CORTEX XXX (2015) 1–14

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation

Corrado Corradi-Dell'Acqua [a,b,*], Francesco Turri [b], Laurence Kaufmann [c,1], Fabrice Clément [a,d] and Sophie Schwartz [a,b,e,2]

[a] Swiss Centre for Affective Sciences, University of Geneva, Geneva, Switzerland
[b] Department of Fundamental Neurosciences, University Medical Center, University of Geneva, Geneva, Switzerland
[c] Institute of Social Sciences, University of Lausanne, Quartier UNIL-Mouline, Lausanne, Switzerland
[d] Cognitive Science Centre, University of Neuchâtel, Neuchâtel, Switzerland
[e] Geneva Neuroscience Center, University of Geneva, Switzerland

ARTICLE INFO

Article history:
Received 18 August 2014
Reviewed 29 September 2014
Revised 28 November 2014
Accepted 17 February 2015
Published online xxx

Keywords:
Theory of mind
Deontic reasoning
Impression formation
dMPFC
Amygdala

ABSTRACT

Forming and updating impressions about others is critical in everyday life and engages portions of the dorsomedial prefrontal cortex (dMPFC), the posterior cingulate cortex (PCC) and the amygdala. Some of these activations are attributed to "mentalizing" functions necessary to represent people's mental states, such as beliefs or desires. Evolutionary psychology and developmental studies, however, suggest that interpersonal inferences can also be obtained through the aid of deontic heuristics, which dictate what must (or must not) be done in given circumstances. We used fMRI and asked 18 participants to predict whether unknown characters would follow their desires or obey external rules. Participants had no means, at the beginning, to make accurate predictions, but slowly learned (throughout the experiment) each character's behavioral profile. We isolated brain regions whose activity changed during the experiment, as a neural signature of impression updating: whereas dMPFC was progressively more involved in predicting characters' behavior in relation to their desires, the medial orbitofrontal cortex and the amygdala were progressively more recruited in predicting rule-based behavior. Our data provide evidence of a neural dissociation between deontic inference and theory-of-mind (ToM), and support a differentiation of orbital and dorsal prefrontal cortex in terms of low- and high-level social cognition.

* *Corresponding author*. NCCR Affective Sciences, University of Geneva – CISA, Campus Biotech, Uni Durfour, 24 rue Général Dufour, CH-1211, Geneva, Switzerland.
E-mail addresses: Corrado.Corradi@unige.ch (C. Corradi-Dell'Acqua), francescoturri@gmail.com (F. Turri), Laurence.Kaufmann@unil.ch (L. Kaufmann), fabrice.clement@unine.ch (F. Clément), sophie.schwartz@unige.ch (S. Schwartz).
[1] Tel.: +41 216923218.
[2] Tel.: +41 223795376.
http://dx.doi.org/10.1016/j.cortex.2015.02.011
0010-9452/

1. Introduction

The ability to predict people's behavior is critical to interact efficiently in complex social environments. Social psychology suggests that others' behaviors are estimated through models (or schemas) which initially rely on first impressions, and are subsequently updated on the basis of new upcoming information (Fiske & Linville, 1980; Srull & Wyer, 1989). Recently, social and cognitive neuroscience research has begun to unveil the neural substrates underlying these abilities, and implicated the dorsomedial prefrontal cortex (dMPFC), the posterior cingulate cortex (PCC), the precuneus and the amygdala in the formation of first impressions (Kuzmanovic et al., 2012; Mende-Siedlecki, Said, & Todorov, 2013; Mitchell, Cloutier, Banaji, & Macrae, 2006; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009; Todorov, 2008). Regions like dMPFC and PCC also contribute to updating such impressions through the experience of new unpredicted behaviors (Baron, Gobbini, Engell, & Todorov, 2011; Cloutier, Gabrieli, O'Young, & Ambady, 2011; Ma et al., 2012; Mende-Siedlecki, Cai, & Todorov, 2013). Interestingly, the role played by these regions (e.g., dMPFC) in learning about others' behavior seems to be social in nature, as it involves only interpersonal settings and it is independent from more general (reward-based) associative learning (Behrens, Hunt, Woolrich, & Rushworth, 2008; Hampton, Bossaerts, & O'Doherty, 2008).

One influential model in social neuroscience relates interpersonal inferences to "theory-of-mind" (ToM) (Amodio & Frith, 2006; Saxe, Carey, & Kanwisher, 2004), that is the ability to ascribe to others mental states such as beliefs/desires. In this framework, our predictions about people are based on representations of their beliefs, desires, and intentions. Consistently, studies mapping the neural correlates of ToM (Corradi-Dell'Acqua, Hofstetter, & Vuilleumier, 2014; Gallagher & Frith, 2003; Mar, 2011; Saxe & Powell, 2006) isolated a network (precuneus, PCC, dMPFC, temporal cortex) partly reminiscent of the one involved in impression formation and learning about others' behavior.

However, developmental studies suggest that not all inferences about others engage ToM. Indeed, children younger than four who have problems at inferring people's mental states (Flavell, 1999; Saxe et al., 2004), grasp quite easily deontic rules or norms, i.e., prescripts on what must (or must not) be done in given circumstances (e.g., washing hands before lunch) (Dunn, 1988; Rubin, Bukowski, & Parker, 1998). Clément, Bernard, and Kaufmann (2011) showed that children who fail in predicting others' behavior on the basis of their beliefs are able to do so on the basis of rules. It is plausible that in adults too interpersonal inferences are not based only on representations of people's mental states, but also on deontic heuristics including the rapid classification of individuals as compliers or cheaters (Cosmides, 1989). From an evolutionary perspective, cheater classification/detection might represent an essential (and ToM-independent) cognitive adaptation (Cosmides & Tooby, 2005; Cummins, 1996).

Neuroimaging studies testing deontic reasoning implicated temporal, cingulate and prefrontal structures, which are often found, not only in learning about others' behavior, but also in ToM tasks (Ermer, Guerin, Cosmides, Tooby, & Miller, 2006; Fiddick, Spampinato, & Grafman, 2005). Such anatomical overlap between these inferential processes might be confounded by the fact that, to our knowledge, deontic reasoning and ToM have never been compared directly in one neuroimaging study.

We engaged 18 volunteers in a prediction task in which, at each trial, evaluation of characters' putative behavior was based on evidence accumulated over preceding trials. Characters could be described either through their desires or as subjected by rules. This experimental setup (Fig. 1) forced volunteers to develop models of the characters' behavior, either in relation to their mental states (which should engage ToM) or to externally-imposed rules (which should engage deontic reasoning). In line with previous studies we expected regions such as PCC and dMPFC to change their activity throughout the experiment as a neural signature of continuous model update. The critical test, however, would be to assess whether these regions update representations of characters' desires, their rules-behavior, or both. Based on the developmental data reviewed above, we expected that predictions based on rules-behavior and those based on characters' desires should be dissociable at the neural level.

2. Material and methods

2.1. Participants

Eighteen participants (10 males, 18–44 years) took part in the experiment. None had any history of neurological or psychiatric illness. Written informed consent was obtained from all subjects, who were naive to the purpose of the experiment. The study was approved by the local ethics committee and conducted according to the declaration of Helsinki.

2.2. Stimuli

We built a database of 438 photographs each depicting one character, out of six males with an age ranging from 20 to 25 years. Six were color photographs (5.43 × 5.43° of visual angle) in which the face of each character was displayed in frontal view (hereafter 'portraits' – see Fig. 1A). The remaining 432 images were color photographs (10.87 × 8.13°) depicting each of the six characters engaged in everyday activities (hereafter 'feedbacks'). Twelve types of activities/themes were chosen: washing oneself, dressing up, eating, drinking, studying, house cleaning, ironing, gardening, board-gaming, drawing, exercising at the gym, and playing an instrument. For each activity and character, six images were created depicting the portrait person engaged in one variant of the theme. For example, playing an instrument could involve a flute, a violin, a tambourine, etc. (Fig. 1A). Critically, although we created a total of 432 feedback photographs, each participant was shown only 210 of these during the scanning session (see below).

2.3. Experimental setup

Participants were first cued with a text describing an everyday activity and a portrait photograph (Fig. 1B). They had to predict

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

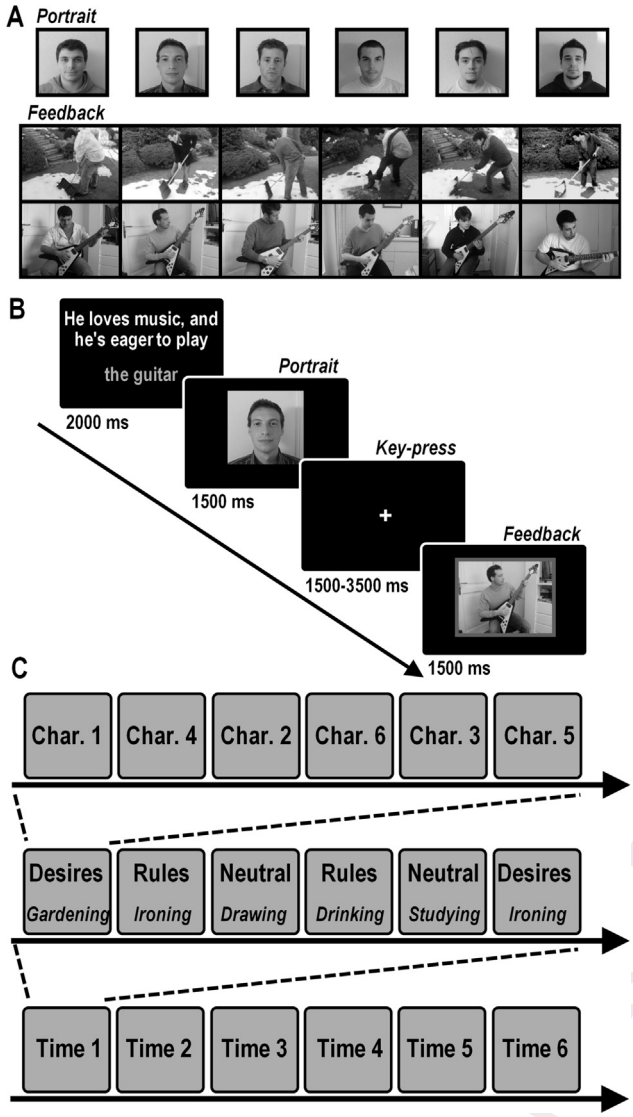

Fig. 1 – (A) Stimuli. The 6 portraits and 12 examples of the feedbacks used in the present study. In feedbacks each of the characters displayed in the portrait is engaged in everyday activities, such as snow-digging in the garden or playing an electric guitar. (B) Trial Structure. Each trial is introduced by a text string describing an everyday activity. After 2000 msec the text is replaced with a portrait of a person. At the point participants have to indicate via key-press whether they believe that the person in the portrait would execute the previously-described activity. The response can be given during the 1500 msec in which the portrait is displayed, but also subsequently when a fixation cross appears for a variable amount of time (max 3500 msec). Finally, a feedback (1500 msec) informs participants on whether or not they were correct. Feedbacks are displayed in a green or red frame indicating whether the person did or did not perform the earlier-described activity. (C) Experimental Setup. The experiment was organized in blocks of six consecutive trials in each of which the six characters were described engaged in variants of a specific activity. Blocks varied throughout the experiment in term of the activity employed (gardening, ironing, etc.) and on how it was edited (SCENARIO: desires, rules, neutral). Finally, six consecutive blocks (two for each scenario) in pseudo-randomized order were clustered together in one time-bin. The whole experimental session comprised 6 time bins. Full details in the text.

whether the character would execute the activity described in the text. Subsequently, participants were shown a feedback depicting the character engaged in the cued activity (consistent) or in a different activity (inconsistent). At the beginning of the experiment participants had no information for making appropriate predictions; however because each character was associated with a systematic consistency/inconsistency profile, the correct answers could be learned *via* feedback throughout the experiment.

Each trial was introduced by a text string on a black background describing a person being about to execute one activity variant (e.g., "He loves music and he's eager to play the guitar", see Fig. 1B). After 2000 milliseconds (msec) the text was replaced by one portrait photograph, which lasted 1500 msec, and was followed by a fixation cross whose duration varied between 1500 and 3500 msec (average = 2500 msec). At that time participants had to report whether they thought that the portrait character would perform the activity variant described by pressing one of two possible keys with the dominant hand. A feedback was then shown in which the character was displayed engaged in the same ("playing the guitar", consistent trials) or different ("playing the flute", inconsistent trials) variant of the everyday activity described in the previous text. Feedbacks lasted for 1500 msec and were displayed within a green or red frame indicating respectively consistency or inconsistency. Each experimental trial lasted on average 7500 msec and was followed by an inter-stimulus-interval ranging from 500 to 2500 msec (average = 1500 msec).

The experiment was organized in blocks of six consecutive trials each focused on the same activity (see Fig. 1C). These trials represented a unique association between the characters and the variants of the activity, so that each character would be seen engaged in one variant only. The association between characters and variants changed across participants.

Blocks changed not only according to the activity described, but also according to the semantic context in which the text was framed (factor: SCENARIO). Indeed, activities could be described in terms of the protagonist's *desires* ("He loves music and he's eager to play the guitar") or as externally-imposed *rules* ("His music professor ordered him to train in playing the guitar"). On the one hand, desires scenarios were overt descriptions of individuals' mental states, which are known to modulate the activity of part of the ToM network (Saxe & Kanwisher, 2003; Saxe & Powell, 2006). On the other hand, rules were plausible sets of prescriptions occurring in dyadic hierarchical relationships, aimed at eliciting considerations about permissions and obligations (thus deontic reasoning) comparably to the experimental materials used in previous studies (e.g., Bucciarelli & Johnson-Laird, 2005). See Table 1 for details. Please note, however, that the notion of deontic reasoning adopted in the present study is partially divergent from that used in previous neuroimaging researches (Ermer et al., 2006; Fiddick et al., 2005), which focused

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

Table 1 – English-translation of the text strings describing 12 everyday activities. Each of these activities can be described in terms of the character's desires, in terms of externally-imposed rules or in a neutral format. Desires and Rules activities appear incomplete as they are each associated with a specific variant.

Activities	Desires	Rules	Neutral
Washing oneself	When he washes himself, he likes to start from …	His mother ordered him to wash himself starting from …	He just woke up and, in the bathroom, he is washing himself.
Dressing Up	He just bought some clothes and he desires to wear …	His boss just ordered him to wear …	This morning the weather is good and he is dressing up.
Eating	He desires a fruit juice and he prepares …	His doctor ordered him to eat more …	It is at home at noon and he's eating.
Drinking	It is hot, and he desires to drink …	At the cooking class, the teacher ordered him to learn the taste of …	It is 5 pm and he is drinking.
Studying	Exams are approaching and he wants to have a good mark in …	Exams are approaching and his parents ordered him to study …	He is reading in his room.
House cleaning	He likes it when all is tidy and so he desires to clean …	His father ordered him to clean …	Holidays are approaching and he is cleaning up.
Ironing	As a surprise for his girlfriend, he desires to iron …	His mother ordered him to iron …	The weather is bad, and he is ironing some clothes.
Gardening	He likes gardening, and today he desires to …	The landlord ordered him to get the garden tidy and to …	Spring arrives and he makes the most of it by taking care of the garden.
Board playing	He took a day off, and he desires to play …	His Game Club ordered him to get better at playing at …	It is the week-end and he is playing some game.
Drawing	He feels in a creative mood and he desires to draw something with …	His teacher gave him as homework a drawing with …	He is in his house in the countryside and he draws.
Exercising at the gym	He is obsessed by his muscles and desires to work on …	His coach ordered him to work on …	It is the end of the day, and he is practicing sport.
Playing music	He loves music, and he's eager to play …	His music professor ordered him to train in playing …	It is Sunday morning and he is playing some musical instrument.

mainly on prescriptions which were conditional to a given benefit, such social contracts/norms (see discussion section). Finally, a third *neutral* condition was included in which the activity was described only in broad terms ("It is Sunday morning and he is playing some musical instrument"). Desires, rules and neutral scenarios were matched for word length.

For each participant, the six characters (Fig. 1A) were randomly assigned to three profile categories (factor: PROFILE), each associated with a systematic consistency/inconsistency pattern in the feedbacks. Two characters were always associated with consistent feedbacks in desires-blocks, and inconsistent feedbacks in rules-blocks (profile: D+R−). Other two characters were instead associated with inconsistent feedbacks in desires-blocks, and consistent feedbacks in rules-blocks (D−R+). Finally, the remaining two characters were always associated with consistent feedbacks, regardless of the scenario (D+R+). All three profiles were always associated with consistent feedbacks in the neutral blocks. In this experimental structure, desires- and rules-blocks were associated with 66% of consistent (and 33% inconsistent) feedbacks, whereas 100% trials in neutral blocks were consistent.

In summary, we conducted a 3 (SCENARIO: desires, rules, neutral) by 3 (PROFILE: D+R−, D−R+, D+R+) factorial design, with different scenarios changing in a block-wise fashion, but different profiles changing in a trial-wise fashion, thus establishing a trial-structure consistent with an event-related fMRI experiment. The overall experiment included 216 experimental trials (3 scenarios × 12 everyday activities × 6 trials per block). The order of the blocks and the order of the trials within each block were pseudo-randomized. In order to avoid an inhomogeneous distribution of the three kinds of scenarios across the whole experiment, we divided the whole experiment into six time bins, each composed of six consecutive blocks (see Fig. 1C); we insured that within each time-bin there were 2 blocks of each scenario. The overall experiment was organized into two functional runs of 108 trials each (corresponding to 18 blocks and 3 time bins).

2.4. *Procedure and apparatus*

The scanning sequence was organized in two functional runs of 16 min each. Subsequently, participants were asked to rate each of the six characters seen during the task in terms of: (a) trustworthiness, (b) likeability, (c) esthetic pleasantness, (d) dominance, (e) familiarity and (*f*) predictability. Ratings were measured on a 5-point Likert scale ranging from −2 to +2.

Stimuli were projected by an LCD projector (CP-SX1350, Hitachi, Japan) on a screen (about 19° × 14°) placed inside the scanner bore. Key-presses were recorded on an MRI-compatible response button box (HH-2 × 4-C, Current Designs Inc., USA). The task was programmed using Cogent 2000 (http://www.vislab.ucl.ac.uk/cogent.php) a Matlab-based toolbox.

2.5. *Imaging processing*

2.5.1. *Data acquisition*

A Siemens Trio 3-T whole-body scanner was used to acquire both T1-weighted anatomical images [repetition time (TR) = 1900 msec, inversion time = 900 msec, echo time

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

(ET) = 2.27 msec, 1 × 1 × 1 mm voxel size] and gradient-echo planar T2-weighted MRI images with blood oxygenation level dependent (BOLD) contrast. The scanning sequence for the functional images was a trajectory-based reconstruction sequence with a TR of 2100 msec, an TE of 30 msec, a flip angle of 90°, in-plane resolution 64 × 64, 32 descending slices of 3 mm thickness and no gap.

2.5.2. *Preprocessing*

Statistical analysis was performed using the SPM8 software (http://www.fil.ion.ucl.ac.uk/spm/). For each subject, all the functional images were realigned to the first image and co-registered to the anatomical image which was in turn used to estimate the deformation field necessary for the normalization to the Montreal Neurological Institute (MNI) template. Finally, all normalized functional images were smoothed with an 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

2.5.3. *First-level analysis*

Data were fed into a first-level analysis using a general linear model. Our main aim was to test for putative condition-specific linear changes in neural activity across the whole experimental session. Such analysis would normally be carried out by testing the parameters estimated by implementing the time-modulation option of SPM. However, because the experimental session was divided into two distinct functional runs, each comprehending half of the trials, these parameters would be informative only about the linear changes within each run, but not of linear changes across both runs. We therefore divided our two sessions in 6 time bins (3 in each run) and carried out a first-level analysis in which for each experimental run, each time-bin, and each of the 9 conditions resulting from our 3 SCENARIO × 3 PROFILE design, and each relevant event within the experimental trial (portrait, feedback), we modeled the event sequence (event duration 1500 msec) as a delta function. The resulting 108 vectors were convolved with a canonical hemodynamic response function and associated with a vector describing its first-order time derivative. Finally, we included, for each session, the six differential realignment parameters as additional regressors of no interest. Low-frequency signal drifts were filtered using a cutoff period of 128 sec.

2.5.4. *Second-level analysis*

The 54 parameters associated with portrait events were fed into second-level flexible factorial analyses, with a factor "condition" with 9 levels (3 SCENARIO × 3 PROFILE), and "subjects" as random factor. Time-bin effects were modeled through a covariate (ranging from 1 to 6) interacting with the factor "condition", which allowed us to investigate linear changes in neural activity across the course of the experiment. A similar second-level model was run for the 54 parameters associated with the feedback-events.

Furthermore, in order to investigate SCENARIO and PROFILE differential activity when participants held a reliable model of the characters' behavior, we conducted an additional flexible factorial analysis on the portrait events based on 9 contrast images, each reflecting condition-specific neural activity in the last two time-bins (associated with the highest accuracy – see Behavioral Data). These 9 images were also fed into a flexible factorial analysis, with a factor "condition" with 9 levels and "subjects" as random factor.

3. Results

3.1. *Behavioral data*

For each subject and condition, the median response times were calculated and fed into 3 (SCENARIO: desires, rules, neutral) × 3 (PROFILE: D+R–, D–R+, D+R+) × 6 (TIME: 1–6 bins) Repeated Measures Analyses of Variance. In a similar fashion, participants' accuracy in each trial was analyzed through a logit regression fitted with the Generalized Estimated Equation method (Hanley, Negassa, Edwardes, & Forrester, 2003). As the accuracy of neutral scenarios was at ceiling (99.45%, Standard Error of the Mean ± .97), with minimal variability and consequent detrimental effects on the model estimation, we focused the logit regression only on desires (78.55% ± 6.55) and rules (82.78% ± 5.91) scenarios.

The analysis of the response times and accuracy revealed a significant main effect of TIME (Response times: $F_{(5,85)} = 16.01$, $p < .001$; Accuracy: Wald's $\chi^2_{(5)} = 76.96$, $p < .001$), which might reflect a progressive easiness in the task over both functional runs (see Fig. 2A). The analysis of response times revealed neither significant main effects of SCENARIO and PROFILE, nor significant interactions ($F < 2.46$). Instead, the analysis of the accuracy revealed a main effect of PROFILE ($\chi^2_{(2)} = 11.30$, $p < .01$) and a PROFILE × TIME interaction ($\chi^2_{(10)} = 84.10$, $p < .001$), reflecting better performance for the D+R+ profile (90.30% ± 4.98), always associated with consistent feedbacks, relative to the others (D–R+: 83.33% ± 6.28; D+R–: 86.55% ± 5.47), especially in the first time-bins. Finally, the SCENARIO × PROFILE ($\chi^2_{(2)} = 15.41$, $p < .001$) and SCENARIO × PROFILE × TIME ($\chi^2_{(10)} = 84.00$, $p < .001$) interactions were significant. Fig. 2A shows how, in the first time-bin, profiles followed by inconsistent feedbacks were associated with an accuracy ≈42%, whereas profiles followed by consistent feedbacks were associated with an accuracy ≈69%. As inconsistent and consistent feedbacks occur in the experiment the 33% and 66% of the trials respectively, participants' behavior in the first time-bins is consistent with that of a *just-above-chance* performance. However, the discrepancy in accuracy between inconsistent and consistent profiles decreases progressively throughout the experiment, leading (in the last two time-bins) to equal accuracy ≈ 92%. On overall, the behavioral data confirm that our prediction task is associated with a slow learning process throughout the experimental session.

The analysis of the post-scanning rating data (see Fig. 2B) revealed that trustworthiness and dominance ratings differed significantly across the three profiles, with the rules-violator profile (D+R–) eliciting significantly less trust (Wilcoxon signed rank test – D+R– > D–R+: W = 23.5, Z = −2.19, $p < .05$ – D+R– > D D+R+: W = 15.5, Z = −2.41, $p < .025$) and appearing much more dominant (D+R– > D–R+: W = 138.5, Z = 2.99, $p < .005$ – D+R– > D+R+: W = 114.5, Z = 3.24, $p < .001$) than the other two profiles. Behavioral data were analyzed with R 2.14.0 (http://cran.r-project.org/) and SPSS 22.0 (IBM Corporation) software.

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

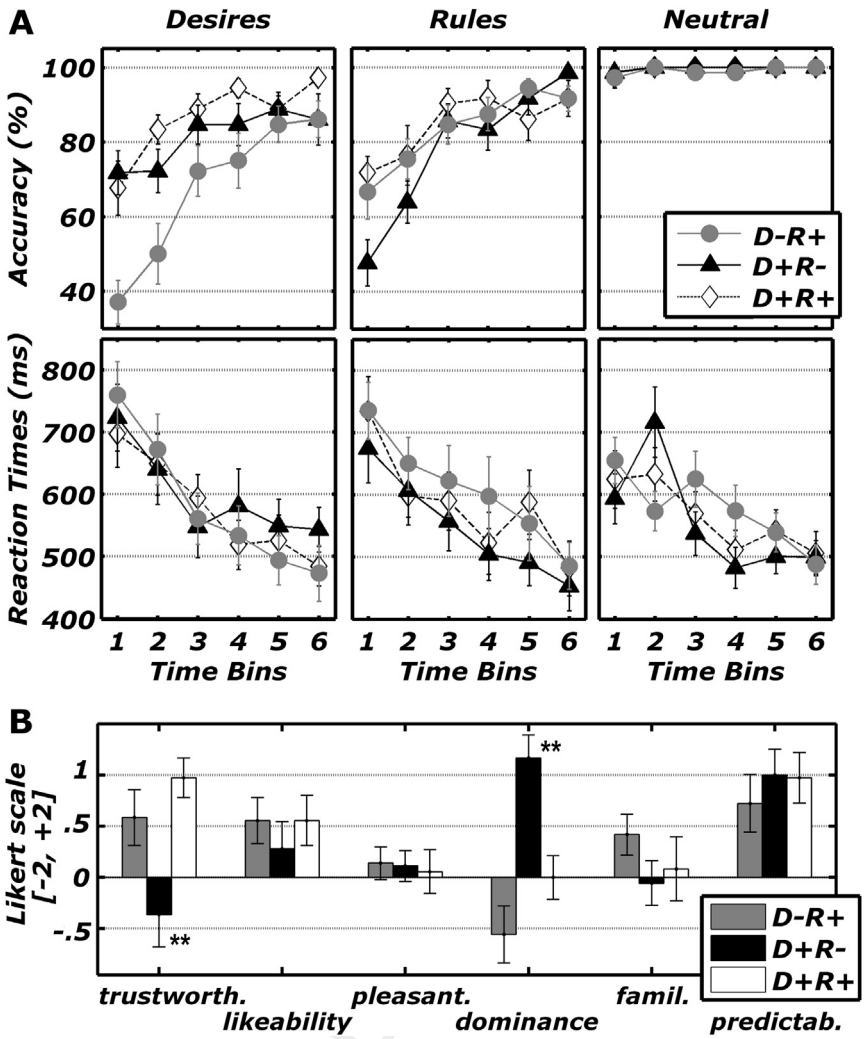

Fig. 2 – Behavioral data. (A) Response times of correct trials (msec) and percentage accuracy from the prediction task plotted against TIME. Left subplots refer to desire scenarios, middle subplot refer to rules scenarios, whereas right subplot refer to neutral scenarios. Gray circles refer to *D*-*R*+, black triangles refer to *D*+*R*-, and white diamonds refer to *D*+*R*+. Error bars correspond to S.E.M. (B) Values from the post-scanning session in which each profile is rated in terms of: trustworthiness, likeability, esthetic pleasantness, dominance, familiarity and predictability.

3.2. *Neural activations*

In this paper we report activations exceeding a cluster-level threshold corresponding to $p < .05$, corrected for multiple comparisons for the whole brain (Friston, Worsley, Frackowiak, Mazziotta, & Evans, 1993), with an underlying height threshold corresponding to $p < .001$ (uncorrected). We also applied small volume correction for those structures previously associated with impression formation (Kuzmanovic et al., 2012; Mende-Siedlecki, Said, et al., 2013; Mitchell et al., 2006; Schiller et al., 2009) and updating (Baron et al., 2011; Cloutier et al., 2011; Ma et al., 2012; Mende-Siedlecki, Cai, et al., 2013), and which have also been implicated in ToM (Mar, 2011) and deontic reasoning (Ermer et al., 2006; Fiddick et al., 2005). We therefore created a volume of interest including medial prefrontal cortex, cingulate cortex, precuneus, and amygdala based on the AAL atlas (Tzourio-Mazoyer et al., 2002) and reported areas of activation within these boundaries if associated with a $p < .05$ corrected for the volume.

3.2.1. *Portrait events*

Table 2 lists those regions whose activity changed linearly across the six time-bins. When focusing on portrait events, and searching for effects common to both rules and desires (i.e., average activity between the two kinds of scenarios), we found increasing activity in PCC and in the ventral striatum bilaterally, involving caudate nucleus, putamen and nucleus accumbens (Fig. 3, red blobs). These regions, however, were not found when testing for a conjunction analysis (i.e., linear increase in activity for rules ∩ desires – see Table 2), which is a more conservative test for common effects between the two kinds of scenarios. No negative trends were found. We then tested for dissociated effects between rules and desires scenarios and found, for the contrast desires > rules, a portion of

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

Table 2 – Regions whose activity was parametrical modulated by the factor TIME for both the portrait and feedback events. All clusters survived correction for multiple comparisons at the cluster level (with an underlying height threshold corresponding to $p < .001$, uncorrected). Coordinates (in standard MNI space) refer to maximally activated foci as indicated by the highest t value within an area of activation: x = distance (mm) to the right (+) or the left (–) of the midsagittal line; y = distance anterior (+) or posterior (–) to the vertical plane through the anterior commissure (AC); z = distance above (+) or below (–) the inter-commissural (AC-PC) line. L and R refer to the left and right hemisphere, respectively. M refers to medial activations.

	SIDE	Coordinates			$T_{(937)}$	Cluster size
		x	y	z		
PORTRAIT – General effects: Rules + Desires > 0						
Lingual Gyrus	M	−12	−102	2	7.60	3260†
Precuneus/Post. Cingulate		2	−56	38	4.32	
Ventral Striatum/Amygdala	L	−16	8	−10	5.18	398‡
Ventral Striatum	R	24	10	−2	4.43	279§
PORTRAIT – Conjoint effects: (Rules > 0) ∩ (Desires > 0)						
Lingual Gyrus	M	−12	−102	2	6.32	326‡
PORTRAIT – Main effect SCENARIO: Desires > Rules						
Medial Prefrontal Cortex (*dorsal aspect*)	M	8	32	40	3.91	111¥
PORTRAIT – Main effect SCENARIO: Rules > Desires						
Medial Prefrontal Cortex (*orbital aspect*)	M	−4	30	−22	4.48*	75
FEEDBACK – General effects: 0 > Rules + Desires						
Middle Cingulate Cortex	M	−4	18	40	4.53	509†
Precuneus	M	0	−80	40	3.83	249§
Midd.-Post. Cingulate Cortex	M	6	−14	34	3.70	117¥
FEEDBACK – Main effect SCENARIO: Desires > Rules						
Medial Prefrontal Cortex (*rostral aspect*)	M	4	52	18	3.67	121¥

†$p < .001$; ‡$p < .01$; §$p < .05$ corrected for the whole brain; ¥$p < .05$ corrected for volume.
*Uncorrected at the cluster level, although local maxima $t_{(937)} = 4.48$, $p < .05$ corrected for volume.

dMPFC. Fig. 4A displays this activation on a medial section of the human brain, mapping it dorsally to the cingulate sulcus over and around the most anterior section of the supplementary motor area. The parameters extracted from this region show how the difference between the two scenarios changed through time (Fig. 4A). No region was found for the opposite contrast (rules > desires) when using whole-brain correction or when correcting for the volume of interest corresponding to those structures previously associated with impression formation. However, within the volume of interest, 75 contiguous voxels (surviving $p < .001$ uncorrected) in the orbital portion of the medial prefrontal cortex (oMPFC) exhibited increasing neural activity throughout the experiment only during rules (see Fig. 4A, yellow blob). Although this cluster did not exceed extent threshold, its local maxima survived correction for multiple comparisons at the voxel-level. No differential linear effects were found when testing PROFILE effects or the interaction between PROFILE and SCENARIO.

3.2.2. *Feedback events*

When searching for effects common to both rules and desires (i.e., average activity between the two kinds of scenarios), we found decreasing activity in the precuneus, and the middle cingulate cortex, in both its more posterior and anterior aspects (Fig. 3, blue blobs). These regions, however, were not found when testing for a conjunction analysis (i.e., linear decrease in activity for rules ∩ desires). No positive trends were found. We then tested for dissociated effects between rules and desires scenarios and found, for the contrast desires > rules, the rostral portion of the medial prefrontal cortex. No region was found for the opposite contrast (rules > desires).

3.2.3. *Analysis of the last two time-bins*

According to behavioral data, at the end of the experiment participants held a reliable model of the characters' behavior, which allowed them to make accurate predictions at the portrait's sight. We therefore carried out an additional analysis on the portrait events, focusing on the last two time-bins. All suprathreshold activations associated with this analysis are listed in Table 3. Shared effects between rules and desires (as opposed to neutral) scenarios were observed in a network comprehending the bilateral inferior frontal sulcus, extending to the anterior insula and the lateral orbital sulcus, and the precuneus, extending to the left intraparietal sulcus. These networks were observed both when testing the main effect of SCENARIO [i.e., (*Desires* + *Rules*)/2 > *Neutral*] and when running a conjunction analysis [i.e., (*Desires* > *Neutral*) ∩ (*Rules* > *Neutral*), see Table 3]. Critically, the contrast desires > rules revealed a portion of the dMPFC overlapping that displayed in Fig. 4A (green blob). Instead, the contrast rules > desires led to no suprathreshold activation when applying correction for multiple comparisons, although 44 contiguous voxels (surviving $p < .001$ uncorrected) were found in oMPFC over and around the cluster displayed in Fig. 4A (yellow blob). Finally, when testing specific increases of neural activity due to the anticipation of inconsistent (relative to consistent) feedbacks [i.e., the SCENARIO × PROFILE interaction: (desires_D–R+ + rules_D+R–) > (desires_D+R– + rules_D–R+)] we found the rostral portion of the medial prefrontal cortex (rMPFC). Fig. 5 suggests that this region was recruited whenever

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

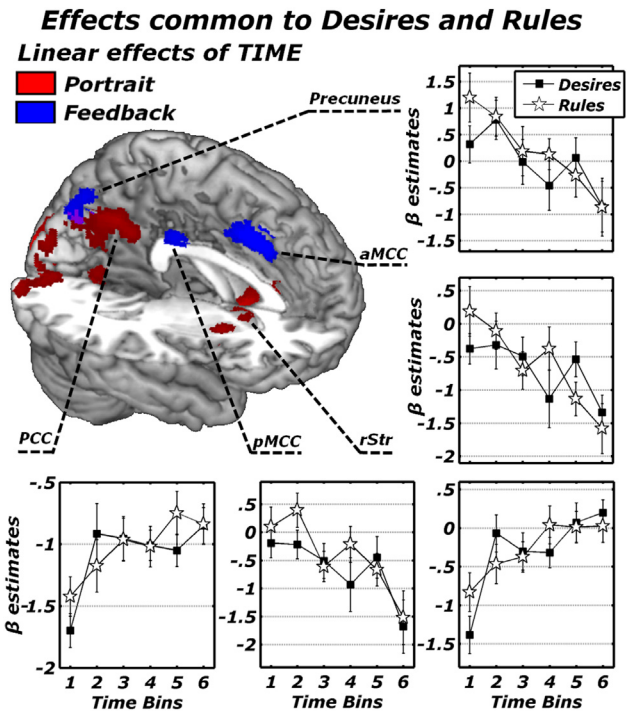

Fig. 3 – Brain regions exhibiting linear changes of neural activity across the six consecutive time bins. Red regions are those exhibiting linear increases of neural activity for the portrait events, whereas blue regions are those exhibiting decreases of neural activity for the feedback events. For each region, the average parameter estimates of portrait/feedback events are also displayed across successive time bins with S.E.M bars. Black squares refer to desires scenarios, whereas white stars refer to rules scenarios. *aMCC* and *pMCC*: anterior and posterior aspect of the middle cingulate cortex. *PCC*: posterior cingulate cortex. *rSrt*: right striatum.

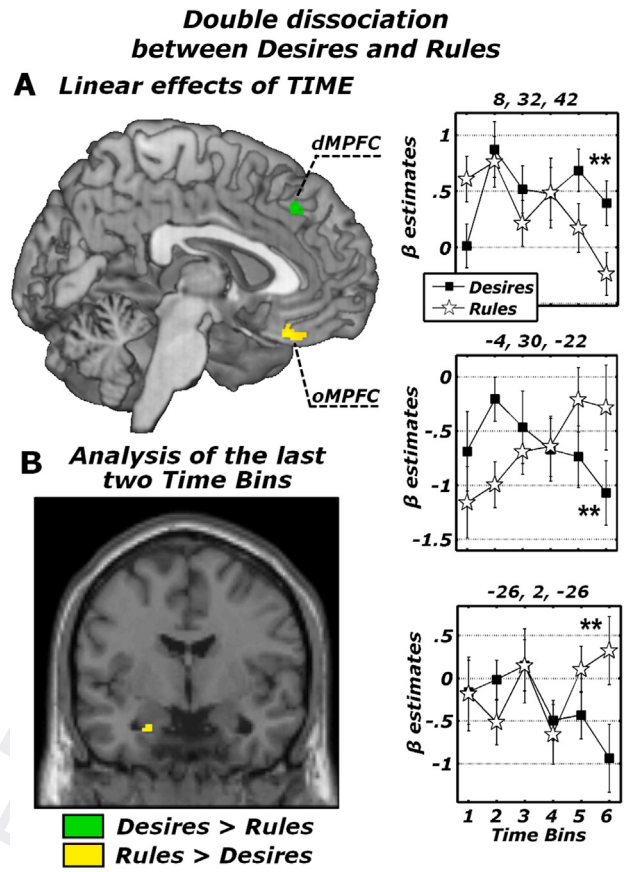

Fig. 4 – Differential brain activity for desires and rules in the analysis of Portrait events. (A) Brain regions showing progressively larger differential activity between desires and rules scenarios across the 6 consecutive time bins. Effects associated with the contrast *desires* > *rules* are displayed in green, whereas effects associated with the contrast *rules* > *desires* are displayed in yellow. (B) Coronal section (y = −4) showing increased left amygdala activity for rules in the last two time-bins. The average parameter estimates from each region are also displayed for successive time bins with S.E.M bars. *dMPFC*: dorsomedial prefrontal cortex. *oMPFC*: orbital aspect of the medial prefrontal cortex. **t = 3.15, p < .001 for the difference among scenarios in the last two time-bins.

participants, in the last time-bins, were exposed to a violating-profile. The opposite contrast [(desires_D+R− + rules_D−R+) > (desires_D−R+ + rules_D+R−)] led to activation of the left frontal eye fields.

3.2.4. *Amygdala*

All previous analyses gave us little clue about the functional role of the amygdala in our experimental paradigm, although earlier studies systematically implicated this region, not only in impression formation (Baron et al., 2011; Kuzmanovic et al., 2012; Schiller et al., 2009), but specifically in trustworthiness evaluations (Mende-Siedlecki, Said, et al., 2013; Todorov, 2008; Todorov & Engell, 2008). We can therefore expect increased amygdala response for the prediction of rules scenarios (in which the character's trustworthiness is indirectly evaluated) particularly for the profile D+R− (associated with the least trust, Fig. 2B). Please note that all previous functional contrasts were tested, not only for whole-brain effects, but also for a volume of interest. However, this volume comprehended all the regions involved in impression formation, but not the amygdala specifically. We therefore carried out small volume analysis focused only on a bilateral amygdala mask (AAL database) and tested specifically the functional contrasts rules > desires and rules_D+R− > rules_D−R+. When focusing the analysis on the last two time-bins, we found three voxels in the left amygdala (x = −26, y = −2, z = −26, $t_{(136)}$ = 3.53, $p < .05$ small volume corrected, see Fig. 4B) associated with the contrast rules > desires; no effect was associated with the contrast rules_D+R− > rules_D−R+. Finally, we tested if the amygdala exhibited increased differential activity for the contrasts rules > desires and rules_D+R− > rules_D−R+ across the six time-bins and found no suprathreshold effects.

Table 3 – Analysis of the PORTRAIT events associated with the last two time-bins.

	SIDE	Coordinates			$T_{(136)}$	Cluster size
		x	y	z		
Main effect SCENARIO: (Rules + Desires)/2 > Neutral						
Inferior Frontal Sulcus	L	−42	20	33	6.17	1435†
Superior Frontal Sulcus	L	−32	0	62	4.97	
Anterior Insula	R	32	20	−4	6.22	1133†
Lateral Orbital Gyrus	R	32	54	−6	4.72	
Anterior Insula	L	−32	20	−8	4.94	1072†
Lateral Orbital Gyrus	L	−38	54	−6	5.24	
Intraparietal Sulcus	L	−40	−56	50	6.00	1068†
Precuneus	M	−2	−70	38	5.04	776†
Midd. Cingulate Cortex	M	2	22	46	5.17	640†
Inferior Frontal Sulcus	R	48	34	24	4.82	509†
Cerebellum	R	10	−82	−26	4.44	410†
Cerebellum	L	−12	−78	−26	4.82	199§
SCENARIO conjoint effects: (Rules > Neutral) ∩ (Desires > Neutral)						
Intraparietal Sulcus	L	−40	−56	50	5.36	672†
Inferior Frontal Sulcus	L	−40	20	32	5.33	646†
Lateral Orbital Gyrus	L	−38	54	−6	4.83	358‡
Precuneus	M	−2	−70	38	4.52	314‡
Anterior Insula	R	30	20	−2	5.04	300‡
Inferior Frontal Sulcus	R	48	34	24	4.36	213§
Main effect SCENARIO: Neutral > (Rules + Desires)/2						
Midd-Post. Cingulate Cortex	M	6	−28	48	4.90	342‡
Main effect SCENARIO: Desires > Rules						
Medial Prefrontal Cortex (*dorsal aspect*)	M	8	30	40	4.06	202§
Superior Frontal Sulcus	R	28	26	34	4.15	195§
Main effect SCENARIO: Rules > Desires						
Medial Prefrontal Cortex (*orbital aspect*)	M	0	36	−24	3.61	44
SCENARIO × PROFILE Interactions: (Desires_D+R− + Rules_D−R+) > (Desires_D−R+ + Rules_D+R−)						
Superior Frontal Sulcus	L	−22	12	42	5.68	193§
SCENARIO × PROFILE Interactions: (Desires_D−R+ + Rules_D+R−) – (Desires_D+R− + Rules_D−R+)						
Medial Prefrontal Cortex (*rostral aspect*)	M	4	48	14	4.16	305‡

†$p < .001$; ‡$p < .01$; §$p < .05$ corrected for the whole brain.

4. Discussion

We engaged participants in a task in which they saw photographs of unknown characters associated with overt descriptions of rules or desires. For each photo, participants had to predict whether the character would behave consistently with the enclosed description. As there were no means, at the beginning of the experiment, to make accurate

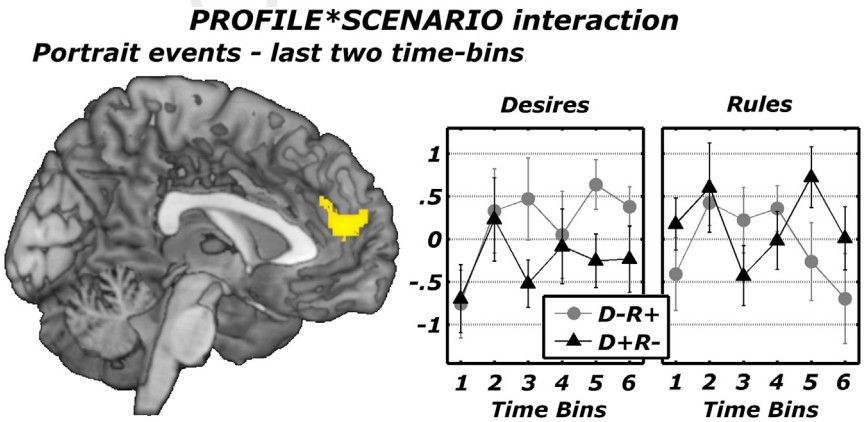

Fig. 5 – Whole brain maps showing the rostral medial prefrontal cortex (rMPFC) associated with the last two time-bins (portrait events). Specifically, in the final part of the experimental session, rMPFC exhibited greater activity when predicting the choice of those characters which were more likely to exhibit violating behavior, for either desires and rules scenarios. The average parameter estimates from the highlighted cluster are also displayed for successive time bins with S.E.M bars. Different scenarios are shown in separate subplots; gray circles refer to *D*-R+ and black triangles refer to *D*+R-.

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

predictions, just-above-chance responses were recorded. However, participants slowly learned (throughout the experiment) how each character would have behaved in different contexts. This gave us the opportunity to identify brain regions involved in the formation and updating of models about others' behavior and to dissociate predictions of actions framed as mental states (desires) from predictions of actions framed as rules. We found a double dissociation between dorsal and orbital portions of the medial prefrontal cortex: on the one hand, dMPFC exhibited progressively larger differential activity for desire-based (relative to rule-based) predictions, presumably reflecting the stronger recruitment of a behavioral model grounded on mental states, and the concurrent inhibition of a behavioral model grounded on rules; on the other hand, oMPFC (together with the left amygdala) displayed the opposite effect (Fig. 4). Please note that, as the two kinds of actions (framed as rules versus desires) were matched in text length (see methods section) and were processed with comparable engagement of cognitive resources (see Reaction Times and Accuracy results), the observed neural dissociation can be interpreted only in relation to differential semantic content. To our knowledge this is the first study showing how the assessment of people's rules-based behavior partly dissociates from the assessment of their mental states, thus challenging prominent views positing deontic reasoning as intimately dependent on ToM (Kalish, 2006; Núñez & Harris, 1998; Wellman & Miller, 2008).

4.1. *Dissociating deontic reasoning from ToM*

Theoretical accounts linking deontic reasoning to ToM have been prevalently justified by two lines of evidence: (a) children's reasoning about rule-behavior emerges usually together with their ToM ability (Wellman & Miller, 2008); (b) reframing abstract problems in terms of social contracts often leads to increased activity in dMPFC and other ToM structures (Ermer et al., 2006; Fiddick et al., 2005). It has been suggested, however, that some of the developmental observations might be confounded by idiosyncratic properties of the experimental design/materials which often emphasize the moral evaluation and the psychological consequences of rule transgression (Clément et al., 2011). Similarly, in neuroimaging studies activation of ToM structures was not observed in all kinds of deontic norms, but in the specific subclass of social contracts as opposed to precautionary rules (Ermer et al., 2006; Fiddick et al., 2005). This heterogeneity in the results obtained with different kinds of norms can be interpreted in relation to domain-specificity in deontic reasoning with precautionary rules triggering mechanisms involved in evaluation of risk and/or anticipation of pain, and social contracts eliciting predominantly moral considerations and evaluations about individuals' intent (Ermer et al., 2006; Fiddick, 2004; Fiddick, Cosmides, & Tooby, 2000; Fiddick et al., 2005). Hence, it was extremely important for the purpose of the present study to refrain from using social contracts and moral rules, as these specific prescriptions might have elicited considerations about the protagonists' mental states which are not intrinsic to all forms of deontic reasoning (Fiddick, 2004). Instead, it was critical to employ carefully matched experimental materials in which the same sets of behaviors were framed either in terms of desires (mental states) or in terms of prescripts which were the least prone to mentalistic biases. Using a similar approach, Clément et al. (2011) suggested that the emergence of deontic reasoning abilities may precede ToM during child development. The present study extends previous findings by showing how in the adult brain too ToM and deontic reasoning rely on partly-dissociated networks.

4.2. *The prediction of desire-related behavior*

We found dMPFC when looking at enhanced activity for desire-based (relative to rule-based) inferences in the final part of the experiment, but also when searching for regions in which the differential activity between desires and rules increased across the experimental session. This allows us to conclude that dMPFC's sensitivity to desires (as opposed to rules) is modulated by the development of a model of people's behavior, leading to a reliable desire-specific neural signal in those time-bins associated with accurate predictions in the task. This dMPFC region is in close proximity to the one reported in previous studies testing the neural structures involved in the formation and the update of impressions (Baron et al., 2011; Cloutier et al., 2011; Kuzmanovic et al., 2012; Ma et al., 2012; Mende-Siedlecki, Cai, et al., 2013; Mitchell et al., 2006; Schiller et al., 2009). For instance, Baron et al. (2011) had participants learn the association between faces and behavioral information, and found that dMPFC activity was correlated with a post-scan measure of learning. Likewise, the activity in dMPFC was found to be greater when people were associated with a behavior that was inconsistent with their known political affiliation (Cloutier et al., 2011), inconsistent with previously implied traits (Ma et al., 2012), or of valence opposite to previous verbal descriptions (Mende-Siedlecki, Cai, et al., 2013). The results from this line of research concord with those associated with strategic decision making tasks, in which dMPFC was not described as generally implicated in any learning process, but specifically in learning about people's behavior (Behrens et al., 2008; Hampton et al., 2008). Our data converge with, but also extend, earlier findings by showing that dMPFC is not broadly involved in predictive learning about others' actions, but specifically in developing/updating behavioral models preferentially grounded on people's mental states.

The sensitivity of dMPFC to characters' desire-based behavior is in line with a wealth of research that consider this region as part of the ToM-network (Gallagher & Frith, 2003; Mar, 2011). However, as shown in Fig. 6, in which our results are compared to those from a recent meta-analysis (Mar, 2011), the dMPFC cluster mapped in the present study is posterior to that identified by standard ToM tasks, such as those using text-based stories (e.g., Corradi-Dell'Acqua et al., 2014; Saxe & Powell, 2006). Interestingly, studies employing graphical materials and relying only partially on verbal stimuli (as in our case) triggered a wider portion of the medial prefrontal cortex, which extends dorsally to our dMPFC cluster (Fig. 6, overlap between red and green blobs). It is therefore likely that the prediction of people's desires-related behavior does not recruit the "core ToM-network" *per se* (common to all experimental materials, Mar, 2011), but rather collateral inferential processes which lead to a representation of the

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

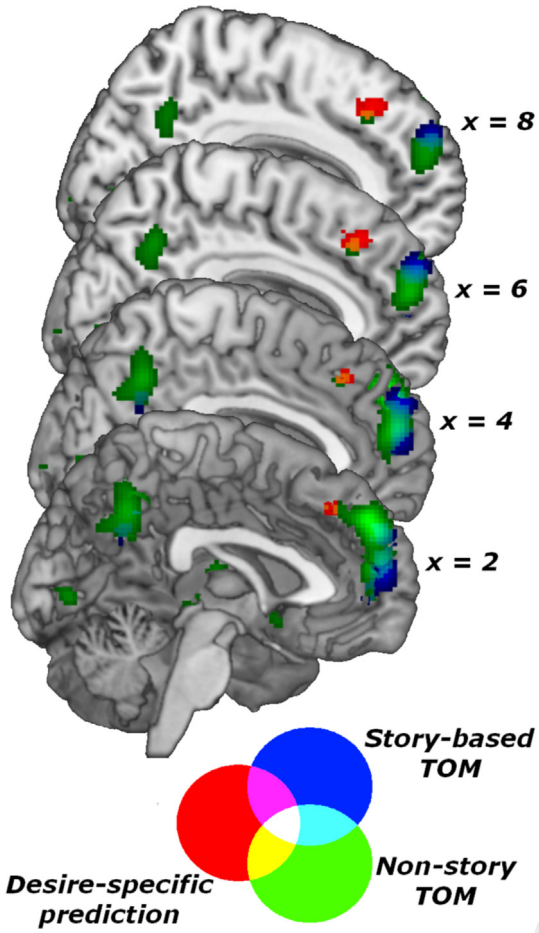

Fig. 6 – Representation of the dMPFC cluster exhibiting a linear increase of neural activity specific for the prediction of desire-related behavior (red blob) displayed together with the medial prefrontal network associated with theory-of-mind processes as described in a recent meta-analysis (Mar, 2011). Blue blobs refer to the activations (false discovery rate $q < .05$) from studies that tested ToM using written storyboards as experimental materials, whereas green blobs refer to the activations from studies using non-story materials (pictures, cartoons, animations, etc.). Non-story ToM recruits a network which extends to the most dorsal and posterior portions of the medial prefrontal cortex, overlapping the cluster found in the present study.

characters' mind on the basis of mixed information. Earlier implications of the dMPFC whilst mentalizing under uncertainty (Jenkins & Mitchell, 2010) or in conditions of high cognitive control (Hartwright, Apperly, & Hansen, 2012, 2013), are consistent with this interpretation. Yet, in the present study, neural activity in dMPFC does not reflect mere task difficulty. Indeed, whereas behavioral measures show that the task was mostly difficult in the first time-bins, regardless of the kind of scenario employed (Fig. 2A), dMPFC activity is clearly modulated by the scenario in the last time-bins (Fig. 4A).

4.3. *The prediction of rule-related behavior*

In line with developmental studies revealing that deontic reasoning dissociates from children ToM abilities (Clément et al., 2011), we found that prediction of people's rule-related behavior triggered in participants differential behavioral and neural responses, relative to desire-related predictions. As behavioral evidence, we found that only rule-based predictions shaped participants evaluations of people in terms of dominance and trustworthiness. Interestingly, although in our experiment both desires and rules assessments were associated with the same outcome likelihood (33% violations, 66% observances), participants' post-scanning ratings were affected exclusively by how each profile behaved in relation to rules. As neural evidence, we found that oMPFC and left amygdala exhibited significant SCENARIO-effects specifically for the last two time-bins, concomitantly with an accurate performance in the task. Similarly to the case of dMPFC, the functional properties of oMPFC and amygdala cannot be related to broad learning about people's behavior, but rather to the specific content of the information learned. However, contrary to the dMPFC which specifically responded to desires (relative to rules) scenarios, oMPFC and amygdala were most sensitive to rules. In line with previous developmental evidence, the combined rating and neural data suggest that in adults too deontic reasoning is processed through cognitive processes which are at least in part independent from ToM. This does not exclude that rule-based behavior might also trigger neural responses related to mentalizing abilities (see below); however these responses are not expected to be rule-specific, but present also during desires scenarios.

The involvement of orbitofrontal cortex and amygdala in rules assessments converges with human and animal studies describing these regions as highly interconnected (Aggleton, Burton, & Passingham, 1980; Ghashghaei & Barbas, 2002) and co-active in many manipulations (Bzdok, Laird, Zilles, Fox, & Eickhoff, 2013; Bzdok, Langner, et al., 2013), specifically in negative/positive evaluations of faces (Mende-Siedlecki, Said, et al., 2013). In particular, consistently with accounts positing a critical role of the amygdala in coding salient (and valence-independent) social information (Sander, Grafman, & Zalla, 2003), recent studies testing the neural correlates of trustworthiness judgments implicated part of the amygdala, not only in linear effects (the less trustworthy, the higher the BOLD signal), but also in quadratic effects in which the activity was high for the perception of positive (trustworthy) or negative (untrustworthy) but not neutral faces (Mende-Siedlecki, Said, et al., 2013; Todorov, 2008; Todorov & Engell, 2008). Furthermore, amygdala response does not seem to be driven exclusively by perceivable face features, but also by their previously-encoded contextual information (Vrticka, Andersson, Sander, & Vuilleumier, 2009). Our data support these observations by showing that the left amygdala responds to acquired knowledge (not only instantaneous impressions) about people's behavior in relation to rules, irrespective of whether rules are violated/obeyed.

Our data are consistent with recent accounts according to which, due to flexible engagement of domain-specific functions, social inference might be achieved through distinct

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

pathways or strategies (e.g., Ames, 2004; Jenkins & Mitchell, 2010, for reminiscent accounts). Furthermore, recent models favored an organization of the medial prefrontal cortex along the *ventral-to-dorsal* axis in terms of: representation of *affective-to-cognitive* states in others (Shamay-Tsoory, Tibi-Elhanany, & Aharon-Peretz, 2006), *outcome-to-goals* of social events (Amodio & Frith, 2006; Krueger, Barbey, & Grafman, 2009), or more generally *automatic-to-controlled* social processes (Bzdok, Langner, et al., 2013; Forbes & Grafman, 2010; Lieberman, 2007). Following our results, we believe that these different accounts might converge into a unique model of medial prefrontal organization, with its dorsal portion representing a high-level mentalistic pathway for interpersonal reasoning (more demanding and focused often on cognitive states such as goals/intentions), and the ventral/orbital part (plus the amygdala) reflecting an early-developing system involved in low-level social inferences (less demanding and focused on people's overt behavior/reactions).

4.4. *Common effects between rules and desires*

Although our experiment served well the purpose of dissociating rules from desires processing, it is less suited for interpreting effects common to both scenarios, such as those identified through main effects and conjunction analyses (see Fig. 3 and Tables 2–3). One reasonable explanation of these effects is related to the predictive nature of the task (see Rushworth & Behrens, 2008, as review): for instance, the activity of PCC and ventral striatum during portrait events increases linearly across the course of the experimental session (Fig. 3, red blobs), presumably reflecting efficient prediction of characters' behavior; likewise, the activity of middle cingulate cortex and precuneus during feedback events linearly decreases across time (Fig. 3, blue blobs), consistently with prediction error signals diminishing during high accuracy.

Furthermore, it should be also acknowledged that our experimental design might be susceptible to asymmetric contaminations between the assessments of rule- and desire-based behavior. In particular, behaviors which violate externally imposed rules can be informative, not only about the characters' dispositions towards norms, but also that their desires do not coincide with the instructions received (e.g., refusing to clean the garden can be interpreted in terms of stronger desires for other activities). Although these contaminations are held to affect minimally the processing of compliant actions (orders can be followed regardless of the underlying desires; see Kaufmann, 2005; Searle, 2001), it is possible that some residual desire-based processing might be associated with the inference of rule-based behavior, especially when non-compliant. This could explain the pattern of rMPFC activity displayed in Fig. 5, which is part of the core-ToM network (Fig. 6), and which is associated with predicting behaviors which violate the premises, regardless of whether these are rules or desires. Although activation of coordinates previously implicated in mentalizing process does not guarantee *per se* an involvement of ToM-abilities (Corradi-Dell'Acqua et al., 2014), it is plausible that the prediction of violating behavior elicits, in addition to the oMPFC/amygdala and dMPFC pathways related respectively to rules and desires, additional processing in the core-ToM network to provide a more comprehensive representation of the characters' intentions ("why doesn't he obey rules?"/"why doesn't he fulfill his desires?"). Keep in mind, however, that the effects associated with rMPFC could also be related to the salient or infrequent ("odd") nature of the inconsistent trials. Future studies will therefore need to better investigate the role of rMPFC in behavior prediction in relation to ToM-related networks.

Acknowledgments

This research was supported by the National Center of Competence in Research (NCCR) for Affective Sciences financed by the Swiss National Science Foundation and hosted by the University of Geneva. Further support comes from the Swiss National Science Foundation grant n. 320030_135653 awarded to SS.

REFERENCES

Aggleton, J. P., Burton, M. J., & Passingham, R. E. (1980). Cortical and subcortical afferents to the amygdala of the rhesus monkey (*Macaca mulatta*). *Brain Research, 190*(2), 347–368. http://dx.doi.org/10.1016/0006-8993(80)90279-6.

Ames, D. R. (2004). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573–585. http://dx.doi.org/10.1037/0022-3514.87.5.573.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews. Neuroscience, 7*(4), 268–277. http://dx.doi.org/10.1038/nrn1884.

Baron, S. G., Gobbini, M. I., Engell, A. D., & Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience, 6*(5), 572–581. http://dx.doi.org/10.1093/scan/nsq086.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*(7219), 245–249. http://dx.doi.org/10.1038/nature07538.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: a theory of meaning, representation, and reasoning. *Cognitive Psychology, 50*(2), 159–193. http://dx.doi.org/10.1016/j.cogpsych.2004.08.001.

Bzdok, D., Laird, A. R., Zilles, K., Fox, P. T., & Eickhoff, S. B. (2013). An investigation of the structural, connectional, and functional subspecialization in the human amygdala. *Human Brain Mapping, 34*(12), 3247–3266. http://dx.doi.org/10.1002/hbm.22138.

Bzdok, D., Langner, R., Schilbach, L., Engemann, D. A., Laird, A. R., Fox, P. T., et al. (2013). Segregation of the human medial prefrontal cortex in social cognition. *Frontiers in Human Neuroscience, 7*, 232. http://dx.doi.org/10.3389/fnhum.2013.00232.

Clément, F., Bernard, S., & Kaufmann, L. (2011). Social cognition is not reducible to theory of mind: when children use deontic rules to predict the behaviour of others. *British Journal of Developmental Psychology, 29*(4), 910–928. http://dx.doi.org/10.1111/j.2044-835X.2010.02019.x.

Cloutier, J., Gabrieli, J. D. E., O'Young, D., & Ambady, N. (2011). An fMRI study of violations of social expectations: when people

are not who we expect them to be. *NeuroImage, 57*(2), 583–588. http://dx.doi.org/10.1016/j.neuroimage.2011.04.051.
Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2014). Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 9*(8), 1175–1184. http://dx.doi.org/10.1093/scan/nst097.
Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*(3), 187–276. http://dx.doi.org/10.1016/0010-0277(89)90023-1.
Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 584–627). Hoboken, NJ: Wiley.
Cummins, D. D. (1996). Evidence for the innateness of deontic reasoning. *Mind & Language, 11*(2), 160–190. http://dx.doi.org/10.1111/j.1468-0017.1996.tb00039.x.
Dunn, J. (1988). *The beginnings of social understanding*. Cambridge, MA: Harvard University Press.
Ermer, E., Guerin, S. A., Cosmides, L., Tooby, J., & Miller, M. B. (2006). Theory of mind broad and narrow: reasoning about social exchange engages ToM areas, precautionary reasoning does not. *Social Neuroscience, 1*(3–4), 196–219. http://dx.doi.org/10.1080/17470910600989771.
Fiddick, L. (2004). Domains of deontic reasoning: resolving the discrepancy between the cognitive and moral reasoning literatures. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 57*(3), 447–474. http://dx.doi.org/10.1080/02724980343000332.
Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*(1), 1–79.
Fiddick, L., Spampinato, M. V., & Grafman, J. (2005). Social contracts and precautions activate different neurological systems: an fMRI investigation of deontic reasoning. *NeuroImage, 28*(4), 778–786. http://dx.doi.org/10.1016/j.neuroimage.2005.05.033.
Fiske, S. T., & Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin, 6*(4), 543–557. http://dx.doi.org/10.1177/014616728064006.
Flavell, J. H. (1999). Cognitive development: children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45. http://dx.doi.org/10.1146/annurev.psych.50.1.21.
Forbes, C. E., & Grafman, J. (2010). The role of the human prefrontal cortex in social cognition and moral judgment. *Annual Review of Neuroscience, 33*, 299–324. http://dx.doi.org/10.1146/annurev-neuro-060909-153230.
Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C. (1993). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping, 1*(3), 210–220. http://dx.doi.org/10.1002/hbm.460010306.
Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind". *Trends in Cognitive Sciences, 7*(2), 77–83. http://dx.doi.org/10.1016/S1364-6613(02)00025-6.
Ghashghaei, H., & Barbas, H. (2002). Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience, 115*(4), 1261–1279. http://dx.doi.org/10.1016/S0306-4522(02)00446-3.
Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746. http://dx.doi.org/10.1073/pnas.0711099105.
Hanley, J. A., Negassa, A., Edwardes, M. D. deB., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology, 157*(4), 364–375.
Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*(4), 921–930. http://dx.doi.org/10.1016/j.neuroimage.2012.03.012.
Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2013). Representation, control, or reasoning? Distinct functions for theory of mind within the medial prefrontal cortex. *Journal of Cognitive Neuroscience*. http://dx.doi.org/10.1162/jocn_a_00520.
Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex, 20*(2), 404–410. http://dx.doi.org/10.1093/cercor/bhp109.
Kalish, C. W. (2006). Integrating normative and psychological knowledge: what should we be thinking about. *Journal of Cognition and Culture, 6*(1–2), 191–208. http://dx.doi.org/10.1163/156853706776931277.
Kaufmann, L. (2005). Self-in-a-Vat: on John Searle's ontology of reasons for acting. *Philosophy of the Social Sciences, 35*(4), 447–479. http://dx.doi.org/10.1177/0048393105282918.
Krueger, F., Barbey, A. K., & Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends in Cognitive Sciences, 13*(3), 103–109. http://dx.doi.org/10.1016/j.tics.2008.12.005.
Kuzmanovic, B., Bente, G., von Cramon, D. Y., Schilbach, L., Tittgemeyer, M., & Vogeley, K. (2012). Imaging first impressions: distinct neural processing of verbal and nonverbal social information. *NeuroImage, 60*(1), 179–188. http://dx.doi.org/10.1016/j.neuroimage.2011.12.046.
Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology, 58*, 259–289. http://dx.doi.org/10.1146/annurev.psych.58.110405.085654.
Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62*, 103–134. http://dx.doi.org/10.1146/annurev-psych-120709-145406.
Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F. V., Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience, 7*(8), 937–950. http://dx.doi.org/10.1093/scan/nsr064.
Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623–631. http://dx.doi.org/10.1093/scan/nss040.
Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience, 8*(3), 285–299. http://dx.doi.org/10.1093/scan/nsr090.
Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience, 1*(1), 49–55. http://dx.doi.org/10.1093/scan/nsl007.
Núñez, M., & Harris, P. L. (1998). Psychological and deontic concepts: separate domains or intimate connection? *Mind & Language, 13*(2), 153–170. http://dx.doi.org/10.1111/1468-0017.00071.
Rubin, K. H., Bukowski, W. M., & Parker, J. G. (1998). Peer interactions, relationships, and groups. In N. Eisenberg (Ed.), *Handbook of child psychology* (pp. 619–700). New York, NY: Wiley.
Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience, 11*(4), 389–397. http://dx.doi.org/10.1038/nn2066.
Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences, 14*(4), 303–316. http://dx.doi.org/10.1515/REVNEURO.2003.14.4.303.

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87–124. http://dx.doi.org/10.1146/annurev.psych.55.090902.142044.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4), 1835–1842.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science: A Journal of the American Psychological Society/APS*, 17(8), 692–699. http://dx.doi.org/10.1111/j.1467-9280.2006.01768.x.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, 12(4), 508–514. http://dx.doi.org/10.1038/nn.2278.

Searle, J. R. (2001). *Rationality in action*. Cambridge, MA: MIT press.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, 1(3–4), 149–166. http://dx.doi.org/10.1080/17470910600985589.

Srull, T. K., & Wyer, R. S. (1989). Person memory and judgment. *Psychological Review*, 96(1), 58. http://dx.doi.org/10.1037/0033-295X.96.1.58.

Todorov, A. (2008). Evaluating faces on trustworthiness. *Annals of the New York Academy of Sciences*, 1124(1), 208–224. http://dx.doi.org/10.1196/annals.1440.012.

Todorov, A., & Engell, A. D. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social Cognitive and Affective Neuroscience*, 3(4), 303–312. http://dx.doi.org/10.1093/scan/nsn033.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1), 273–289. http://dx.doi.org/10.1006/nimg.2001.0978.

Vrticka, P., Andersson, F., Sander, D., & Vuilleumier, P. (2009). Memory for friends or foes: the social context of past encounters with faces modulates their subsequent neural traces in the brain. *Social Neuroscience*, 4(5), 384–401. http://dx.doi.org/10.1080/17470910902941793.

Wellman, H. M., & Miller, J. G. (2008). Including deontic reasoning as fundamental to theory of mind. *Human Development*, 51(2), 105–135. http://dx.doi.org/10.1159/000115958.

Please cite this article in press as: Corradi-Dell'Acqua, C., et al., How the brain predicts people's behavior in relation to rules and desires. Evidence of a medio-prefrontal dissociation, Cortex (2015), http://dx.doi.org/10.1016/j.cortex.2015.02.011