SCIENTIFIC REPORTS

OPEN

Neural Representations Behind 'Social Norm' Inferences In Humans

Felipe Pegado[1,2,3], Michelle H. A. Hendriks[1,3], Steffie Amelynck[1], Nicky Daniels[1], Jessica Bulthé[1], Haemy Lee Masson[1], Bart Boets[2,3] & Hans Op de Beeck[1]

Received: 8 December 2017
Accepted: 8 August 2018
Published online: 28 August 2018

Humans are highly skilled in social reasoning, e.g., inferring thoughts of others. This mentalizing ability systematically recruits brain regions such as Temporo-Parietal Junction (TPJ), Precuneus (PC) and medial Prefrontal Cortex (mPFC). Further, posterior mPFC is associated with allocentric mentalizing and conflict monitoring while anterior mPFC is associated with self-reference (egocentric) processing. Here we extend this work to how we reason not just about what one person thinks but about the abstract shared social norm. We apply functional magnetic resonance imaging to investigate neural representations while participants judge the social congruency between emotional auditory utterances in relation to visual scenes according to how 'most people' would perceive it. Behaviorally, judging according to a social norm increased the similarity of response patterns among participants. Multivoxel pattern analysis revealed that social congruency information was not represented in visual and auditory areas, but was clear in most parts of the mentalizing network: TPJ, PC and posterior (but not anterior) mPFC. Furthermore, interindividual variability in anterior mPFC representations was inversely related to the behavioral ability to adjust to the social norm. Our results suggest that social norm inferencing is associated with a distributed and partially individually specific representation of social congruency in the mentalizing network.

Humans have an extraordinary capacity to understand their conspecifics. This 'social reading' in natural environments involves the processing of visual cues – e.g., face expressions[1], auditory cues – e.g., prosody[2,3], and other sensory information that is usable to infer others' feelings, desires and thoughts[4–6]. Although our mentalizing capacity (or Theory of Mind) typically relies on sensory cues of concrete targets, it can also be performed with more abstract cues such as verbal information about a person[7]. Mentalizing tasks systematically activate the so-called mentalizing brain network, including the temporo-parietal junction (TPJ), precuneus (PC) and medial prefrontal cortex (mPFC)[4–6,8]. However it is unclear how the human brain mentalizes at a more abstract level, for instance, when targeting not only the thinking of one particular person but instead how the general population 'thinks'? In other words, how does the human brain infer what 'most people' think, for instance concerning appropriate social behavior?

From a behavioral point of view, it is now known that the development of such abstract inferences of social norms relies on active learning during concrete social interactions at very early ages (at least 3 years-old)[9] (but see also[10] for a passive social learning alternative). Learning social norms in a particular culture and in a particular family ultimately generates personal references of what most people think about appropriate reactions in different contexts (personal bias for social norms)[5,11,12]. Imagine for a moment that you are presenting your holiday pictures to an audience of relatives. Upon displaying a positive valence image (e.g., a photograph of a beautiful scene), one observer could react by expressing a positive valence reaction (e.g., admiration, by using a vocal utterance such as "uaaaauu"). In this situation, the reaction will probably be perceived by most observers as appropriate (i.e., "congruent"). Whereas in the case of a negative reaction to the same positive picture (e.g., a disgust reaction, expressed by an utterance such as "uuuurg"), this response will generally be perceived as "incongruent". Note however that in more nuanced or ambiguous situations, it can be challenging to judge the congruency of social responses and to further estimate the 'common sense' or the 'social norm' (i.e., what most people would think about the social congruency).

Here we present a new behavioral and neuroimaging paradigm which implements this latest example, requesting people to infer how most people would judge the congruency of vocal reactions to visual scenes. Here we will

[1]Department of Brain and Cognition, KU Leuven, 3000, Leuven, Belgium. [2]Center for Developmental Psychiatry, Department of Neurosciences, KU Leuven, 3000, Leuven, Belgium. [3]Leuven Autism Research consortium, KU Leuven, 3000, Leuven, Belgium. Correspondence and requests for materials should be addressed to F.P. (email: felipepegado@yahoo.com) or H.O.B. (email: hans.opdebeeck@kuleuven.be)

focus on the neural representations underlying this 'social norm' inferencing, an unexplored aspect in the social cognitive neuroscience literature.

We expect to find social congruency information (i.e., a distinction between congruent vs incongruent trials in neural representations) represented in the Theory of Mind (ToM) network but not in sensory areas in the visual or auditory systems. In addition, the mPFC might show further dissociations, as self-reference processing (egocentric) has been associated with brain activity in more *anterior* parts of mPFC, while mentalizing about others (allocentric) has been related to activity in more *posterior* parts of mPFC[7,12,13], as confirmed in a meta-analysis of more than 100 studies[14]. Furthermore, conflict monitoring is also hosted in posterior parts of mPFC[15]. Thus, monitoring conflict of social congruency itself and/or between egocentric and allocentric (social norm) responses, could also engage the posterior part of mPFC. We therefore predict to find stronger social congruency representations in posterior mPFC than in anterior mPFC.

Results

Behavioral results during the fMRI. During the experiment in the fMRI scanner, binary judgments of social congruency (i.e., congruent vs incongruent) relative to the inferred social norm were collected from each participant and run for the 96 Audio-Visual (A-V) conditions (12 visual X 8 auditory; see Fig. 1a). The most common response across runs was calculated at the individual level and then averaged at the group level (Fig. 1a). The group result reflects the 'shared social norm' pattern of response among participants and follows essentially (in 94 out of 96 conditions) a cross-modal valence congruency pattern, i.e., visual and auditory stimuli with the same valence (both positive or both negative) are considered congruent while with different valences are considered incongruent.

We further analyzed if the participant's response agreement varied as a function of valence in the two sensory modalities. We thus calculated the '% of incongruent responses' across runs for each of the 96 cells, as presented in Fig. 1b (see also Fig. 1c for individual results) and compared the responses across the four main quadrants. Results show that responses did not vary as a function of valence in the visual (two upper quadrants vs two lower quadrants: $T(47) = 0.9$; $p = 0.37$) or auditory (two left vs two right quadrants: $T(47) = 0.52$; $p = 0.61$) modalities. Nevertheless, a difference in the response agreement was found between positive vs negative *congruent* trials (quadrants upper left vs down right in Fig. 1b): $T(23) = 3.58$; $p = 0.002$, but no difference in *incongruent* trials (quadrants upper right vs down left): $T(23) = 0.81$; $p = 0.42$, suggesting that valence played a role in subjects' agreement only in congruent trials.

Further, for each individual run, we created dissimilarity matrices of behavioral responses by pairwise comparing the responses for each A-V cell (resulting in binary 96×96 matrices) using 1 minus correlation (Spearman's correlation coefficient) [the same procedure used for non-binary ratings; see following section and Fig. 2a–c]. We then correlated these dissimilarity matrices across the 144 recorded runs (24 subjects X 6 runs each; one subject excluded for missing run; see Fig. 3a left) to calculate both the individual within-subject correlations (15 subject-specific run-combinations, averaged) and between-subject correlations (correlations of each subject-specific run with each of the other runs, except from her/himself, averaged) (see Fig. 3a, right). Results revealed a high degree of variability in the between-subject correlations, demonstrating that some participants were better than others in the ability to match the 'shared social norm'. Note also the important variability of the "internal noise" of participant's responses, i.e., consistency of responses for the same trial type across runs (within-subject correlations). The within-subject consistency represents the ceiling of between-subject correlations that can be expected for a particular participant. We will later use the between-subject variability in task performance (normalized by the participant's internal SNR, i.e., reliability). The same rationale is used in typical MVPA research: the within-measurement correlation is used as the realistic ceiling, and ROIs are only compared with regard to their between-measurement correlations after taking into account this within-measurement correlation. For instance, comparing correlations in low level visual areas such as BA 17 and 18 (which present low noise levels) with correlations in a noisy frontal or parietal ROI does only make sense after taking reliability (within-measurement correlations) into account (see SNR ceiling as background bar in Figure 4 in[16]). For further explanations see the discussion on the seminal work of Cronbach (basic measurement theory) applied to fMRI data in[17].

Note that in our data (Fig. 3a) all between-subject correlation values are smaller than the within-subject correlation values. This is an expected result given that subjects would naturally agree better with themselves than with others. Note also that the values of the between/within correlations ratio are never close or above 1 (see Fig. 3c). Thus, potential extreme cases that could make our index problematic, i.e., if within-subject correlations would be much lower than between-subject correlations, or extremely high between/within ratio, do not apply to our data. In conclusion, in a realistic example, participants showing equivalent reliability levels (in Fig. 3c, subjects 13 and 14 exhibit similar average within-subject correlation, in the range of 0.6) can present different levels in the capacity of inferring what others will answer (between-subject correlations, in this example in the range of 0.3 to 0.4). Thus, in this case, subject 13 will have a slightly higher index than subject 14.

We then analyzed the response times (RTs). The global mean of participant's median RTs was 1.98 seconds (see Supplementary Figure 1). No difference was found for positive ($RT = 1.99 +/- 0.03$ secs) vs negative ($RT = 2.01 +/- 0.05$ secs) valence *images* ($T(46) = -0.36$; $p = 0.72$; Supplementary Figure 1a). Also, equivalent RTs were found for positive ($RT = 1.99 +/- 0.03$ secs) vs negative ($RT = 2.01 +/- 0.02$ secs) *vocalizations* ($T(46) = -0.12$; $p = 0.22$; Supplementary Figure 1b). Importantly, no difference was found between congruent ($RT = 2.00 +/- 0.22$ secs) vs incongruent trials ($RT = 1.96 +/- 0.23$ secs) ($T(46) = 0.43$; $p = 0.66$; Supplementary Figure 1c). Indeed, RTs were quite equivalent across all audio-visual combination trials (Supplementary Figure 1d).

Behavioral results using a fine-grained scale (outside scanner). Outside the scanner, 2 runs of the same task were administered, but instead of binary responses (congruent vs incongruent), subjects used a

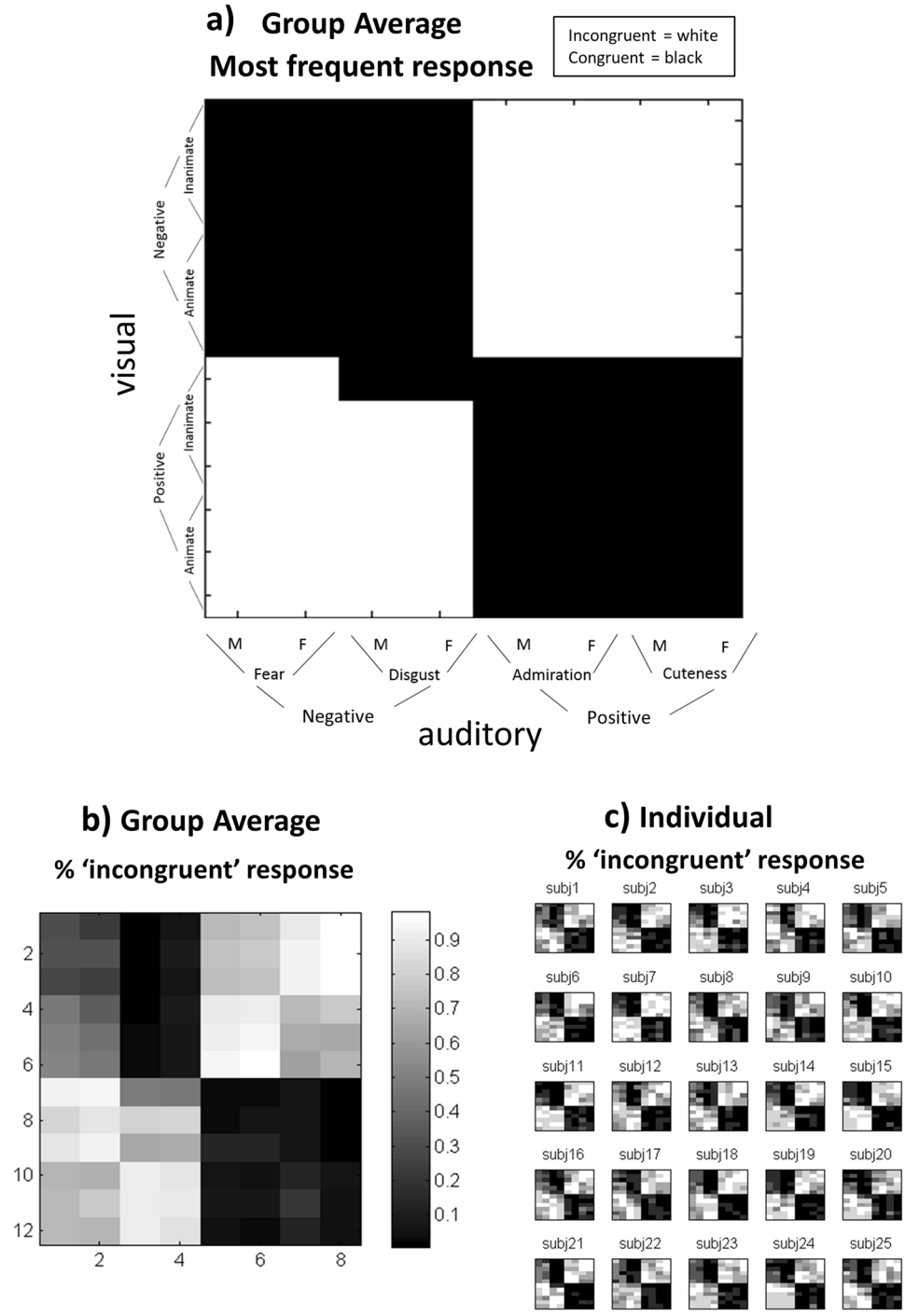

Figure 1. Behavioral responses during the scanner. Both for visual and auditory domains, half of the stimuli presented a positive valence and the other half a negative valence. **(a)** Most frequent binary (congruent vs incongruent) response per Audio-Visual (A-V) stimuli combination (8 audio X 12 visual = 96) across the six runs, at the group level, representing thus the 'shared social norm' among participants. **(b)** % of 'incongruent' responses, at the group level. **(c)** % of 'incongruent' responses at the individual level.

finer-grained 9-level scale to rate the congruency level (see Methods). Using the same procedure as before (see previous Results section), we calculated the similarity of response patterns and the results show again variations between subjects in how much they correlate with other subjects in terms of which congruency rating they give to specific visual-auditory stimulus combinations (Fig. 2a). Thus, both using fine-grained scales and binary decisions the task was sensitive to capture variations in the subject's capactity to guess what others (their peers) would answer.

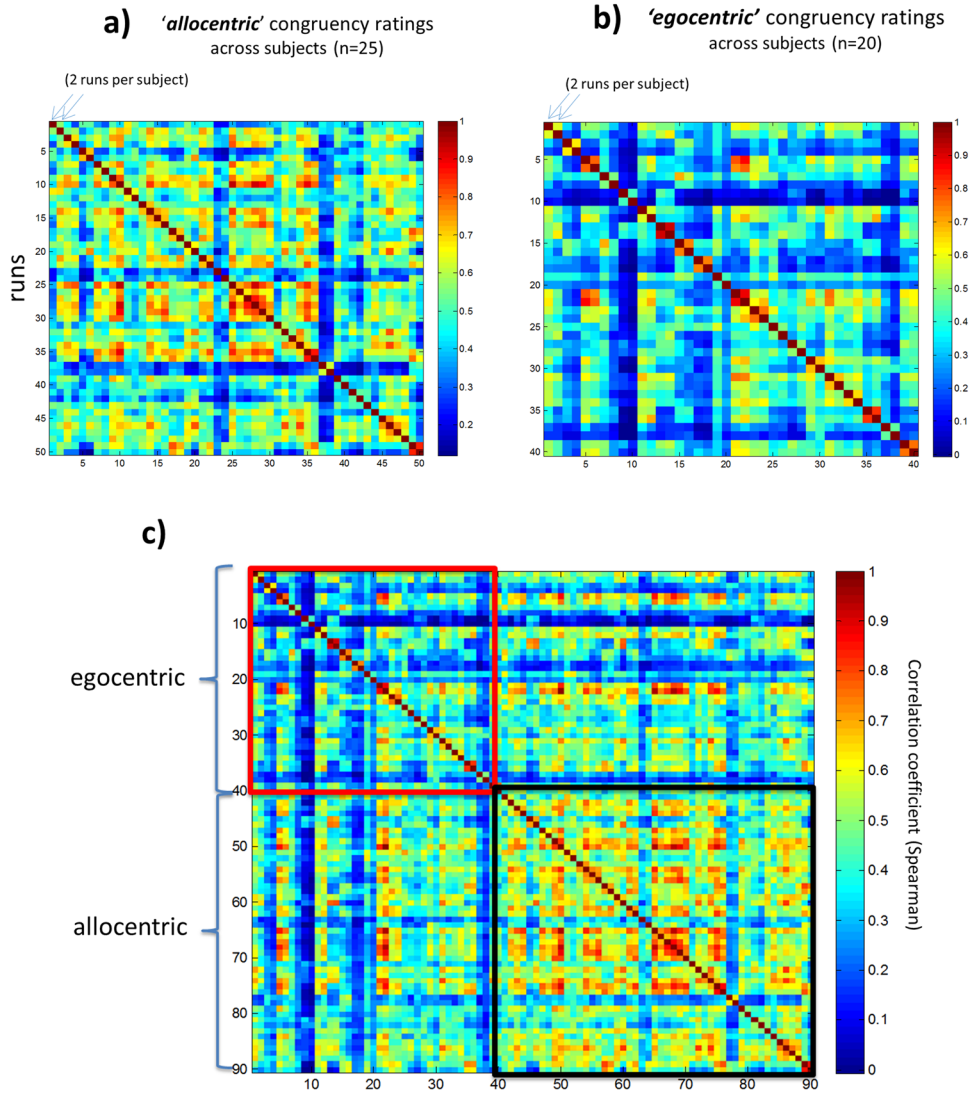

Figure 2. Behavioral responses outside the scanner using a fine-grained scale. **(a)** *Same task as in the scanner (allocentric reference)*. Outside the scanner, subjects performed 2 runs of the same social norm perspective task but instead of binary responses, they used here a more fine-grained scale (9 levels). **(b)** *Control task (egocentric reference)*. A separate group of subjects performed a control task, where the judgements were based on their own perspective of social congruency, using again a 9 level scale (see Methods for details). **(c)** *Comparing egocentric versus allocentric reference judgements*. Comparison of the results of the two tasks, so that each of the 45 participants in total is correlated with each other. Colorbar = Spearman's correlation coefficient. Run = one recording block with all audio-visual combinations presented once.

Behavioral validation of the social norm perspective relative to egocentric perspective. We assumed here that between-subject correlations are a quantitative measure of the extent that individual participants can identify the shared social norm, which they were required to do through explicit task instructions. We tested this assumption by asking an independent group of participants to perform the same task with the fine-grained rating scale but now judging according to their own perspective (egocentric perspective) (see Methods). Results revealed that *within*-subject correlations were equivalent to those in the social norm perspective task ($r = 0.65$ vs 0.67 respectively; $t(43) = -0.3$) (see Fig. 2), suggesting an equivalent level of internal noise (within-subject correlations across runs) in both tasks. More importantly however, egocentric perspective responses showed significantly lower *between*-subject correlations than in the social norm perspective ($r = 0.33$ vs 0.50; $t(43)$ −7.1; $p < 0.0001$). In other words, by comparing the two fine-grained tasks, we find an objective increase in the similarity of the response patterns among participants when they are requested to judge social congruency under the 'social norm' perspective relative to when they use their personal perspective. This result demonstrates the sensitivity of our paradigm for subjective changes in the adopted perspective.

Neural representations of social norm processing. We used correlational MVPA to investigate whether there is a systematic difference between the multi-voxel pattern associated with congruent trials and

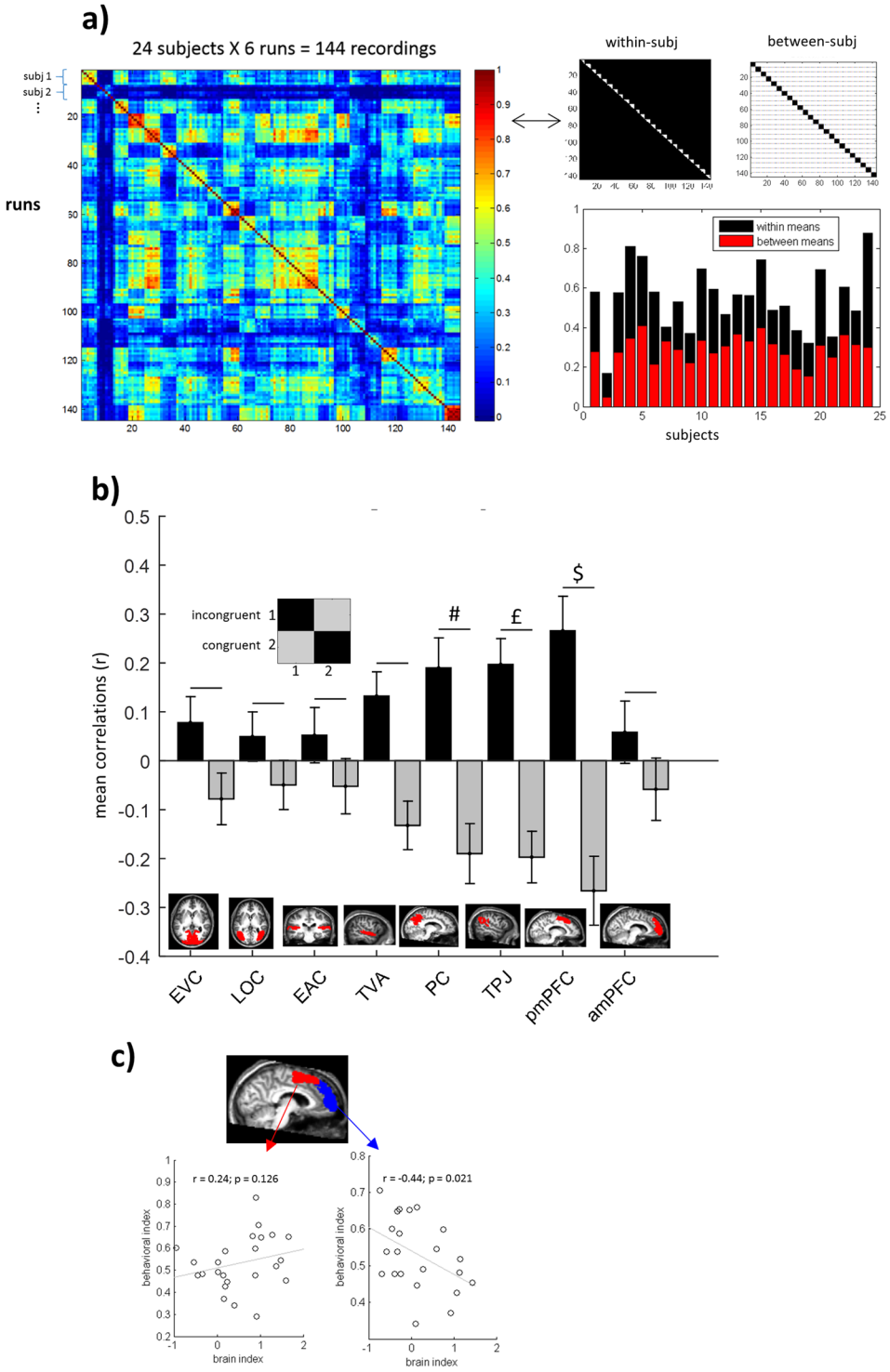

Figure 3. Similarity of Social Congruency Representations. (**a**) *Similarity of behavioral response patterns.* Similarity of behavioral response patterns across runs and subjects (left panel) calculated for each pair of runs (Spearman correlations). Within and between subject correlations (right panel upper: cells in white for within [left] and between [right] subject correlations (each line delineates the between-subject correlations for each subject). A behavioral indication of the ability to infer what most people would answer (a social norm mentalizing 'performance'), was indexed by the individual agreement with their peers (between-subject correlation), normalized by the participant internal noise, i.e., consistency of response across runs (within-subject correlation): a between/within ratio. (**b**) *ROIs with social congruency informatio*n. Two conditions (congruent vs incongruent) were defined in a GLM model of the fMRI data and then 2 × 2 neural similarity matrices were created (inset). None of the sensory areas (EVC = Early Visual Cortex; LOC = Lateral Occipital Complex; EAC = Early Auditory Cortex; TVA = Temporal Voice Area) show significant social congruency information content, while three of the mentalizing network ROIs did show (PC = Precuneus; marginally

significant; TPJ = Temporo-Parietal Junction and mPFC = medial Prefrontal Cortex in its posterior part). Inset (upper left): neural similarity matrix of congruency. P-values (Bonferroni-corrected): #p = 0.057; $^{£}$p = 0.013; $^{\$}$p = 0.011. (**c**) *Linking neural and behavioral data*. To test if the social norm mentalizing performance (behavioral index) could be explained by differences in social congruency neural data in subparts of mPFC, brain x behavior correlations were performed. Two subjects didn't meet the ROI criteria in anterior mPFC (see ROIs section in Methods).

the multi-voxel pattern associated with incongruent trials (see Methods), thus revealing a neural representation of (in)congruency with the social norm. We calculated the correlation between multi-voxel patterns elicited by the same condition, r(same) (e.g., correlating congruent with congruent), and compared it with the correlation between patterns from different conditions, r(diff) (e.g., correlating congruent with incongruent) (see Methods section 4.8). Using a set of a priori defined ROIs we were able to capture social congruency information. As expected and shown in Fig. 3b, most brain regions of the mentalizing network represent social congruency as revealed by r(same) being higher than r(diff), i.e., r(same) – r(diff) one sample t-test: PC marginally significant (p = 0.057; all p's Bonferroni-corrected for multiple comparisons, i.e., the number of ROIs), TPJ (p = 0.013) and posterior mPFC (p = 0.011) (see Fig. 3b). However, this was not the case for the anterior mPFC (p > 0.9) nor for any of the sensory regions: EVC (p > 0.9), LOC (p > 0.9), EAC (p > 0.9), and TVA (p = 0.12). Using small spheres searchlight analysis (see Methods section 4.11), we did not find brain regions capturing the social congruency information, even uncorrected for multiple tests (p = 0.001). Accordingly, previous results suggest that larger ROI analysis can be more sensitive than small spheres searchlight analysis[18].

Correlation between neural x behavioral data. The dissociation between posterior and anterior mPFC is directly relevant with regard to the perspective taken by our participants during the social congruency task. Despite differences in the absolute location (see mPFC activations in a meta-analysis of different Theory of Mind tasks[6]), previous studies using different tasks have related with good consistency allocentric perspective-taking with activity in a relatively more posterior portions of mPFC, and egocentric perspective taken with activity in more anterior portions of mPFC[7,12–14]. In addition, other studies using allocentric perspective tasks more closely related to our task, such as the 'social conformity' task[19–21], mentalizing about groups[22] and the 'simulating other's decisions' task[23,24] found an engagement of posterior portions of mPFC. We asked subjects to take an allocentric perspective, which might be the explanation for the presence of the neural representation in posterior mPFC and not in anterior mPFC. However, not all subjects performed this task in the same manner. As mentioned before, the size of between-subject correlations in behavioral response patterns, which we have shown to be modulated by the adoption of an allocentric vs egocentric perspective (see Section 2.3), showed a lot of inter-subject variability. Hence, we wondered whether participants who were less able to adopt the requested allocentric social norm perspective may show a stronger social congruence representation in anterior mPFC (a portion of mPFC linked to egocentric processing). To test this hypothesis, we computed a behavioral index quantifying to what extent the response pattern of a particular subject is similar to the other participants. For this, we calculated the behavioral individual between-subject correlations, normalized by the participant internal noise (i.e., consistency of responses for the same trial type [one of the 96 audio-visual combinations] across runs): a between/within subject correlations ratio. A lower ratio indicates that a participant was less able to incorporate the shared social norm response. Results indeed revealed the predicted direction of the association: a negative correlation between social congruency information (i.e., difference between correlation values for the same trials (congruent or incongruent) vs different trials in the congruency neural matrix; see Results section 2.4 above and Methods section 4.8 and 4.10) in *anterior* mPFC and the behavioral index of social norm inference: Pearson's r = −0.44; one-tailed p = 0.021 (Fig. 3c). In contrast, *posterior* mPFC shows a nonsignificant trend in the opposite direction (r = 0.24; p = 0.126). Comparing the differences in correlation between anterior and posterior mPFC, we found a significant difference (Z = −2.89; p = 0.002).

Discussion

To shed light upon the neural basis of 'social norm' inferences, we used a naturalistic audio-visual fMRI paradigm that mimics social reactions (vocalizations) in different visual contexts while asking subjects to imagine what most people would answer concerning the social congruency.

Behavioral data enabled us to (1) assess to what extent our paradigm is sensitive to the adopted perspective, with reference to an allocentric "social norm" versus an egocentric perspective, and (2) quantify the degree to which individual participants adhere to the shared social norm. First, behavioral response patterns revealed that when using the social norm perspective, participants objectively show higher between-subject correlations than when an egocentric perspective was adopted (Fig. 2). Further, as there is no unequivocal 'correct response' on these subjective judgments of social congruency, similarity of response patterns across subjects was used to quantify task 'performance', i.e., the ability of each participant to match the shared social norm (Fig. 1), revealing a good degree of variability across subjects (Fig. 3a).

At the brain level, we found that social congruency processing was not hosted in low and high-level visual and auditory sensory areas. In contrast, three regions often associated with mentalizing and theory of mind[4–6,8] were engaged: Precuneus (PC; marginally significant), Temporo Parietal Junction (TPJ) and medial Prefrontal Cortex. Note that our results only points to the posterior (but not the anterior) part of mPFC. Our allocentric mentalizing task could have favored the posterior instead of the anterior mPFC site for social information processing. Accordingly, previous work have shown a dissociation between egocentric and allocentric mentalizing processing in anterior and posterior mPFC respectively[7,12–14]. In summary, the recruitment of the ToM network and the

dissociation in mPFC found here for the *abstract* entity 'most of people' (social norm inferences) are equivalent to what is typically found for inferences about *concrete* individuals, potentially suggesting a common neural substrate[6,14]. Accordingly, common areas for mentalizing about concrete groups and individuals (photographs) was previously demonstrated[22]. It is thus not implausible that the same concept extends for more abstract inferences.

We further tested if the social norm perspective used during scanning would be related to the lack of social congruency information in anterior mPFC. If so, we would expect that participants who are less able to adopt an allocentric social norm perspective are the ones who show the strongest evidence for egocentric social congruency representations in anterior mPFC. And indeed, this is exactly the pattern that was observed in anterior mPFC (Fig. 3c). The pattern was clearly different in posterior mPFC.

We should acknowledge that the mere observation of a social congruency representation does not pinpoint which processes are reflected in this representation. For the same reason we cannot be sure to what extent the significant congruency representation in the three mentalizing regions reflects similar or different functions. As one example of potentially different functions, the posterior part of mPFC has also been associated with conflict monitoring[15], including conflict monitoring in the context of social information processing[25–27]. Assuming that conflict monitoring would have played a role in posterior mPFC activity, it remains unclear whether posterior mPFC may have tracked conflicts concerning the social congruency itself (congruent vs incongruent) or rather conflicts between egocentric and allocentric responses (or even both). Finally, the effects observed could have been driven by surprise (violations of expected social reactions). Further investigations should clarify this issue.

Thus, our results suggest that when making inferences about the social norm, three regions of the ToM network regions show a significant representation of social congruency. The fourth region, anterior mPFC, shows considerable inter-subject variability in this congruence representation which is related inversely to the ability to take the allocentric perspective. Taken together, the recruitment of the ToM network and the anterior/posterior mPFC dissociation can potentially indicate an equivalent neural basis for *abstract* social norm inferences, as found in many previous studies requiring mentalizing about *concrete* single persons and groups[6,14,22].

Material and Methods

Participants. Twenty-five healthy subjects (7 female, 23.16 ± 3.32 years old, 7 left-handed) took part in the fMRI study. They all reported normal or corrected-to-normal vision, normal hearing and no neurological or psychiatric disorders. They received a financial compensation for their participation. The study was approved by the Medical Ethics Committee UZ/ KU Leuven University and all methods were performed in accordance with the relevant guidelines and regulations. All participants provided written informed consent prior to scanning.

Visual stimuli. Twelve images were selected from the standardized and widely used emotional pictures set IAPS (International Affective Picture System)[28], based on extreme valence ratings (positive vs negative) and on animacy categorization (animate vs inanimate). The six animate (e.g., humans, animals…) and the six non-animate pictures (e.g., landscapes, objects…) were orthogonal to the image valence, with half of them rated positively (e.g., happy baby), and half of them negatively (e.g., people being threatened with a gun). By including these sub-categories of visual stimuli we could investigate the structure of neural representations in the visual domain, which is reported elsewhere[29]. Based upon the quality of the brain responses evoked by a larger dataset of 24 IAPS images present in the pilot fMRI study (cf. infra), this final set of 12 images was selected: numbered 2341, 1710, 1750, 5760, 5825, 7492, 3530, 1300, 1930, 9290, 9300, 9301 in the IAPS database. IAPS policy requests not to publish the original images.

Auditory stimuli. Eight different non-verbal vocal utterances were used, inspired by previous work[30]. They express four different emotional reactions that could be more or less congruent with the pictures previously selected. Utterances expressing disgust, fear, admiration and cuteness, were recorded in an expressive but still natural manner (not an exaggerated caricature). Each emotional vocalization was performed by one male and one female actor. Including these sub-categories allow us to analyze the structure of auditory representations, which is reported elsewhere[29].They were recorded in a sound-proof room at 96 kHz sampling rate and 32-bit resolution, and were down-sampled to 44 kHz and 16-bit mono-recordings to reduce the size of the audio files. All stimuli had a fixed duration of 700 ms and an equivalent total Root Mean Square (RMS) power (−17.4 +/− 0.17 dB). Stimuli were slightly manipulated in Cool Edit Pro software and Adobe Audition CC 2015 software. Identical 600 ms silent periods were added before the onset of each auditory stimulus to create a natural delay from the visual stimulus onset, and a 100 ms silent period was added after the end of the utterance to provide stable ending transitions. Stimuli can be found here: https://osf.io/t7xp9/.

Behavioral Task in the scanner. Participants were lying in the scanner while watching a visual display and hearing auditory input through headphones. They were instructed to imagine they were seeing images (photographs) together with other unknown people. For each image, after a short delay, participants heard a vocal reaction (emotional utterance) that could be more or less congruent with the particular scene. The task was to evaluate the congruency of the vocal reaction in relation to the visual context. Yet, they did not have to judge this congruency from their own personal perspective, but instead they were explicitly instructed to evaluate whether most people would consider this vocal response appropriate or not (making inferences about the 'social norm') and respond accordingly. The two assigned buttons (congruent vs incongruent) were switched after 3 runs out of 6, and the assigned order was balanced across subjects.

Experimental fMRI runs. The fMRI session consisted of 6 runs, each with 96 pseudo-randomly presented experimental trials, i.e., all 12 visual stimuli paired with all 8 auditory stimuli. Additionally, 10 silent fixation trials were included among them, as well as 3 initial and 3 final dummy trials, making a total of 112 trials, with

4.5 seconds of Stimulus Onset Asynchrony (SOA), summing to 504 seconds of duration per run. Each experimental trial started with a visual image for 2.5 secs during which an auditory utterance was played via headphones (from 0.6 to 1.3 secs relative to the onset of the visual image) to simulate a natural delay before the vocal reaction. A 2 secs fixation cross was then displayed until the end of the trial. Subjects could respond any time within the trial and were instructed to press the buttons as soon as they know the answer.

fMRI data acquisition. Imaging data were acquired using a 3 T Philips Ingenia CX scanner (Department of Radiology of the University of Leuven) with a 32-channel head coil. Each functional run consisted of T2*-weighted echoplanar images (EPIs), with voxel size = 2. 2 × 2.2 × 2.7, interslice gap 0.2 mm, TR = 2550 ms, TE = 30 ms, matrix = 84 × 82, 45 slices, field of view (FOV) = 211 × 211 × 121. In addition to the functional images we collected a high-resolution T1-weighted anatomical scan for each participant (182 slices, voxel size = 0.98 × 0.98 × 1.2 mm, TR = 9.6 ms, TE = 4.6 ms, 256 × 256 acquisition matrix). Stimuli were presented using Psychtoolbox 3 (Brainard, 1997). Visual stimuli were displayed via an NEC projector with a NP21LP lamp that projected the image on a screen the participant viewed through a mirror. Viewing distance was approximately 64 cm. Auditory stimuli were presented through headphones at a comfortable hearing level.

fMRI preprocessing. Imaging data were preprocessed and analyzed using the Statistical Parametrical Mapping software package (SPM 8, Welcome Department of Cognitive Neurology, London, UK) and MATLAB. Functional images underwent slice timing correction (ascending order; first image as reference), motion correction (3rd degree spline interpolation), co-registration (anatomical to functional images; mean functional image as reference), and spatial normalization to the standard MNI (Montreal Neurological Institute) brain space. Functional images were spatially smoothed by convolution of a Gaussian kernel of 5 mm full-width at half-maximum[31]. One run of one subject was not considered due to excessive head movement (cut-off: 1 voxel size for two successive images).

General linear model (GLM). We applied a general linear model focusing upon the representation of social congruence here (but see[29] for results in the visual and auditory dimensions. For each participant and run, pre-processed images were modeled for each voxel using GLMs. They included regressors for each experimental condition and the 6 motion correction parameters (x, y, z for translation and rotation). Each predictor's time course was convolved with the canonical hemodynamic response function (HRF) in SPM. The social congruency GLM had two conditions (congruent vs incongruent) based on cross-modal valence congruency (e.g., a positive image combined with a negative utterance is considered 'incongruent') (see[29] Methods section "General Linear Model (GLM)" for further explanation on the chosen GLM approach). This approach guaranteed a perfect balance between 'congruent' and 'incongruent' conditions, while avoiding potential visual or auditory biases, because all visual and auditory stimuli occurs an equal number of times for congruent and incongruent trials. As we used extreme ends of the valence continuum (very positive or very negative), both for the visual and auditory domains, we assumed that in binary choices this social congruency criterion would reflect quite well the participants' responses. Indeed, group average responses showed a large overlap with this a priori definition of social congruency, as it was the most common response in 94 out of 96 A-V combinations (see Fig. 1a). Thus, this social congruency modeling also represents to a very large extent, the 'shared social norm' pattern, among the participant sample. The two social congruency conditions were modeled in relation to the time-window during which social congruency judgements could be performed, i.e., from the beginning of the auditory presentation (0.6 secs from trial onset) until the end of the trial (4.5 secs).

Regions of interest (ROIs). As primary ROIs we targeted the mentalizing neural network: medial Prefrontal Cortex (mPFC), Temporo-Parietal Junction (TPJ) and Precuneus (PC). Following the same approach as a recent meta-analysis of different mentalizing tasks[6], we used the same parcels that were obtained in functional connectivity studies, both for mPFC[32] and TPJ[33]. Note that only right hemisphere parcels are available for these two ROIs. To better account for individual anatomical and functional variability, the size of the ROIs should be sufficiently large. Thus, in some cases we grouped sub-divisions of smaller ROIs. As we did not have particular hypotheses for the two subdivisions of TPJ (anterior and posterior parcels) we grouped them together in a single ROI. In contrast, for the mPFC, we kept this distinction as the literature shows a clear functional dissociation between anterior vs posterior parts for self-related vs others-related mentalizing processes respectively[7,12–14], by integrating the four original parcels into two. Finally, for PC, we used the anatomical mask in WFU Pickatlas (SPM).

In addition, we included two visual and two auditory ROIs, functioning as control regions here (but see their neural patterns for visual and auditory dimensions in[29]). Our approach was to select the best available templates to delineate brain areas corresponding to low-level as well as high-level processing in each modality, avoiding in that way both manual delineation of ROIs and the use of several functional localizers. Low-level processing is localized in primary sensory cortices, of which we know that anatomical landmarks provide a proper approximation. Thus, we used anatomical masks from the anatomical atlas WFU PickAtlas Toolbox (Wake Forrest University PickAtlas, http://fmri.wfubmc.edu/cms/software). The low level visual ROI (Early Visual Cortex-EVC) was defined based on Brodman's areas (BA) 17 and 18 as they are widely accepted landmarks for low level visual processing. The low-level auditory ROI (Early Auditory Cortex- EAC) was composed by BA 41 and 42. The resulting EVC and EAC ROIs presented very thin configurations. This would lead to unrealistic delimitations of early processing cortex given the spatial uncertainty involved when comparing brains across subjects. We thus made them thicker by 1 voxel in all three directions to accommodate the spatial uncertainty in the probabilistic map. This procedure is nowadays already incorporated in PickAtlas through the 3D dilatation function.

Pure anatomical delimitation is less appropriate for high-level sensory regions, thus we used functional parcels obtained independently by other laboratories. As a high-level visual ROI, a functional parcel of the Lateral Occipital Complex (LOC) from the Kanwisher lab[34] was used, as it encompass functional specialized regions known to process visual features presented in our stimuli such as faces and places. The high-level auditory ROI was based upon the 'Temporal Voice Area'- TVA probabilistic map from Belin's lab[35] concerning more than 200 subjects, available at neurovault (http://neurovault.org/collections/33/).

These general masks were combined (by means of a conjunction analysis) with individual functional data that specify voxels modulated by our task: the F-contrast of all task trials against fixation trials, at a threshold of 0.0001 (uncorrected for multiple comparisons), using a separate 'neutral GLM' where all task trials were modeled as a single condition (fixation was implicitly modeled). ROIs with at least 20 active voxels were included. If a given participant ROI did not meet these criteria, his/her data was not used in the group analysis for this ROI. This situation only took place for two subjects in the anterior mPFC.

To ensure that no overlap occurs between ROIs, we visually inspected ROI borders of each ROI pair and restricted the ROIs to avoid the overlap. As a first measure, we restricted the TVA probabilistic map to the most significant voxels by imposing an arbitrary threshold of 20%, which restricted the ROI to the temporal cortex (where voice-sensitive voxels are more typically reported), and reduced considerably its overlap with other regions (e.g., TPJ). The resulting map was then transformed in a binary mask. As an additional measure, for this and all the other ROIs (all binary masks), we excluded the remaining overlapping voxels from the largest ROI of each pair. Only the following ROI intersections presented some overlap: EVC x LOC, EVC x PC, EAC x TVA, EAC x TPJ and TVA x TPJ (the first of each pair being the largest one). This procedure ensured a complete separation of the ROIs.

Given its role in emotional processing, we have also considered using an amygdala ROI. Yet, we are not confident that our standard imaging protocol at 3 T was sensitive to pick up multi-voxel patterns in amygdala, which was further confirmed by null results in other (here unreported) analyses focusing upon other dimensions such as the properties of the visual/auditory stimuli. The ROI did not show any significant representation of social congruence, but given our doubts about the data quality in this ROI we do not think we can interpret this result and we preferred to not include the ROI further.

Correlation-based multivoxel pattern analysis. We used correlation-based multivoxel pattern analysis (MVPA) to explore how the spatial response pattern in individual ROIs differs between experimental conditions[36]. Smoothed data were used here as spatial smoothing does not hurt MVPA and sometimes even improves it[31,37]. For each participant, we extracted the parameter estimates (betas) for each condition (relative to baseline) in each voxel of the ROIs. These obtained values for each run were then normalized by subtracting the mean response across all conditions (for each voxel and run separately), to emphasize the relative contribution of each condition beyond global activation level, as previously done in the literature[16,38]. The full dataset (6 runs) were randomly divided into two independent subsets of runs (using 'randperm' function in matlab). Thus, typically three runs were randomly assigned to set 1 and three other runs to the set 2 of the classification procedure. In the single case of incomplete data (5 runs instead of 6), only two runs were assigned as set 1. The multi-voxel patterns associated with each condition (congruent and incongruent) in set 1 (runs averaged) were pairwisely correlated with the activity patterns in set 2 (runs averaged) by using the 'corrcoef' function in matlab (Persons' r correlation coefficient). This procedure of splitting the data in two parts followed by correlating the multi-voxel patterns was repeated 100 times. The final 2×2 neural similarity matrix for each ROI was obtained by averaging these 100 matrices.

To test whether a certain region contained information about social congruency, we applied the following procedure. First, we calculated for each ROI, the mean correlations in the diagonal (correlation of the same condition across runs) and non-diagonal cells (correlation of different conditions across runs) of the neural similarity matrix of congruency (see inset in Fig. 3b). Then, we used a Fisher's transform (r to z) and performed a two sample two-tailed t-test across participants for diagonal vs non-diagonal mean correlations. This procedure is based on the fact that the same condition will typically show higher similarity across runs relative to different conditions, i.e., higher correlations for diagonal minus non-diagonal cells (one sample t-test)[39]. Lastly, a Bonferroni correction for multiple comparisons (i.e., the number of ROIs) was applied.

To investigate the potential relationship between brain information level of social congruency with behavioral 'performance' on the social norm inference task, i.e., the ability to match the response chosen by their peers, we performed a correlational analysis using a brain index (diagonal minus non-diagonal values) against a behavioral index (individual between/within subject correlations ratio; see Fig. 3c).

Searchlight MVPA analysis. Searchlight MVPA analysis was used as a complementary way to check for potential missing anatomical areas outside our a priori ROIs. We used the searchlight scripts of the cosmo MVPA toolbox[40] to search for local neighborhoods that revealed significant congruence representations (diagonal > non-diagonal), using the default parameters (e.g., spherical neighborhood of 100 voxels). The searchlight analysis was performed for each individual participant using cosmo MVPA toolbox, after which the individual subject maps were smoothed to 8 mm full-width at half-maximum and a 2nd level model was performed (both using SPM). One participant was excluded from this analysis because of a missing run (excluded for excessive movement).

Outside the scanner. Prior to scanning, participants performed valence rating judgments for the stimuli of each modality (visual and audio) separately, using a 9-level scale. This was used to familiarize participants with the stimuli set used in the fMRI runs and to create group averaged perceived valence models.

Additionally, participants also performed two runs of the main task (as inside the scanning, i.e., using an allocentric 'social norm' perspective to judge congruency) but instead of binary responses (congruent vs

incongruent) they used here a 9-level scale rating. One run was performed before and one after the fMRI scans. Beyond the familiarization with the task, this was helpful to check reproducibility of response patterns within and between-subjects, in a more fine-grained way.

A behavioral control experiment: egocentric perspective. To validate the sensitivity of our behavioral paradigm to the 'social norm' perspective taken during the main behavioral and neuroimaging experiments, we compare these results with a behavioral control experiment. The control experiment included the same task but instead of asking participants to judge 'what most people would answer' (allocentric perspective), they were instructed to answer as a function of their own perspective (egocentric perspective). Twenty other subjects (age: 19.9 +/− 2.02 yrs old) performed this task with a 9 level rating of congruence. The visual stimulus set was larger (double) but the analysis presented here was restricted to the exact same stimulus set used in the main fMRI experiment and in the task performed outside the scanner previously described. The auditory set was perfectly equivalent at the perceptual level, but slightly differed in terms of low-level auditory features such as duration and total RMS.

Data availability. The files needed to replicate the analyses are available on the Open Science Framework (e.g., ROI definitions and individual subject representational similarity matrices) (https://osf.io/t7xp9/). Other aspects of the data (raw data files, other steps in the analyses) are available from the corresponding author on reasonable request.

References

1. Vuilleumier, P. & Pourtois, G. Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia* **45**, 174–194 (2007).
2. Bestelmeyer, P. E., Maurage, P., Rouger, J., Latinus, M. & Belin, P. Adaptation to vocal expressions reveals multistep perception of auditory emotion. *J. Neurosci.* **34**, 8098–8105 (2014).
3. Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P. E. G. & Belin, P. Dorsal and Ventral Pathways for Prosody. *Curr. Biol.* **25**, 3079–3085 (2015).
4. Frith, C. D. & Frith, U. The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
5. Mitchell, J. P. Inferences about mental states. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 1309–1316 (2009).
6. Schurz, M., Radua, J., Aichhorn, M., Richlan, F. & Perner, J. Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **42**, 9–34 (2014).
7. Saxe, R., Moran, J. M., Scholz, J. & Gabrieli, J. Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Soc. Cogn. Affect. Neurosci.* **1**, 229–234 (2006).
8. Amodio, D. M. & Frith, C. D. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **7**, 268–277 (2006).
9. Schmidt, M. F. H., Butler, L. P., Heinz, J. & Tomasello, M. Young Children See a Single Action and Infer a Social Norm: Promiscuous Normativity in 3-Year-Olds. *Psychol. Sci.* **27**, 1360–1370 (2016).
10. Pegado, F., Vankrunkelsven, H., Steyaert, J., Boets, B. & Op de Beeck, H. Exploring the Use of Sensorial LTP/LTD-Like Stimulation to Modulate Human Performance for Complex Visual Stimuli. *PloS One* **11**, e0158312 (2016).
11. Locke, K. D. *et al*. Cross-Situational Self-Consistency in Nine Cultures: The Importance of Separating Influences of Social Norms and Distinctive Dispositions. *Pers. Soc. Psychol. Bull.* **43**, 1033–1049 (2017).
12. Mitchell, J. P., Macrae, C. N. & Banaji, M. R. Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron* **50**, 655–663 (2006).
13. Sul, S. *et al*. Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proc. Natl. Acad. Sci.* **112**, 7851–7856 (2015).
14. Denny, B. T., Kober, H., Wager, T. D. & Ochsner, K. N. A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *J. Cogn. Neurosci.* **24**, 1742–1752 (2012).
15. Botvinick, M. M., Cohen, J. D. & Carter, C. S. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* **8**, 539–546 (2004).
16. Bracci, S. & de Beeck, O. H. Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* **36**, 432–444 (2016).
17. Op de Beeck, H. P., Deutsch, J. A., Vanduffel, W., Kanwisher, N. G. & DiCarlo, J. J. A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. *Cereb. Cortex N. Y. N 1991* **18**, 1676–1694 (2008).
18. Bulthé, J., Hurk, J. van den, Daniels, N., Smedt, B. D. & Beeck, H. P. O. de. A validation of a multi-spatialscale method for multivariate pattern analysis. In *2014 International Workshop on Pattern Recognition in Neuroimaging* 1–4 https://doi.org/10.1109/PRNI.2014.6858513 (2014).
19. Wu, H., Luo, Y. & Feng, C. Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neurosci. Biobehav. Rev.* **71**, 101–111 (2016).
20. Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. Reinforcement Learning Signal Predicts Social Conformity. *Neuron* **61**, 140–151 (2009).
21. Izuma, K. The neural basis of social influence and attitude change. *Curr. Opin. Neurobiol.* **23**, 456–462 (2013).
22. Contreras, J. M., Schirmer, J., Banaji, M. R. & Mitchell, J. P. Common brain regions with distinct patterns of neural responses during mentalizing about groups and individuals. *J. Cogn. Neurosci.* **25**, 1406–1417 (2013).
23. Suzuki, S. *et al*. Learning to Simulate Others' Decisions. *Neuron* **74**, 1125–1137 (2012).
24. Nicolle, A. *et al*. An Agent Independent Axis for Executed and Modeled Choice in Medial Prefrontal Cortex. *Neuron* **75**, 1114–1121 (2012).
25. Mende-Siedlecki, P., Cai, Y. & Todorov, A. The neural dynamics of updating person impressions. *Soc. Cogn. Affect. Neurosci.* **8**, 623–631 (2013).
26. Ma, N. *et al*. Inconsistencies in spontaneous and intentional trait inferences. *Soc. Cogn. Affect. Neurosci.* **7**, 937–950 (2012).
27. Cloutier, J., Gabrieli, J. D. E., O'Young, D. & Ambady, N. An fMRI study of violations of social expectations: when people are not who we expect them to be. *NeuroImage* **57**, 583–588 (2011).
28. Lang, P., Bradley, M. & Cuthbert, B. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. (2008).
29. Pegado, F. *et al*. A Multitude of Neural Representations Behind Multisensory 'Social Norm' Processing. *Front. Hum. Neurosci.* **12**, 153 (2018).
30. Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci.* **107**, 2408–2412 (2010).

31. Op de Beeck, H. P. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses. *NeuroImage* **49**, 1943–1948 (2010).
32. Sallet, J. *et al*. The Organization of Dorsal Frontal Cortex in Humans and Macaques. *J. Neurosci.* **33**, 12255–12274 (2013).
33. Mars, R. B. *et al*. Diffusion-Weighted Imaging Tractography-Based Parcellation of the Human Parietal Cortex and Comparison with Human and Macaque Resting-State Functional Connectivity. *J. Neurosci.* **31**, 4087–4100 (2011).
34. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).
35. Pernet, C. R. *et al*. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *NeuroImage* **119**, 164–174 (2015).
36. Haxby, J. V. *et al*. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* **293**, 2425–2430 (2001).
37. Hendriks, M. H. A., Daniels, N., Pegado, F. & Op de Beeck, H. P. The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Front. Neurol.* **8**, 222 (2017).
38. Op de Beeck, H. P., Torfs, K. & Wagemans, J. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci. Off. J. Soc. Neurosci.* **28**, 10111–10123 (2008).
39. Ritchie, J. B., Bracci, S. & Op de Beeck, H. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage* **148**, 197–200 (2017).
40. Oosterhof, N. N., Connolly, A. C. & Haxby, J. V. CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Front. Neuroinformatics* **10** (2016).

Acknowledgements

Stefania Bracci and Jean Steyaert for helpful discussions, Anne de Vries and Ninke De Schutter for their help in behavioral data acquisition and The Department of Radiology of the University Hospital in Leuven for their support. F.P. was funded by FWO (Fonds Wetenschappelijk Onderzoek) postdoc fellowship (12Q4615N) and research grant (1528216N). H.O.B. was supported by ERC-2011-Stg-284101 and IUAP P7/11. B.B. and H.O.B. were supported by Interdisciplinary Research Fund (IDO/10/003). H.O.B., F.P., B.B. were supported by FWO research grant (G088216N). J.B. was funded by FWO PhD fellowship (11J2115N). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

F.P., B.B. and H.O.B. design; F.P., M.H., S.A., N.D. performed, F.P., J.B., H.L.M., H.O.B. analyzed and F.P., J.B., B.B., H.O.B. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-31260-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.