Neural correlates of the motivation to be moral

Nunspeet, F. van

Chapter 3

Emphasizing moral task implications influences visual attention: An fMRI study

Collaborators on the research described in this chapter are:
Naomi Ellemers, Anna van Duijvenvoorde, Belle Derks, Eveline Crone, and Serge Rombouts. Their contribution can be specified as follows: Design of the study: FvN, NE, EC. Performing the experiment: FvN. Data analyses: FvN, AvD. Additional data processing: EC, SR. Writing of the paper: FvN, NE, AvD, BD.
Manuscript in preparation.

Morality, having the knowledge of and behaving according to what is right and wrong, is seen as key to human social life: It is one of the hallmarks of society since it is the basis for people's individual choices, their social interactions and group functioning. To gain understanding of how (im)moral behavior is initiated, much research has focused on the development of people's individual level of rational decision making: Knowing what is and what is not moral and considering what would be the best thing to do in particular situations. Using neuroscientific research methods, researchers have been able to reveal the neural networks involved in such moral cognition. Specifically, participants in those studies are often asked to take the observers perspective and judge the (im)moral content of phrases or pictures (e.g., Cope et al., 2010), to decide on different moral dilemmas (e.g., Christensen & Gomila, 2012), or to imagine behaving in line with or opposed to moral norms (e.g., Decety & Porges, 2011). However, as Casebeer (2003) noted, thinking about (doing) moral things is different from actually doing moral things –and imagined compared to real moral decision making is even associated with different neural networks (FeldmanHall et al., 2012). In the current research, we therefore aimed to extend previous research on moral cognition by examining how people's motivation to behave morally affects their actual performance on a task said to be indicative of their moral values. Moreover, we investigated how such moral motivation affects the cognitive processes involved in this task performance.

In neuroscientific research on moral psychology the social significance of morality is often underemphasized or even excluded (Casebeer, 2003). Nevertheless, moral choices and behaviors are inherently social: They often imply taking care of others or treating others well. In fact, some analyses consider morality and sociability as representing one evaluative domain, although they encompass different characteristics and behaviors (Leach, Ellemers, & Barreto, 2007). Indeed, judging other people's moral integrity and trustworthiness is important in social interactions (e.g., Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Delgado, Frank, Phelps, 2005). Moreover, being perceived as a moral person is important for one's social identity: People experience pride in being a member of a group with high morality (Leach et al., 2007), and they indicate a willingness to adhere to moral group norms (Ellemers, Pagliaro, & Barreto, 2008) because they expect to receive respect from their fellow group members in this way (Pagliaro, Ellemers, & Barreto, 2011). Being moral thus encompasses more than intrapsychological processes associated with cold moral reasoning. Even when we know the moral thing to do and certain brain regions may be associated with coming to that decision, it is not self-evident that such mechanisms are also associated with actual behavior. It is therefore important, in addition to the investigation of moral cognition, to increase our understanding of the neural processes involved in the motivation to display moral behavior.

Prejudice Control as an Indicator of Moral Behavior

As mentioned above, being moral often has social implications: Defining what is right or wrong may depend on what others value as the moral thing to do and on how others are affected by our actions. In the current research, we examine moral behavior in the context of intergroup relations and prejudice: Fairness towards and the equal treatment of different groups in society are seen as core moral values. There thus tends to be a general motivation to be or to appear unprejudiced. Because of those moral and social implications, prejudice is often measured on an implicit level, for instance with an implicit association test (IAT). The IAT (Greenwald, McGhee, & Schwartz, 1998) was first designed to assess people's positive versus negative associations with particular social groups: Their implicit social bias. Stimuli in this reaction time test consist of target concepts –representing members of social groups, such as faces of Black and White men, or Muslim and non-Muslim women– and positive and negative attributes. On prejudice-congruent trials, participants are asked to categorize the stimuli representing their own (in-)group using the same response key as positive attributes, and stimuli representing another (out-)group and negative attributes with another key. On prejudice-incongruent trials they are asked to categorize stimuli representing their ingroup and negative attributes with the same key, as well as stimuli representing the outgroup and positive stimuli. To the extent that people are more inclined to associate their ingroup with positivity and the outgroup with negativity, they should respond more quickly and easily to the congruent as compared to the incongruent trials. The IAT assesses this difference in response latencies on incongruent compared to congruent trials, as an indicator of implicit bias.

Recent research has revealed that people are able to influence their performance on an IAT if they are motivated to do so. For instance, Fiedler and Bluemke (2005) have shown that participants can reduce their negative bias when they are aware of how the IAT bias is computed and when they are encouraged to find out effective strategies to adjust their performance. Moreover, Van Nunspeet, Ellemers, Derks, and Nieuwenhuis (2014) showed that people's motivation to control prejudice was higher when the moral implications of the IAT were emphasized, resulting in a weaker bias against Muslim women.

In the current research, we will adopt the same paradigm as used by Van Nunspeet et al. (2014) to create circumstances that amplify the motivation to behave morally (i.e., to control expressions of implicit bias). In addition, we use functional magnetic resonance imaging to examine the neural processes underlying such moral behavior.

Neural Correlates of Social Bias Control

In their study, Van Nunspeet et al. (2014) measured brain activation with an electroencephalogram (EEG) to examine the cognitive processes underlying moral IAT performance. Their results revealed that participants who had been reminded of the moral implications of the IAT (as compared to a control condition) showed increased perceptual attention to the different types of targets in the IAT (both in terms of group membership and individuating facial features, as indicated by increased N1 and P150 modulations in response to viewing the pictures of outgroup and ingroup targets; Van Nunspeet et al., 2014). The present study aims to further examine these processes using fMRI, by examining how the motivation to perform in line with one's moral values affects patterns of brain activation associated with performance on an IAT.

In fMRI research, face perception is often located in the inferior part of the occipital lobe. More specifically, within the inferior occipital gyrus (also called the occipital face area, OFA; for a review see Pitcher, Walsh, & Duchaine, 2011) and the fusiform gyrus (FG; but see Haxby, Hoffamn, & Gobbini, 2000; and Ishai, 2008, for more complete overviews of the cortical network involved in face processing). Activation in the OFA is associated with facial recognition (i.e., the establishment that a face is a face), which occurs at an early stage of visual perception (Pitcher et al., 2011). In contrast, activation in the FG is associated with the subsequent and deeper processing of higher-level facial features. For example, activation in the FG is greater when people view ingroup compared to outgroup members (e.g., Kubota, Banaji, & Phelps, 2012; Van Bavel, Packer, & Cunningham, 2011). Since previous research revealed that when the moral implications of the IAT are emphasized, perceptual attention towards and social categorization of ingroup and outgroup faces is increased (Van Nunspeet et al., 2014), we hypothesized that participants who performed the moral (as compared to the control) IAT in the current research would show increased activation in the FG when viewing ingroup as compared to outgroup targets. Moreover, since the process of social categorization was found in early event-related brain potentials (i.e., around 100 and 150ms after stimulus-onset), we also wanted to examine whether we could find any evidence for increased social categorization of ingroup and outgroup targets in the OFA given its association with early facial processing.

Another finding in the EEG study was that the inhibition of social bias on the IAT was associated with increased modulations of the error-related negativity (ERN; Van Nunspeet et al., 2014). The ERN is associated with response-monitoring (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993; Nieuwenhuis, Blom, Band, & Kok, 2001) and the significance of making errors (Hajcak, Moser, Yeung, & Simons, 2005). Specifically, results of the research of Van Nunspeet et al. (2014) showed that conflict- and error-monitoring was enhanced for participants to whom the moral implications of the IAT were emphasized, indicating the increased significance of making errors on a task indicative of their moral values.

The conflict- and response monitoring processes found by Van Nunspeet et al. (2014) are in line with patterns of brain activation found in fMRI studies on social bias: In studies using the IAT, brain activation associated with performance on incongruent IAT trials is contrasted to brain activation associated with performance on congruent IAT trials. Results reveal that performance on the incongruent compared to congruent trials is associated with increased activation in the dorsolateral prefrontal cortex (dlPFC) and anterior cingulate cortex (ACC; e.g., Chee, Sriram, Soon, & Lee, 2000; Stanley, Phelps, & Banaji, 2008). These brain regions are known to be involved in conflict monitoring and control of behavior

(Botvinick, Braver, Barch, Carter, & Cohen, 2001; MacDonald, Cohen, Stenger, & Carter, 2000) and specifically, in the area of prejudice, considered as regulating (implicit) bias (Stanley et al., 2008). In the current research, we aim to extend the findings of Van Nunspeet et al. (2014), namely that participants tend to be highly vigilant while performing an IAT framed as a measure of their morality –as compared to a control condition in which the IAT is framed as a measure of their competence. Accordingly, we examined whether the motivation to perform in line with moral values affected brain activation in regions associated with cognitive control when participants perform incongruent versus congruent IAT trials.

Triangulation

Whereas cognitive processes associated with people's concerns to behave in line with their moral (e.g., egalitarian) values have been revealed in previous EEG research, our current goal is to expand these insights using fMRI. Both methodologies have their advantages: EEG has a high temporal resolution, making it is possible to examine the onset and time course of different cognitive processes, including very early and immediate responses. In addition, fMRI has a high spatial resolution which gives us the opportunity to locate the brain regions involved in moral task performance. Thus, whereas previous research has revealed that perceptual attention to different types of faces is increased (as seen in the N1 and P150 potentials, measured at the frontocentral sites of the scalp; Van Nunspeet et al., 2014), the current research will examine whether this is also evident in patterns of brain activation in the visual cortex. Moreover, we can investigate whether the enhanced error-detection and conflict-monitoring processes found in EEG research are also evident in brain areas associated with cognitive control. By using such a triangular approach (i.e., combining insights from behavioral, EEG and fMRI research) we will get a better understanding of people's motivation to *be* moral, in addition to our knowledge of brain processes and networks involved and necessary for moral cognition.

Study 3

Method

Participants

Twenty-six, non-Muslim, right-handed students from Leiden University participated in the study. None of the participants reported a history of psychiatric or neurological disorders, or current use of any medications. One participant was excluded from the data because of a software failure during the scanning session; two other participants were excluded from the fMRI data analyses because of technical problems. The remaining twenty-three participants (8 males, M_{age} = 21.0 years, *SD* = 4.9) were randomly divided across the two conditions of the between-participants design (i.e., the morality [*N* = 11] or control [*N* = 12] task domain). All procedures were approved by the medical ethical committee of the Leiden University Medical Center (LUMC) and all participants gave informed consent for the study.

Morality Framing of the IAT

While in the scanner, before the start of the IAT, half of the participants read that the computer task they were going to perform could indicate their endorsement of *moral values* concerning egalitarianism and discrimination (the morality condition). The other half of the participants was informed that the test could indicate their *ability* to process new information and to learn new tasks (the control condition). All participants were instructed to respond as quickly and accurately as possible. They were reminded about the test implications before the start of each test run (i.e., runs 3 and 5; see also Van Nunspeet et al., 2014).

Instruments and Procedure

Participants performed the five steps (blocks) of the IAT as designed by Greenwald et al., (1998). We used an event-related block design: Each scanning run consisted of one IAT block, but each stimulus was preceded by a fixation cross creating a jittered interstimulus interval (min. = 1100 ms, max. = 6600 ms) in order to model the hemodynamic response function for each stimulus type. After the fixation cross, the stimulus was presented (with a maximum duration of 680 ms in which participants were asked to respond [e.g., Beer et al., 2008]; more details about the stimuli below) which was followed by a feedback screen (1650 ms minus the reaction time on the stimulus; see Figure 3.1). The feedback screen consisted of a green checkmark (i.e., correct response), a red cross (i.e., incorrect response), or the words "too late" when the participant did not press a key on time. Only trials on which participants responded correctly and on time were analyzed.

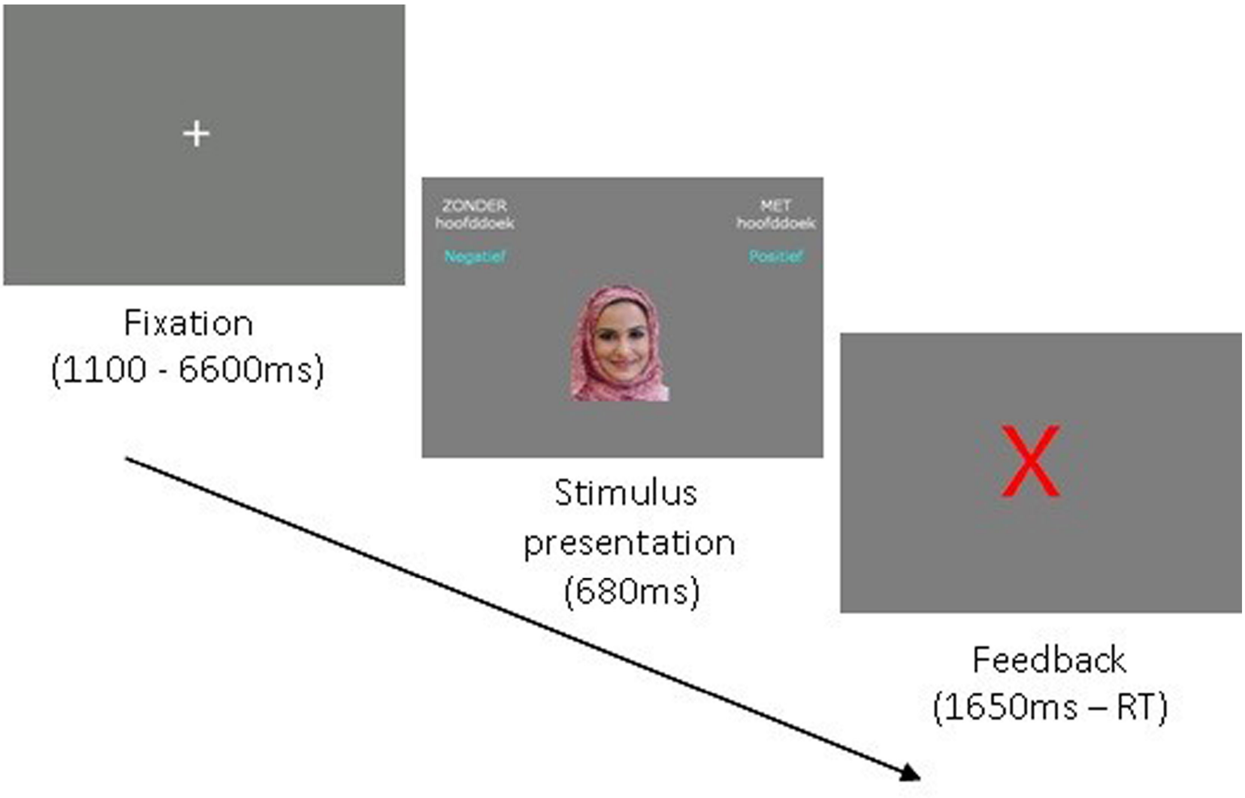

Figure 3.1. Example of an IAT trial. RT = reaction time.

In run 1, stimuli consisted of 10 faces of Muslim women (wearing a headscarf; outgroup pictures) and 10 faces of non-Muslim women (not wearing a headscarf; ingroup pictures) which participants were asked to categorize using a right (index finger) or left (index finger) response key. In run two, stimuli consisted of 5 pictures of positive scenes, and 5 pictures of negative scenes (International Affective Picture System; Lang et al., 2005). In run three, both picture types were presented and participants responded either with one key to outgroup pictures and negative scenes and with the other key to ingroup pictures and positive scenes (i.e., congruent trials). Or they responded with one key to outgroup pictures and positive scenes and with the other key to ingroup pictures and negative scenes (i.e., incongruent trials). Run four was similar to run one except for the fact that the response keys for the ingroup and outgroup pictures were switched. Finally, run five was similar to run three: Both ingroup/outgroup pictures and pictures of positive/negative scenes were presented. However, when congruent trials (i.e., 'ingroup + positivity' and 'outgroup + negativity) were presented in run3, then run 5 consisted of incongruent trials (i.e., 'outgroup + positivity' and 'ingroup + negativity'), and vice versa. The order of the runs was thus counterbalanced between participants. Training runs 1, 2, and 4 consisted of 20 trials each and lasted approximately two minutes. Testing runs 3 and 5 consisted of 120 trials each and lasted approximately six minutes. All IAT instructions were presented on the screen in the scanner bore before the start of each run. Since the experiment was part of a larger study, participants spent approximately 2 hours in the laboratory, and received 20 euros as a compensation for their participation.

fMRI Data Acquisition and Analysis

Scanning was performed at the Leiden University Medical Centre (LUMC) with a standard whole-head coil on a 3.0 Tesla Philips Achieva scanner. Using E-prime 1.0 software, the IAT was projected onto a screen at the back of the scanner bore, which participants could view via a window attached to the top of the head coil. Participants could respond by pressing keys on boxes attached to their legs. The IAT consisted of five event-related runs, of which we only analyzed test runs 3 and 5 (consisting of congruent and incongruent trials). Functional data were obtained using T2*-weighted echo-planar imaging ([EPI], repetition time (TR) = 2200 ms, echo time (TE) = 30 ms, slice matrix = 80 x 80, slice thickness = 2.75 mm, slice gap = 0.28 mm, field of view [FOV] = 220 mm). A high-resolution 3D T1-weighted anatomical image (TR = 9.751 ms, TE = 4.59 ms, flip angle = 8°, 140 slices, 0.875 mm x 0.875 mm x 1.2 mm, and FOV = 224.000 x 168.000 x 177.333) was collected at the end of the scanning session.

Data were preprocessed and analyzed using SPM8 software (Welcome Department of Cognitive Neurology, London) implemented in MATLAB (Mathworks, Sherborn, MA). The functional time series were realigned to compensate for small head movements. Translational movement parameters never exceeded 1 voxel (< 3 mm) in any direction for any subject or scan. Functional volumes were spatially normalized to EPI templates. The normalization algorithm used a 12 parameter affine transformation together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 3 mm cubic voxels. Functional volumes were spatially smoothed using an 8 mm full-width half-maximum Gaussian kernel. Templates were based on the MNI305 stereotaxic space (Cocosco, Kollokian, Kwan, Pike, & Evans, 1997), and the Montreal Neurological Institute (MNI) atlas was used to refer to the coordinates.

To analyze the data, a canonical hemodynamic response function was convolved at the onset of the stimulus and modeled as a zero-duration event. We distinguished between eight different types of stimuli: The IAT consisted of pictures of ingroup targets, outgroup targets, positive scenes, and negative scenes. Moreover, these stimuli were presented in a congruent as well as an incongruent run. Invalid trials were included in the model as a regressor of no interest. Two main contrast analyses were distinguished: To examine brain activation associated with visual perception of ingroup and outgroup targets, we investigated a contrast of viewing faces of non-Muslim women versus Muslim women, collapsed over the two IAT test runs (i.e., congruent/incongruent ingroup targets and congruent/incongruent outgroup targets, measured *within* participants). Moreover, to test whether this activation interacted with the emphasis on the moral implications of the task (measured *between* participants), we conducted a 2 (ingroup/outgroup targets) X 2 (morality/control) full factorial ANOVA.

To examine activity associated with the IAT effect (measured *within* participants), brain activation for the incongruent IAT run (for all ingroup / outgroup / positive / negative pictures) was compared to brain activation during the congruent IAT run (also for all ingroup / outgroup / positive / negative pictures). Moreover, to test whether the activation associated with the IAT effect interacted with the emphasis on the moral compared to the competence implications of the task (measured *between* participants), we conducted a 2 (incongruent/congruent) X 2 (morality / control) full factorial ANOVA.

The analyses were carried out using the general linear model in SPM8. For each individual, contrast parameter images were computed and the resulting contrast images were submitted to second-level group analyses. Only effects of at least 10 continuous voxels that exceeded a False Discovery Rate (FDR) corrected threshold of $p < .05$ are reported.

Moreover, since we were interested in the –perhaps quite subtle– difference between the emphasis on the moral compared to the competence implications of the task, we also extracted parameter estimates from the regions of interest (ROI) that were identified in the whole brain analyses to explore the pattern of the activation across our conditions. These regions were extracted using the Marsbar toolbox (Brett, Anton, Valabregue, & Poline, 2002) for SPM8.

Results

Behavioral Results

The IAT effect is indicated by the *D* score, and measured as the difference in reaction times on incongruent and congruent trials divided by a pooled *SD* of all correct trials (Greenwald, Nosek, & Banaji, 2003). We included all trials, replaced error latencies with a replacement value (M + 2 SD_{correct}; Greenwald et al., 2003) and replaced latencies exceeding the maximum response time with the maximum response time of 680 ms. The resulting positive *D* scores are an indication of people's evaluative bias against the outgroup (i.e., Muslim women).

To test whether participants showed an IAT effect overall, we conducted a one-sample t-test with *D* score as the dependent variable and a comparison test score of zero. As expected, results revealed the standard IAT effect; $M = 0.18$, $SD = 0.33$, $t(24) = 2.66$, $p = .01$, indicating that participants showed bias against Muslim women. Subsequently, we tested whether the task domain manipulation influenced the IAT effect. Specifically, whether emphasizing the moral implications of the IAT caused participants to show a smaller bias against Muslims. However, contrary to our hypothesis, an ANOVA with *D* score as dependent variable and task domain and the order of IAT test blocks as independent factors showed no main effect of task domain, nor an interaction effect between task domain and order; *F*'s < 1. There was only a main effect of order, F(1,21) = 9.52, $p = .006$, $\eta_p^2 = .31$, indicating that participants who performed the congruent block first showed a smaller bias against Muslim women ($M = 0.01$, $SD = 0.22$) than participants who performed the incongruent block first ($M = 0.36$, $SD = 0.34$). Perhaps, this effect is due to starting the task with the relatively difficult trials which could increase response latencies and thus the difference between responses on incongruent and congruent trials.

Even though we observed no differences at the overt behavioral response level, it is still of interest to see whether different brain areas are involved in displaying these responses dependent on experimental conditions.

Imaging Results

Face perception.

To examine the neural activation associated with viewing faces of outgroup members (Muslim women) and ingroup members (non-Muslim women), we first conducted a 2 (Target identity: ingroup/outgroup faces) x 2 (Task Domain: morality/control) full factorial ANOVA at the whole brain level. Results revealed no main effects, nor an interaction. One-sample t-tests –averaged across the task domain conditions– showed no significant patterns of activation for the outgroup > ingroup targets contrast. However, as expected, the ingroup > outgroup targets contrast showed a significant difference in activation in the bilateral fusiform gyrus (see Table 3.1 and Figure 3.2A), indicating that –in line with previous research (e.g., Kubota et al., 2012; Van Bavel et al., 2011)– activation was greater when participants viewed faces of ingroup members (non-Muslim women) as compared to faces of outgroup members (Muslim women).

In addition to the whole-brain analyses, we extracted parameter estimates from the regions of interest (ROIs) that were identified in the whole brain analyses to further examine the patterns of activity between participants in the morality and the control condition. Specifically, we localized ROIs in two area's in the visual cortex known to be associated with processing faces: The fusiform gyrus (FG, Brodmann area 37) and the occipital face area (OFC, Brodmann area 19). Both ROIs were based on the contrast of 'All faces' (i.e., congruent/incongruent ingroup targets and congruent/incongruent outgroup targets) > 'fixation' (FDR corrected p < .05, 10 continuous voxels). Within this contrast, we located the FG and OFA bilaterally and the peaks of the activation (MNI coordinates FG: +39, -49, -26 and -36, -43, -26; MNI coordinates OFA: -36, -67, -20 and +42, -64, -23) defined the centers of four 10 mm diameter sphere-shaped ROIs (Figure 3.2; see Ratner, Kaul, & Van Bavel, 2013, for a similar approach). Parameter estimates from these ROIs were included as the dependent variable in a 2 (Hemisphere: left/right) x 2 (Target identity: ingroup/outgroup faces) x 2 (Congruency: congruent/incongruent) repeated measures ANOVA with Task Domain (morality/control) and order of the IAT test blocks (congruent/incongruent first) as independent factors. Relevant to our interest in face perception, we did not find a main effect of, nor any interaction effects with task domain in the FG. Consistent with the whole-brain analysis, only the effect of target identity was significant indicating that activation in the FG was greater for viewing ingroup compared to outgroup faces, $F(1,19) = 7.49, p = .01, \eta_p^2 = .28$.

Results concerning face perception in the OFA also showed the main effect of target identity, $F(1,19) = 7.41, p = .01, \eta_p^2 = .28$, indicating that activation was greater for viewing ingroup as compared to outgroup faces. There was also an interaction effect between congruency and order, $F(1,19) = 4.29, p = .05, \eta_p^2 = .18$: For congruent trials, activation in the OFA was greater when the congruent (rather than the incongruent) run was presented first, $F(1,19) = 10.41, p = .004, \eta_p^2 = .35$. The other simple main effects were not significant, F's ≤ 2.39, p's $\geq .14$. Additionally and more interestingly, we observed a marginally significant interaction effect between target identity, congruency, and task domain, $F(1,19) = 3.56, p = .07, \eta_p^2 = .16$. To interpret this complex interaction, we conducted separate analyses for the control and morality conditions separately. Results revealed that there were no main effects of target identity or congruency, nor an interaction effect in the control condition; F's ≤ 2.39, p's $\geq .15$. However, in the morality condition, there was a significant main effect of target identity; $F(1,11) = 13.90, p = .003, \eta_p^2 = .56$, indicating greater activation for ingroup compared to outgroup targets. There was also a marginally significant main effect of congruency; $F(1,11) = 4.22, p = .07, \eta_p^2 = .28$, showing that activation in the OFA was greater on congruent compared to incongruent trials (see Figure 3.3). This findings is consistent with previous research, showing that participants adjusted their behavioral responses on prejudice-congruent trials when the moral implications of the IAT were emphasized (Van Nunspeet et al., 2014). There was no interaction effect; F's ≤ 1.62, p's $\geq .23$.

IAT effect. To examine the neural correlates of the IAT effect (i.e., bias against Muslim women), we examined neural activation associated with participants' performance on congruent versus incongruent trials. We first conducted a 2 (Congruency: incongruent/congruent) x 2 (Task Domain: morality/control) full factorial ANOVA at the whole brain level. Results revealed no significant main effects nor an interaction effect. Additionally, we conducted one-sample t-tests, averaged across the task domain conditions. Whole-brain contrasts (congruent > incongruent and incongruent > congruent) revealed no significant differences.

In addition to the whole-brain analyses, we again identified ROIs to further examine the patterns of activity between participants in the morality and the control conditions. Specifically, we localized ROIs based on activation in the contrast 'All stimuli' (i.e., congruent/incongruent ingroup targets; congruent/incongruent outgroup targets; congruent/incongruent positive scenes; and congruent/incongruent positive scenes) > 'fixation' (FDR corrected $p < .05$, 10 continuous voxels). Results of this contrast did not reveal activation in the ACC. However, we were able to detect activation within the right DLPFC and the peak of this activation (MNI coordinates: +45, +32, +28) defined the center of a 10 mm diameter sphere-shaped ROI. Parameter estimates from this ROI were included as the dependent variable in a repeated measures ANOVAs with congruency (i.e., congruent/incongruent trials, averaged over picture type) as within-participants factor and Task Domain (morality/control) and order of the IAT test blocks (congruent/incongruent first) as independent factors. Results of this analysis showed however no effects of congruency, task domain, or an interaction effect, F's ≤ 2.87, p's ≥ .11.

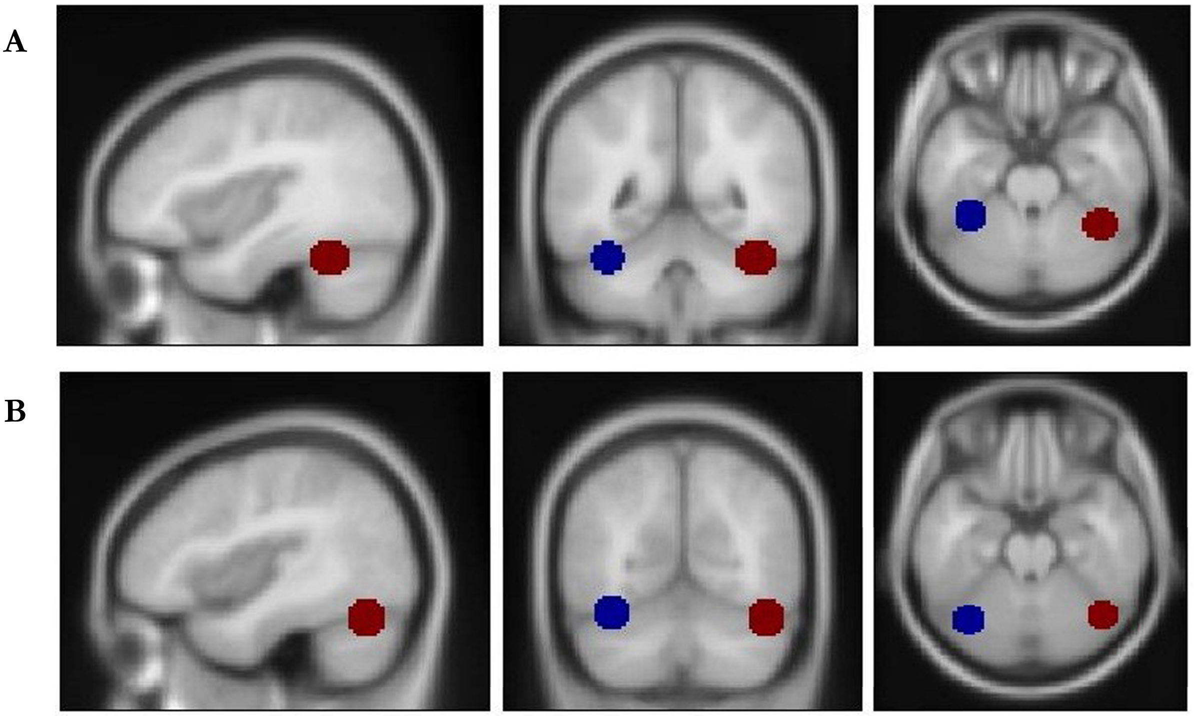

Figure 3.2. Activation was found in the bilateral fusiform gyrus (FG; Brodmann area 37) and occipital face area (OFA; Brodmann area 19) in the *faces* > *fixation* contrast (FDR corrected p < .01, 20 continuous voxels). Spheres were built around peak voxels at X = +39, Y = -49, Z = -26 and X = -36, Y = -43, Z = -26 for the FG (Figure A). And around peak voxels at X = -36, Y = -67, Z = -20 and X = +42, Y = -64, Z = -23 for the OFA (Figure B).

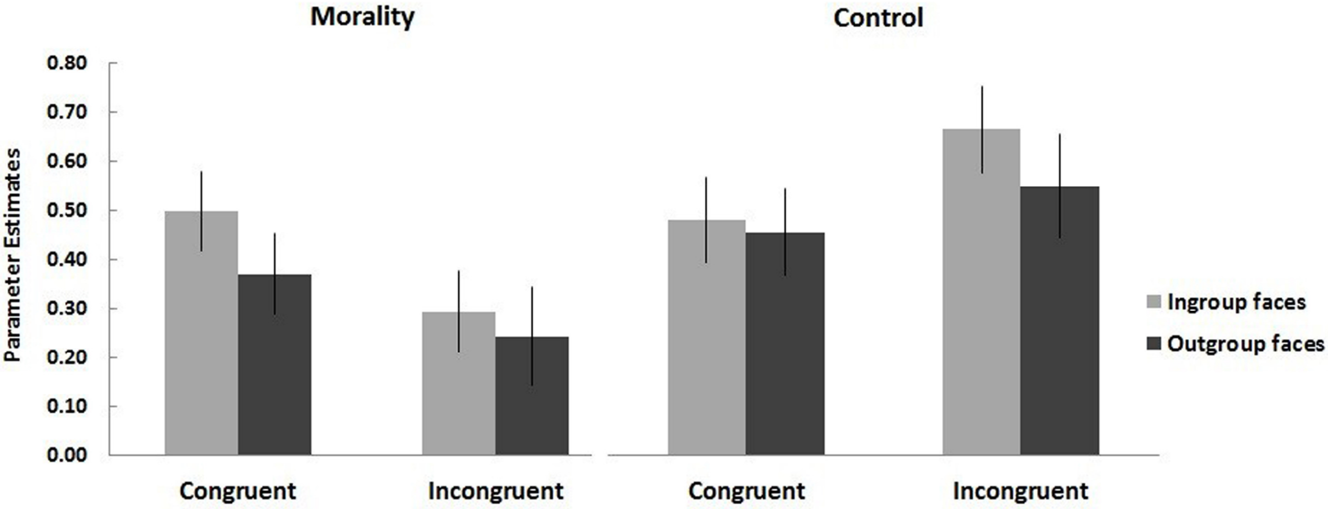

Figure 3.3. Within the ROI of the occipital face area (OFA) there was a significant interaction between target identity, congruency and task domain. Within the morality condition, activation was greater for viewing ingroup compared to outgroup faces and on congruent compared to incongruent trials.

Discussion

The goal of the current study was to examine the neural correlates of the motivation to behave in line with moral values. Complementing research on moral cognition –examining the cognitive processes involved in *thinking* about morality (i.e., moral reasoning and decision making) – we investigated the neural underpinnings of people's *behavior* on a task said the be indicative of their moral values. Specifically, we tested whether and how emphasizing the moral (compared to competence) implications of an implicit association test (IAT) would cause participants to inhibit their evaluative bias against Muslims. We used fMRI to study whether such an emphasis affects activation in brain areas implemented in visual attention towards facial stimuli and cognitive control – which would complement recent EEG research revealing enhanced perceptual attention and response-monitoring when the moral implications of an IAT are made salient (Van Nunspeet et al., 2014).

Our results revealed that visual attention towards the different facial stimuli in the IAT was dependent upon the emphasis on participants' morality: Participants in this condition showed greater activation in the occipital face area (OFA) when viewing ingroup compared to outgroup targets. Additionally, OFA activation was somewhat increased on congruent as compared to incongruent trials. These findings are consistent with the expectation that emphasizing the moral implications of the IAT affects participants' focus towards stimuli and perhaps their approach towards the task. As was shown by Van Nunspeet et al. (2014), participants who had read the moral test implications inhibited their responses on congruent trials. These are trials in which the easy (automatic) associations between Muslims and negative attributes and non-Muslims and positive attributes become evident.

To inhibit these prepotent responses, participants may need to be even more focused on the facial stimuli in the congruent compared to the incongruent trials to respond in line with their moral values. Furthermore, the fact that there was only a (marginally significant) effect of emphasizing moral concerns on initial (i.e., early) visual attention to faces and not on later and deeper facial processing in the fusiform gyrus, is in line with research showing that stressing moral test implications is associated with increased social categorization on very early event-related brain potentials (i.e., the N1 and P150, occurring around 100 and 150ms after stimulus-onset; Van Nunspeet et al., 2014).

Table 3.1

Brain regions revealed by the Ingroup > Outgroup targets contrast.

Anatomical Region	L/R	voxels	Z	MNI coordinates		
				x	y	z
Fusiform Gyrus	R	3161	4.71	39	-49	-23
			4.70	36	-43	-23
			4.17	33	-61	-20
	L	3161	4.28	-36	-49	-20
			4.07	-33	-73	-11
			3.96	-36	-64	-14
(anterior) Medial Cingulate Cortex	R	486	4.42	9	5	34
			4.17	12	-7	52
			3.80	30	-25	46
Supramarginal Gyrus	L	160	4.76	-45	-1	10
			3.31	-33	8	19
			3.26	-57	2	7
Temporal Parietal Junction	L	141	4.39	-54	-25	22
			3.56	-42	-22	16

MNI coordinates for main effects, peak voxels reported with an FDR-corrected threshold of *p* < .05, with an extent threshold of 10 continuous voxels (voxels size was 3.0 x 3.0 x 3.0 mm).

Although visual attention was affected by the emphasis on the moral implications of the task, this did not affect participants' bias against Muslim women. This is different from previous studies (e.g., Van Nunspeet et al., 2014; Van Nunspeet et al., *under review*) in which participants who had read the moral test implications showed a smaller bias than participants who had read the implications concerning their competence. There it was argued that the emphasis on morality caused participants to inhibit their prepotent (automatic) prejudiced responses which resulted in the increased response times on congruent trials. However, compared to the study of Van Nunspeet et al. (2014) in which the interstimulus interval (ISI) lasted for just 500 milliseconds, the duration of the current ISI was around two seconds. The inhibition of prepotent prejudiced responses may thus have occurred previous to stimulus onset or may have been undermined since participants had the time to prepare their response on the upcoming trial. In other words, the task could have become too easy to reveal implicit bias. Indeed, the amount of errors in the current research (4.5%) was only half of the error rates (8.3%) in research of Van Nunspeet et al. (2014). This explanation could also account for why we did not find greater activation in the neural regions associated with the regulation of implicit bias: The relatively easy IAT may have prevented participant from worrying about their performance. This is in line with research of Bengtsson, Lau, and Passingham (2009) who asked participants to perform either a significant (i.e., assessing their intelligence) or an insignificant (pilot test) experimental task. Their results revealed no differences in neural activation in prefrontal areas between the different types of tasks for correct responses. However, they did find that participants who performed the significant (compared to the insignificant) task showed increased neural activity on errors (Bengtsson et al., 2009). This is somewhat related to the study of van Nunspeet et al. (2014) in which it was shown that participants showed increased error-monitoring (i.e., greater error-related negativity modulations to *in*correct responses) when the moral (compared to the competence) implications of the IAT were emphasized. It is therefore possible that the difference between the motivation to perform in line with moral values as compared to one's competence is more evident on incorrect than correct responses. (Artificially) increasing the amount of errors during such an IAT and analyzing these events may thus reveal the differential neural activation we were aiming to find in the current research. Moreover, instructing participants to "clear their minds" when they see the fixation point may be crucial to overcome the effects of the increased ISIs (as was done in research by Beer et al., 2008). That is, to prevent participants to prepare their response on the upcoming trial.

Although we did not find an effect of the task instruction manipulation on the behavioral results, participants did show the typical IAT effect: A negative bias against Muslims. They responded more slowly on incongruent as compared to congruent IAT trials, indicating that associating outgroup members with positivity and ingroup members with negativity was more difficult for them than associating outgroup members with negativity and ingroup members with positivity. However, the expected neural activation in regions associated with cognitive conflict and control –the ACC and DLPFC– was not evident for the incongruent > congruent contrast. It should be noted that not all fMRI studies that used an IAT have found these activation patterns. For example, Knutson et al. (2007) neither showed significant patterns of activation for incongruent compared to congruent trials when analyzing their *single* IATs (i.e., a gender and race IAT separately). Nevertheless, another experimental design (for example, in which IAT test blocks are presented repeatedly, alternating between blocks of congruent and incongruent trials) may have improved the BOLD response supposedly associated with the task demands.

Another aim of the current research was to extend previous behavioral and EEG research on the motivation to display moral behavior, by adding insights from an fMRI study revealing the particular brain areas involved in that motivation. Unfortunately, we were unable to expand current insights concerning increased cognitive control in case of an emphasis on moral concerns. This may have been due to the restrictions of the current research design mentioned previously (i.e., increased ISIs needed in an event-related fMRI experiment, and analyzing only correct responses since errors were too scarce), and illustrates the difficulty of optimizing an experimental paradigm for different scientific research methods (see also Scheepers, Ellemers, & Derks, 2013). On the other hand, we did find some additional support for increased visual attention towards targets when an IAT is presented as a measure of individual morality. And it may be the combination of such findings from different scientific research methods that can strengthen our knowledge of the underlying cognitive and neural mechanisms of moral motivation.

Conclusion

The current research revealed that when the moral implications of an IAT are emphasized, participants show greater activation in the occipital face area when they view pictures of ingroup compared to outgroup targets. Moreover, activation in this region was greater on (prejudice-) congruent compared to incongruent trials. In addition to previous research, these findings may suggest that especially people's (visual) attention to a task increases once they have an opportunity to show their moral side.

Acknowledgements

We thank Gert-Jan Lelieveld for his help with the data collection and Jay van Bavel for helpful discussions about data analyses.