OXFORD

Social Cognitive and Affective Neuroscience, 2019, 435–445

doi: 10.1093/scan/nsz016
Advance Access Publication Date: 11 March 2019
Original article

Moral identity relates to the neural processing of third-party moral behavior

Carolina Pletti,[1] Jean Decety,[2] and Markus Paulus[1]

[1]Developmental Psychology Unit, Department of Psychology, Ludwig Maximilian University Munich, 80539, Munich, Germany and [2]Department of Psychology, and Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL 60637, USA

Correspondence should be addressed to Carolina Pletti, Developmental Psychology Unit, Department of Psychology, Ludwig Maximilian University Munich, Leopoldstrasse 13, 80802 Munich, Germany. E-mail: carolina.pletti@psy.lmu.de

Abstract

Moral identity, or moral self, is the degree to which being moral is important to a person's self-concept. It is hypothesized to be the 'missing link' between moral judgment and moral action. However, its cognitive and psychophysiological mechanisms are still subject to debate. In this study, we used event-related potentials to examine whether the moral self-concept is related to how people process prosocial and antisocial actions. To this end, participants' implicit and explicit moral self-concept were assessed. We examined whether individual differences in moral identity relate to differences in early, automatic processes [i.e. Early Posterior Negativity (EPN), N2] or late, cognitively controlled processes (i.e. late positive potential) while observing prosocial and antisocial situations. Results show that a higher implicit moral self was related to a lower EPN amplitude for prosocial scenarios. In addition, an enhanced explicit moral self was related to a lower N2 amplitude for prosocial scenarios. The findings demonstrate that the moral self affects the neural processing of morally relevant stimuli during third-party evaluations. They support theoretical considerations that the moral self already affects (early) processing of moral information.

Key words: moral identity; moral self; ERPs; prosociality

Introduction

Moral identity, defined as 'the degree to which being a moral person is important to a person's self concept' (Hardy and Carlo, 2011), has received increased attention in psychological research (Aquino and Reed, 2002; Lapsley and Narvaez, 2004; Johnston *et al.*, 2013; Hertz and Krettenauer, 2015; Paulus, 2018). It has been hypothesized to be the 'missing link' between moral judgment and moral action (Blasi, 1983). It is a matter of simple observation that people have moral beliefs and convictions, and yet occasionally behave in immoral ways. This mismatch between moral judgment and moral actions has been known to psychologists and laymen since years: whether individuals judge an action as right or wrong does not always predict whether they decide to perform such action or not (Tassy *et al.*, 2013; Pletti *et al.*, 2017). For instance, a professional might know that paying taxes is the right thing to do, but might nevertheless decide to let some clients pay 'under the table' in order to have a few extras at the end of the month. Or, one might not be motivated to donate to the poor, or volunteer for aid agencies, despite deeming such activities praiseworthy. In order to understand why some people, notably moral exemplars, seem to behave according to their moral judgements, whereas others do not, moral psychologists have introduced the concepts of moral self and moral identity (e.g. Blasi, 1983; Colby and Damon, 1992; Frimer and Walker, 2009).

According to moral identity theory, if someone strongly identifies with their own moral values, they will behave accordingly to their moral judgment, because doing otherwise would create

Received: 12 October 2018; **Revised:** 20 February 2019; **Accepted:** 26 February 2019

an unpleasant situation of cognitive dissonance (Blasi, 1983). Thus, people who have a strong moral identity would be more motivated to behave morally, because this is necessary for them in order to maintain a coherent image of themselves. Moreover, for people with a strong moral identity, there is a high correspondence between moral goals and personal goals, and thus these people often are not even faced with the choice between the two (Colby and Damon, 1992; Frimer and Walker, 2009).

The link between moral self-concept[1] and moral action has received empirical support by several studies (e.g. Aquino *et al.*, 2009; Perugini and Leone, 2009). Moreover, a recent meta-analysis found a consistent effect of moral self-concept on moral behavior (Hertz and Krettenauer, 2015).

Despite the wealth of research investigating the link between moral self and moral action, however, the psychological mechanisms behind the moral self have remained unclear (for an overview, see Lapsley and Narvaez, 2004). The picture is made more complex by the fact that there are different aspects of the moral self that may relate to different psychological processes. For instance, several accounts of moral identity focus predominantly on the conscious and explicit concept that people have of themselves (Blasi, 1983; Colby and Damon, 1992; Krettenauer, 2005). According to these accounts, individuals with a high moral self-concept have integrated their moral values into their own self-concept. They chose to follow moral goals over other conflicting interests, as their personal goals coincide with moral goals. Their self-concept is in principle consciously accessible and thus based on explicit reasoning and self-reflection. However, other authors have suggested that another aspect of the moral self does not necessarily have to be based on deliberate thinking and does not always have to be accessible to conscious reasoning. For instance, according to Lapsley and Narvaez (2004), a person with a moral identity '*would be one for whom moral constructs are chronically accessible and easily activated for social information-processing*'. Such a person would filter their perception of the world and their memories through chronically active moral schemas. The concept of schemas does not imply consciousness, on the contrary, schemas are supposed to be activated implicitly and in an automatic fashion (e.g. Bargh *et al.*, 1988). In fact, one study by Perugini and Leone (2009) has successfully measured an implicit moral self-concept by using an implicit association test (IAT). In this study, the implicit and explicit moral self were not correlated with each other and were shown to predict different kinds of moral decisions (but see also Aquino and Reed, 2002). The implicit moral self related to cheating behavior and donations, and the explicit one correlated with decisions in hypothetical moral dilemmas (Perugini and Leone, 2009). Furthermore, a study by Johnston *et al.*, 2013 found that the moral self IAT—but not the explicit moral self—predicted physiological arousal in response to moral violations, whereas the explicit moral self—but not the implicit one—predicted religiosity. For this reason, in order to understand the mechanisms behind the moral self-concept, it seems to be important to distinguish between explicit and implicit moral self-concept.

Despite the wealth of evidence on the relation between the moral self and actual behavior, the related information processing and neurocognitive mechanisms have remained unclear. One theoretical view proposes that the moral self already affects how information about morally relevant situations is processed (Lapsley and Narvaez, 2004). Here, early stages of information processing can be contrasted with later stages. This question is difficult to tap into with behavioral studies, but can be addressed using neuroscience methods such as event-related potentials (ERPs). ERPs allow to quantify ongoing neural responses with excellent temporal resolution (millisecond) and do not require explicit reporting of psychological operations, as most behavioral measures do. The deflections in the waveform, referred to as components, are thought to reflect discrete information processing operations, with latency varying with stimulus evaluation time. The high temporal resolution allows to make inference about information processing mechanisms. In particular, ERPs allow to identify information processing from very early stages, thus tapping into automatic/rapid processes to controlled/slow ones that are supposed to be related to later stages. This characteristic is especially suited to address our research question, since it allows to investigate which stages of processing are influenced by the moral self: earlier and automatic ones or later and controlled ones?

An answer to this question would provide insight concerning the neurocognitive mechanisms of the moral self, allowing to find out whether it is more based on controlled and deliberative mechanisms as traditionally thought (Blasi, 1983; Colby and Damon, 1992; Krettenauer, 2005) or also on fast and automatic/automatized mechanisms as more recently proposed by Lapsley and Narvaez (2004). This question can be answered using ERPs. Moreover, ERPs can also allow greater insight on the differences between implicit and explicit moral self: are the two constructs relying on separate mechanisms? Previous literature has indicated that the implicit and explicit moral selves do not correlate with each other and relate to different types of moral behavior (Perugini and Leone, 2009; Johnston *et al.*, 2013). This study aims at testing whether the implicit and the explicit moral selves might be related to different stages or forms of processing of morally relevant information. One possibility is that the implicit moral self relates to early automatic stages, and the explicit moral self to later stages of moral processing.

As a means to answer these questions, we can capitalize on findings regarding the electrocortical processes related to the perception of moral content. These have been investigated by studies using both pictorial and text stimuli (Lahat *et al.*, 2013; Yoder and Decety, 2014; Cowell and Decety, 2015b; Gan *et al.*, 2016; Gui *et al.*, 2016). These studies show that moral content is detected early, and processed throughout several stages. In particular, studies using pictorial stimuli with moral content reported effects on three different ERP components: the N1 or EPN, the N2 and the late positive potential (LPP; Yoder and Decety, 2014; Cowell and Decety, 2015b; Gui *et al.*, 2016).

The N1 (Yoder and Decety, 2014; Gui *et al.*, 2016) or EPN (Cowell and Decety, 2015b) is an early negativity detected at parietal electrodes with a peak around 100–150 ms. It was reported to be greater for stimuli depicting prosocial versus antisocial actions in both children (Cowell and Decety, 2015b) and adults (Yoder and Decety, 2014), a result which might reflect greater early attentional capture by prosocial scenes (cfr. Schupp *et al.*, 2003). Outside the moral domain, the N1/EPN has been reported as being sensitive to valence, being greater for positive than negative stimuli (Schupp *et al.*, 2004; Weinberg and Hajcak, 2010).

[1] Given that the majority of scholars use the terms 'moral identity', 'moral self' and 'moral self-concept' as synonyms (e.g. Aquino and Reed, 2002; Hardy and Carlo, 2011; Jennings *et al.*, 2015; Reed *et al.*, 2007), for the purpose of this paper, we are also not differentiating between the two.

The N1/EPN was also reported to be greater for pictures representing moral violations as compared to negative pictures matched by arousal but devoid of moral content (Gui et al. 2016). Thus, the N1/EPN might also reflect moral intuition or the quick detection of moral content in a visual scene (Gui et al. 2016).

A second negative component, peaking around 200–300 ms after stimulus onset, was observed on frontal electrodes and labeled N2 (Yoder and Decety, 2014; Cowell and Decety, 2015b; Gui et al., 2016). In adults, the N2 is greater for prosocial vs antisocial actions (Yoder and Decety, 2014) and sensitive to arousal (Gui et al. 2016). Interestingly, in children, the N2 has been reported to be greater for antisocial as compared to prosocial actions (Decety and Cowell, 2018), which was interpreted as evidence that third-party implicit evaluations of antisocial actions elicit cognitive conflict (cfr. Folstein and Van Petten, 2008) or a violation of expected rules of social interaction (Decety and Cowell, 2018). Since the N2 is sensitive to violation of expectations (Folstein and Van Petten, 2008), one might think that it should always be greater to immoral acts as compared to moral or neutral ones (since immoral acts violate social norms and thus should be unexpected). However, Yoder and Decety (2014) reported a greater N2 for prosocial compared to antisocial actions in adults. They suggested that this reflects a greater personal relevance of the prosocial actions compared to the antisocial actions.

Finally, effects related to moral content were also reported as a late positive deflection (LPP) on midline electrodes starting around 400 ms after stimulus onset (Yoder and Decety, 2014; Cowell and Decety, 2015b; Gui et al., 2016). The LPP has been reported to be greater for prosocial versus antisocial acts in adults and children alike (Yoder and Decety, 2014; Cowell and Decety, 2015b). Gui et al. (2016) found that an early LPP (350–420 ms) was greater for high-arousing stimuli versus low-arousing stimuli irrespective of moral content, while a later slow wave (450–650 ms) was also affected by moral content. The LPP is known to reflect cognitively controlled allocation of processing resources and appraisal of motivationally salient stimuli (Hajcak et al., 2010). Thus, a greater amplitude of this component might indicate that the corresponding stimulus category is being allocated more resources and is processed longer.

The interpretation of the exact psychological mechanisms that these three components correspond to is still partially a matter of debate. However, what clearly emerges from the aforementioned studies is that moral content affects both automatic processes, starting at 100 ms with the EPN, and controlled ones, reflected by the LPP. Capitalizing on these previous results, the present research aimed at using ERPs in order to investigate whether and how the moral self is related to the neural processing of third-party moral scenarios. This will clarify the neurocognitive mechanisms underlying the moral self, especially regarding whether and how the moral self relates to controlled processes only, or also to automatic processes as proposed by Lapsley and Narvaez (2004). Furthermore, the use of both explicit and implicit measures of the moral self can elucidate the relationship between the two and distinguish whether the two constructs are related to different stages of processing of moral content. We hypothesized that the implicit moral self, being based on automatized associations between concepts, should modulate the early, automatic processing of moral content. More specifically, it should affect early components such as the EPN, which is related to automatic attentional capture by salient stimuli (Schupp et al., 2004), or the N2, which is related to conflict detection, violation of expectation (Folstein and Van Petten, 2008) or personal relevance (Yoder and Decety, 2014). The explicit moral self, on the other hand, being rooted on deliberate reasoning, might affect the more controlled appraisal of moral content. We would therefore expect a modulation of the amplitude of the LPP, which reflects the sustained, cognitively influenced appraisal of motivationally significant stimuli (Hajcak et al., 2010).

Next to our main research question, we can also advance hypotheses on the direction of the effects for each ERP component (although more speculatively given the mixed pattern of previous research): in particular, if moral content captures the attention of people with high implicit moral self, then their EPN should be increased for both antisocial and prosocial scenes as compared to people with low implicit moral self. If people with high implicit moral self expect prosocial actions to appear, then their N2 for prosocial actions should be of reduced amplitude as compared to people with low implicit moral self. If they expect antisocial actions to appear less, then their N2 for antisocial actions should be increased. However, if we assume the N2 to be related to personal relevance, then the opposite effect should appear, that is, an increased N2 to prosocial actions and possibly a reduced N2 to antisocial actions, for people with high moral self as compared to people with low moral self. Finally, as concerns the LPP, we can expect people with high explicit moral self to allocate more processing resources to moral content, irrespective whether prosocial or antisocial, and thus have a greater LPP for both stimuli as compare to people with low moral self.

Materials and methods

Participants

Seventy-five adult participants (40 female) took part in the study (mean age, 24.6; s.d., 5.5). All were right handed, German speakers, had normal or corrected-to-normal vision and no neurological or psychiatric disorder. Participants were recruited through flyers distributed in the university areas of a large German city, and their participation was compensated either with 10 euro per hour or, when applicable, with course credit. The study was approved by the local ethics committee. The sample size was determined through a power analysis hypothesizing a medium-to-large-sized effect of the moral self on ERP amplitudes ($r = 0.40$), which is typical for studies reporting correlations between ERPs and individual differences in the moral domain (e.g. Sarlo et al., 2014; Yoder and Decety, 2014) and setting the alpha level to 0.5 and the beta level to 0.90. According to this analysis, the estimated sample size was $N = 61$. To account for attrition, we collected data from 75 participants. From these, four had to be excluded from all the analyses because of equipment failures or too many movement and sweat artifacts in the electroencephalogram (EEG) [final sample size for ERP data analyses: 71 (38 F); mean age, 24.63; s.d., 5.55] Furthermore, the implicit moral self measure could not be completed due to technical errors by one participant, who had to be excluded from all the analyses concerning the implicit moral self [final sample size for the analyses including the implicit measure: 70 (38 F); mean age, 24.81; s.d., 5.53]. Similarly, five participants had to be excluded from all the analyses concerning the explicit moral self because of problems with the computerized version of the questionnaire [final sample size for the analyses including the explicit measure: 66 (34 F); mean age, 24.75; s.d., 5.56].

Stimuli and measures

Chicago moral sensitivity task. The Chicago moral sensitivity task (CMST) consists of three-picture vignettes portraying two

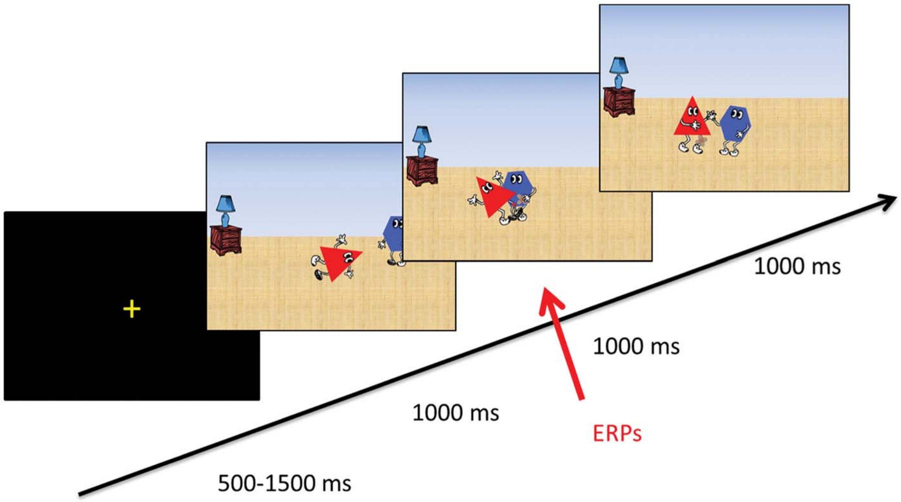

Fig. 1. Time course of one trial of the CMST. The ERPs were time locked to the onset of the second picture in the sequence. Here an example of a prosocial action: the blue character is helping the red one to stand up.

characters, depicted as geometric shapes, interacting in either a prosocial or an antisocial manner (Cowell and Decety, 2015a, 2015b). In the present study, the task comprised 60 prosocial and 60 antisocial actions, presented in random order for 3 s each (1 s per each single picture; see Figure 1) and with a jittered inter-trial interval of 500–1000 ms. The first picture introduces the scene, the second pictures portrays the crucial prosocial or antisocial action and the third picture shows the outcome of the action and finishes the trial. The ERPs were time locked to the onset of second picture, since a validation study found that participants are able above chance to distinguish between conditions from the second picture (Cowell and Decety, 2015b). The task was administered using Presentation® software (Version 19.0; Neurobehavioral Systems, Inc., Berkeley, CA; www.neurobs.com).

Moral self IAT. We used a German translation of the moral self IAT devised by Perugini and Leone (2009). Participants were asked to categorize, by pressing one of two keys as quickly and accurately as possible, a series of words based on whether they were moral or immoral adjectives, or pronouns referring to the self or to others (see Table 1 for a translated list of stimuli). The IAT was composed of seven total blocks. In the first block (*moral/immoral discrimination*, 20 trials), participants had to press a key with the right hand or another with the left hand in order to categorize adjectives as moral or immoral. In the second (*self/other discrimination*, 20 trials), they had to categorize pronouns as pertaining to the self or to others. In the third and fourth blocks (*first paired*, 20 trials; *second paired*, 40 trials), the two pairs of labels were presented simultaneously, so that to one side corresponded two labels (e.g. moral and self on the right and immoral and others on the left or vice versa). The fifth block (*moral/immoral discrimination reversed*, 40 trials) was identical to the *moral/immoral discrimination* block, but the location of the labels was reversed. Finally, the sixth and seventh block (*first paired reversed*, 20 trials; *second paired reversed*, 40 trials) were identical to the *first-paired* and *second-paired* block, but with the *moral/immoral* labels reversed as in the fifth block. The initial location of the categories and the order by which categories were paired (moral with self first or moral with others first) were counterbalanced between participants. The 10 moral/immoral words were presented 12 times each, the 6 self/other words were presented on average 13 times each. The words were presented in random order within blocks. See Figure 2 for two example trial of a *paired* block. The task was administered using Presentation® 19.0.

Self-importance of moral identity questionnaire. In this measure, created and validated by Aquino and Reed (2002), participants are asked to imagine a person characterized by the following traits: concerned, honest, fair, hardworking, friendly, generous, helpful, loving and compassionate. Then, participants are asked to reply to a list of items referring to how much being such a person is part of the participant's own self-concept (internalization scale) and to how much the participant tries to appear to others like such a person (symbolization scale), using a 7-point Likert scales ranging from 1 (not at all) to 7 (completely). In this study, we used the 10 item version (Reed *et al.*, 2007), translated in German following Pohling *et al.*, 2018.

G recordingsThe EEG was acquired using 64 active electrodes (ActiCap, Brain Products GmbH, Gilching, Germany) referenced to Cz, a BrainAmp MR amplifier and recorded through the software Brainvision recorder (Brain Products GmbH), with an high-pass filter set at 0.016 Hz and a low-pass filter set at 1000 Hz, 500 Hz sampling rate and 0.1 μV resolution per least significant bit. All impedances were kept below 25 KOhm as recommended by Brain Products for the ActiCap.

Procedure

Upon arrival in the laboratory, participants read and signed the informed consent form. Subsequently, they sat in a dimly lit, sound-attenuated cabin, 90 cm from a 19″ computer monitor (60 Hz refresh rate) where the tasks would be presented, and they were applied the EEG cap. Afterwards, the experimenter left the cabin and the participants started the CMST. Then, they were given a keyboard and, after receiving instructions from the experimenter, they completed the moral self IAT and finally the self-importance of moral identity questionnaire. Finally, they were debriefed and received compensation for participating in the study. Participants were not told until the experiment ended that the tasks were about morality or their self-concept. This, together with the order of the tasks, made sure that participants were not primed on moral contents before starting the CMST, or focused on their self-concept before starting the IAT.

Table 1. Words used in the IAT in German, with English translation in brackets

Moral	Immoral	Self	Others
Ehrlich (honest)	Betrüger (cheater)	Ich (I)	Andere (others)
gewissenhaft (faithful)	unerhlich (dishonest)	mein (my)	sie (they)
aufrichtig (sincere)	täuschend (deceptive)	mich (me)	ihnen (them)
bescheiden (modest)	arrogant (arrogant)		
Altruist (altruist)	überheblich (pretentious)		

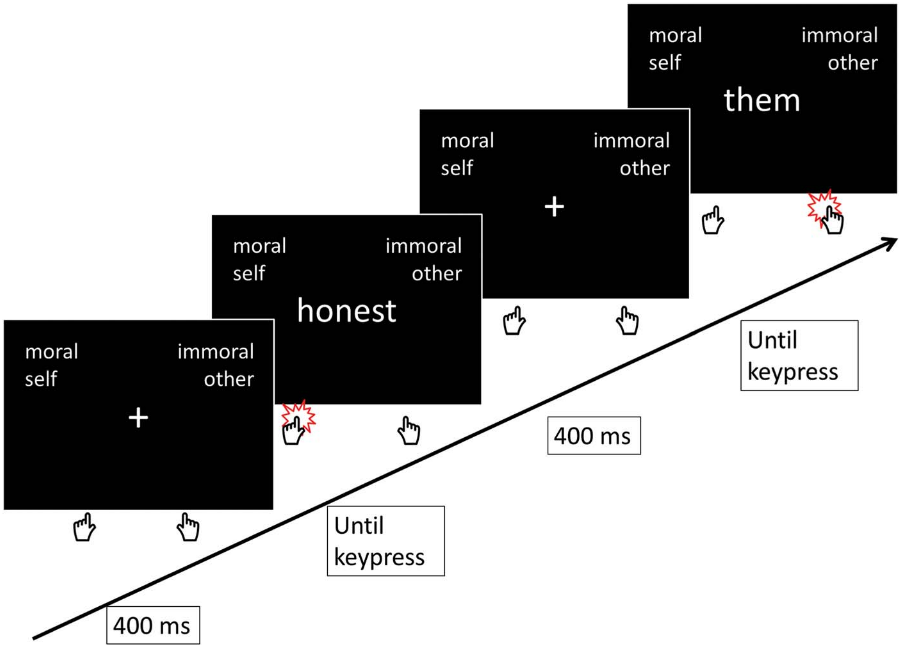

Fig. 2. Example of two trials of a paired block of the IAT. In this example, participants have to respond to words related to self or moral with the left hand and to words related to other or immoral with the right hand.

Data analysis

G data reduction and analysisThe EEG was analyzed using EEGlab 14_1_1b (Delorme and Makeig, 2004) and ERPlab 7.0 (Lopez-Calderon and Luck, 2014) on Matlab 2017a (The MatWorks®, Inc). The signal was resampled at 250 Hz and visually scored for artifacts. Bad channels (i.e. resulting from broken electrodes) were interpolated, and moments with gross movement artifacts were excluded from further data processing. Then, data were re-referenced to the average reference and high-pass filtered at 0.1 Hz. Eye movements and blinks were corrected using Independent Component Analysis (ICA). Then, data were low-pass filtered at 30 Hz and epoched at −200 to 1000 ms from the onset of the second picture, baseline corrected for the mean signal between −200 and 0 ms. Finally, all epochs including amplitudes exceeding ±70 μV were rejected, and the remaining data were averaged separately per condition.

Based on the study by Cowell and Decety (2015b), we examined the EPN, N2 and LPP components. However, the latencies and scalp distribution of these components in our data did not exactly match those reported by Cowell and Decety with children participants, as was to be expected given the fact that we tested an adult sample. Thus, we adjusted the clusters of electrodes to be analyzed and the time windows of interest based on the inspection of the grand average (see Figure 3). We calculated clusters instead of choosing single electrodes in order to maximize the signal-to-noise ratio and to reduce the impact of variability in scalp distribution of the components between participants. As a result, we quantified the EPN as the peak amplitude between 100 and 200 ms in two lateralized parietal clusters (parietal left: P5, P3, P1; parietal right: P2, P4, P6); the N2 as the peak amplitude between 250 and 350 ms in three midline fronto-central clusters (frontal: Fz, F1, F2; frontocentral: FCz, FC1, FC2; central: Cz, C1, C2); and the LPP as the mean amplitude between 400 and 1000 ms in two midline centro-parietal clusters (centroparietal: CPz, CP1, CP2; parietal: Pz, P1, P2).

Each component was analyzed with a separate repeated measure ANOVA with cluster and condition as factor.

Moral self measure data reduction and analysis

The IAT was scored based on the improved algorithm by Greenwald et al. (2003). The self-importance of moral identity questionnaire was scored in the two subscales internalization and symbolization.

Correlations were performed between the two scales of the questionnaire and the IAT scores in order to assess associations between measures. To answer our main research questions, correlations were performed between each moral self measure and each ERP component separately by cluster and condition in order to assess associations between the moral self and the processing of prosocial and antisocial scenes.

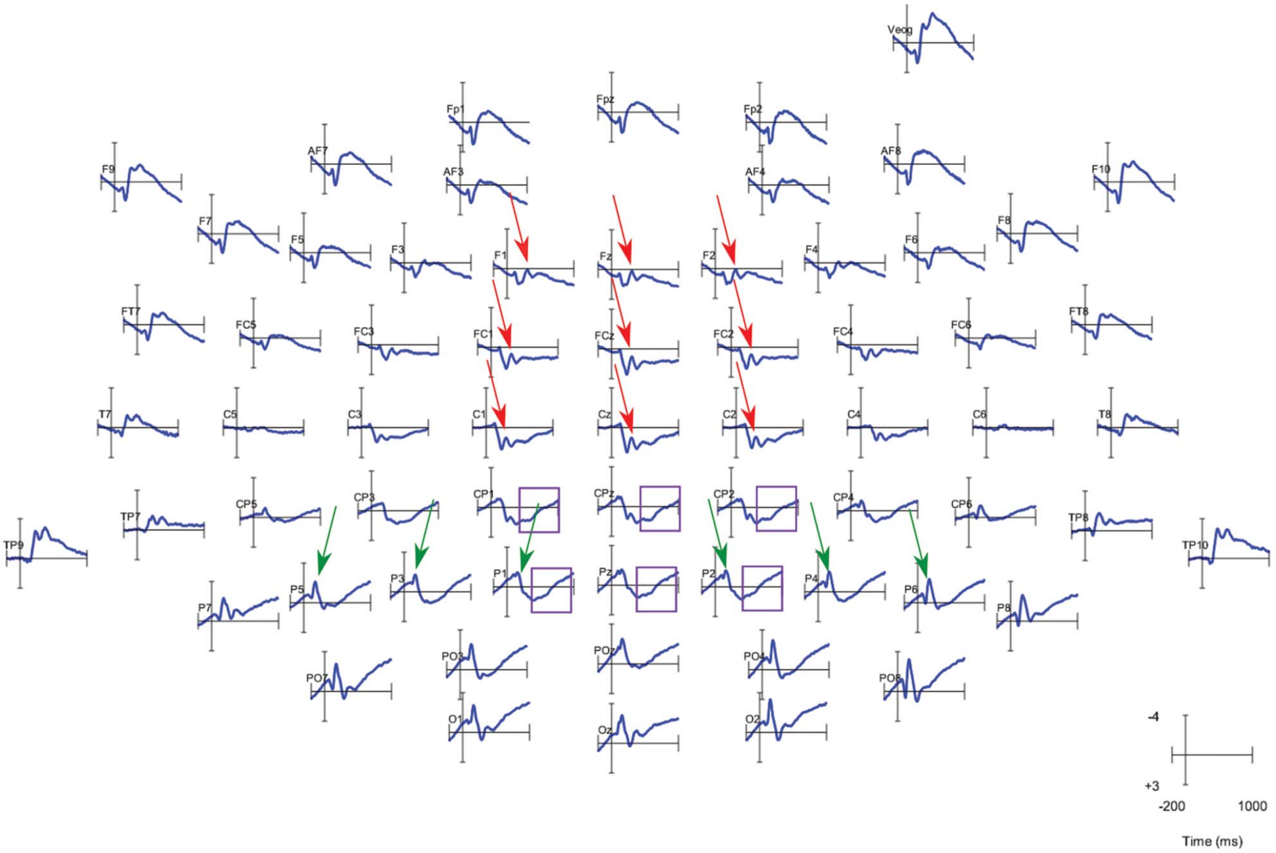

Fig. 3. Grand-averaged ERPs across conditions, with components of interest marked with arrows and squares. The red arrows show the N2 peak, the green arrows the EPN peak and the purple squares the LPP time window.

Results

Moral self measures: associations between explicit and implicit moral self

The symbolization and internalization scale of the self-importance of moral identity questionnaire were correlated: $r = 0.56$, $P < 0.001$. Neither of the two scales was correlated with the moral self IAT score ($rs < 0.24$, $Ps > 0.06$). See Table 2 for descriptive statistics regarding the moral self measures.

ERP analyses: difference between prosocial and antisocial and associations with explicit and implicit moral self

EPN. An analysis of variance (ANOVA) with cluster and condition as factors revealed a significant main effect of condition (antisocial versus prosocial). Prosocial actions elicited greater EPN amplitudes (greater negativity) compared to antisocial actions: $F(1,70) = 6.3$, $P = 0.01$, $\eta^2_p = 0.08$ (see Figure 4).

The EPN amplitude elicited by prosocial scenes at both clusters was also significantly correlated with the IAT score, so that participants with greater IAT score (greater implicit moral self) had a reduced EPN amplitude (lower negativity) for prosocial scenes (see Figure 5). Parietal right: $r = 0.28$, $P = 0.02$; parietal left: $r = 0.24$, $P = 0.04$. No other correlation was significant ($rs < 0.15$, $ps > 0.26$).

N2. The ANOVA yielded a significant main effect of condition, as prosocial scenes elicited a greater negativity as compared to antisocial scenes: $F(1,70) = 4.25$, $P = 0.04$, $\eta^2_p = 0.06$ (see Figure 6).

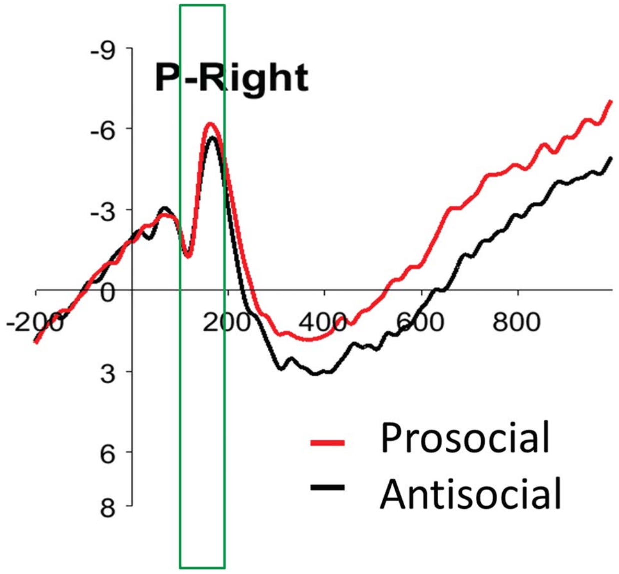

Fig. 4. Grand average of the ERP wavefor on the right parietal cluster. The green rectangle shows the time window in which the EPN peak was measured.

Furthermore, there was a main effect of cluster, as the N2 was greater on the frontal cluster, followed by the frontocentral and central clusters: $F(2,140) = 47.92$, $P < 0.001$, $\eta^2_p = 0.41$; frontal versus frontocentral cluster mean difference, −1.97, SE, 0.19,

Table 2. Means and standard deviations for the moral self measures. For all three variables, higher scores indicate a higher moral self-concept. The internalization and symbolization scores are calculated by summing up the scores given to each item in the scale. Every scale is composed of 10 items with scores ranging from 1 to 7. The IAT score is calculated as adjusted reaction time difference following Greenwald *et al.* (2003). Positive numbers indicate that participants were faster to respond when moral adjectives were associated to the self. Negative numbers indicate that they were faster to respond when immoral adjectives were associated to the self

Variable	Mean (s.d.)
Self-importance of moral identity—internalization	30.19 (3.79)
Self-importance of moral identity—symbolization	20.51 (5.51)
Moral self IAT	0.59 (0.32)

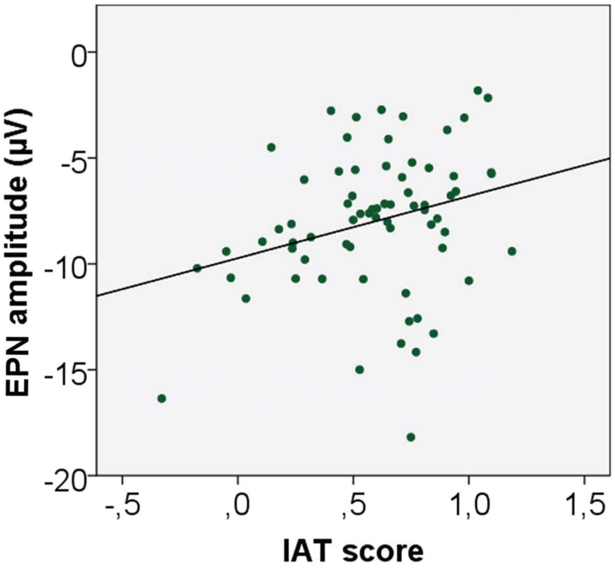

Fig. 5. Correlation between IAT Score and EPN amplitude in microvolt for the prosocial condition. A high positive IAT score indicates a strong implicit moral self. Since the EPN is a negative component, a more negative value indicates a greater amplitude.

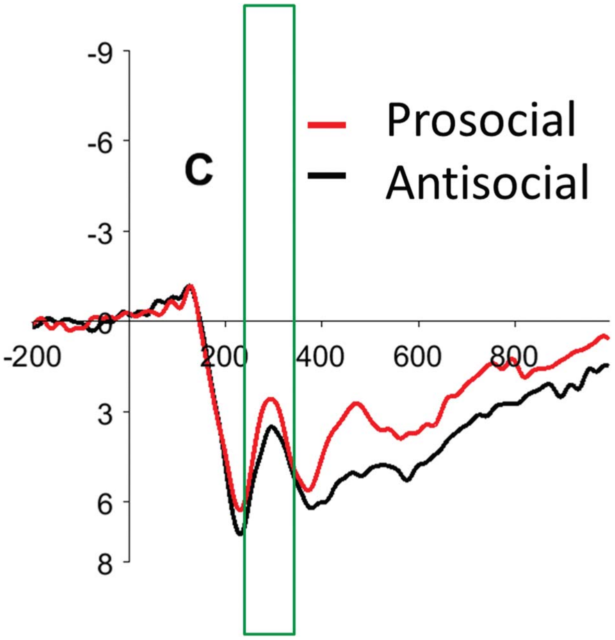

Fig. 6. Grand average of the ERP waveform on the central cluster. The green rectangle shows the time window in which the N2 peak was measured.

$P < .001$; frontal versus central cluster mean difference, −2.76, SE, 0.39, $P < 0.001$; frontocentral versus central cluster mean difference, −.79, SE, 0.26, $P = 0.004$.

The N2 amplitude elicited by prosocial scene at the central cluster correlated with the internalization score, so that to a greater internalization score corresponded a lower N2 amplitude (lower negativity) for prosocial scenes (see Figure 7): $r = 0.26$, $P = 0.03$. No other correlation was significant (rs < 0.21, Ps < 0.10).

LPP. The main effect of condition was significant, $F(1,70) = 41.59$, $P < 0.001$, $\eta^2{}_p = 0.37$, as well as the main effect of cluster, $F(1,70) = 100.54$, $P < 0.001$, $\eta^2{}_p = 0.59$, and the interaction between cluster and condition, $F(1,70) = 7.89$, $P = 0.006$, $\eta^2{}_p = 0.10$. *Post hoc* pairwise comparisons showed that antisocial scenes elicited greater LPP amplitudes in both centroparietal and parietal clusters, but this difference was greater in parietal clusters (see Figure 8). Centroparietal cluster: antisocial versus prosocial mean difference, 1.88, SE, 0.32, $P < .001$; parietal cluster: antisocial versus prosocial mean difference, 2.55, SE, 0.40, $P < 0.001$. The LPP amplitude did not correlate with any moral self measure, all rs < 0.17, all Ps > 0.18.

Discussion

The present study aimed at investigating the neurocognitive mechanisms related to the moral self-concept. In particular, we examined whether and how the moral self-concept relates to the neural processing of scenes depicting morally laden scenarios. To this aim, we collected ERPs while participants perceived third-party prosocial and antisocial scenarios, and measured both the implicit and explicit moral self-concept in the same participants. We focused on three ERP components marking three different processing stages: the EPN, the N2 and the LPP. We chose these three components as they have been found to discriminate between prosocial and antisocial scenarios in both children and adults (Yoder and Decety, 2014; Cowell and Decety, 2015b). Based on the idea that the implicit moral self-concept should be based on the automatic activation of moral schemas and the explicit moral self-concept on deliberate moral reasoning, we hypothesized that the implicit moral self-concept would be related to the early processing of moral scenes, and the explicit moral self-concept to the later processing of moral scenes. The results we found were generally in line with our first hypotheses, but not with the second. We interpret our results as providing

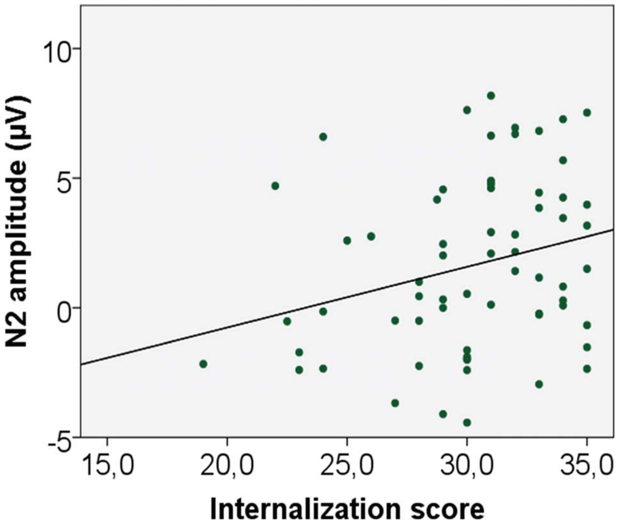

Fig. 7. Correlation between internalization score and N2 amplitude in microvolt for the prosocial condition. A high internalization score indicates a high explicit moral self. Since the N2 is a negative component, a more negative value indicates a greater amplitude.

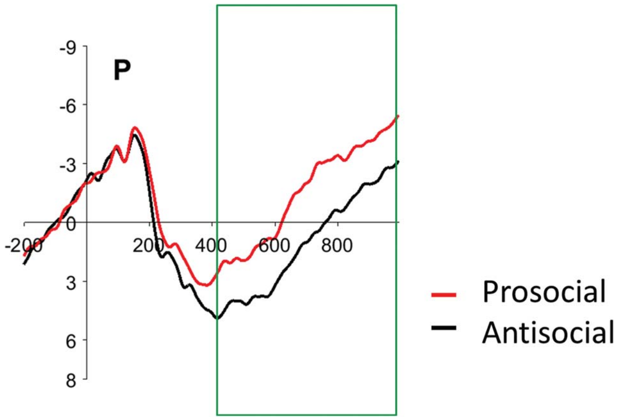

Fig. 8. Grand average of the ERP wavefor on the parietal cluster. The green rectangle shows the time window in which the LPP mean amplitude was measured.

first neurophysiological evidence for the theoretical claim that people with a high moral identity perceive the world through moral schemas that affect early stages of social information processing (Lapsley and Narvaez, 2004).

The neurophysiological computations associated with the moral self

To be more specific, we found differences in the ERPs elicited by prosocial versus antisocial scenes in all the three expected components (EPN, N2 and LPP). Of these components, the EPN and the N2 were correlated to the moral self. In particular, the higher the implicit moral self, the lower the amplitude of the EPN for prosocial actions; the higher the explicit moral self, the lower the amplitude of the N2 for prosocial actions. There was no such relation for the perception of antisocial actions.

The EPN is related to attentional capture by salient/relevant stimuli. It has been reported to have a larger amplitude for emotional stimuli as compared to neutral ones (Schupp *et al*., 2003, 2004; Codispoti *et al*., 2006). Furthermore, studies using morally laden stimuli found a greater EPN amplitude for prosocial scenes as compared to antisocial scenes in both children and adults (Yoder and Decety, 2014; Decety *et al*., 2015). According to our hypothesis, we expected people with high implicit moral self to have a greater EPN for both prosocial and antisocial actions, since moral content should be more salient for them. We found instead that people with a high implicit moral self show a reduced amplitude of this component for prosocial scenarios. These results could be explained by recent predictive coding frameworks. In a predicting coding framework, the processing of an expected stimulus is reduced as compared to that of an unexpected stimulus (Murray *et al*., 2002; Aoyama *et al*., 2005; Friston, 2005). Since prosocial stimuli match the prosocial schemas that are already active for people with a high moral self (Lapsley and Narvaez, 2004), one could thus speculate that they are expected and thus require less processing resources, hence the reduced EPN. Yet, how to explain the lack of an enhanced EPN for antisocial action for people with a high moral self? This could be explained considering that morality can be divided in two dimensions—one regarding avoidance of antisocial action and one regarding motivation for prosocial action (also called negative and positive duties; see, for instance, Belliotti, 1981; Lichtenberg, 2010). For an adult, it is not enough to avoid antisocial action in order to be considered a moral person. The avoidance of antisocial action is expected by society, is recognized as a moral obligation from childhood on (e.g. Killen and Smetana, 2015) and does not per se make someone a moral exemplar. What makes the difference, instead, is especially free-willingly engaging in prosocial actions, which can be costly and elevate someone's moral character (e.g. Colby and Damon, 1992). Thus, adults' moral self-concepts might be based especially on a greater importance given to prosociality and not to a greater sensitivity to antisocial actions, which should be clearly perceived as wrong irrespective of the moral self-concept. Another potential explanation of the EPN findings is related to valence processing. The EPN component has been shown to be greater for stimuli of positive valence (Schupp et al., 2004; Weinberg and Hajcak, 2010). It might be that people with a high moral self, who are more dedicated to prosocial action, find them more neutral since they are habituated to them. Thus, these actions might be seen as more positive by people with lower moral self. Antisocial actions, on the other hand, might be perceived as likewise negative by people with a low moral self as well as by people with a high moral self-concept, thus producing a 'floor effect' where a correlation does not emerge.

Moving to a later time window, we found that the explicit moral self, as measured through the internalization scale of the self-importance of moral identity questionnaire, was associated with a reduced N2 amplitude for prosocial scenarios. The frontal N2 component has been related to detection of novelty and interpreted as signaling mismatch from a perceptual template (see Folstein and Van Petten, 2008, for a review). It was also proposed to relate to personal relevance (Yoder and Decety, 2014). We thus advanced two hypotheses for this component: if the N2 signals violation of expectations, then the moral self should be negatively correlated with the N2 amplitude for prosocial scenes. If, on the other hand, the N2 signals personal relevance, then a positive correlation should appear. Our results are in line with the first hypothesis: similarly to the interpretation given for the EPN, also in this case we can assume that prosocial scenes match the active perceptual template for people with greater moral identity, who thus show a reduced N2 to them. However, as opposed to what predicted, this effect emerges for the explicit moral self and not for the implicit one, showing how even the explicit moral self relates to early stages of processing and relatively automatic mechanisms.

As concerns the antisocial scenes, an effect of the moral self on the N2 did not emerge. This is in line with our results for the EPN component. Again, one might hypothesize that the effect does not emerge in the antisocial actions because these actions are deviant for everybody, and thus the moral self does not influence their processing.

Taken together, these results show how both the implicit and the explicit moral self relate to the processing of prosocial scenes in the early, automatic stages. This in line with information processing approaches to the moral self (Lapsley and Narvaez, 2004): people with a strong moral self have chronically active moral schemas, and this influences the early processing of moral events.

Our results also show that the implicit and explicit moral self-concepts are associated with different neural responses. This, together with the lack of significant correlations between the two measures, provides further evidence indicating that the two concepts are separate constructs relying on different cognitive mechanisms. This finding extends previous behavioral studies (Perugini and Leone, 2009; Johnston *et al.*, 2013) by revealing distinct neurophysiological processes that may relate to the implicit and explicit moral self. This study thus makes a first step in gaining a deeper understanding on the difference between the implicit and explicit moral self, a topic which should be further investigated in future research.

In contrast to our predictions, we did not find any correlations between the explicit moral self-concept and the later stages of processing, as measured through the LPP. Conversely, even the explicit moral self was related to early differences in information processing, in stages that can still be considered relatively automatic. We can speculate that this effect could be the result of automatization: adults with high moral identity might have automatized the processing of prosocial actions due to the time spent acting and thinking in a prosocial way. To validate this hypothesis, it would be interesting to examine the influence of the moral self on the processing of moral content in childhood. Children are still developing their moral identity (e.g. Krettenauer and Hertz, 2015) and thus the processing of moral content might not be automatized yet. One might hypothesize that, at an earlier testing age, it would be less likely to find an effect on early components and more likely to find an effect on late components. This will be investigated in future studies.

Neurophysiological responses to prosocial and antisocial scenarios

Beyond our main research question on the moral self, our study also speaks to the issue of how people process antisocial and prosocial scenes. We found differences between prosocial and antisocial stimuli in similar time windows and locations that had been reported in previous research (e.g. Yoder and Decety, 2014; Cowell and Decety, 2015b). In addition, our results differ in some aspects from those found with children using the same stimuli or from those found with adults using different stimuli.

Coherently with what has been reported in both children and adults, the EPN had a greater amplitude for prosocial as compared to antisocial scenes. This seems to indicate that prosocial scenes are generally more salient (cfr. Schupp *et al.*, 2003).

In a later time window, we found a greater frontal N2 for prosocial scenes as compared to antisocial scenes. This replicates previous findings with adults (Yoder and Decety, 2014). Since it has been argued that the N2 reflects deviation from expectation (Folstein and Van Petten, 2008), this result could be interpreted as a sign that participants were expecting antisocial actions to appear and were thus considering the prosocial ones as 'deviant'. This appears surprising, since antisocial actions present a deviation from social norms. They could thus be hypothesized to violate participants' expectations and as a consequence should elicit a greater N2. This might be explained if we take contextual issues into account. In this specific study, prosocial scenes displayed a wider variety of behaviors (helping, consoling, giving presents, cooperating) as compared to the antisocial ones (mainly physical harm). Thus, antisocial actions might have become more predictable than prosocial ones, which led to a reduced N2 for these actions. Given the mixed findings regarding this component that emerged in previous literature (Yoder and Decety, 2014; Cowell and Decety, 2015b; Gui *et al.*, 2016), the N2 does not have an univocal explanation yet and should be further investigated.

Finally, we found a greater centro/parietal LPP for antisocial scenes as compared to prosocial scenes. This is the opposite of what was previously observed in both preschoolers (Cowell and Decety, 2015b) and in adults (Yoder and Decety, 2014). However, this is coherent with other findings related to the LPP: for instance, the LPP is sensitive to arousal, since arousing stimuli have a higher motivational salience and thus demand more processing resources (Codispoti *et al.*, 2006; Hajcak et al., 2010). Antisocial scenes can be hypothesized to be more arousing, since they involve harm and threat. As the LPP is interpreted as indicating sustained attention to intrinsically motivating stimuli (Hajcak *et al.*, 2010), this indicates that participants allocate more attentional resources to the processing of antisocial scenes, which again might be due to the threatening value of these stimuli. The discrepancy between our results and those from some previous studies might be due to differences in the locations where the LPP was measured. For instance, the study by Yoder and Decety (2014) reported a higher LPP amplitude for prosocial actions in a frontal cluster, whereas we found centroparietal clusters (in line with the literature describing the LPP as a centropariental component (Hajcak *et al.*, 2010; Weinberg and Hajcak, 2010), which might have caused the switch in polarity of the effects. Another possible explanation of this discrepancy might be related to differences in the task that participants had to complete in the current study as compared to the one by Yoder and Decety (2014). In the present study, participants were asked to look at the scenarios without having to perform any evaluation task. In the study by Yoder and Decety (2014), participants had to alternatively focus on the intention behind a prosocial/antisocial action or on its outcome. Since the LPP is sensitive to task demands more than to mere stimulus properties (Hajcak *et al.*, 2010), this factor might explain the discrepancy.

Overall, while we partly found different directions of the effects, our results are in line with previous research in showing that the difference between prosocial and antisocial scenes is detected early on and that these scenes are processed differently through several stages. This contributes to furthering the knowledge regarding the neural correlates of implicit moral evaluations.

Limitations and conclusions

Although the current study adds to our understanding of the neurophysiological processes related to the moral self, some limitations should be mentioned. One point regards our choice of, using cartoon stimuli instead of naturalistic ones. This choice had a number of advantages, inter alia close experimental control, the elimination of confounding factors (e.g. age, race, gender, emotional facial expressions) and comparability with earlier studies (Cowell and Decety, 2015b). Yet, it leaves open the question how they generalize to everyday life experiences. Thus, future research should extend the investigation of how the moral self-concept influences the processing of moral content by using different kind of stimuli or different types of (im)moral behaviors. Furthermore, this study is only the first step toward understanding the neurocognitive mechanisms subserving the moral self given that it assessed the neural correlates in one specific paradigm. More research is needed in order to gain a deeper understanding of what processes are involved in the moral self-concept and how they develop. Future research, for instance, might use other neuroimaging techniques (e.g. fMRI) to investigate if the moral self relates to the activity of 'emotional' or 'cognitive' brain areas while perceiving moral situations.

Taken together, notwithstanding some limitations, this study provides novel insights into the nature of the moral self. Importantly, the results suggest that the moral self-concept influences the early processing of morally relevant contexts. Moreover, the implicit and the explicit moral self-concepts have different neural correlates, influencing respectively early and intermediate processing stages. Overall, the findings inform theoretical approaches on how the moral self informs social information processing (Lapsley and Narvaez, 2004).

Funding

This work was funded by the European Research Council (ERC) Starting Grant "Unravelling the moral self" (No. 679000) to M.P.

References

Aoyama, A., Endo, H., Honda, S., Takeda, T. (2005). Neuromagnetic analysis of effect of audition-based prediction on visual information processing. *International Congress Series*, **1278**, 219–22.

Aquino, K., Freeman, D., Reed, A.I., Felps, W., Lim, V.K.G. (2009). Testing a social-cognitive model of moral behavior: the interactive influence of situations and moral identity centrality. *Journal of Personality and Social Psychology*, **97**(1), 123–41.

Aquino, K., Reed, A.I. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, **83**(6), 1423–40.

Bargh, J.A., Lombardi, W.J., Higgins, E.T. (1988). Automaticity of chronically accessible constructs in person x situation effects on person perception: it's just a matter of time. *Journal of Personality and Social Psychology*, **55**(4), 599–605.

Belliotti, R.A. (1981). Negative and positive duties. *Theoria*, **47**(2), 82–92.

Blasi, A. (1983). Moral cognition and moral action: a theoretical perspective. *Developmental Review*, **3**(2), 178–210.

Codispoti, M., Ferrari, V., De Cesarei, A., Cardinale, R. (2006). Implicit and explicit categorization of natural scenes. *Progress in Brain Research*, **156**, 53–65.

Colby, A., Damon, W. (1992). *Some Do Care: Contemporary Lives of Moral Commitment*, New York: Free Press.

Cowell, J.M., Decety, J. (2015a). Precursors to morality in development as a complex interplay between neural, socioenvironmental, and behavioral facets. *Proceedings of the National Academy of Sciences*, **112**(41), 12657–62.

Cowell, J.M., Decety, J. (2015b). The neuroscience of implicit moral evaluation and its relation to generosity in early childhood. *Current Biology*, **25**(1), 93–7.

Decety, J., Cowell, J.M. (2018). Interpersonal harm aversion as a necessary foundation for morality: a developmental neuroscience perspective. *Development and Psychopathology*, **30**(1), 153–64.

Decety, J., Cowell, J.M., Lee, K., *et al.* (2015). The negative association between religiousness and children's altruism across the world. *Current Biology*, **25**(22), 2951–5.

Delorme, A., Makeig, S. (2004). EEGLAB: an open sorce toolbox for analysis of single-trail EEG dynamics including independent component anlaysis. *Journal of Neuroscience Methods*, **134**, 9–21.

Folstein, J.R., Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, **45**(1), 152–70.

Frimer, J.A., Walker, L.J. (2009). Reconciling the self and morality: an empirical model of moral centrality development. *Developmental Psychology*, **45**(6), 1669–81.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, **360**(1456), 815–36.

Gan, T., Lu, X., Li, W., *et al.* (2016). Temporal dynamics of the integration of intention and outcome in harmful and helpful moral judgment. *Frontiers in Psychology*, **6**, 1–12.

Greenwald, A.G., Nosek, B.A., Banaji, M.R. (2003). Understanding and using the implicit association test: an improved scoring algorithm. *Journal of Personality and Social Psychology*, **85**(2), 197–216.

Gui, D.Y., Gan, T., Liu, C. (2016). Neural evidence for moral intuition and the temporal dynamics of interactions between emotional processes and moral cognition. *Social Neuroscience*, **11**(4), 380–94.

Hajcak, G., MacNamara, A., Olvet, D.M. (2010). Event-related potentials, emotion, and emotion regulation: an integrative review. *Developmental Neuropsychology*, **35**(2), 129–55.

Hardy, S.A., Carlo, G. (2011). Moral identity: what is it, how does it develop, and is it linked to moral action? *Child Development Perspectives*, **5**(3), 212–8.

Hertz, S.G., Krettenauer, T. (2015). Does moral identity effectively predict moral behavior? A meta-analysis. *General Review of Psychology*, **20**(2), 129–40.

Jennings, P.L., Mitchell, M.S., Hannah, S.T. (2015). The moral self: a review and integration of the literature. *Journal of Organizational Behavior*, **36**(S1), S104–68.

Johnston, M.E., Sherman, A., Grusec, J.E. (2013). Predicting moral outrage and religiosity with an implicit measure of moral identity. *Journal of Research in Personality*, **47**(3), 209–17.

Killen, M., Smetana, J.G. (2015). Origins and development of morality. In Lerner, R.M., editor, *Handbook of Child Psychology and Developmental Science*, pp. 701–49, New York, NY: Wiley-Blackwell.

Krettenauer, T. (2005). Revisiting the moral self-construct: developmental perspectives on moral selfhood. In: *Self-Regulation and Autonomy: Social and Developmental Dimensions of Human Conduct*, 115–40. Cambridge: Cambridge University Press.

Krettenauer, T., Hertz, S.G. (2015). What develops in moral identities? A critical review. *Human Development*, **58**(3), 137–53.

Lahat, A., Helwig, C.C., Zelazo, P.D. (2013). An event-related potential study of adolescents' and young adults' judgments of moral and social conventional violations. *Child Development*, **84**(3), 955–69.

Lapsley, D.K., Narvaez, D. (2004). A social-cognitive approach to the moral personality. In: *Lapsley, D.K., Narvaez, D. (Eds.), Moral development, self, and identity (pp. 201–224)*, New York: Psychology Press.

Lichtenberg, J. (2010). Negative duties, positive duties, and the 'new harms'. *Ethics*, ***120***(3), 557–78.

Lopez-Calderon, J., Luck, S.J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, ***8***(April), 1–14.

Murray, S.O., Kersten, D., Olshausen, B.A., Schrater, P., Woods, D.L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, ***99***(23), 15164–9.

Paulus, M. (2018). The multidimensional nature of early prosocial behavior: a motivational perspective. *Current Opinion in Psychology*, ***20***, 111–6.

Perugini, M., Leone, L. (2009). Implicit self-concept and moral action. *Journal of Research in Personality*, ***43***(5), 747–54.

Pletti, C., Lotto, L., Buodo, G., Sarlo, M. (2017). It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations. *British Journal of Psychology*, ***108***(2), 351–68.

Pohling, R., Diessner, R., Strobel, A. (2018). The role of gratitude and moral elevation in moral identity development. *International Journal of Behavioral Development*, ***42***(4), 405–15.

Reed, A., Aquino, K., Levy, E. (2007). Moral identity and judgments of charitable Behaviors. *Journal of Marketing*, ***71***(1), 178–93.

Sarlo, M., Lotto, L., Rumiati, R., Palomba, D. (2014). If it makes you feel bad, don't do it! Egoistic rather than altruistic empathy modulates neural and behavioral responses in moral dilemmas. *Physiology & Behavior*, ***130C***, 127–34.

Schupp, H.T., Junghöfer, M., Weike, A.I., Hamm, A.O. (2003). Attention and emotion: an ERP analysis of facilitated emotional stimulus processing. *NeuroReport*, ***14***(8), 1107–10.

Schupp, H.T., Junghöfer, M., Weike, A.I., Hamm, A.O. (2004). The selective processing of briefly presented affective pictures: an ERP analysis. *Psychophysiology*, ***41***(3), 441–9.

Tassy, S., Oullier, O., Mancini, J., Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, ***4***, 250.

Weinberg, A., Hajcak, G. (2010). Beyond good and evil: the time-course of neural activity elicited by specific picture content. *Emotion*, ***10***(6), 767.

Yoder, K.J., Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. *Neuropsychologia*, ***60***, 39–45.