associated with more activity than assertions in a region of the mPFC that was identified in both automated and manual meta-analyses of studies investigating the neural bases of ToM. We can see at least three potential explanations for the involvement of this brain region in argumentative reasoning.

First, it has long been argued that brain regions supporting mentalizing might contribute to discourse-level processing (for a review, see Mar, 2011). For instance, previous neuroimaging studies have found increased activity in several ToM regions (particularly at the level of the mPFC) when participants are presented with coherent narratives (Ferstl & von Cramon, 2002; Jacoby & Fedorenko,

2018). These regions have also been found activated in tasks that require pragmatic inferencing, such as in metaphor and irony processing (Bohrn, Altmann, & Jacobs, 2012; Prat, Mason, & Just, 2012; Spotorno et al., 2012). Prat and colleagues, for example, found stronger recruitment of ToM regions (including the rTPJ and mPFC) when participants processed a sentence in a metaphorical than literal context (activity in the mPFC also increased with contextual difficulty) (Prat et al., 2012). Here, the fact that we found enhanced activity in the mPFC might suggest that evaluating an argument presented in a social setting also requires understanding the mental states of the character stating the argument. However, this possibility is undermined by studies showing that activity in the mPFC during discourse processing can be observed even when narratives involve inanimate entities (and therefore minimally rely on mentalizing, e.g., expository texts; Jacoby & Fedorenko, 2018). It is also undermined by the lack of enhanced activity in other brain regions that are thought to support mentalizing, such as the PC or TPJ (Van Overwalle & Baetens, 2009).

Second, although numerous studies have implicated the mPFC in ToM (e.g., Van Overwalle & Baetens, 2009), it is critical to acknowledge that this region also supports other cognitive processes. These include short-term and long-term memory, executive control, reward-guided learning, as well as decision making (Euston, Gruber, & McNaughton, 2012). Thus, enhanced activity in the mPFC may in theory index any of these processes. It is also possible that increased activity in the mPFC during argumentative reasoning reflects greater demands in domain-general inference-making processes that are also not specifically related to ToM (Ferstl et al., 2008; Kuperberg et al., 2006). To some extent, these possibilities are supported by the fact that we did not find extensive activation in regions of the TPJ and PC that are typically co-activated with the mPFC in ToM tasks. However, it is also important to consider that the region of the mPFC found activated in the present study overlapped with previous meta-analyses of ToM. Thus, there might be a third hypothesis, which is that the mPFC supports computations that contribute to both ToM and argumentative reasoning.

Specifically, studies have found that functional specialization for mentalizing in the mPFC decreases from adolescence to adulthood, whereas it increases in the rTPJ (Blakemore, den Ouden,

Choudhury, & Frith, 2007; Burnett, Bird, Moll, Frith, & Blakemore, 2009; Pfeifer, Lieberman, & Dapretto, 2007; Pfeifer et al., 2009). Therefore, it has been proposed that the mPFC "may play a specific role in the metarepresentational component of mentalizing" (Lombardo, Chakrabarti, Bullmore, Baron-Cohen, & Consortium, 2011, p.1837). In line with this proposal, we speculate that a general role of the mPFC in metarepresentational abilities may explain why this region is involved in argumentative reasoning (which is metarepresentational in nature) as well as in tasks that involve representing speakers' intents (Bašnáková et al., 2013; Paunov et al., 2019; Spotorno et al., 2012; Van Ackeren et al., 2012) and our own mental states (Vaccaro & Fleming, 2018). More broadly, a relatively specific role of the mPFC in metarepresentational abilities would also account for the relatively selective involvement of this region in discourse processing (Ferstl & von Cramon, 2002; Jacoby & Fedorenko, 2018; Lin et al., 2018).

Although the present findings are in keeping with studies on discourse processing (see above), they are at odds with a number of studies that have investigated the neural bases of logical reasoning. Overall, these studies have identified a wide reasoning-related brain network involving lateral regions in the frontal, parietal and temporal cortices (Goel, 2007; Prado et al., 2011). However, perhaps the most salient finding from that literature is that brain activity associated with reasoning appears to be largely modulated by characteristics of the task (Goel, 2007; Prado, 2018; Prado et al., 2011; Wertheim & Ragni, 2018). For example, studies have found that the neural substrates of logical reasoning depend upon the presence or absence of concrete content (Goel, Buchel, Frith, & Dolan, 2000; Goel, Makale, & Grafman, 2004), the presence or absence of conflicting information (Goel & Dolan, 2003; Prado, Kaliuzhna, Cheylus, & Noveck, 2008; Prado & Noveck, 2007; Stollstorff, Vartanian, & Goel, 2012), the amount of information available to evaluate a conclusion (Goel, Stollstorff, Nakic, Knutson, & Grafman, 2009), the difficulty of the argument (Coetzee & Monti, 2018; Monti, Osherson, Martinez, & Parsons, 2007; Noveck, Goel, & Smith, 2004), or the logical form of the premises (Prado, Mutreja, & Booth, 2013; Prado, Van Der Henst, & Noveck, 2010; Reverberi et al., 2010). In the present study, we provide evidence that evaluating simple, relevant arguments similar to those used in this literature but presented in a discourse context may be

associated with brain regions that largely differ from the neural network that has been previously reported (Prado et al., 2011).

To some extent, our results could be interpreted as lending support to the claim that the neural bases of logical reasoning are task-dependent (Goel, 2007). However, our results may also question the ecological validity of tasks that have been used to investigate the brain substrates of reasoning (and by extension the validity of tasks that are used in the cognitive literature on reasoning). Specifically, most tasks used to investigate the neural bases of logical reasoning involve abstract (and artificial) reasoning problems and capture brain activity when subjects explicitly evaluate the logical validity of a conclusion, a task which is arguably rarely done in everyday life (see also, Prado et al., 2015). Such tasks are likely to require some cognitive effort and it may be difficult to disentangle reasoning-related activity from activity associated with specific demands in terms of executive control and working memory. In other words, previous neuroimaging studies might have missed reasoning-related activity in important brain regions because they exclusively studied logical reasoning in a context that was not social and of relatively poor ecological validity (as is also the case for most laboratory tasks used to study reasoning in the cognitive literature; Frederick, 2005; Khemlani & Johnson-Laird, 2012; Wason, 1966).

Finally, it is important to consider two potential limitations of our work. First, sample size for the behavioral analyses was relatively large (n=54, which leads to 80% power to detect an effect size of d=0.39 in the comparison of reading times between conclusions of arguments and assertions). However, sample size for the fMRI analyses was smaller (n=25, which only leads to 80% power to detect an effect size of d=0.58 in the comparison of brain activity between conclusions of arguments and assertions in a given region). Thus, it is possible that the fact that we did not observed a difference between conclusions of arguments and assertions in posterior ToM regions results from this relatively low power. Because low power may also reduce the likelihood that a significant result reflects a true effect (Button et al., 2013), the present results would need to be replicated in future experiments. Second, the mPFC identified in the contrast of arguments versus assertions overlapped with a region

often found in studies on ToM (as demonstrated by meta-analyses). However, the precise locations of brain regions involved in ToM may vary between participants. Therefore, it is possible that our analyses might have missed some larger overlap between activity associated with ToM and argumentative reasoning because they were performed at the group rather than at the individual level. The use of localizer tasks identifying the neural bases of ToM in each individual may be helpful in future studies investigating the relationship between metarepresentational abilities and discourse processing (see, e.g., Jacoby & Fedorenko, 2018).

This investigation opens the way for research on an understudied aspect of reasoning: argumentative reasoning, arguably the most common form of reasoning tapped in everyday life (Hahn & Oaksford, 2007; Kuhn, 1991; Mercier & Sperber, 2017). Moreover, our efforts to develop a simple reasoning task with increased ecological validity, and for which implicit responses can be recorded, might prompt others researchers studying the neuroscience of reasoning to use similar tasks instead of the more difficult and abstract tasks generally used.

Acknowledgments

We thank the MRI engineers (Franck Lamberton and Danielle Ibarrola) at the CERMEP-Lyon platform for their assistance in collecting the fMRI data.

Funding

This work was supported by a grant from the Agence Nationale de la Recherche (ANR-16-TERC-0001-01) to H.M.

References

Ashburner, J., Friston, K. J. (2005). Unified segmentation. *NeuroImage* 26 (3), 839–851.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2013). Beyond the language given: the neural correlates of inferring speaker meaning. *Cerebral Cortex, 24*(10), 2572-2578.

Blakemore, S. J., den Ouden, H., Choudhury, S., & Frith, C. (2007). Adolescent development of the neural circuitry for thinking about intentions. *Soc. Cogn. Affect. Neurosci., 2*, 130-139.

Bohrn, I. C., Altmann, U., & Jacobs, A. M. (2012). Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia, 50*(11), 26692683.

Braine, M. D. S., & O'Brien, D. P. (1998). *Mental logic*. Mahwah: Lawrence Erlbaum Associates Ltd.

Burnett, S., Bird, G., Moll, J., Frith, C., & Blakemore, S. J. (2009). Development during adolescence of the neural processing of social emotion. *J. Cogn. Neurosci., 21*, 1736-1750.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365-376.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*, 191–258.

Coetzee, J. P., & Monti, M. M. (2018). At the core of reasoning: Dissociating deductive and non-deductive load. *Hum Brain Mapp, 39*(4), 1850-1861. doi:10.1002/hbm.23979

Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, 76(6), 1057-1070.

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1495–1513.

Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends Cogn Sci, 18*(3), 120-126.

Ferstl, E., Neumann, J., Bogler, C., & von Cramon, D. (2008). The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Hum Brain Mapp, 29*(5), 581-593. doi:10.1002/hbm.20422

Ferstl, E. C., Rinck, M., & von Cramon, D. Y. (2005). Emotional and temporal aspects of situation model processing during text comprehension: An event-related fMRI study. *J Cogn Neurosci, 17*, 724-739.

Ferstl, E. C., & von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Brain Res Cogn Brain Res, 11*, 325-340.

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: Coherence or Theory of Mind? *Neuroimage, 17*(3), 1599-1612.

Fletcher, P. C., F., H., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of ''theory of mind'' in story comprehension. *Cognition, 57*, 109-128.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Friese, U., Rutschmann, R., Raabe, M., & Schmalhofer, F. (2008). Neural indicators of inference processes in text comprehension: An event-related functional magnetic resonance imaging study. *Journal of cognitive neuroscience, 20*(11), 2110-2124.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of "Theory of Mind" in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11-21.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*(3), 223–251.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends Cogn. Sci. (Regul. Ed.), 11*(10), 435-441. doi:10.1016/j.tics.2007.09.003

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage, 12*(5), 504-514. doi:10.1006/nimg.2000.0636
S1053-8119(00)90636-0 [pii]

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87*(1), B11-22.

Goel, V., Makale, M., & Grafman, J. (2004). The hippocampal system mediates logical reasoning about familiar spatial environments. *J Cogn Neurosci, 16*(4), 654-664. doi:10.1162/089892904323057362

Goel, V., Stollstorff, M., Nakic, M., Knutson, K., & Grafman, J. (2009). A role for right ventrolateral prefrontal cortex in reasoning about indeterminate relations. *Neuropsychologia, 47*(13), 2790-2797. doi:10.1016/j.neuropsychologia.2009.06.002

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.

Jacoby, N., & Fedorenko, E. (2018). Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Language, Cognition and Neuroscience*, 1-17.

Johnson-Laird, P. N., & Byrne, R. M. J. (1996). Mental models and syllogisms. *Behavioural and Brain Sciences*, *19*, 543–546.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci, 12*(5), 535-540. doi:10.1038/nn.2303

Kuhn, D. (1991). *The skills of arguments*. Cambridge: Cambridge University Press.

Kuperberg, G. R., Lakshmanan, B. M., Caplan, D. N., & Holcomb, P. J. (2006). Making sense of discourse: An fMRI study of causal inferencing across sentences. *Neuroimage, 33*(1), 343-361.

Laughlin, P. R. (2011). *Group problem solving*. Princeton: Princeton University Press.

Lin, N., Yang, X., Li, J., Wang, S., Hua, H., Ma, Y., & Li, X. (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cogn Affect Behav Neurosci, 103*(2), 1-11.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Baron-Cohen, S., & Consortium, M. A. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *Neuroimage, 56*(3), 1832-1838.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annu Rev Psychol, 62*(1), 103-134.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge: Harvard University Press.

Mitchell, J. P. (2009). Inferences about mental states. *Philos. Trans. R. Soc. Lond. B Biol. Sci., 364*(1521), 1309-1316.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews, 65*, 276-291.

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage, 37*(3), 1005-1016. doi:10.1016/j.neuroimage.2007.04.069

Noveck, I. A. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge, UK: Cambridge University Press.

Noveck, I. A., Goel, V., & Smith, K. W. (2004). The neural basis of conditional reasoning with arbitrary content. *Cortex, 40*(4-5), 613-622.

Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *J Neurophysiol, 121*(4), 1244-1265.

Pfeifer, J. H., Lieberman, M. D., & Dapretto, M. (2007). "I know you are but what am I?": neural bases of self- and social knowledge retrieval in children and adults. *J. Cogn. Neurosci., 19*(1323-1337).

Pfeifer, J. H., Masten, C. L., Borofsky, L. A., Dapretto, M., Fuligni, A. J., & Lieberman, M. D. (2009). Neural correlates of direct and reflected self-appraisals in adolescents and adults: when social perspective-taking informs self-perception. *Child Dev., 80*, 1016-1038.

Piaget, J. (1928). *Judgment and reasoning in the child*. London: Routledge and Kegan Paul.

Politzer, G. (2010). Solving natural syllogisms. In K. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reason: A festschrift for Jonathan St. BT Evans* (p. 19). London: Psychology Press.

Politzer, G., Bosc-Miné, C., & Sander, E. (2017). Preadolescents solve natural syllogisms proficiently. *Cognitive Science*, *41*(5), 1031–1061.

Prado, J. (2018). The relationship between deductive reasoning and the syntax of language in Broca's area: A review of the neuroimaging literature. *L'année psychologique/Topics in Cognitive Psychology, 118*(3), 289-315.

Prado, J., Chadha, A., & Booth, J. R. (2011). The brain network for deductive reasoning: a quantitative meta-analysis of 28 neuroimaging studies. *Journal of cognitive neuroscience, 23*(11), 3483-3497. doi:10.1162/jocn_a_00063

Prado, J., Kaliuzhna, M., Cheylus, A., & Noveck, I. A. (2008). Overcoming perceptual features in logical reasoning: an event-related potentials study. *Neuropsychologia, 46*(11), 2629-2637. doi:10.1016/j.neuropsychologia.2008.04.017

Prado, J., Mutreja, R., & Booth, J. R. (2013). Fractionating the Neural Substrates of Transitive Reasoning: Task-Dependent Contributions of Spatial and Verbal Representations. *Cereb Cortex, 23*(3), 499-507.

Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: a parametric functional magnetic resonance imaging study. *J Cogn Neurosci, 19*(4), 642-657. doi:10.1162/jocn.2007.19.4.642

Prado, J., Spotorno, N., Koun, E., Hewitt, E., Van Der Henst, J. B., Sperber, D., & Noveck, I. A. (2015). Neural interaction between logical reasoning and pragmatic processing in narrative discourse. *Journal of cognitive neuroscience, 27*, 692-704.

Prado, J., Van Der Henst, J.-B., & Noveck, I. A. (2010). Recomposing a fragmented literature: how conditional and relational arguments engage different neural systems for deductive reasoning. *Neuroimage, 51*(3), 1213-1221. doi:10.1016/j.neuroimage.2010.03.026

Prat, C. S., Mason, R. A., & Just, M. A. (2012). An fMRI investigation of analogical mapping in metaphor comprehension: The influence of context and individual cognitive capacities on processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(2), 282.

Proust, J. (2007). Metacognition and metarepresentation: Is a self-directed theory of mind a precondition for Metacognition? *Synthese*, *159*(2), 271–295.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction*, *11*(3/4), 347–364.

Reverberi, C., Cherubini, P., Frackowiak, R. S. J., Caltagirone, C., Paulesu, E., & Macaluso, E. (2010). Conditional and syllogistic deductive tasks dissociate functionally during premise integration. *Hum Brain Mapp*. doi:10.1002/hbm.20947

Schwartz, F., Epinat-Duclos, J., Noveck, I., & Prado, J. (2018). The neural development of pragmatic inference-making in natural discourse. *Dev Sci, 21*(6), e12678. doi:10.1111/desc.12678

Sperber, D. (2000). Metarepresentations: A multidisciplinary perspective. Oxford University Press.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. New York: Wiley-Blackwell.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J. B., & Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: the case of irony. *Neuroimage, 63*(1), 25-39. doi:10.1016/j.neuroimage.2012.06.046

Stollstorff, M., Vartanian, O., & Goel, G. (2012). Levels of conflict in reasoning modulate right lateral prefrontal cortex. *Brain Research, 1428*, 24-32.

Stone, V. E., & Gerrans, P. (2006). What's domain-specific about theory of mind? *Social Neuroscience, 1*(3-4), 309-319.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain and neuroscience advances, 2*, 1-14.

Van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., & Rueschemeyer, S. A. (2012). Pragmatics in action: indirect requests engage theory of mind areas and the cortical motor network. *Journal of cognitive neuroscience, 24*(11), 2237-2247.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage, 48*(3), 564-584. doi:10.1016/j.neuroimage.2009.06.009

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.

Wertheim, J., & Ragni, M. (2018). The Neural Correlates of Relational Reasoning: A Meta-analysis of 47 Functional Magnetic Resonance Studies. *J Cogn Neurosci, 30*(11), 1734-1748. doi:10.1162/jocn_a_01311

Figures

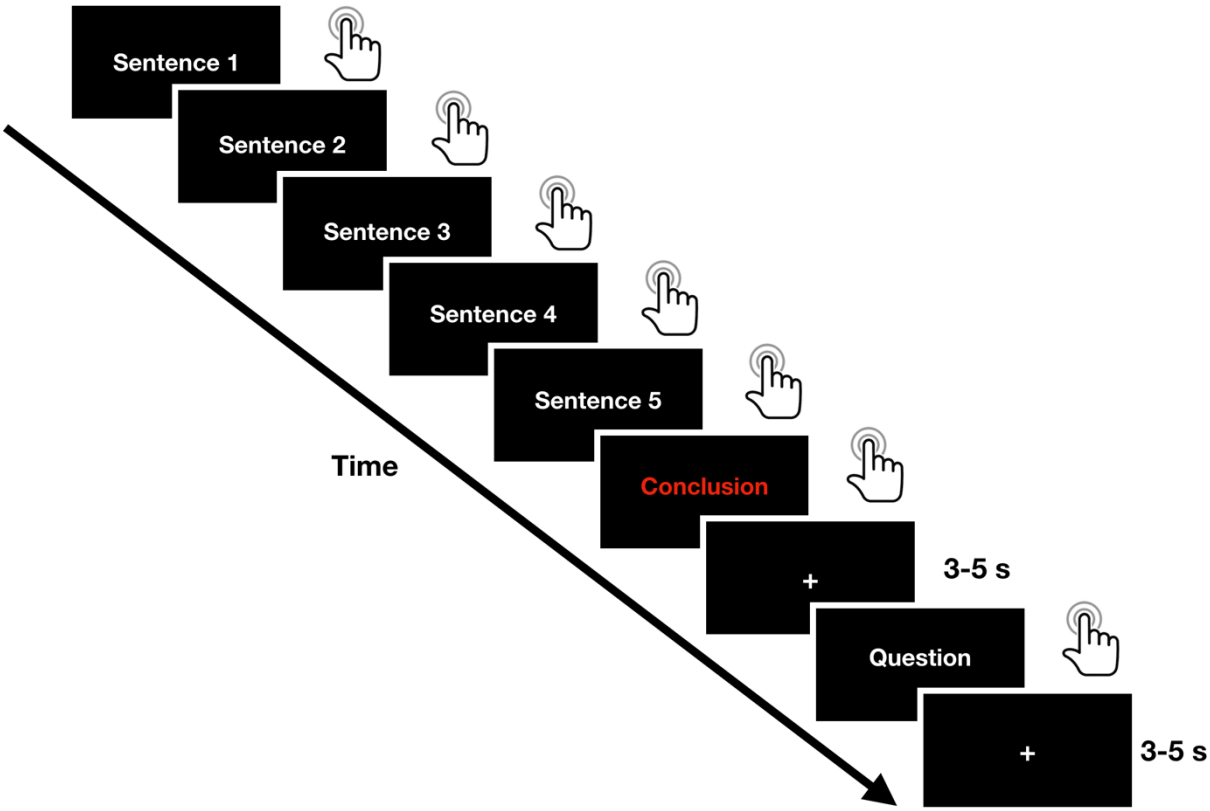

Figure 1. Timeline for a sample story. Each of the 6 lines (5 sentences and 1 conclusion) was displayed sequentially on the screen. The task was entirely self-paced. Participants pressed on a button to indicate that they were ready to read the next sentence, which was shown after a 500 ms delay (not shown). The scenario ended with a question. This question was preceded and followed by a jittered interval ranging from 3 to 5 seconds. The line of interest considered in the analyses was the conclusion (in red).

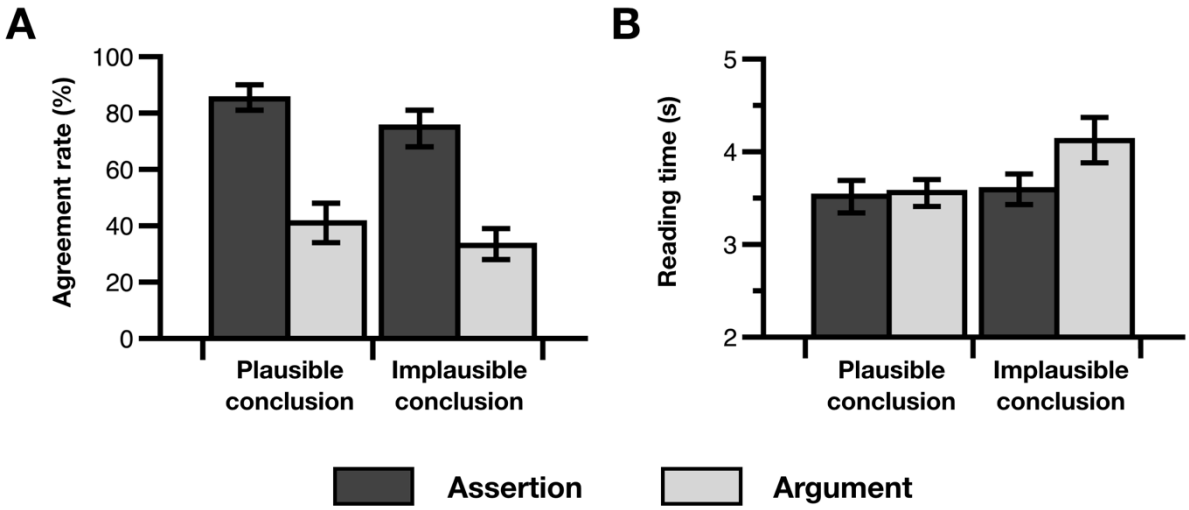

Figure 2. Behavioral results. (A) Rates of agreement with the conclusion of Arguments and Assertions as a function of plausibility. (B) Conclusion reading times for Arguments and Assertions as a function of plausibility. Error bars represent within-subjects 95% confidence intervals.

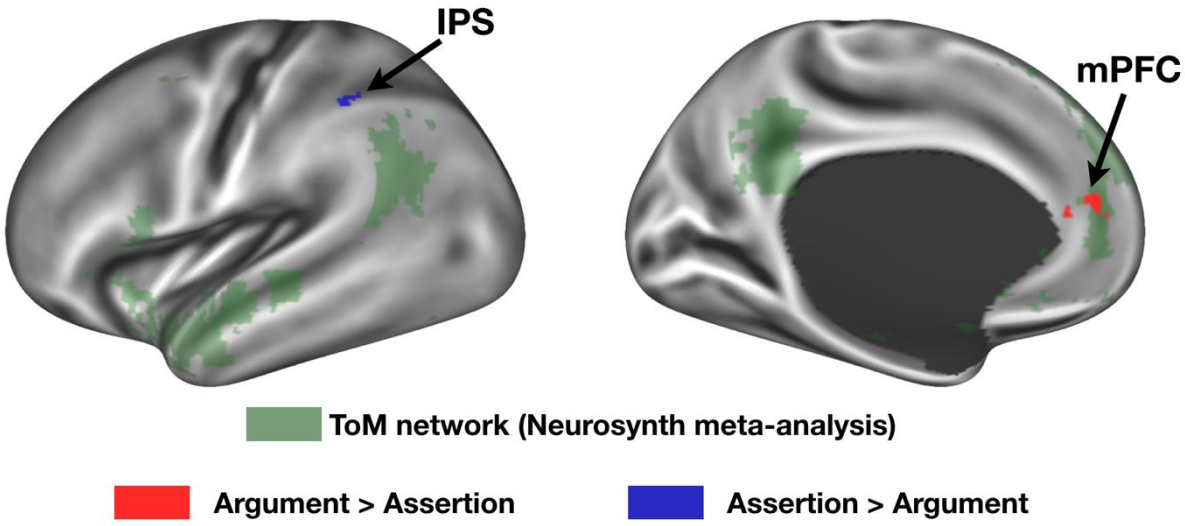

Figure 3. Whole-brain analyses. Brain regions in which activity differed between the conclusion of Arguments and Assertions are superimposed on clusters identified in the meta-analysis of studies investigating Theory of Mind. All activations are overlaid on an inflated 3D rendering of the MNI-normalized anatomical brain (lateral and medial views of the left hemisphere).

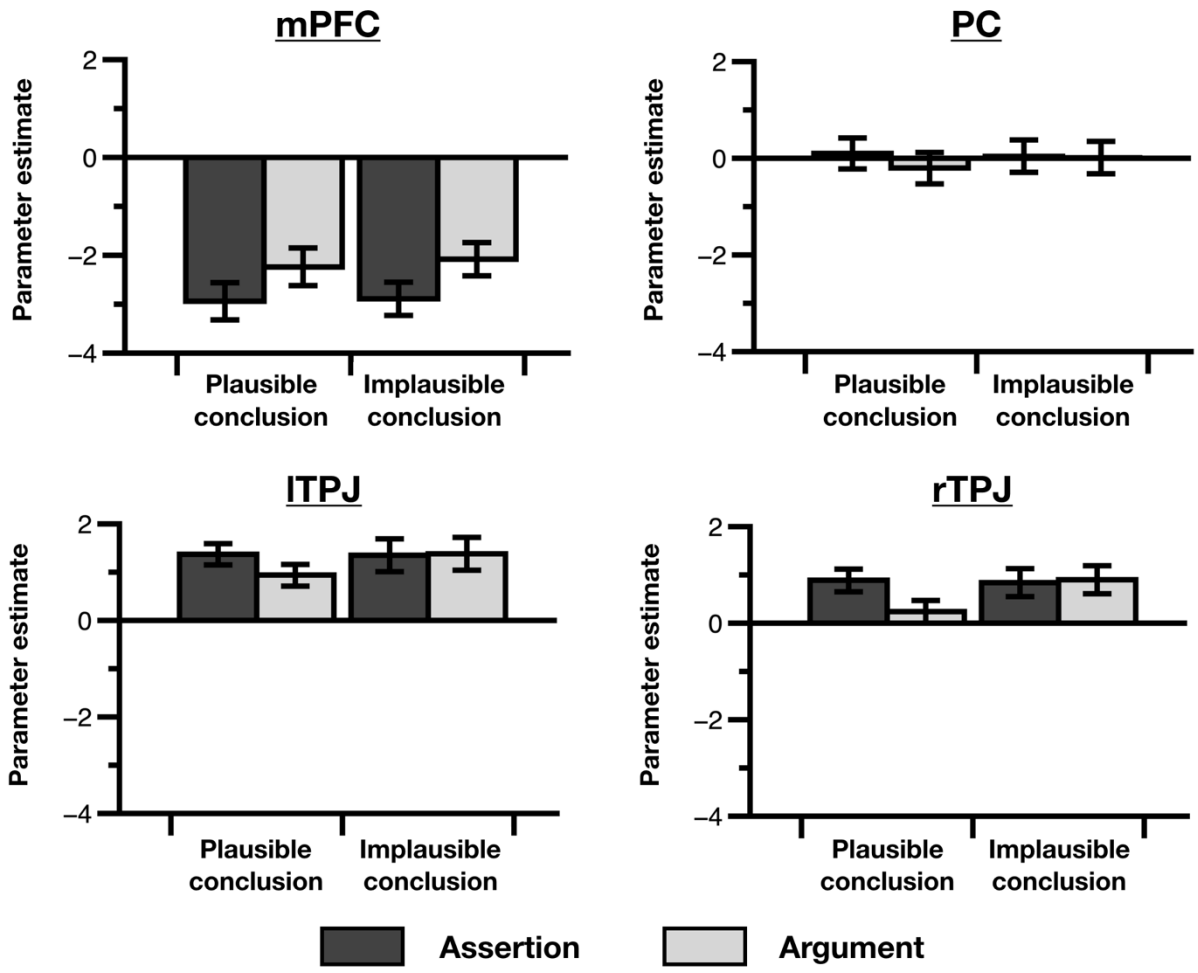

Figure 4. Main ROI analyses. Brain activity (i.e., parameter estimate) associated with the conclusion of Arguments and Assertions as a function of plausibility in each ROI defined based on the meta-analysis of (Van Overwalle & Baetens, 2009). Error bars represent within-subjects 95% confidence intervals. mPFC: medial prefrontal cortex, PC: Precuneus, lTPJ: left Temporo-Parietal junction, rTPJ: right Temporo-Parietal junction.

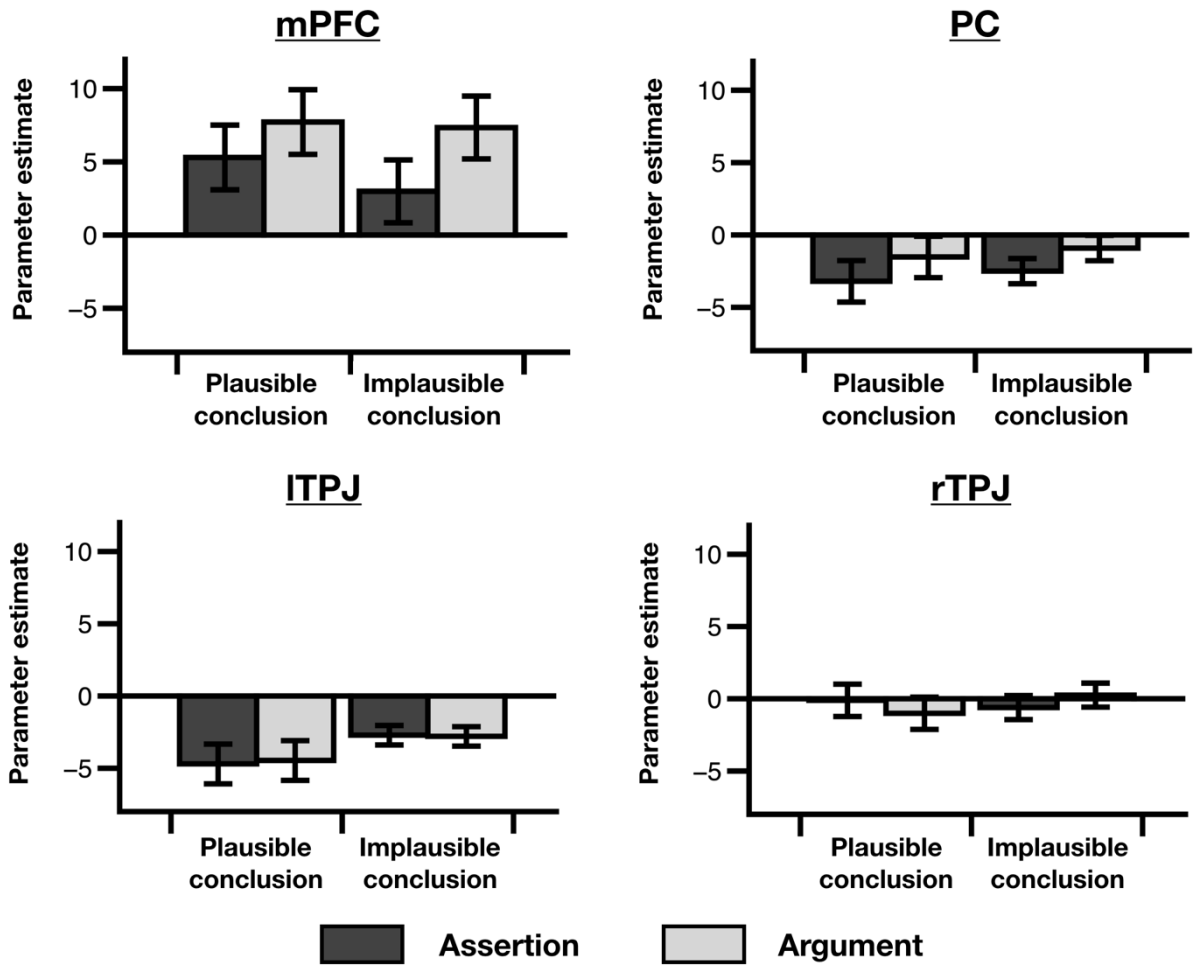

Figure 5. Exploratory ROI analyses. Brain activity (i.e., parameter estimate) associated with the question of Arguments and Assertions as a function of plausibility in each ROI defined based on the meta-analysis of (Van Overwalle & Baetens, 2009). Error bars represent within-subjects 95% confidence intervals. mPFC: medial prefrontal cortex, PC: Precuneus, lTPJ: left Temporo-Parietal junction, rTPJ: right Temporo-Parietal junction.