NeuroImage 59 (2012) 1752–1764

Contents lists available at SciVerse ScienceDirect

NeuroImage

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

Large scale brain activations predict reasoning profiles

Carlo Reverberi [a,b,c,*], Luca L. Bonatti [d], Richard S.J. Frackowiak [b,e], Eraldo Paulesu [a,f], Paolo Cherubini [a], Emiliano Macaluso [b]

[a] Department of Psychology, Università Milano Bicocca, Milano, Italy
[b] Neuroimaging Laboratory, Santa Lucia Foundation, Roma, Italy
[c] Bernstein Centre for Computational Neuroscience, Charité-Universitätsmedizin, Berlin, Germany
[d] ICREA and Universitat Pompeu Fabra, Barcelona, Spain
[e] Département de Neurosciences Cliniques, Centre Hospitalier Universitaire Vaudois, 1011-Lausanne, Switzerland
[f] IRCCS Galeazzi, Milano, Italy

A R T I C L E I N F O

Article history:
Received 10 January 2011
Revised 31 May 2011
Accepted 9 August 2011
Available online 24 August 2011

A B S T R A C T

Deduction is the ability to draw necessary conclusions from previous knowledge. Here we propose a novel approach to understanding the neural basis of deduction, which exploits fine-grained inter-participant variability in such tasks. Participants solved deductive problems and were grouped by the behavioral strategies employed, i.e., whether they were sensitive to the logical form of syllogistic premises, whether the problems were solved correctly, and whether heuristic strategies were employed. Differential profiles of neural activity can predict membership of the first two of these groups. The predictive power of activity profiles is distributed non-uniformly across the brain areas activated by deduction. Activation in left ventro-lateral frontal (BA47) and lateral occipital (BA19) cortices predicts whether logically valid solutions are sought. Activation of left inferior lateral frontal (BA44/45) and superior medial frontal (BA6/8) cortices predicts sensitivity to the logical structure of problems. No specific pattern of activation was associated with the use of a non-logical heuristic strategy. Not only do these findings corroborate the hypothesis that left BA47, BA44/45 and BA6/8 are critical for making syllogistic deductions, but they also imply that they have different functional roles as components of a dedicated network. We propose that BA44/45 and BA6/8 are involved in the extraction and representation of the formal structure of a problem, while BA47 is involved in the selection and application of relevant inferential rules. Finally, our findings suggest that deductive reasoning can be best described as a cascade of cognitive processes requiring the concerted operation of several, functionally distinct, brain areas.

Introduction

Every cognitive activity requires the ability to process inferential relations among representations. While the mechanisms underlying such abilities are complex and varied, one of the most important aspects of inferential ability is deduction. Deduction is the ability to draw necessary conclusions from previous items of knowledge. Thanks to deductive reasoning, information can be left dormant within previously encoded knowledge to be made explicit only when needed: it is not necessary to represent Socrates mortality given the knowledge that Socrates is a man and all men are mortals. Thus, deduction offers substantial advantages in supporting an efficient and flexible cognitive architecture.

In recent years, increasing efforts have been devoted to clarify the neural underpinnings of this fundamental human ability. Several brain areas have been reported to activate consistently across studies e.g., regions near the left intraparietal sulcus, the left inferior frontal gyrus and in the basal ganglia (e.g. Goel and Dolan, 2003a; Goel et al., 2000; Monti et al., 2007; Prado et al., 2010; Reverberi et al., 2007, 2010). Nevertheless, an unexpectedly large and varied set of other brain areas, such as bilateral occipital cortex, anterior cingulate, medial frontal cortex and right dorsolateral frontal cortex, have also been activated in some studies, but were not found in others (Fangmeier et al., 2006; Goel, 2007; Goel and Dolan, 2003a; Goel et al., 2000; Knauff et al., 2003; Monti et al., 2007; Monti et al., 2009; Noveck et al., 2004; Prado et al., 2010; Reverberi, et al., 2007; Reverberi et al., 2010; Rodriguez-Moreno and Hirsch, 2009). Several explanations have been proposed for this apparent inconsistency. Some authors have suggested that different types of deductive inference may rely on different sets of cognitive processes, leading to the involvement of different brain areas for different types of inference (Goel and Dolan, 2003a; Goel et al., 2000; Monti et al., 2007; Prado et al., 2010; Reverberi et al., 2007, 2010). Thus, for example, it has been shown that relational syllogisms activate the right temporo-parieto-occipital junction while complex conditional inferences activate the left inferior frontal lobe (Prado et al., 2010). A second, related explanation is that the brain areas activated during

* Corresponding author at: Department of Psychology, Università Milano Bicocca, Piazza Ateneo Nuovo, 1, 20126 Milano, Italy.
E-mail address: carlo.reverberi@unimib.it (C. Reverberi).

1053-8119/$ – see front matter
doi:10.1016/j.neuroimage.2011.08.027

deduction have, in fact, different functional roles. One hypothesis postulates that a "core" set of brain structures responsible for critical deductive processes (namely, posterior BA10 and BA8) is distinct from "accessory" areas that are involved in other non-core supporting functions such as attention (Monti et al., 2007; Monti et al., 2009). Different functions within the reasoning network have also been associated with a multi-stage processing model of deductive reasoning (Reverberi et al., 2007, 2009c). According to this hypothesis, even simple logical problems undergo a series of processing steps manipulating the information contained in the premises, either implicitly or explicitly. Such processes compute subcomponents of deductive reasoning, such as the retrieval of the logical form of the premises, the application of rules of inferences to those premises and their consequences, or the overall monitoring of the demonstration structure constructed in searching for a conclusion. It is plausible that such cognitive components recruit different brain regions. Asymmetric sampling of those various components in different experiments may also explain some of the variability of results across studies. Finally, experimental issues may also have been responsible, for example the use of inadequate baseline tasks (Monti et al., 2007; Reverberi et al., 2009b). Notwithstanding the progress made in the last decade, knowledge about the neural structures involved in deductive reasoning and an understanding of the operations carried out by component areas of the reasoning network, is still incomplete.

In this study we adopted a novel approach to identify the brain areas involved in deduction. We examined specific relationships between brain activity and behavioral performance during reasoning across participants, exploiting in this way the inter-participant variance for the same task conditions. We asked participants to solve several types of categorical syllogisms with abstract premises and conclusions. Participants had to respond by combining information contained in the two premises, thus generating a necessary, logically valid, conclusion. Categorical syllogisms have a property that makes them particularly well suited for this study. Despite similar surface characteristics, all involving two premises with one quantifier each, such syllogisms generate a range of response patterns (Ford, 1995; Johnson-Laird and Bara, 1984; Reverberi et al., 2009b). We measured and classified participant performances by reference to three basic features. First, we sought evidence for any individual tendency in generation of valid or invalid deductions, thus discriminating participants who more frequently made correct as opposed to incorrect responses (validity index). Second, we quantified any tendency to select the same conclusion for superficially different, but formally identical problems. In this way we discriminated participants who systematically selected the same conclusion (be it correct or not) for the same syllogisms, from those who chose from several alternatives. The former rely on the formal structure of problems to reason, whether or not they derive a correct conclusion (consistency index). Third, we looked for any preference to select a conclusion that is consistent with the "atmosphere" of the given premises (Chapman and Chapman, 1959; Reverberi et al., 2009b; Woodworth and Sells, 1935). "Atmosphere" refers to a simple heuristic shortcut frequently noted in studies on syllogisms (heuristic index). The first two indices allowed us to probe neural structures involved in different processes of deductive reasoning. The consistency index targeted brain areas involved in detection and representation of the formal structure of the given premises, regardless of any logically irrelevant features. The validity index targeted brain areas critical for generating a logically valid response from the given premises. The heuristic index allowed us to check whether heuristic reasoning relies on cognitive systems different from those involved in other types of reasoning, and thus on dissociable brain networks.

We explored whether it is possible to predict the behavior of participants on the basis of activation levels in brain areas involved in the solution of deductive problems. A finding that the pattern of activation in a single region, or across multiple brain areas, predicts strategies for syllogistic reasoning would advance understanding of deductive reasoning in several ways. First, it would provide independent and more direct evidence that areas activated during deductive reasoning are indeed involved in performance-critical inferential processes. Second, it would clarify whether there is functional specialization within the larger deduction network. For example, the observation that a subset of activated areas predicts generation of logically valid answers, whereas another subset predicts sensitivity to the formal structure of given premises would suggest that only the former areas are specifically involved in the generation of a valid response, while the latter have a different functional role (e.g., encoding the formal structure of a syllogistic problem). Finally, such a result would help explain some of the observed between-study inconsistencies previously reported, suggesting a biological underpinning for the variability of activation patterns found that is based on differential reasoning strategies.

Methods

Participants

Twenty-six healthy participants (average age 24.9 years, SD = 5.0; 15 males) participated in the experiment. After instruction about the procedure, all participants gave written informed consent. All were right-handed, with normal vision and no neurological or psychiatric history. The study was approved by the Santa Lucia Foundation (Scientific Institute for Research Hospitalization and Health Care) Independent Ethics Committee.

Stimuli

Seventy-two categorical syllogisms were administered during fMRI scanning. Forty-eight of these were integrable and 24 non-integrable, as explained below. Besides syllogistic problems, we also administered 40 memory trials and 60 conditional problems. Categorical syllogisms, conditional problems and memory trials were presented randomly intermixed. The order of administration was different across participants. For the aims of this study the critical stimuli were the categorical syllogisms. We have previously reported accessory behavioral analyses on memory trials (see also Reverberi et al., 2010). Each deductive problem consisted of two or three premises and a set of four alternative conclusions (Table 1 and Fig. 1). All sentences described the qualities of an unspecified "thing" by means of

Table 1
Example of the different type of problems used in the study.

	Integrable	Non-integrable
Solvable with heuristics	P_1 Every thing *a* is *b* P_2 Every thing *b* is *c* C Every thing *a* is *c*	P_1 Every thing *a* is *b* P_2 Every thing *c* is *d* P_3 Some thing *c* is *e* C Some thing *d* is *e*
Non-solvable with heuristics	P_1 Every thing *b* is *a* P_2 Every thing *b* is *c* C Some thing *a* is *c*	

P_1 first premise; P_2 second premise; P_3 third premise; C correct conclusion. Letters written in italics stand for a bi-syllabic Italian non-word such as "rufa". The non-words are in an adjectival position and agree with the gender of the word "thing" (feminine in Italian). The integrable problems have always a common term between P_1 and P_2. The non-integrable sentences can be made of two or three sentences. When present, the third sentence is always integrable with the second. In the examples only the correct conclusion is reported, but during the experiment the correct conclusion was presented along with three other non-valid conclusions. Example 1 (top left) is solvable also by using the atmosphere heuristic. By contrast, in Example 2 (bottom left) the answer suggested by the atmosphere heuristic ("Every thing *a* is *c*") is wrong, i.e. is not logically valid. It should be noticed that some arguments, such as Example 2, are not valid in standard logic unless one assumes a presupposition of existence. Such an assumption is common in studies of deductive reasoning and was endorsed by participants.

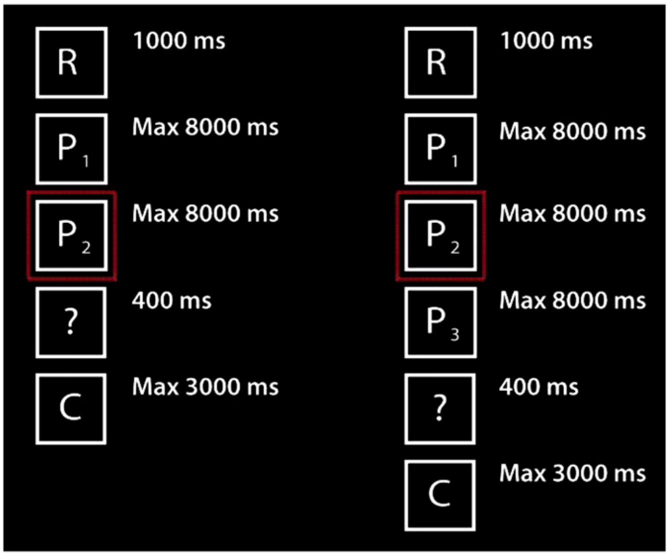

Fig. 1. Schema of stimulus presentation. P: Premises; C: Conclusions. Subjects pressed a button after completion of each task phase: P_1, P_2, P_3 and C. Each key-press was followed by the presentation of a blank screen for 2 s. Most trials contained only two premises (left), while some non-integrable problems also contained a third (right). Examples of stimuli can be found in Table 1. We modeled the BOLD signal time-locked to each event. However, only the signal associated with the presentation of the second premise was used for statistical inference at the group level.

non-existent two-syllable adjectives with a legal phonological structure in Italian (e.g. "every rufa thing is tenna").

Memory trials contained only one sentence (P_1), followed by a set of four alternative answers. The single premise was either a conditional (n = 20) or a quantified (n = 20) statement, with the same sentence structure as used in deductive problems.

Integrable vs. non-integrable premises

The critical premise for fMRI analyses was the second premise (P_2). It could either share (48 problems) or not share (24 problems) an adjectival term with the first premise (P_1). This feature defined whether the premise (and hence the problem containing it) was *integrable* or *non-integrable*. For example, the premises

P_1 Every thing *b* is *a*
P_2 Every thing *b* is *c*

share the common term "*b*". Hence, the problem is considered *integrable*: the presence of a common term allows the generation of a deductive conclusion from P_1 and P_2.

By contrast, in the premises

P_1 Every thing *a* is *b*
P_2 Every thing *c* is *d*

P_2 does not share any common term with P_1: the problem is *non-integrable* and no inference can be drawn.

In a subset of these non-integrable problems (10 problems) a third premise was added which was always integrable with P_2. For example:

P_1 Every thing *a* is *b*
P_2 Every thing *c* is *d*
P_3 Some thing *c* is *e*

This problem is not integrable in P_2. However it is integrable in P_3, with a valid conclusion such as "Some thing *d* is *e*". It is important to clarify, that since our main fMRI analyses only consider activity during processing of P_2 (see below), we defined problems as integrable or non-integrable *only* depending on whether P_2 is integrable or not, irrespectively of the possible presence of an integrable third premise.

After the presentation of the premises, every problem was followed by the presentation of four alternative conclusions (presented simultaneously), from which one had to be rapidly chosen. For the integrable problems, P_2 was directly followed by the conclusions. For the non-integrable problems, the conclusions could either directly follow P_2 (in which case the correct answer to the problem was "nothing follows"), or else a third premise (P_3) could appear before the conclusion. The third premise always had a term in common with P_2, so that a deductive conclusion could follow from the combination of P_2 and P_3. In trials with a third premise, the four target conclusions followed P_3.

The presentation of problems integrable in P_3 allowed us to control that participants fully processed P_2 in the non-integrable problems. Without having fully processed and memorized P_2, participants could not generate a deductive conclusion by combining P_2 and P_3 in problems with three premises. Thus, because participants could not know which problem had two or three premises before the presentation of the third premise, their performance on the non integrable problems with three premises allowed us to control that P_2 was fully processed at the moment of its presentation.

In short, for all the integrable problems (i.e. problems integrable in P_2), a conclusion could *always* be generated by combining P_1 and P_2, whereas in non-integrable problems this was never possible. In a subset of non-integrable problems it was possible to generate a conclusion by combining P_2 and P_3. Therefore, the design allowed us to always compare integrable and non-integrable premises in P_2 while ensuring that, even in non integrable problems, participants paid due attention to P_2, because it might be needed for a possible integration with a P_3. This excluded the possibility that any difference at the P_2 stage could be due to differences in attentional resource allocation between integrable and non-integrable problems.

Types of syllogism

We used fifteen types of syllogism (see Table S1, supplementary material online for the full list of the syllogisms used). Nine of them were drawn from the easy end of the spectrum of syllogism types (Dickstein, 1978). In order to solve them, a simple non-logical heuristic (Reverberi et al., 2009b) could suffice. For example, consider the following two premises:

P_1 Every thing *a* is *b*
P_2 Every thing *b* is *c*

This problem could readily suggest the conclusion

C Every thing *a* is *c*.

This conclusion is indeed valid, but a simple strategy matching the quantifier words in the premises would suggest the same conclusion without supposing that participants engaged in a logical reasoning process. By contrast, the remaining six problems could not be solved correctly with simple heuristics, such as the atmosphere heuristic illustrated above. For example, from the two premises:

P_1 No thing *a* is *b*
P_2 Every thing *b* is *c*

there is no logically correct conclusion: thus, the right answer would be "nothing follows". However, somebody applying simple heuristics could accept the (wrong) conclusion that shares the surface form of one premise, for example:

C No thing *a* is *c*.

Thus, wrong answers to such problems allowed us to disentangle whether participants gave logically valid answers, or were using simple non-logical heuristics. This manipulation had a twofold aim. Firstly, during training (with feedback at the end of each trial), participants became aware that the use of the atmosphere heuristic was not a viable strategy for correctly solving all problems. Secondly, during fMRI scanning performance on the non-heuristic set allowed us to measure the extent to which participants used a heuristic strategy to solve syllogisms (Reverberi et al., 2009b).

Overall structure of the problems analyzed

Overall, during fMRI scanning, 48 integrable syllogistic problems were administered. Of these 48 problems, 24 were integrable-easy and 24 belonged to the non-heuristic set. Furthermore of the 24 syllogistic non-integrable problems, 10 were followed by a third premise P_3. When P_3 was present, a deductive conclusion could always be generated by combining P_2 and P_3. Thus altogether, a deductive conclusion could be generated in 58 problems: 48 integrable problems plus 10 non-integrable problems with a P_3. The set of problems included multiple repetitions of syllogisms with exactly the same formal structure but with different superficial features (i.e. different non-words). Specifically, ten problem types were administered four times and one six times. This paradigm feature was important for computing the consistency index (see below). The number of repetitions of each type of syllogism was not the same for all problem types. The reason for this was that, given the aim of the study, we balanced the difficulty levels of the problems, as well as their validity and their proneness to induce atmosphere biases. Furthermore we did this while maintaining a high variability of syllogistic forms and matching the integrable/non-integrable status of P_2. Satisfying all these constraints and, at the same time, presenting the identical number of repetitions of each problem type was not possible, thus we opted to control for features of the material that were most important for our specific aims.

Procedure

The experiment was carried out in Italian using Presentation™ software (www.neurobs.com). Participants were required to solve a deductive problem about imaginary features of some objects. All the premises were to be assumed true.[1] Participants were asked to read each premise and – whenever possible – to draw a new conclusion as promptly and accurately as possible (without uttering it). At the end of each trial, participants were asked to recognize the conclusion they inferred. They were informed that if they were unable to make a new deduction on P_2, they should nevertheless read it conscientiously because it was critical for establishing a conclusion with a P_3 on certain trials. Each trial started with a central cue lasting a second (Fig. 1): "R" introduced reasoning trials, while "M" introduced memory trial (see below). After an average delay of 3 s (range 2–4 s), premises were shown one at a time. Premises and conclusions never appeared together on the screen. The presentation rate was in part controlled by participants, who were required to press a key as soon as they were ready to proceed to the next premise or to the conclusions. Once the key pressed, a blank screen (the inter-stimulus delay) was presented for 2 s. The maximum time available for processing each premise was 8 s. If a participant failed to press the key within the available time, the trial was interrupted and scored as incorrect.

After the final premise a question mark was presented for 0.4 s, anticipating the presentation of four alternative conclusions. The four conclusions were presented simultaneously on the screen. Participants had to recognize the sentence representing their final conclusion as rapidly as possible and indicate it by pressing the relevant buttons with the index and middle fingers of both hands on an MR compatible button box. Only one out of four conclusions was correct, corresponding to a chance level of 25%. Three seconds were permitted for answering at this stage. It is important to realize that this short time was barely enough to recognize the target sentence among three distractors. Due to these strict time limitations it is a plausible assumption that no further reasoning was possible at the conclusion evaluation stage. Participant debriefing and previous evidence with the same procedure (Reverberi et al., 2007) confirm this assumption. Our experimental paradigm forced participants to produce an inference during processing of P_2 or P_3. The average duration of a trial across participants and problem types was 19.8 s (SD = 3.0 s). The 72 syllogisms of the experimental phase were divided into four fMRI runs comprising 18 syllogisms each. The fMRI scanning lasted on average 55 min, partially depending on participant problem processing speed.

Before fMRI scanning, all participants underwent a training session. During training, we presented problems similar to those used during scanning. Training problems presented the same logical formal structures. Unlike experimental fMRI sessions, participants received feedback at the end of each training trial. The training phase ended either after at least eight correct responses were made out of 10 consecutive trials of easy syllogistic problems, or after 45 min regardless of performance. A minimum of 40 training trials was administered.

As noted above, the experimental protocol also included some memory trials. In these trials, subjects were told to read and remember sentences carefully for fast recognition from among four subsequently presented sentences. The memory trials began with a central cue ("M") presented for 0.4 s followed by a delay lasting on average 3 s (2–4 s range). Participants were then presented with either a conditional or a quantified statement (P_1). They had to press a key as soon as they were ready to proceed to the next phase. Again, a maximum of 8 s was allowed after which the trial was interrupted and marked incorrect. Once participants pressed a key, a question mark was shown for 0.4 s. Four alternative and numbered sentences followed. The task was to choose the sentence identical to P_1 and to press the corresponding key (maximum response time 5 s). The overall duration of each memory trial ranged from a theoretical minimum of 4.8 s to a maximum of 20.8 s, depending on how fast participants responded to premises and drew conclusions. The average duration of a trial across participants and problem types was 12.1 s (SD 2.8 s).

Dependent variables

The following behavioral variables were considered.

(i) Average accuracy across all syllogistic problems. Accuracy was assessed separately for integrable and non-integrable problems.
(ii) Reaction times on integrable and non-integrable sentences for both conditional and syllogistic problems.
(iii) Consistency index. This index measures whether participants answer identically to problems that are superficially different, but share the same formal structure. As this index is not used in the literature, we explain its meaning in detail. For each problem type we first assessed how many different responses each participant chose across all repetitions of the same problem. For example, consider a problem with the following two premises:
P_1: "Every thing *b* is *a*";
P_2: "Every thing *b* is *c*".
Sometimes a participant may select "Every thing *a* is *c*" as a conclusion; at other times s/he may select "No thing *a* is *c*", thus providing two different responses to the same problem

[1] While assuming that a nonsense sentence is true may be different from assuming that a meaningful sentence is true, participants had no difficulty in assuming the putative truth of a nonsense premise for the sake of argument. Here we choose nonsense material to make sure that our experimental procedure would activate selectively processes related to logical form analysis, rather than any process related to meaning retrieval.

type. We thus compute the average number of different response types given by a participant across all eleven problem types with at least four replications (see Table S1): This can be thought of as the average dispersion for answers to formally identical stimuli. It ranges from a theoretical minimum of 1 (no dispersion of answers), to the maximum of 4 (full dispersion). We then transformed the average dispersion number into a consistency index ranging from 0 (no consistency) to 1 (maximum consistency) by applying the following transformation:

$$C = 1-[(D-1)/(4-1)]$$

where C is the consistency index, D is the average dispersion, and 4 is the maximum dispersion of the answers. A consistency index equal to 1 indicates that a participant always gave the same answer to problems with the same formal structure. Trials with no answer were excluded from analysis.

(iv) Validity index: This index assesses the tendency of participants to use a valid procedure in order to generate conclusions. It is similar to accuracy but trials with no answers were excluded from its computation. This is because, in those cases, we could not judge whether following P_2 participants generated a correct (or incorrect) answer but failed to respond within the relatively short response time-window, or failed to generate any answer at all.

(v) Heuristic index: This index measures the proportion of responses to problems in the non-heuristic set that were wrong, but consistent with premise "atmosphere". For example, in a problem like:
P_1 "every thing *b* is *a*";
P_2 "every thing *b* is *c*",
the correct answer would be "some thing *a* is *c*". However, the atmosphere heuristic would incorrectly lead to a different choice, namely, "every thing *a* is *c*" (Table 1). Out of the four alternative conclusions proposed for each problem, only one was consistent with the atmosphere heuristic. Thus, participants could also choose answers that were both not valid and failed to follow a heuristic strategy. For example, in the problem above subjects could answer: "nothing follows". Therefore, the heuristic index signals how frequently participants selected a wrong answer on the basis of a heuristic strategy. The index ranges from 0, meaning a bias toward non-heuristic responses, to 1, meaning the systematic use of atmosphere heuristics in all problems.

Identification of the behavioral subgroups: partitioning criteria

On the basis of the consistency, validity and heuristic indices we identified three pairs of subgroups. We used the consistency index to separate participants providing the same answers to formally identical syllogisms from those who did not. We classified participants as highly consistent with a consistency score *below* the group median. Similarly, participants with a validity index above the median were classified as using valid procedures for solving deductive problems, while participants below the median were considered not. Two participants were unclassifiable according to validity because they had a validity index identical to the median. Finally, we identified participants solving problems by the atmosphere heuristic by comparing their individual heuristic indices to chance level (0.25) with a binomial test. When the heuristic index was reliably higher than chance ($p<0.05$), a participant was classified as showing a bias toward the heuristic-driven response.

Image acquisition

Imaging was carried out in a 3T Siemens Allegra head scanner (Siemens, Erlangen, Germany). BOLD contrast was obtained using echo planar T2*-weighted imaging (EPI). The acquisition of 32 transverse slices, in ascending order, provided coverage of the whole cerebral cortex. Repetition time was 2.08 s and in-plane resolution was 3×3 mm; slice thickness and gap were 2.5 mm and 1.25 mm, respectively. Time-to-Echo was 30 ms, and the flip angle was 70°.

Data analysis

Behavioral data were analyzed with the SPSS statistical package. We used a significance threshold of $p<0.05$, two-tailed if not explicitly stated otherwise. Imaging data were analyzed using SPM8 (www.fil.ion.ucl.ac.uk/spm). The first four image volumes of each run were discarded to allow for stabilization of longitudinal magnetization. The overall number of volumes available partly depended on the average speed of each participant. Thus, for each participant, we obtained an average of 1516 volumes (SD = 120), ranging from a minimum of 1320 to a maximum of 1748 volumes. Pre-processing included rigid-body transformation (realignment) and slice timing to correct for head movement and slice acquisition delays. The images were then normalized non-linearly into MNI space using the mean of the functional volumes and smoothed with a Gaussian filter of 8 mm FWHM. The time series for each participant were high-pass filtered at 128 s and pre-whitened by means of an autoregressive model AR(1) (Friston et al., 2002). Statistical inferences were based on a random effects approach (Friston et al., 1999; Penny et al., 2004) that comprised two steps. First, the data were best fitted at every voxel for each participant using a combination of effects of interest. The effects of interest were the onset times of the considered event types. Onsets corresponded to the time of appearance on the screen of the specific stimulus type, delayed by 1 s to take account of the initial reading of sentences (Goel and Dolan, 2003a; Reverberi et al., 2010). Given that validity is one of our predictors, we did not exclude trials with incorrect answers. We also modeled events that, while not considered in second-level analyses, may have produced specific hemodynamic responses such as the presentation of the first premise and the conclusions. All events were modeled as mini-blocks with the duration corresponding to the presentation time of stimuli on the screen (mean = 3.02 s, SD = 1.33 s). All stimulus functions were convolved with the standard SPM8 hemodynamic response function. Linear compounds (contrasts) were used to determine responses for the integration effect (P_2, integrable > non-integrable sentences). This resulted in the generation of one contrast image per participant. The contrast images then underwent a second step comprising three one-way ANOVAs (for validity, consistency and heuristics). Each ANOVA modeled the average integration effect in each of the subgroup pairs described in the section "partitioning criteria". Correction for non-sphericity (Henson and Penny, 2003) was applied to account for possible differences in error variance across subgroups. Linear compounds using between-participant variance were devised to assess both the simple effects of integration in each subgroup (supplementary material), and the interaction integration × subgroup. For the latter we considered these contrasts: high > low consistency subgroup, high > low validity subgroup, and subgroup using atmosphere heuristic > subgroup not using it. The interaction integration × subgroup was tested only in the subsets of voxels that were active for the integration effect over all subjects. Thus a Small Volume Correction (SVC) was applied using the volume shown in Fig. 2. In voxel-level analyses, we considered effects as being significant at $p<0.05$, corrected for multiple comparisons with the Family Wise Error (FWE) procedure.

We identified a reasoning network comprising six different nodes. Each node was one of the main clusters of activation in the map of the integration effect across all participants, irrespectively of subgroup membership. These comprised three nodes in the frontal cortex (nodes number 2, 3, and 4), one in the basal ganglia, and two in the parietal cortex, the medial parietal node and lateral occipital node (Fig. 2 and

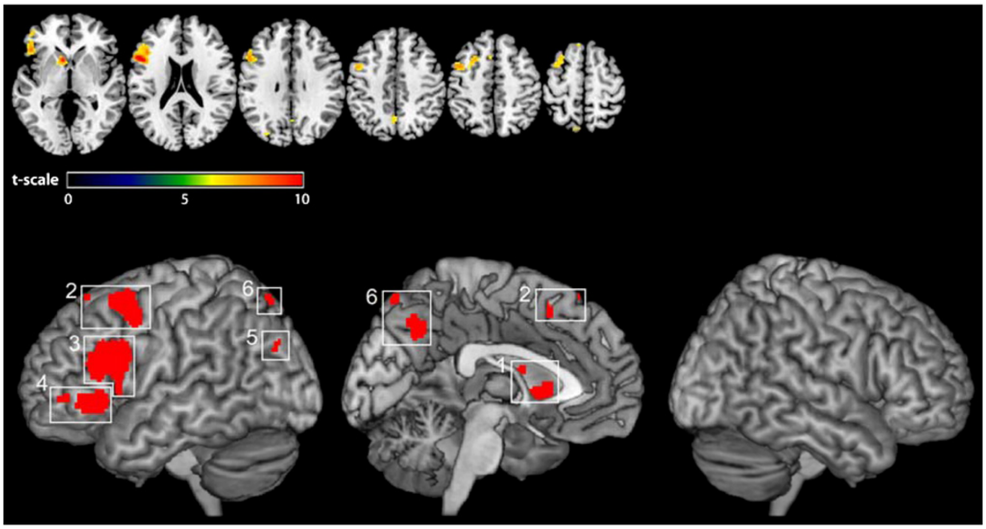

Fig. 2. Top: Active brain areas during generation of new deductive conclusions. This analysis included the whole group of participants ($n = 26$; $p < 0.05$, corrected for multiple comparisons). Middle: The activated areas are rendered onto a single-subject standard T1-weighted brain image (from left to right: left, medial and right surface of the brain). Bottom: The six nodes derived from the preceding analysis. These comprised three nodes in the frontal cortex (nodes 2, 3, and 4), basal ganglia (node 1), medial parietal cortex (node 6) and lateral occipital cortex (node 5). The average activity of these nodes was used to predict individual behavior.

Table 3). For each node we computed the mean integration effect across all voxels in the node with the SPM toolbox "marsbar" (marsbar.sourceforge.net). The integration effect was calculated by contrasting the activation related to integrable P_2 to that related to non-integrable P_2. Overall, for every participant, for each node, we obtained one index of the neural activity related to the integration of premises. We evaluated whether the large-scale pattern of activity across the six nodes allowed prediction of how each participant performed a given deductive task. Three sets of logistic analyses were carried out; for each participant, group membership (consistency, validity, heuristics) was considered as the critical dependent variable. The predictors consisted of average integration effect in each of the six nodes. To identify brain regions that were independently informative for classification we used a stepwise forward procedure. The use of a stepwise procedure also prevented errors in logistic regression classification from the inclusion of irrelevant variables. The stepwise logistic regression was implemented following a leave-one-out cross-validation procedure, in order to avoid using the same data for both model estimation and class membership allocation (Kohavi, 1995). Thus, we estimated the optimal logistic model and its regression parameters on a training subset of 25 participants (23 in the case of validity classification). We then used this model to predict the class membership of the remaining participant (the test sample). The same procedure was repeated iteratively using a different training subset until every participant had been used as a test sample once. Classification accuracy for each test sample was then evaluated against chance (50% classification accuracy) with a chi-squared test. All logistic analyses were performed using MATLAB.

Results

Behavioral performance

Average accuracy was 55% (SD 11%) for integrable and 96% (SD 6%) for non-integrable syllogistic problems (not considering non-integrable problems with a third, integrable sentence). Both values were significantly higher than the chance level (i.e. 25%; integrable: $t(25) = 14.5$, $p < 0.001$; non-integrable: $t(25) = 58.2$, $p < 0.001$). In the subset of problems with a third integrable sentence (P_3) the accuracy was 64% (SD 22%). The good accuracy on problems with a non-integrable P_2 followed by an integrable P_3 strongly suggests that even non-integrable P_2 are adequately processed. Reaction times on second premises with integrable sentences were slower than with non-integrable sentences (respectively, 3866 ms and 3161 ms), producing a significant integration effect of 695 ms (paired $t(25) = 6.59$, $p < 0.001$). Given the relatively long duration of the experiment (on average 55 min), we also checked for practice/fatigue effects by considering accuracy and reaction times on the second premise of integrable problems over time. We found no systematic change over time ($R < 0.02$; $p > 0.1$; when serial position over time was used to predict either accuracy or reaction with linear regression).

The average consistency index was 0.69. This value is significantly lower than 1 ($t(25) = 14.07$, $p < 0.001$), the value expected in participants always answering consistently. However, the consistency index was also higher than 0.41 ($t(25) = 12.79$, $p < 0.001$), the value expected were all problems given random answers. Thus, overall, our participants show a pattern of response that is different from chance but that is not perfectly consistent.

The average validity index was 0.60, slightly more than the average accuracy to which it is related. This is due to exclusion from the computation of the validity index of trials without answer; which were on average 4.4 per participant. The average heuristic index (HI) was 0.44 ($SD = 0.17$). This value is different from the three theoretical values expected in a behaviorally homogeneous group. Namely, HI was different from 0 ($t(25) = 13.37$, $p < 0.001$), the value expected in participants who systematically avoid heuristic answers; it was different from 0.25 ($t(25) = 5.78$, $p < 0.001$), expected in participants answering randomly and thus unbiased toward atmosphere-consistent conclusions; finally, it was different from 1 ($t(25) = 17.01$, $p < 0.001$), expected in participants always employing the "atmosphere" heuristic. Our participants were heterogeneous in their propensity to use the heuristic, as demonstrated by additional analyses of individual performance. Each participant was compared to chance (0.25) by means of a binomial test. Ten of twenty-six participants showed an HI biased toward non-heuristic responses or insignificantly different from chance (average HI in this subset is 0.26), indicating that they avoided heuristic strategies. In contrast, sixteen participants showed an HI reliably higher than chance (i.e. >0.25), but still less than the expected value – i.e. 1 – for consistent use of heuristics (average = 0.55; range 0.46–0.75). This pattern suggests a significant but inconsistent use of heuristic strategies.

Consistency, validity and heuristic indices are theoretically related. A completely inconsistent participant should not score highly on validity or heuristic indices and a participant with a high validity index will also obtain a high consistency score. Participants

Table 2
Main behavioral variables in each of the six subgroups identified by means of the consistency, validity and heuristic indices. For each subgroup we report the mean and standard deviation of the consistency, validity and heuristic indices, the simple Reaction Times (ms) on the premises (P_1 and P_2) and conclusion, the Integration time (RT on integrable minus RT on non-integrable premises), and the accuracy on memory trials. RTs on conclusion phase are reported for integrable problems.

	Consistency		Validity		Heuristic	
	High	Low	High	Low	Yes	No
Group size	13	13	12	12	16	10
Consistency	0.78 (0.03)	0.60 (0.09)	0.73 (0.07)	0.64 (0.13)	0.71 (0.11)	0.66 (0.12)
Validity	0.62 (0.09)	0.57 (0.10)	0.67 (0.06)	0.52 (0.06)	0.57 (0.06)	0.64 (0.12)
Heuristic	0.49 (0.19)	0.39 (0.13)	0.35 (0.15)	0.51 (0.15)	0.55 (0.09)	0.26 (0.09)
RT P_1	3027 (977)	2781 (440)	2760 (823)	2865 (554)	2944 (753)	2839 (788)
RT P_2	3651 (812)	3618 (744)	3521 (655)	3617 (877)	3603 (848)	3684 (643)
RT conclusion	2630 (492)	2900 (376)	2503 (387)	2942 (354)	2839 (451)	2645 (446)
Integration time	666 (478)	722 (609)	862 (625)	573 (420)	608 (415)	832 (693)
Memory trial accuracy	96% (5)	94% (6)	94% (6)	94% (6)	96% (4)	92% (7)

principally using heuristic strategies would tend to have low validity and high consistency indices.[2] Accordingly, we evaluated the weight of these relationships between variables empirically. Correlation patterns in our group partially confirmed these theoretically driven expectations. We observed a significant correlation between consistency and validity ($r=0.56$, $p<0.01$), and between heuristic and validity ($r=-0.55$, $p<0.01$). Thus, the more consistent participants were, the higher validity they attained; and the more heuristic a strategy they used, the less valid their answers tended to be. By contrast we failed to find any significant correlation between consistency and the use of heuristics (0.20, $p=0.34$).

Following the procedures detailed in the Methods section, the consistency, heuristic and validity indices were used to generate three pairs of subgroups (Table 2). Each of the subgroups was relatively homogeneous with respect to one of the three dimensions explored: consistency, heuristic and validity. As expected, the grouping variable was always reliably different between related pairs of subgroups (Consistency: $t(24)=6.94$, $p<0.001$; Validity $t(22)=6.23$, $p<0.001$; Heuristic: $t(24)=8.03$, $p<0.001$). However, the validity subgroups differed also in consistency ($t(22)=2.28$, $p=0.032$) and heuristic ($t(22)=2.70$, $p=0.013$), which is consistent with the correlation analyses reported above. By contrast, overall reaction times on P_1, P_2, the conclusion phase, and the integration time on P_2 were similar in all subgroup pairs (Table 2, all $p>0.1$). The only exception to this general RT pattern was the difference in reaction times at the conclusion stage between the two validity subgroups: participants who generated valid responses were faster than participants who did not ($t(22)=2.9$, $p=0.008$). All subgroups were similar in terms of accuracy on non-integrable sentences without a P_3 (all $p>0.1$); and no accuracy differences were found on memory trials (all $p>0.1$).

One possible worry is that the criteria used to identify the different subgroups may not selectively highlight differences in problem solving strategies. For example, it could be argued that, because we define consistency as response-constancy across variation of nonsense words in an argument, we may simply distinguish subjects who paid more or less attention to the tasks. Likewise, it could be argued that participants who we classified in the heuristics group were simply paying less attention to the premises, relying on quick answers based on the form of the conclusions. Finally, given that other studies have indicated that heuristic strategies are often implemented at the conclusion phase (Reverberi et al., 2009b), it could be argued that our analyses focusing on P_2 may have failed to correctly identify subjects using heuristics. However, the set of behavioral control analyses presented above indicates that these objections are unlikely.

First, systematic changes in attention levels between pairs of a subgroup should result in differential RTs and a fall in accuracy in non-integrable sentences and memory trials too. None of these effects were observed. Second, a shift of the processing load from P_2 to the conclusion phase in the heuristic subgroup should cancel out the integration effect on P_2 and should increase the processing time in the conclusion phase. Again none of these behaviors was found in our dataset.

All together the behavioral data show that the differences between pairs in the three subgroups (consistency, validity and heuristics) were reasonably specific along the dimension measured by each index, while other performance indices were overall well-matched.

Identification of the reasoning network

We identified brain areas associated with the integration of premises by contrasting the activation related to non-integrable sentences with that related to integrable sentences during P_2 processing (Reverberi et al., 2007, 2010). It is important to note that this analysis is blind to subgroup divisions, which we only considered in successive analyses. The network of areas active during the generation of deductive sentences was left-lateralized, involving mainly left frontal areas ($p<0.05$, corrected). Besides activations in the frontal cortex, foci of activation were also observed in the left caudate nucleus, in the precuneus and in the lateral occipital cortex (Fig. 2 and Table 3). The observed reasoning network is compatible with findings of previous studies (e.g. Goel et al., 2000; Monti et al., 2007; Prado et al., 2010; Reverberi et al., 2007). Six nodes were identified in the reasoning network. These six nodes were used in the following analyses (Fig. 2).

Predicting behavior from the large-scale pattern of activity

We evaluated whether the large-scale pattern of activity across the six nodes predicted how a participant approached deductive tasks. Three sets of logistic analyses were performed. The dependent variable was group membership (consistency, validity, heuristics). By using average activity in the six nodes, we successfully classified participants into consistency and validity subgroups, but we failed to predict heuristic subgroup membership.

Specifically, for consistency 73% of participants were correctly classified into high or low consistency subgroups ($\chi^2(1)=5.57$, $p=0.02$). The node selected by the stepwise logistic procedure, the left BA44/45, was the same in all 26 iterations (Fig. 2, node 3). The regression coefficient estimated for the frontal node was positive, meaning that the more BA44/45 was active, the higher the chance of belonging to the high consistency group. We further checked

[2] Notice however that a deterministic association between these indices is true only at extreme values. For example, a participant with a validity index of 1 (fully valid) will have a consistency index of 1. By contrast, a participant with a validity index of 0.5, could have a consistency score ranging between 1 (fully consistent) and 3, depending on how invalid responses are distributed. As a consequence, a participant with a validity index of 0.5 could be classified as being either consistent or inconsistent.

Table 3
Peak activations for the reasoning network over all participants (n = 26). Coordinates [x, y, z] in space of Montreal Neurological Institute (MNI) template.

	Brodmann area	x	y	z	t scores
Frontal node 2, 394 voxels					
Precentral gyrus	6	−46	6	48	8.10
Frontal middle gyrus	8	−26	16	54	7.74
Frontal middle gyrus	6	−38	8	60	6.91
Supplementary motor area	6	−6	16	51	6.96
Frontal superior medial gyrus	8	−8	32	58	6.68
Frontal node 3, 820 voxels					
Inferior frontal gyrus	44	−52	16	20	12.00
Inferior frontal gyrus	45	−44	24	18	8.25
Inferior frontal gyrus	44	−54	14	6	7.04
Frontal node 4, 232 voxels					
Inferior frontal gyrus	45	−48	46	0	6.89
Inferior frontal gyrus	47	−48	28	−2	8.63
Basal ganglia (node 1), 157 voxels					
Caudate nucleus		−8	14	2	10.08
Caudate nucleus		−16	0	16	6.62
Medial parietal (node 6), 110 voxels					
Precuneus	7	−8	−74	58	7.38
Precuneus	7	0	−60	36	6.90
Lateral occipital (node 5), 12 voxels					
Occipital middle gyrus	19	−32	−78	30	6.57

how specific the distribution of information was to BA44/45. Thus, we systematically evaluated the performance of models resulting from all possible combinations of the six predictors (i.e. activity in the six nodes). Sixty-three cross-validated logistic models were evaluated in this way. This analysis confirmed that the model selected by the stepwise procedure resulted in the highest accuracy. Furthermore, in all models with accuracy greater than 60% (n = 27) frontal nodes 3 and 2 were also predictors, either alone or in combination with other nodes. The best model that excluded these two nodes performed at chance (50%). Finally, when both frontal nodes were introduced into the same model, the independent contribution of the frontal node 2 was negligible. That is, all the information provided by frontal node 2 is already contained in frontal node 3. Overall, these analyses show that left BA44/45 (frontal node 3) is the brain area whose activation profile is maximally informative about the consistency of responses. Some information is also available from frontal node 2, while all the other nodes, either alone or combined, provide no further information of significance.

These results must be considered in the light of the definition of consistency and the specific material we used. In our study, arguments changed only because they contained different nonsense words. Hence, the non-logical, non-formal content of an argument was factored out by design. For our purposes the material characteristics are advantageous because they highlight activations specifically associated with the formal structure of problems. Our findings show that when a paradigm emphasizes the role of the formal structure of a problem, e.g. by removing any possible confounds due to problem content, activation of BA44/45 predicts sensitivity to the problem's formal structure. We acknowledge that other tasks, including problems with content variations, may produce different results. Nevertheless, our paradigm detects any reasoning process related to the formal structure of premises. An analogous experimental strategy has been applied in previous reasoning studies (e.g. Goel et al., 2000; Monti et al., 2007; Noveck et al., 2004; Reverberi et al., 2007) and in studies of syntactic processing (e.g. Pallier et al., 2011).

In the case of validity, 79% of participants were classified correctly ($\chi^2(1) = 9.91$, $p = 0.002$). Frontal node 4 (Fig. 2, mostly BA 47) and the lateral occipital node (Fig. 2, node 5) were identified by stepwise logistic analysis in all 26 iterations. The coefficient estimated for BA47 was positive; thus, the higher the activity in BA47, the greater the probability of belonging to the high validity group. In contrast, the coefficient estimated for the lateral occipital node was negative; that is, the lower the lateral occipital cortex activation (i.e. tending towards *no* activation), the greater the chance of belonging to the high validity group. Again, we checked whether group assignment depended specifically on inferior frontal and lateral occipital areas. The systematic exploration of all possible models confirmed that the model selected by the stepwise procedure provided the best cross-validated accuracy. In all models with an accuracy level higher than 60% (n = 21) frontal node 4 was identified as a predictor, in most cases together with the lateral occipital region (n = 16). The best model containing neither node was uninformative (accuracy at chance, smaller or equal to 50%). When both nodes were included in the model, the contribution of each remained significant. All the other nodes, alone or in combination, provided no further information on performance validity.

Finally, for heuristics the stepwise procedure selected only the constant as a predictor. Thus, all participants were classified as belonging to the larger subgroup, corresponding to an accuracy of 16/26 = 62%. The introduction of all variables in the model failed to significantly improve accuracy (65%, $p > 0.1$ when compared to the outcome of the constant model or to chance). Together these findings show that the pattern of activation in none of the nodes contained information predictive of use of heuristic strategies.

Activations across subgroups

Next we tested for the presence of differential activations between each pair of the three behaviorally identified subgroups (consistency, validity and heuristics). For this, we ran three two-sample t-tests, one for each pair of subgroups. We explored whether and where subgroups selected for high consistency, high validity or the frequent use of heuristics had greater activations than subgroups with lower consistency, validity and use of heuristics. These analyses focused on areas associated with deductive reasoning as identified by an analysis of all 26 participants (cf. Fig. 2), using a Small Volume Correction procedure (Worsley et al., 1996). For completeness, we also verified reasoning-related activations in each subgroup separately. This analysis highlighted that all subgroup activations lie within the reasoning network detected by the main analysis (see online supplementary material). Thus a reduction of the volume interrogated implicit in the use of a Small Volume Correction is very unlikely to miss significant brain areas (e.g. areas active in one subgroup only). The only comparison showing a significant effect of subgroup involved the consistency groups. The high-consistency subgroup activated left inferior frontal gyrus (peak in −42 14 20, cluster size 22 voxels, $p < 0.05$, corrected) more than the low-consistency one; there was also a statistical trend in the left middle frontal gyrus (peak in −42 6 58, cluster size 6 voxels, statistical trend: $p = 0.08$, corrected). The two regions lay, respectively, within frontal node 3 (Fig. 2, BA 44/45) and frontal node 2 (BA 6/8).

To increase statistical power, we ran the same comparisons using a more focused approach, exploring the average integration activity in each node across subgroups (Fig. 4). In accordance with the voxel-based analysis, frontal nodes 2 and 3 (Fig. 2) were more active in the high-consistency than low-consistency subgroups (respectively, $t(24) = 2.96$, $p = 0.007$; $t(24) = 2.40$, $p = 0.024$). This analysis also revealed differential activation between the two validity subgroups. Specifically, we found that the valid subgroup activated frontal node 4 significantly more than the invalid subgroup ($t(22) = 2.61$, $p = 0.016$). In the heuristic subgroups, analyses centered on nodes showed no significant effect in any of the six nodes of the deductive network (all $p > 0.1$). Finally, we directly compared activations in the three nodes associated with classification into consistency and

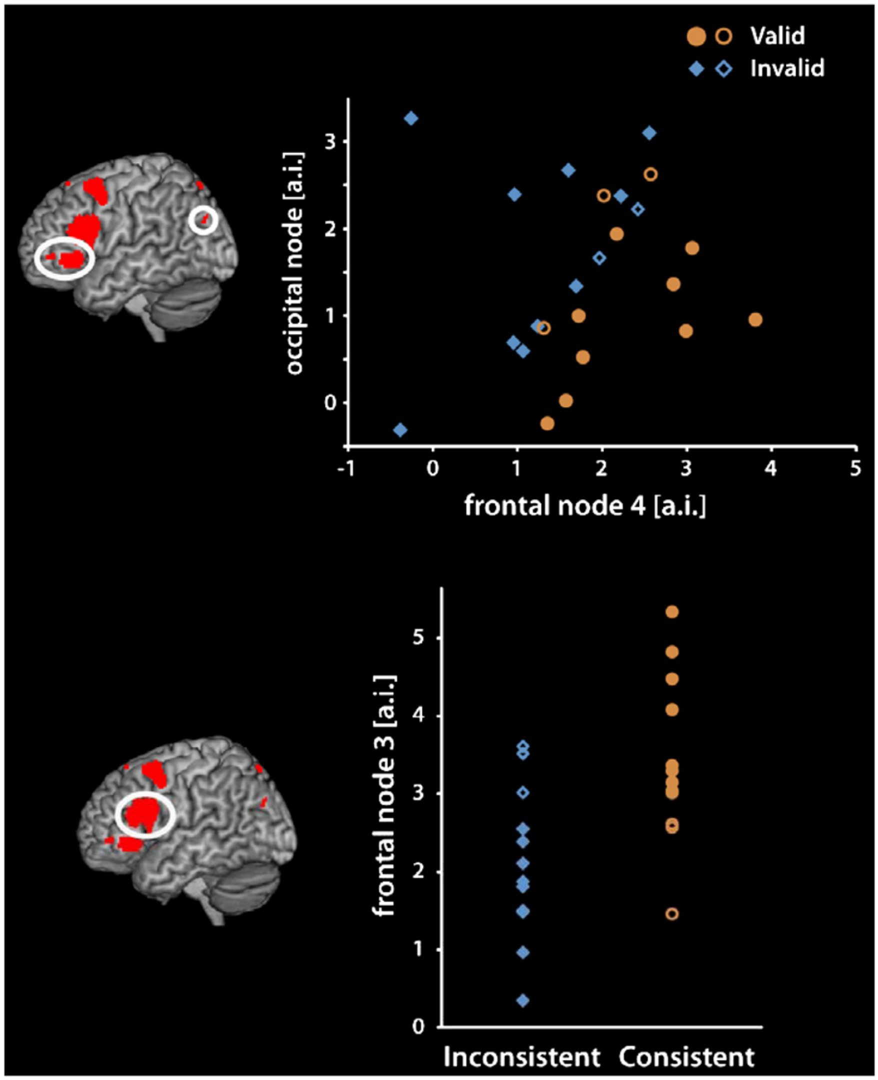

Fig. 3. Prediction of behavioral performance by large-scale patterns of brain activity. Top: The combined integration effect (integrable>non-integrable premises) in frontal node 4 (BA47) and in the lateral occipital node predicted whether a participant belonged to the low or high validity groups with 79% accuracy ($p<0.05$, cross-validated). The higher the integration effect of frontal node 4, the higher the probability that the participant belonged to the high validity group; the higher the integration effect of lateral occipital node, the lower the probability of being classified in the valid group. Bottom: The activity in frontal node 3 predicted whether a participant belonged to the low or high consistency groups with 73% accuracy ($p<0.05$, cross-validated). The higher the integration effect in frontal node 3, the higher the probability of being classified in the coherent group. The filled symbols in the plots represent participants correctly classified by logistic analysis. The activation always refers to the contrast integrable minus non-integrable sentences during the presentation of the second premise.

validity subgroups. We ran a 3 (node, within participant) × 2 (consistency) × 2 (validity) ANOVA. This revealed a main effect of node ($F(2,40)=11.63$, $p<0.001$), showing that the average activation level was different across the three nodes. Most importantly, there was both a node × consistency interaction ($F(2,40)=4.69$, $p=0.015$) and a node × validity interaction ($F(2,40)=8.39$, $p=0.001$). These results are consistent with the preceding analyses of simple effects: The pattern of activity is different across nodes in both the consistency and validity subgroups, even when accounting for co-variation between validity and consistency.

Comparison with preceding studies

In order to better relate our results to findings reported by previous studies of deductive reasoning, we evaluated whether other brain areas, previously associated with deductive reasoning, were able to predict participant behavior. We considered two previous imaging experiments on deductive propositional reasoning (Monti et al., 2007; Reverberi et al., 2007). Even though the material and procedure of the previous studies are different from those used here, the brain regions associated with deduction in these previous studies may in principle contribute to the prediction of participant behavior also in the present experiment. In Reverberi et al. (2007) two regions were found active during inference making. We evaluated one of those, the left lateral parietal region (BA40, local maximum −44 −46 50), as the other area reported in that study largely overlaps with the left frontal nodes already considered in the present experiment. The entire left lateral parietal cluster originally reported was considered for the new analyses here (see Fig. S4). Average activity in left lateral parietal node fails to predict any of the behavioral groups we consider here (all $p>0.1$). Next, we explored whether the contribution of this additional lateral parietal node adds to information provided by the two nodes we identified as informative for validity in this study (i.e. the lateral occipital and the inferior frontal areas). Introduction of the lateral parietal node in a logistic model containing only the inferior frontal area improved the predictive power for validity

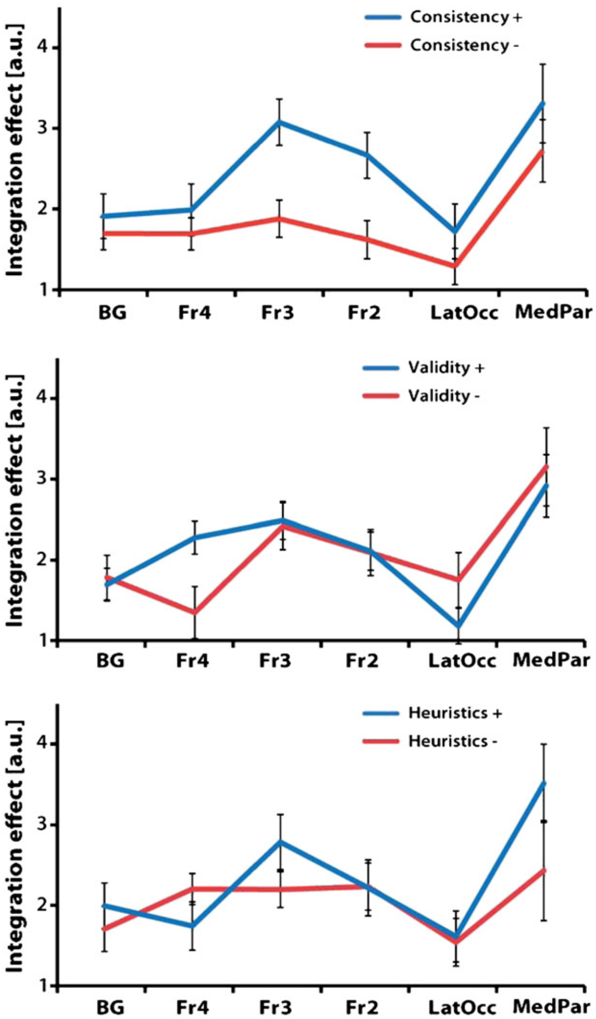

Fig. 4. Average integration effect (integrable>non-integrable premises) across different subgroups in the six nodes examined. They were created on the basis of an analysis of the integration effect over the whole group (n = 26, see Fig. 2). Error bars represent the standard error of the mean. BG: Basal Ganglia; Fr3-5: Frontal nodes 3–5; LatOcc: Lateral Occipital node; MedPar: Medial Parietal node.

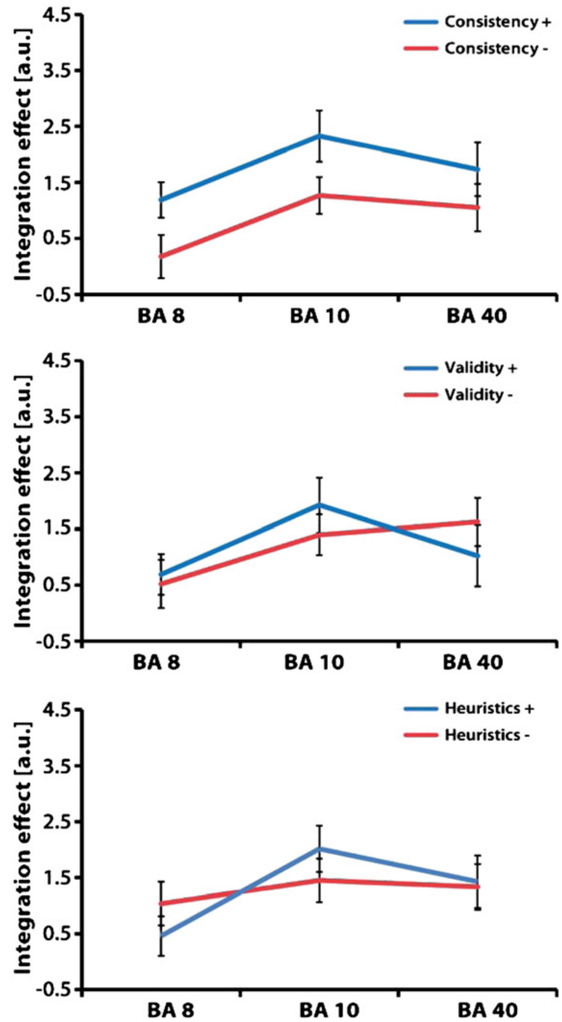

Fig. 5. Average integration effect (integrable>non-integrable premises) across different subgroups in three additional nodes derived from two previous studies of deductive reasoning (Monti et al., 2007; Reverberi et al., 2007). BA 8 and BA10 were derived from the study by Monti and collaborators, while BA40 was derived from the study by Reverberi and collaborators. We found only a statistical trend for higher activation in BA6/8 and B10 in the high consistency subgroup. Error bars represent the standard error of the mean.

classification ($\chi^2(1) = 4.77$, $p = 0.03$). In contrast, introducing the parietal node in the model containing only the lateral occipital area did not ($p > 0.1$). The coefficient of the lateral parietal node contribution in the former model was negative, meaning the less active the area the greater the probability of belonging to the high-validity subgroup. Next, we ran the same analyses using nodes derived from the study of Monti et al. (2007). In this case, we considered BA10p and BA8m, which they consider the “core regions of deduction” (Monti et al., 2007, pp. 1010–1012). We defined two spherical nodes of radius = 8 mm, centered on the maxima of activation reported (−46 50 −4 for BA10p, and −4 36 48 for BA8m). BA8m significantly predicted membership of the high-consistency group ($\chi^2(1) = 4.07$, $p = 0.04$). The coefficient was positive, indicating that the greater the activation of BA8m the higher the probability of belonging to the high consistency group. BA10p showed a near trend in the same direction as BA8m ($\chi^2(1) = 3.59$, $p = 0.058$). Neither of the two nodes was able to predict group membership for the validity groups ($p > 0.1$), both when considered alone and together with either the lateral occipital node or frontal node 4. The same was true in the case of the heuristic group ($p > 0.1$).

Finally, we compared the level of activity in the three additional nodes across the three subgroup-pairs (Fig. 5). BA8 and BA10 showed average activity that tended to be higher in the high-consistency than low-consistency subgroup (BA8: t (24) = 2.02, $p = 0.06$; BA10: t (24) = 1.88, $p = 0.08$). No other significant relationship was found ($p > 0.1$) for these two nodes.

Discussion

Deduction is a fundamental cognitive ability, but its cognitive and neural bases are still largely unknown. Many efforts have been devoted to clarifying how it relates to other cognitive functions and to understand its neural underpinnings. In this paper, we present a novel approach to this controversial area. We exploited differences in solutions to deductive problems to test whether they were associated with specific activation profiles, predictive of individual performance.

We defined subgroups of participants according to the response profiles they displayed in solving identical syllogisms. We considered three dimensions: tendency to consistently respond to the formal structures of problems, (which defined the consistency subgroups), ability to find valid conclusions (the validity subgroups) and reliance on non-logical, heuristic response strategies (the heuristic subgroups). Such group differences are unlikely to be due to the fact that participants allocated differential memory or attentional resources while reasoning. We show that in two cases the large-scale pattern of activation across brain areas involved in inference making can be used to predict individual response profiles. Specifically, we show that brain activation can be used to correctly classify participants into those using consistency and validity criteria. We show that neural information (indexed by changes in regional activation) is differentially distributed across the six areas that constitute the reasoning-associated brain network. In contrast, brain activity does not predict the use of heuristics, at least not for the specific heuristic strategy we evaluated.

Brain areas related to formal consistency

We presented participants repeatedly with formally identical syllogisms with different superficial characteristics in the non-words that were used for their formulation. Thus, we were able to assess how consistently participants answered to problems that were identical in logical structures. Participants differed in their ability to answer consistently (Table 2). We therefore examined brain activity across the brain network engaged by problem solving to see if it was possible to predict individual's tendency to produce consistent responses. The best logistic model, selected by a stepwise procedure, accurately predicted to which consistency subgroup belonged participants in 73% of the cases. Prediction depended on activations of frontal area that we call node 3 in the region of BA44/45 (Fig. 2). Other models, dependant on activity in frontal node 2 (BA6/8), were also good predictors of membership of the consistency group (accuracy 60% or higher). However, models combining information from both regions were no better than using BA44/45 activity alone. Finally, models in which neither frontal node 3 nor 2 were included all failed to predict consistency group membership. Further analyses comparing activations between high and low consistency subgroups (Fig. 4) confirmed the involvement of frontal nodes 3 and 2 in consistent responders. Overall, these analyses show that BA 44/45 and BA6/8 are involved in the successful extraction and representation of the formal structure of premises during processing of deductive problems. Furthermore, together with behavioral findings related to the validity index (see below), they show that the different brain regions engaged by generation of a deductive conclusion are not functionally homogeneous. Different parts of the deductive network are differentially involved in different cognitive operations associated with the processing of deductive problems. Thus, while our study highlights the involvement of BA44/45 and BA6/8 in the extraction of the formal structure of premises, a similar involvement of the inferior frontal (BA47), lateral occipital or medial parietal regions could not be demonstrated.

Several studies of deductive reasoning have found Brodmann areas 44/45 implicated in deductive reasoning (e.g. Goel and Dolan, 2003a; Goel et al., 1998, 2000; Reverberi et al., 2007, 2010). However, critically, in none of these studies it has been possible to dissociate the functional role of BA44/45 from other brain areas activated in the same experiments. The functional role for BA 44/45 that we propose is consistent with the recognized involvement of this brain region in syntax processing (Ben-Shachar et al., 2004; Dapretto and Bookheimer, 1999; Friederici et al., 2006; Grodzinsky and Santi, 2008; Makuuchi et al., 2009; Pallier et al., 2011). Because syntactic representations are hierarchically organized, our results are also compatible with theories attributing BA 44/45 a role in the hierarchical processing of information (Koechlin and Jubault, 2006; Tettamanti and Weniger, 2006).

Besides areas BA44/45, previous studies of deductive reasoning have also reported activation in BA8. Indeed, it has been claimed that BA8 should be considered one of a few “core regions” for deductive reasoning, along with BA10 (Monti et al., 2007, 2009). Monti and colleagues have also shown that the activation of these two regions dissociates deduction from inferences based purely on syntax processing. In our experiment, we tested whether these two regions are predictors of membership of behaviorally identified groups. We find that BA8 activity is indeed a predictor of consistency, but not of validity or heuristics. A trend in the same direction is also present in BA 10. These findings are compatible with the reports implicating BA8/10 in deductive reasoning in general, because the correct encoding and representation of premises is a critical step for the generation of correct conclusions. Such a conclusion is also compatible with the proposed role of BA8 and BA10 in selection, coordination and representation of multiple sub-goals, in the integration of relational information and in working memory (Charron and Koechlin, 2010; Christoff et al., 2001; Koechlin et al., 2000; Owen et al., 2005; Ramnani and Owen, 2004). However, our findings suggest that in deductive reasoning the differences in functional roles between BA8/10 and BA44/45 are less important than previously proposed, because both predicted the same response profiles, characterized primarily by the extraction of logical form rather than the search for logical conclusions.

Brain areas related to validity

The second main result of our study is that the combined level of activity of frontal node 4 (Fig. 2, mainly BA47) and the lateral occipital node (BA19) was the best predictor of validity group membership. A logistic model including activity in these two regions generated accurate predictions for 79% of participants. Other models were also good predictors of membership (60% or higher), but all also included frontal node 4. Interestingly, frontal node 4 and the lateral occipital node had opposite effects on prediction. The probability of high validity group membership increased with greater BA47 activation, it decreased with greater BA19 activation. That is, the optimal scenario associated with achieving high validity includes activation of BA47 and no activation of lateral occipital cortex (Fig. 3). Consistent with the central role of BA47 in generating valid answers, we find that BA47 was the only area with differential activation across subgroups (Fig. 4). Overall, our findings demonstrate that BA47 is critical for the generation and selection of valid conclusions in deductive problems. We find that the functional role of BA47 is unique and distinct from that of all other brain regions considered because they show no correlation with validity (basal ganglia, frontal nodes 2 and 3, medial parietal node), or because they exhibit a negative correlation (lateral occipital node, but see below).

Several studies of deductive reasoning have reported deduction associated activity in BA 47 (Goel et al., 1998; Monti et al., 2007, 2009; Noveck et al., 2004; Prado et al., 2010; Reverberi et al., 2010; Rodriguez-Moreno and Hirsch, 2009). None were able to assign a specific functional role to it, because it was often co-activated with other brain regions. BA47 is also involved in rule-guided behavior, particularly in the representation of an active rule-set and in the selection and implementation of appropriate task rules (Bode and Haynes, 2009; Bunge et al., 2003; Reverberi et al., 2011; Sakai and Passingham, 2003, 2006). It is also involved in the controlled retrieval of semantic concepts (Badre and Wagner, 2007). On the basis of these findings and of our results, we speculate that BA47 is critical for the selection and implementation of the relevant inferential rules for generating the valid conclusions of an active set of premises.

Apparently challenging our results and conclusions, some studies have failed to find BA47 activations in deductive reasoning (e.g. Goel and Dolan, 2003b; Reverberi et al., 2007). Absence of BA47 activation in these studies may be explained in two ways. One possibility follows directly from our findings. If for any reason invalid strategies are used to solve deductive problems, then BA47 may not be activated. This explanation may be responsible for the result of Goel and Dolan (2003b) who found no BA47 activation in their main reasoning contrast for categorical syllogisms with no emotional content (similar to those used in this study). Their participants responded with an accuracy level of about 66% (chance level 50%), which corresponds roughly to a performance of 50% on our task (chance level 25%) and is equivalent to that of our low-validity group in which we failed to demonstrate BA47 activation too (Fig. S2). Thus, the data in Goel and Dolan (2003b) are potentially compatible with our hypothesis given that their participants were insufficiently involved in the logical nature of their problems.[3]

A second possible interpretation refers to the relation between reasoning and selection processes. In studies of memory it has been argued that BA47 is involved only when there is a weak association between cues and target knowledge (Badre and Wagner, 2007). Likewise, in our case, it is plausible that very simple deductive tasks may not require selection and so BA47 may not be significantly activated during simple reasoning. We propose that deduction recruits selection procedures only when several derivation rules can plausibly be applied to the same premises. When the premises strongly cue a single rule, rule selection or scheduling is unnecessary and accordingly BA47 may not be significantly engaged. For example, in one of our previous experiments (Reverberi et al., 2007) conditional sentences always cued the application of a Modus Ponens derivation rule (i.e., if a then b; a, therefore b). Despite the fact that solutions to these problems were provided with high accuracy (93%) BA47 was not significantly activated. Because Modus Ponens is a basic, possibly automatic, rule without difficulty or ambiguity (Reverberi et al., 2009a), a lack of BA47 activation is to be expected if our hypothesis is correct.

Besides BA47, activity in the lateral occipital node also predicted membership of the validity subgroup. As noted above, the greatest probability was associated with the lowest lateral occipital activity (i.e. tending towards no activation). Given that a minor deactivation in lateral occipital cortex was observed in only two cases, we could not reliably assess the effect of lateral occipital cortex deactivation on validity. However, low or absent activation in lateral occipital cortex increases the probability of producing valid responses. This pattern suggests that lateral occipital cortex activation is not critical for the generation of valid deductive conclusions. A similar line of reasoning applies to BA40, examined because previous studies have often reported its activation.

The use of heuristic strategies

Our third result is that no specific pattern of activation involving the six areas of the reasoning network specifically identified the use of heuristic strategies. This finding was confirmed by a direct comparison of activation in the network nodes between low and high heuristic subgroups (Fig. 4). The lack of differential activation was further corroborated by a whole-brain analysis of reasoning-related activations, separately in the two subgroups. The brain areas activated by reasoning were similar in the two subgroups, and confined to the reasoning network highlighted in the main analyses (see Fig. S3). These observations suggest that heuristic strategies, particularly the “atmosphere” heuristic, do not involve activation of qualitatively different networks from that active in syllogistic reasoning (Reverberi et al., 2009b). That is, syllogistic problems activate the same network of brain areas regardless of whether participants use heuristics or other strategies. Apparently, reasoning does not recruit different cognitive mechanisms for “pure” logical and heuristic reasoning.

A well-known position among scholars of reasoning is that heuristics hold a prominent role in human deduction (e.g. Evans, 2003; Kahneman and Frederick, 2002; Sloman, 1996; Stanovich and West, 2000). Although the theories differ in several respects, they all hypothesize the existence of two cognitive systems. One system (often called System 1) is supposed to have evolved first. It is fast and mostly driven by associative links and heuristics. It drives responses when people answer on the basis of non-logical problem properties, such as the “atmosphere” of the presented premises (Reverberi et al., 2009b; Woodworth and Sells, 1935). By contrast, a second, evolutionarily more recent system (System 2) is responsible for logically valid but effortful cognitive procedures. This system considers the logical structure of the premises of a problem and thus can assess the validity of an inference, but in a slow, brittle and resource-demanding manner. It is natural to interpret these “dual process” theories to imply that the two systems depend on largely independent neural structures, predicting a potential anatomo-functional dissociation between people who rely for reasoning mainly on System 1 or 2. We found no such dissociation; the brain areas involved in both heuristic (System 1) and non-heuristic (System 2) subgroups were functionally indistinguishable.

Our findings can be interpreted in two ways. A first possibility is that no evolutionarily primitive system provides pre-analytic answers to deductive problems, thus questioning a main tenet of dual process theories. A second, more cautious interpretation hinges upon the nature of the heuristic we investigated, i.e., the “atmosphere” heuristic. Like other deductive heuristics such as the “matching principle” for syllogisms (Wetherick and Gilhooly, 1995), or the matching bias for conditional problems (Evans, 1998), the atmosphere effect is grounded in procedures that abstract out the content of the premises (Beggs and Denny, 1969; Chapman and Chapman, 1959). Thus, unlike other heuristics such as the availability and the representativeness heuristics (Tversky and Kahneman, 1974), the atmosphere heuristic does not process the “emotional” content or the truth of the premises. Hence, the lack of any specific activation pattern associated with the heuristic group may depend on this particular aspect of the atmosphere heuristic. Whether the first or the second interpretation is correct, our findings suggest that at a minimum, dual process theories need to devise more diverse accounts of the cognitive processes involved in heuristic reasoning, taking into account the fact that some heuristics rely at least in part on the same cognitive processes as those implicated in analytical reasoning.

Conclusion: towards a multi-componential analysis of deduction

We find that different patterns of activation predict different profiles of behavioral response to the same problems. Furthermore, we find that predictive power is not evenly distributed across all brain areas activated during deduction. These results corroborate the idea that left BA47, BA44/45 and BA6/8 are critical areas for inference making. They demonstrate differential functional roles for these brain areas, which with others constitute a deductive reasoning network. In particular, we suggest that BA44/45 and BA6/8 are involved in extraction and representation of the formal structure of problems, while BA47 is involved in the selection and application of the relevant inferential rules to generate correct solutions. In contrast with many behavioral models of human reasoning that consider deduction a unitary process, our findings suggest it is composed of a set of cognitive processes, including the retrieval of logical forms of the premises; the application of relevant elementary logical rules and the monitoring of reasoning pathways from the premises to a conclusion (Bonatti, 1994, 1998; Reverberi et al., 2009c). These processes require the concerted co-operation of several, functionally distinct brain areas. Therefore, our findings suggest a multi-component view of reasoning, according to which the proper unit of analysis of deduction is not, generically, “logical reasoning”, but that of several subcomponents that interact in complex ways. Careful

[3] The task phase considered for fMRI analyses represents a further difference between Goel and Dolan (2003b) and our study. We considered the integration phase. By contrast, Goel and Dolan considered the evaluation of the conclusion phase. Thus, the two studies may not necessarily explore the same cognitive processes. This interpretation could also explain the observed difference.

consideration of the subcomponents of elementary deduction may offer a fruitful perspective to clarify the neural basis of human reasoning.

Acknowledgments

Part of the data presented here was examined in a preceding work by our group (Reverberi et al., 2010). The Neuroimaging Laboratory of the Santa Lucia Foundation is supported by the Italian Ministry of Health. This work was supported by MICINN Grant PSI2009-08232PSIC to L.L.B.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.neuroimage.2011.08.027.

References

Badre, D., Wagner, A.D., 2007. Left ventrolateral prefrontal cortex and the cognitive control of memory. Neuropsychologia 45 (13), 2883–2901.

Beggs, I., Denny, J.P., 1969. Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. J. Exp. Psychol. 81, 351–354.

Ben-Shachar, M., Palti, D., Grodzinsky, Y., 2004. Neural correlates of syntactic movement: converging evidence from two fMRI experiments. Neuroimage 21 (4), 1320.

Bode, S., Haynes, J.D., 2009. Decoding sequential stages of task preparation in the human brain. Neuroimage 45 (2), 606–613.

Bonatti, L., 1994. Why should we abandon the mental logic hypothesis? Cognition 50 (1–3), 17–39.

Bonatti, L.L., 1998. Possibilities and real possibilities for a theory of reasoning. In: Pylyshyn, Z.W. (Ed.), Essays on Representations. Ablex Publishing Corp, Norwood, NJ, US, pp. 85–119.

Bunge, S.A., Kahn, I., Wallis, J.D., Miller, E.K., Wagner, A.D., 2003. Neural circuits subserving the retrieval and maintenance of abstract rules. J. Neurophysiol. 90 (5), 3419–3428. doi:10.1152/jn.00910.2002.

Chapman, L.J., Chapman, J.P., 1959. Atmosphere effect re-examined. J. Exp. Psychol. 58, 220–226.

Charron, S., Koechlin, E., 2010. Divided representation of concurrent goals in the human frontal lobes. Science 328 (5976), 360–363.

Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J.K., Holyoak, K.J., 2001. Rostrolateral prefrontal cortex involvement in relational integration during reasoning. Neuroimage 14 (5), 1136–1149.

Dapretto, M., Bookheimer, S.Y., 1999. Form and content: dissociating syntax and semantics in sentence comprehension. Neuron 24 (2), 427–432.

Dickstein, L.S., 1978. The effect of figure on syllogistic reasoning. Mem. Cognit. 6 (1), 76–83.

Evans, J.S.B.T., 1998. Matching bias in conditional reasoning: do we understand it after 25 years? Think. Reason. 4 (1), 45–82.

Evans, J.S.B.T., 2003. In two minds: dual-process accounts of reasoning. Trends Cogn. Sci. 7 (10), 454–459.

Fangmeier, T., Knauff, M., Ruff, C.C., Sloutsky, V., 2006. FMRI evidence for a three-stage model of deductive reasoning. J. Cogn. Neurosci. 18 (3), 320–334.

Ford, M., 1995. Two modes of mental representation and problem solution in syllogistic reasoning. Cognition 54 (1), 1–71.

Friederici, A.D., Fiebach, C.J., Schlesewsky, M., Bornkessel, I.D., von Cramon, D.Y., 2006. Processing linguistic complexity and grammaticality in the left frontal cortex. Cereb. Cortex 16 (12), 1709–1717.

Friston, K.J., Holmes, A.P., Price, C.J., Buchel, C., Worsley, K.J., 1999. Multisubject fMRI studies and conjunction analyses. Neuroimage 10 (4), 385–396.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. Neuroimage 16 (2), 465–483.

Goel, V., 2007. Anatomy of deductive reasoning. Trends Cogn. Sci. 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.

Goel, V., Dolan, R.J., 2003a. Explaining modulation of reasoning by belief. Cognition 87 (1), B11–B22.

Goel, V., Dolan, R.J., 2003b. Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. Neuroimage 20 (4), 2314–2321.

Goel, V., Gold, B., Kapur, S., Houle, S., 1998. Neuroanatomical correlates of human reasoning. J. Cogn. Neurosci. 10 (3), 293–302.

Grodzinsky, Y., Santi, A., 2008. The battle for Broca's region. Trends Cogn. Sci. 12 (12), 474–480.

Henson, R.N.A., Penny, W.D., 2003. ANOVAs and SPM. Wellcome Department of Imaging Neuroscience, London, UK.

Johnson-Laird, P.N., Bara, B.G., 1984. Syllogistic inference. Cognition 16 (1), 1–61.

Kahneman, D., Frederick, S., 2002. Representativeness revisited: attribute substitution in intuitive judgment. In: Gilovich, T., Griffin, D., Kahneman, D. (Eds.), Heuristics and biases: The psychology of intuitive judgment. Cambridge University Press, New York, pp. 49–81.

Knauff, M., Fangmeier, T., Ruff, C.C., Johnson-Laird, P.N., 2003. Reasoning, models, and images: behavioral measures and cortical activity. J. Cogn. Neurosci. 15 (4), 559–573.

Koechlin, E., Corrado, G., Pietrini, P., Grafman, J., 2000. Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. Proc. Natl. Acad. Sci. U. S. A. 97 (13), 7651–7656.

Koechlin, E., Jubault, T., 2006. Broca's area and the hierarchical organization of human behavior. Neuron 50 (6), 963–974.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the Proceedings of the 14th international joint conference on Artificial intelligence, Montreal, Quebec, Canada, vol. 2.

Makuuchi, M., Bahlmann, J., Anwander, A., Friederici, A.D., 2009. Segregating the core computational faculty of human language from working memory. Proc. Natl. Acad. Sci. U. S. A. 106 (20), 8362–8367.

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.

Monti, M.M., Parsons, L.M., Osherson, D.N., 2009. The boundaries of language and thought in deductive inference. Proc. Natl. Acad. Sci. U. S. A. 106 (30), 12554–12559.

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

Owen, A.M., McMillan, K.M., Laird, A.R., Bullmore, E., 2005. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. Hum. Brain Mapp. 25 (1), 46–59.

Pallier, C., Devauchelle, A.D., Dehaene, S., 2011. Cortical representation of the constituent structure of sentences. Proc. Natl. Acad. Sci. U. S. A. 108 (6), 2522–2527.

Penny, W.D., Holmes, A.P., Friston, K.J., 2004. Random-effects analysis, In: Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W.D. (Eds.), Human Brain Function, 2nd Edition. Academic Press, San Diego (CA), pp. 843–850.

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.

Ramnani, N., Owen, A.M., 2004. Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. Nat. Rev. Neurosci. 5 (3), 184–194.

Reverberi, C., Burigo, M., Cherubini, P., 2009a. A subliminal premise can automatically trigger an elementary deductive inference. In: Taatgen, N., van Rijn, H. (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX, Cognitive Science Society.

Reverberi, C., Cherubini, P., Frackowiak, R.S., Caltagirone, C., Paulesu, E., Macaluso, E., 2010. Conditional and syllogistic deductive tasks dissociate functionally during premise integration. Hum. Brain Mapp. 31 (9), 1430–1445.

Reverberi, C., Cherubini, P., Rapisarda, A., Rigamonti, E., Caltagirone, C., Frackowiak, R.S., 2007. Neural basis of generation of conclusions in elementary deduction. Neuroimage 38 (4), 752–762.

Reverberi, C., Gorgen, K., Haynes, J.-D., 2011. Compositionality of Rule Representations in Human Prefrontal Cortex. Cerebral Cortex. doi:10.1093/cercor/bhr200.

Reverberi, C., Rusconi, P., Paulesu, E., Cherubini, P., 2009b. Response demands and the recruitment of heuristic strategies in syllogistic reasoning. Q. J. Exp. Psychol. (Colchester) 62 (3), 513–530.

Reverberi, C., Shallice, T., D'Agostini, S., Skrap, M., Bonatti, L.L., 2009c. Cortical bases of elementary deductive reasoning: Inference, memory, and metadeduction. Neuropsychologia 47 (4), 1107–1116.

Rodriguez-Moreno, D., Hirsch, J., 2009. The dynamics of deductive reasoning: an fMRI investigation. Neuropsychologia 47 (4), 949–961.

Sakai, K., Passingham, R.E., 2003. Prefrontal interactions reflect future task operations. Nat. Neurosci. 6 (1), 75–81.

Sakai, K., Passingham, R.E., 2006. Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. J. Neurosci. 26 (4), 1211–1218.

Sloman, S.A., 1996. The empirical case for two systems of reasoning. Psychol. Bull. 119 (1), 3–22.

Stanovich, K.E., West, R.F., 2000. Individual differences in reasoning: implications for the rationality debate? Behav. Brain Sci. 23 (5), 645–665 (discussion 665–726).

Tettamanti, M., Weniger, D., 2006. Broca's area: a supramodal hierarchical processor? Cortex 42 (4), 491–494.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185 (4157), 1124–1131.

Wetherick, N.E., Gilhooly, K.J., 1995. 'Atmosphere', matching, and logic in syllogistic reasoning. Current Psychology, vol. 14. Transaction Publishers, p. 169.

Woodworth, R.S., Sells, S.B., 1935. An atmosphere effect in syllogistic reasoning. J. Exp. Psychol. 18, 451–460.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4 (1), 58–73.