Original Articles

The selfless mind: How prefrontal involvement in mentalizing with similar and dissimilar others shapes empathy and prosocial behavior

Jasminka Majdandžić a,b,c,⇑, Sandra Amashauer a, Allan Hummer d, Christian Windischberger d, Claus Lamm a,c

a Social, Cognitive and Affective Neuroscience Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria
b Institute of Normal and Pathological Physiology, Slovak Academy of Sciences, Bratislava, Slovakia
c Cognitive Science Research Platform, University of Vienna, Austria
d Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Austria

Article info

Article history:
Received 12 January 2016
Revised 20 July 2016
Accepted 4 August 2016
Available online 29 August 2016

Keywords:
Mentalizing
Similarity
Self-inhibition
Empathy
Prosocial
fMRI

abstract

Engaging in mentalizing, i.e., reflecting on others’ thoughts, beliefs and feelings, is known to facilitate later empathy and prosocial behavior. Activation in dorsomedial prefrontal (dmPFC) areas during mentalizing has been shown to predict the extent of prosocial behavior. It is unclear, however, what cognitive process drives these effects: a simulation process in which the own mental states are used as a proxy for those of others (self-projection) or an effortful other-enhancement process in which one’s own perspective is overridden. In this fMRI study we examined the effects of mentalizing with similar and dissimilar others on behavioral and brain measures of empathy and prosocial behavior, to assess which cognitive process mediates the facilitative effects of mentalizing. Participants had to mentalize with two fictitious target persons, one of whom was manipulated to have similar thoughts and beliefs as the participant, while the other had dissimilar mental states. We then assessed participants’ behavioral and neural responses during an empathy for pain task and a prosocial behavior task. Similarity between participant and target person increased empathy and affiliation ratings, and mentalizing with dissimilar persons evoked increased activation in ventrolateral prefrontal cortex, the extent of which was inversely related with bias towards the similar person in empathy. Responses in two dmPFC regions were also predictive of later variations in subsequent empathy and prosocial behavior, either predicting overall prosociality and empathic concern (lateral dmPFC), or predicting reduced empathic bias towards the similar person and a lower response to self-related stressors in pain matrix areas (medial dmPFC). This pattern of results suggests that generating and enhancing other-related representations while overcoming one’s own perspective, rather than enhanced recruitment of self-projection processes, is driving the facilitative effects of mentalizing on later empathic and prosocial responses.

1. Introduction

Humans have the capacity, and spontaneous tendency, to reflect on the thoughts, beliefs and feelings of their conspecifics, known as mentalizing. Although the mental states of others cannot be directly observed, making inferences about them greatly increases our ability to predict others’ future behavior, which is a crucial advantage for individuals living in complex social groups.

It has been shown that engaging in mentalizing with others may also intensify feelings of empathy for these persons, as well as the willingness to engage in prosocial behavior towards them. For instance, explicitly instructing participants to take the cognitive or affective perspective of others increases both self-reported empathy and helping behavior (Oswald, 1996). These effects can also be induced by more implicit forms of mentalizing: in a previous fMRI study we have shown that having to take others’ affective or cognitive perspective to solve questions about them in a priming task impedes subsequent decisions to harm them in hypothetical moral dilemmas where one individual could be sacrificed to save the lives of several others. These behavioral effects were accompanied by increased activity in brain areas associated with aversive emotions and empathy during the decision phase (Majdandžić et al., 2012). Others have shown that the magnitude of the BOLD response in several parts of the dorsomedial prefrontal cortex (dmPFC), an area reliably associated with mentalizing, during a
mentalizing task predicts later altruistic helping (Waytz, Zaki, & Mitchell, 2012). This led the authors to conclude that prosociality is associated with a propensity to engage in social-cognitive thought, i.e. the consideration of others’ subjective experiences.

However, despite clear evidence for a link between mentalizing and diverse measures of prosocial attitude and behavior, it is unclear what cognitive process, as reflected in medial prefrontal activation during mentalizing, drives these effects. Presumably, mentalizing forces us to perceive others as full-blown individuals with beliefs and desires. Therefore, the effects of harmful or benefi-
cial acts on the mental states of these persons might also be repre-
sented in a more salient way. Yet, this altered, more full-blown “mental” representation of others could be achieved in several ways.

One candidate mechanism is based on the hypothesis that men-
talizing engages simulation mechanisms in which the self is used
as a “template” or anchor for making inferences about others
(Epley & Gilovich, 2001; Epley, Keysar, Van Boven, & Gilovich,
2004; Tamir & Mitchell, 2010). That is, inferring someone's
thoughts and feelings is achieved by mentally projecting oneself
onto his or her position, and imagining how one would think
and feel then. Support for a role of such self-projection processes
in mentalizing comes from behavioral studies showing that people
tend to assume that others hold the same opinions (Krueger & Clement, 1994; Ross, Greene, & House, 1977) and have the same
knowledge (Epley et al., 2004), as they themselves do. People are
worse at understanding others’ thoughts and feelings when those
mental states differ from their own (Niedenthal, Halberstadt,
Margolin, & Innes-Ker, 2000; Wellman, Cross, & Watson, 2001).
In addition, self-referential processing (i.e., reflecting on one's
own mental states) and mentalizing with others seem to rely on
largely overlapping activation patterns, especially in the medial
prefrontal cortex, indicating that they involve a common process
also on the functional level (Frith & Frith, 1999; Gallagher, Jack,
Roepstorff, & Frith, 2002; Gallagher et al., 2000; Jenkins, Macrae,
& Mitchell, 2008; Kelley et al., 2002; Mitchell, 2009; Mitchell,
Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006;
Northoff et al., 2006; Zaki & Ochsner, 2011). Activation in several
subregions of the medial prefrontal cortex is moreover modulated
by the perceived similarity between the self and the person men-
talized with (Mitchell et al., 2006). This might be seen as a further
indication that mentalizing involves self-projection processes,
given the assumption that people will engage in more self-
projection when thinking about similar persons, for which the self
a priori seems to be a more suitable model.

Evidence on how similarity exactly affects these activation pat-
terns is far from conclusive, though: while involvement of some
subareas seems to increase with self-other similarity (Jenkins
et al., 2008; Mitchell et al., 2005), other regions show a stronger
response for dissimilar persons (Mitchell et al., 2006; Tamir & Mitchell,
2010), or show inconsistent effects across different studies
(Mitchell et al., 2006; Tamir & Mitchell, 2010). Yet, in spite of
these inconsistencies, it is widely assumed that self-referential
simulation processes play at least some role in mentalizing.

Self-projection processes, then, would drive the effects of men-
talizing on subsequent prosocial behavior by increasing the extent
of “overlap” between mental representations of one’s own inner
states and those of the target person (Aron, Aron, & Smolian,
1992; Davis, Conklin, Smith, & Luce, 1996; Galinsky, Ku, & Wang,
2005). Such merging of self-other representations may facilitate
empathic responses (Chambers & Davis, 2012), since empathy also
seems to rely on shared representations: it involves activation of
the same neural networks as those involved in the first-hand ex-
perience of the affective state (Bernhardt & Singer, 2012; Lamm,
Decety, & Singer, 2011; Lamm & Majdandžić, 2015; Rügten,
Seidel, Riečanský, & Lamm, 2015; Rügten, Seidel, Silani, et al.,
2015; for reviews see Bernhardt & Singer, 2012 and Lamm & Majdandžić, 2015). Enhanced empathy may in turn motivate prosocial behaviors (Batson et al., 1997; Cialdini et al., 1987). Thus, according to this view, the effects of mentalizing on prosocial behavior can be explained on a functional level by the engagement of self-projection processes, with more self-projection being reflected in enhanced recruitment of medial prefrontal areas clas-
sically involved in mentalizing.

Yet, reflecting on the mental states of others evidently also
involves other processes than basic self-projection mechanisms.
From a theoretical point of view, successfully evaluating others’
mental states not only requires the ability to mentally project one-
self into their position, but also to inhibit one's own perspective
if needed, and to enhance a potential alternative perspective. In view
of this, it has been proposed that mentalizing involves a two-stage
process, in which the own perspective is used as a starting point or
“anchor”, which then gets adjusted in a serial and effortful process
(Epley & Gilovich, 2001; Tamir & Mitchell, 2010). Support for this
notion comes from studies showing that assessing perspectives
different from one's own is time-consuming, and that time con-
straints increase egocentric biases (Epley et al., 2004). This adjust-
ment process seems to rely on general self-inhibition abilities
(Launay et al., 2015). For instance, the ability to reason about men-
tal states different from one's own is strongly correlated with inhibi-
tory control in children (Carlson & Moses, 2001). Notably, medial
prefrontal brain areas tend to show higher activation during men-
talizing with dissimilar others, which is consistent with a serial
adjustment process from a self-based starting point, with more
adjustment requiring stronger involvement of these areas (Tamir
& Mitchell, 2010). Along with this self-inhibition process, a repre-
sentation of the other’s perspective should be generated and
actively enhanced. This fits with notions that the cognitive opera-
tions carried out by the brain network involved in mentalizing can
be framed in more general terms as a “fuzzy” mode of processing,
characterized by generating and manipulating approximate mental
estimates based on inexact, internally retrieved information
(Mitchell, 2009). Accordingly, mentalizing can be seen as an
instance of a more general cognitive ability in which “past experi-
ences are used adaptively to imagine perspectives and events
beyond those that emerge from the immediate environment”
(Buckner & Carroll, 2007, p. 49).

Thus, although mentalizing may involve setting an initial, self-
based “anchor”, it may be the subsequent effortful process of
inhibiting this own perspective so as to enhance an alternative per-
spective that is decisive for its facilitative effects on prosocial
behavior. In this view, making the own perspective less salient to
“make room” for the other's perspective implies a transition from
a self-centered to an other-directed orientation, which may foster
empathy and prosocial behavior. This is in line with earlier notions
that while empathy entails both self-related emotions (personal
distress) and other-related emotions (empathic concern) (Batson,
Fultz, & Schoenrade, 1987), the latter seem to better predict altru-
istic behavior (Batson, Duncan, Ackerman, Buckley, & Birch, 1981;
Batson et al., 1989). Notably, this suggestion does not dismiss the
possibility that mentalizing involves a form of self-projection; rather, it argues that this initial, automatic self-projection is insen-
sitive to the degree of similarity of the person mentalized with, and
is not driving changes in prosocial orientation. Instead, it is the
effectiveness of subsequent self-inhibition and other-enhancement that is mediating these effects.

With the present fMRI-study we aimed to clarify which of the
abovementioned cognitive processes: (1) self-projection or (2)
self-inhibition / other-enhancement, as reflected in increased con-
tributions by prefrontal “mentalizing” areas, is driving the enhanc-
ing effects of mentalizing on prosocial behavior, and how these
effects are mediated by empathy. To this end, we manipulated
the similarity of two allegedly real persons and assessed the effect of this manipulation on brain activity during mentalizing with these persons, and on brain and behavioral indices of empathy and prosocial behavior in two subsequent tasks.

Overall, we predicted similarity to affect empathy and prosocial behavior, with the more similar person evoking stronger empathic responses and prosocial tendencies than the dissimilar person. This prediction is in line with both proposed mechanisms: mentalizing with similar persons may elicit more initial self-projection but would also require less self-inhibition or adjustment, which in both cases leads to a more full-blown representation of the other as a human being with mental states, worthy of being treated well (Majdandžić et al., 2012; Waytz, Epley, & Cacioppo, 2010). However, our manipulation yielded different predictions about the pattern of responses across tasks, depending on the cognitive processes implied.

Assuming that we would replicate Waytz et al.’s (2012) finding that stronger activity in certain subregions of medial prefrontal cortex during mentalizing predicts greater overall prosociality, we formulated the following two opposing hypotheses: (1) If mentalizing-related activity in these medial prefrontal subregions reflects different extents of self-projection, then the activity should be higher for similar than for dissimilar persons, and the extent of this differential activity should predict (i.e. be positively correlated with) the extent of differentiating in empathy, prosocial behavior and other indices of affiliation (i.e. the extent to which the similar person is favored over the dissimilar person). (2) If, in contrast, the activity in these mPFC subregions during mentalizing reflects engaging in effortful self-inhibition / other-enhancement, then the activity should be higher for dissimilar than for similar persons, and the extent of this differential activity should counteract (i.e. be negatively correlated with) the extent of differentiating in empathy, prosocial behavior and affiliation measures (i.e. the extent of favoring the similar person).

In addition to assessing neural responses in medial prefrontal regions derived from previous literature, we also explored how response modulations in other brain areas during mentalizing with similar and dissimilar persons predicted later empathic and prosocial biases. Specifically, more ventrolateral prefrontal areas known to be specifically involved in self-inhibition aspects of mentalizing (Hartwright, Apperly, & Hansen, 2012; Samson, Apperly, Kathirgamanathan, & Humphreys, 2005; Samson, Houthuys, & Humphreys, 2015) can be expected to be sensitive to our similarity manipulation, and may also play a role in overcoming behavioral biases towards favoring more similar persons.

Thus, our study was designed to extend previous work on the links between behavioral and brain measures of mentalizing, empathy and prosocial behavior, by pinpointing the cognitive mechanism by which these links are mediated. To overcome the limitations of relying on associations between presumably invariant behavioral, brain and personality measures in individuals, we used a within-subject experimental design to tap into flexibly recruited processes across tasks and interaction partners.

2. Material and methods

2.1. Participants

We analyzed MRI data from 32 healthy right-handed female volunteers (age 22.7 ± 3.2 years, mean ± standard deviation). Only female participants were included to increase the homogeneity of our sample of MRI measurements. Data from two further participants were discarded due to excessive head-movement (i.e., exceeding 2 mm) during the scanning procedure. Due to technical problems, behavioral and/or imaging data from several participants were not available for the Mentalizing Task. This resulted in a final sample of 30 MR datasets and 30 behavioral datasets for this task and a full sample of 32 participants for the other tasks.

All participants had normal or corrected-to-normal vision, and gave written informed consent. Participants were paid €35 for their participation. The study was performed in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Medical University of Vienna.

2.2. Experimental procedures and design

The fMRI experiment consisted of four consecutive tasks: the Similarity Induction Task (I), the Mentalizing Task (II), the Empathy Task (III), and the Prosocial Task (IV). During all tasks, participants were led to believe that they were in a live webcam-based interaction with two other participants, who were in fact confederates who had been filmed in advance.

In the Similarity Induction Task, participants were presented with pairs of statements about mental states and had to select which statement described them best. They also saw which statements were selected by the other alleged participants, one of whom appeared to select mostly similar, and the other mostly dissimilar statements as the participant. In the Mentalizing Task participants were presented with new pairs of mental statements and had to decide, which statement would apply most to one of the alleged other participants. In the Empathy Task, both the participant and the two other alleged participants received a series of painful or non-painful shocks applied via an electrode to the hand. The Prosocial Task involved an ostensibly gambling and decision game in which the participant had to decide how a further series of painful shocks was to be distributed between her and the other presumed participants.

All tasks were performed in the MR scanner; however, the Similarity Induction Task did not involve functional MRI measurements and was performed at the beginning of the scanning session during acquisition of the MRI calibration and structural scans.

2.2.1. Preparation session (outside the MR scanner)

2.2.1.1. Cover story. The experiment was preceded by an instruction session outside the scanner. During this session, participants read and signed an informed consent form, and received written and oral instructions about the tasks. They were informed that the study was about how people get to know each other, and that throughout the experiment they would interact with two other participants over a live network connection. The other two participants were said to be in a different part of the research institute to undergo EEG recordings. Participants were then given a yellow coat to wear over their normal clothing, ostensibly to make them easily recognizable for the other two participants, who in turn would wear a red or blue coat for the same reason.

2.2.1.2. Pain calibration. The Empathy Task involved electrophysical stimulation applied via electrodes to the hand. In preparation for this, we assessed participants’ individual detection and pain thresholds. Stimulation was delivered using a Digitimer D55 Iso- lated Bipolar Constant Current Stimulator (Digitimer Clinical and Biomedical Research Instruments) via two concentric surface elec- trodes with 7 mm diameter and a platinum pin (WASP electrode, Specialty Developments) attached to the back of the left hand using adhesive plaster. This set-up was motivated by previous empathy studies using a similar paradigm (Hein, Silani, Preuschhof, Batson, & Singer, 2010; Rütgen, Seidel, Riečansky, et al., 2015; Singer et al., 2006), with one stimulation train lasting 1500 ms. During the calibration procedure the stimulation...
intensity was gradually increased to a level that participants reliably indicated as either “clearly detectable but not unpleasant” (threshold for non-painful stimulations) or “very painful but tolerable over multiple trials” (threshold for painful stimulations).

2.2.1.3. Webcam interaction. Participants were then seated behind a computer to engage in a short feigned Skype webcam interaction with the other two participants. They were told that all three participants could shortly meet each other over a one-direction webcam connection. They were instructed that talking was not possible but that they could briefly wave to each other. In two consecutive feigned video calls, participants then viewed short videos of the presumed other participants. The videos lasted 11 s and showed two female confederates, one of whom was wearing a red coat (the “red participant”) and the other a blue coat (the “blue participant”). Both confederates were silently looking into the camera with a friendly face, and then briefly waved. The participants were then ostensibly recorded in a similar way.

After the simulated webcam interaction, participants were photographed; the resulting photos were used in the subsequent experimental tasks, which also included photos of the two other alleged participants. Assignment of the two confederates and coat colors (red and blue) to the conditions (SIMILAR, DISSIMILAR) was counterbalanced across the 32 participants.

2.2.2. Experimental tasks (inside the MR scanner)

During the scanning session, participants were lying supine in the MR scanner. They could see the stimuli via a back projection system and a mirror that was attached to the head coil. An optical response button box, positioned on the upper right thigh, was used to record participants’ responses. Stimulus presentation and response recording were carried out using the MATLAB Toolbox Cogent 2000 v1.32 (http://www.vislab.ucl.ac.uk/cogent_2000.php).

2.2.2.1. Similarity Induction Task. In the Similarity Induction Task participants were presented with pairs of statements about mental states and had to select the statement that they felt applied most strongly to them. They subsequently saw which statements were selected by the presumed other participants, one of whom selected mostly (i.e. in 10 out of 12 cases) similar, the other mostly dissimilar statements.

All statements involved a description of some mental state (i.e. a cognitive style, preference, intention or desire). They were developed according to the following requirements: (1) statements in a pair referred to highly differing or mutually exclusive mental states, ensuring that persons who selected statement A would not agree with statement B, and vice versa; (2) statements that might seem socially desirable or suggestive of sensitive or prosocial (e.g. pain-sensitive, caring) personality characteristics were omitted, to avoid that participants’ empathic responses and social orientation towards the target persons would be affected by these presumed qualities, rather than by similarity per se. Furthermore, to ensure that statements in a pair were about equally likely to be selected, we tested a set of 71 statement pairs in advance in an online selection procedure involving 223 participants (121 female; mean age 22.5) and selected statement pairs of which each statement was selected with roughly equal frequency. This resulted in a battery of 48 statement pairs, of which for each participant a random selection of 24 was used in the Similarity Induction Task, while the remaining 24 were used in the Mentalizing Task (see below). An example of a statement pair is: (A) I prefer to take clear decisions (B) I prefer to leave all options open. A complete list of all statements is included in the Supplementary Material (Table S1).

Participants were instructed to select the statement that applied more to them in a spontaneous, immediate way, and to pay close attention to the statements selected by the other participants, since the goal of the task was to get to know each other. After two practice trials the task started. A trial started with a grey screen on which a photo of the participant was shown centrally, along with the instruction to select the statement that applied more strongly to her (see Fig. 1A). The two statements were shown on the left and right side of the photo, and could be selected by pressing the corresponding left (index finger) or right (middle finger) response button. The chosen statement was marked by a yellow selection frame that disappeared after 4 s, after which a photo of one of the other alleged participants was shown along with the same statements. This person seemingly also selected a statement, which was marked by a red or blue frame that remained on the screen for 4 s. After an intertrial interval (ITI) of 3 s, during which a fixation cross was shown, the next trial was presented.

The task consisted of 24 trials, with 12 trials showing the response of the similar person after the participant’s response and 12 trials showing the response of the dissimilar person. The order of similar and dissimilar trials was pseudo-randomized. To avoid a too-obvious response pattern, the similar person selected the same statement in only 10 of these 12 trials (83%), and the dissimilar person selected the opposite statement in only 10 of the 12 trials. The task lasted approximately 8–10 min, depending on individual decision times. No functional MRI measurements were acquired during the task.

2.2.2.2. Mentalizing Task. In this task, participants were presented with new statement pairs and had to choose which statement best described the mental states of their presumed co-participants. After detailed on-screen instructions and two practice trials, the task started. A trial started with a grey screen on which a photo of one of the presumed other participants was shown, along with the instruction to select the statement that would apply, according to the participant’s estimation, more strongly to this person (see Fig. 1B). After 3500 ms the two statements were shown on the left and right side of the photo, one of which the participant could select using the left or right response button. There were no response time constraints; response time typically varied between 3.5 and 8.5 s. The selected statement was marked with a red or blue frame that remained on screen for 2700 ms. After a variable inter-trial interval with fixation cross of 4–9 s, which was balanced between conditions, the next trial was presented.

The task consisted of 24 mentalizing trials, with 12 trials requiring mentalizing with the similar person (SIMILAR condition) and 12 trials requiring mentalizing with the dissimilar person (DISSIMILAR condition). The order of SIMILAR and DISSIMILAR trials was pseudo-randomized. After the mentalizing part, participants were presented with the 24 statement pairs again and from each pair had to select the statement that applied most strongly to themselves. This was done to obtain a measure of the extent to which the participants believed their co-participants to have similar or dissimilar mental states to themselves. The time course of these SELF trials was the same as for mentalizing trials. Total task duration depended on individual decision times and varied from 12 to 16 min, with an average of 14 min.

2.2.2.3. Empathy Task. In the Empathy Task, both the participants and their presumed co-participants received a number of painful and non-painful electrodermal stimulation trains. Behavioral ratings and functional MRI were acquired as measures of empathy for the alleged co-participants’ pain. The task was preceded by detailed on-screen instructions and presentation of two probe trials without stimulation to familiarize participants with the paradigm. At the start of each trial it was indicated which person
would receive stimulation, and what type of stimulation would be delivered. A trial started with the presentation of a photo of the person who would receive stimulation (see Fig. 1C), followed by a variable waiting interval of 2500–4000 ms. Then a lightning symbol was projected onto the photo, to indicate that the stimulation would start in 1 s (the anticipation cue). The color of this lightning symbol indicated the type of stimulation about to be delivered, with a blue symbol indicating non-painful stimulation and an orange one indicating painful stimulation. After 1000 ms the color of the lightning symbol changed to green (for non-painful stimulation) or red (for painful stimulation), to indicate that stimulation was applied. After 1500 ms, corresponding to stimulation duration, the lightning symbol and photo disappeared and a grey screen with fixation cross was shown. A randomly selected 20% of trials was followed by a visual analogue scale, on which participants had to indicate, on a five-point scale that ranged from 1 ("not at all") to 5 ("very much"), how unpleasant the preceding stimulation had been for them (in case of a trial in which they received stimulation...
themselves) or how unpleasant it had been for them to witness the preceding stimulation (in case of a trial in which one of the other persons had seemingly been stimulated). The rating scale stayed on screen until a response was given. Trials were separated by a variable ITI with a duration of 3.4–8.6 s, which was balanced across conditions.

The task consisted of 90 trials, structured in a $3 \times 2$ factorial design with factors Target Person (Self, Similar, Dissimilar) and Stimulation Type (Painful, Non-Painful), amounting to 15 trials per condition. For each condition a rating was collected three times, resulting in a total of 18 rating epochs. Order of the conditions was pseudo-randomized, omitting repetition of a condition in consecutive trials. Task duration was dependent on individual rating times and varied from 24 to 26 min, with an average of 25 min.

2.2.2.4. Prosocial Task. The Prosocial Task involved a decision paradigm designed to assess participants' prosocial orientation towards their alleged co-players, as well as the extent to which they differentiated in prosocial behavior towards the similar and dissimilar person. Participants were told that in the last part of the experiment a further series of painful stimulations would be delivered, and that in the upcoming gambling and decision task the three participants could influence the number of stimulations, as well as how these shocks would be distributed among them. The task was explained in on-screen instructions and preceded by two practice trials. In each trial the participants ostensibly played with either the red or the blue participant. A trial started with a photo of the participant they played with, remaining on screen for 4000 ms, followed by a row of colored shapes from which both participants had to select one (see Fig. 1D). Seemingly based on this choice, it was determined how many shocks would be delivered, and whether the participant, the other player, or the computer would decide about how these shocks would be distributed over the two players, which was shown on the next screen. Then three potential distributions of shocks were shown on the screen; if the participant was selected as the divider, she had to choose one of these distributions, which was then marked by a yellow frame that remained on the screen for 2500 ms. If the decision had to be made by the other player or the computer, the participants merely waited until a distribution was allegedly selected. Participants were not informed about the “chosen” distribution and were told that the other participant could not see their decision either. They were furthermore told that at the end of the task, a random trial would be delivered. Trials were separated by a variable ITI of 9–17 s.

The number of trials in which the participant had to decide was fixed to 16, with 8 trials involving the Similar and 8 trials involving the Dissimilar person. The other alleged participants were selected to decide in 6 trials each, and the computer decided in 4 trials for each condition. This resulted in a total of 36 trials. The distributions shown on the screen were pre-specified and balanced across trials involving the Similar and Dissimilar person. The total number of shocks varied from 14 to 17. The distributions that the participant could select denoted either (1) an equal distribution of the shocks (E); (2) a prosocial allocation (P) in which the participant target would receive 2 shocks more than the target person; (3) a highly prosocial allocation (HP) in which the target participant would receive 4 shocks more than the target person; (4) a selfish allocation (S) in which the target person would receive 2 shocks more than the participant; and (5) a highly selfish (HS) allocation in which the target person would receive 4 shocks more than the participant. Each of the following four combinations of distributions was shown twice per target person: [E-S-HS]; [E-P-HP]; [S-P-HP]; [HS-S-P]. The positions of the distributions on the screen were randomized to prevent the participants from noticing these regularities and from giving automatized responses. Thus, the paradigm allowed us to assess both general prosocial orientation (operationalized by the total number of shocks taken over) and potential decision biases in prosociality or selfishness in decisions affecting the Similar or Dissimilar person. Total duration of the task varied from 20 to 24 min, with an average of 21 min.

2.2.3. Post-experimental questionnaires

Following the Prosocial Task, we presented participants with a brief rating task in order to examine the effects of our manipulation on explicit measures of affiliation and interpersonal closeness. While still lying in the MR scanner, participants saw a photo of each co-participant; using a five-point scale, ranging from 1 (“not at all”) to 5 (“very much”), they had to rate them on the following attributes: (1) “How likable do you find this person?”; (2) “How connected do you feel to this person?”; (3) “How similar do you find this person to yourself?”; (4) “How cooperative do you find this person?”; and (5) “How familiar does this person seem to you?”. Order of the to-be-rated person and of the items was randomized across participants.

Participants were then removed from the scanner. Outside the scanner, they filled out a structured debriefing form and the German version of the Interpersonal Reactivity Index (IRI; Davis, 1983), after which they were debriefed and thanked for their participation.

2.3. Analysis of behavioral data

Behavioral data were analyzed using PASW 18.0 (SPSS Inc., Chicago, IL, USA), using parametric tests, or equivalent non-parametric tests if the assumption of normality was violated, as assessed with Shapiro–Wilks tests. For the Mentalizing Task we obtained the statements the participants selected for the Similar and Dissimilar person, and for themselves. From this we could calculate the occurrence of choices similar to participants’ own choices for the Similar and Dissimilar condition, and test for a difference between these frequencies using a paired t-test. The ratings obtained during the Empathy Task were entered into a $2 \times 2$ repeated measures ANOVA with factors Target Person (Similar, Dissimilar) and Stimulation Type (Painful, Non-Painful), after which post hoc t-tests were applied to assess specific effects of in more detail. For the Prosocial Task we obtained, both across conditions and for the Similar and Dissimilar person separately, the overall number of shocks taken over by the participant (rather than passed to the other player) as a measure of prosociality. Effects of Target Person (Similar, Dissimilar) on this prosociality measure were assessed with paired t-tests. The explicit affiliation and interpersonal closeness ratings of the Target Persons were entered into a $2 \times 5$ repeated measures ANOVA with factors Target Person (Similar, Dissimilar) and Scale (Likeability, Connectedness, Similarity, Cooperativeness, Familiarity); specific effects were assessed with paired t-tests. In addition, we calculated difference scores for each of these measures, by subtracting the values for the Dissimilar Target Person from those of the Similar Target Person. The size of these difference scores was used as a measure of the extent to which the similar person was favored over the dissimilar person, which we used in correlational analyses. Correlation analyses were performed using Spearman’s rank correlation coefficient ($r_s$), due to the ordinal scaling of the data. Alpha level was set to 0.05 for all analyses of behavioral data.

When multiple tests were conducted to test one hypothesis, we used Bonferroni-correction to correct for multiple comparisons. Greenhouse-Geisser corrections were applied to ANOVA P values.
if the assumption of homogeneity of covariances was violated (as determined by Mauchly tests of sphericity).

2.4. MRI acquisition

MRI data were acquired with a 3 T Siemens Tim Trio MRI system (Siemens Medical, Erlangen, Germany) using a 32-channel head coil for signal reception. Blood oxygen level-dependent (BOLD) sensitive functional imaging was performed using a multi-band accelerated echoplanar imaging (EPI) sequence with the following parameters: echo time [TE]/repetition time [TR] = 34/1300 ms, flip angle 65°, interleaved acquisition, 30 axial slices co-planar to the line connecting the anterior and posterior commissure, FOV 192 × 192 mm, matrix size 96 × 96, voxel size 2.0 × 2.0 × 3.5 mm, interslice gap 0.4 mm. Structural images were acquired before functional scanning using a magnetization-prepared rapid gradient-echo (MPRAGE) sequence (TE/TR = 4.21/2300 ms, 160 sagittal slices, voxel size = 1.0 × 1.0 × 1.1 mm, field of view = 256 mm).

2.5. Analysis of fMRI data

MRI data were analyzed using SPM12 (Statistical Parametric Mapping, http://www.fil.ion.ucl.ac.uk/spm). As part of data pre-processing, the time series for each voxel were first realigned temporally to TR/2 to correct for differences in slice time acquisition (Sladky et al., 2011). The image time series were then spatially realigned using a sinc interpolation algorithm that estimates rigid body transformations (translations, rotations) by minimizing head-movements between each image and the reference image. Subsequently, each participant’s mean EPI image was segmented into gray matter (GM), white matter (WM), and cerebral spinal fluid (CSF) using GM, WM, and CSF tissue probability maps provided by SPM12 and then spatially normalized to the International Consortium for Brain Mapping (ICBM) space templates (European brains) using both linear and nonlinear transformations. The other EPI images were then also spatially normalized by using the same transformation matrix as applied to the mean EPI image. Finally, the images were spatially smoothed using an isotropic 6 mm full-width-at-half-maximum Gaussian kernel.

The fMRI time series were analyzed using an event-related approach in the context of the General Linear Model (GLM). Trial-by-trial measures of the timing of the different epochs within each trial were either derived from stimulus presentation times or extracted from the button responses collected during the experiment.

Single-subject (“first level”) models of the Mentalising Task consisted of multiple regressors describing: (1) presentation of the target person photo before the statements appeared on the screen (collapsed over SIMILAR and DISSIMILAR trials); (2) and (3) the time period from presentation of the statements until response (i.e., the mentalizing phase), separate for SIMILAR and DISSIMILAR trials; (4) presentation of the just-chosen statement (collapsed over SIMILAR and DISSIMILAR trials); (5) the time period from presentation of the statements until response during SELF trials (6) presentation of the just-chosen statement during SELF trials. Single-subject models of the Empathy Task consisted of three regressors describing the trial phase in which the photo of the Target Person (SELF, SIMILAR, DISSIMILAR) was shown, before stimulation; 6 regressors describing, announcement (1000 ms) and delivering (1500 ms) of stimulation, separate for Target Person (SELF, SIMILAR, DISSIMILAR) and Stimulation Type (PAINFUL, NON-PAINFUL); and one regressor describing the rating epochs from onset till response. Single-subject models of the Prosocial Task consisted of regressors describing (1) presentation of the target person photo with whom the upcoming trial was played (collapsed over all conditions); (2) presentation of the shapes comprising the gambling task, until a shape was selected (collapsed over all conditions); (3) presentation of the screen showing which player had to decide (collapsed over all conditions); (4) and (5) for trials in which the participant had to choose a distribution of shocks: the time period from presentation of the distributions until response (i.e., the decision phase), separate for SIMILAR and DISSIMILAR target person; (6) the period in which the participant waited till the decision by the other person or the computer was ostensibly made; (7) for trials in which the participant had to choose a distribution of shocks: presentation of the just-selected distribution. Each effect was modelled on a trial-by-trial basis as a concatenation of square-wave functions. Each of these square-wave functions was then convolved with a canonical hemodynamic response function, as implemented in SPM12, in order to generate the regressors modeling the main effects described above. Head movement effects were accounted for by including the six rigid-body motion parameters (translation and rotation) as nuisance covariates into the models.

The statistical significance of the estimated evoked hemodynamic responses was assessed using t-statistics in the context of general linear model-based analyses, as implemented in SPM12. We were specifically interested in assessing effects of the mentalizing manipulation (SIMILAR, DISSIMILAR) on brain activity during the different tasks.

For the Mentalizing Task, we calculated first-level contrasts of the parameter estimates for the Mentalizing phase, separately for SIMILAR (i.e., mentalizing with SIMILAR person > baseline) and DISSIMILAR trials (i.e., mentalizing with DISSIMILAR person > baseline), where the IFI was considered the implicitly modelled baseline. These contrasts were entered into a random-effects second-level analysis using paired t-tests (i.e., SIMILAR > DISSIMILAR Mentalizing and the reverse contrast), in order to enable inferences on the population level (Penny & Holmes, 2004). Likewise, for the Empathy Task, we calculated first-level contrasts of the parameter estimates for the stimulation phase with respect to baseline, separately for Target Person (i.e., SELF, SIMILAR, DISSIMILAR) and Stimulation Type (PAINFUL, NON-PAINFUL). These six contrasts were entered into a random-effects second-level 3 × 2 within-subjects ANOVA. For the Prosocial Task we calculated first-level contrasts of the parameter estimates for the time phase in which a distribution had to be chosen, separately for decisions involving the SIMILAR and DISSIMILAR person, and entered these contrasts into a random-effects paired t-test on the second level.

Statistical inference was performed using a threshold of P = 0.05 corrected for multiple comparisons over the whole brain, using the Gaussian random fields approach at cluster-level with a voxel-level intensity threshold of P = 0.001 (Friston, Holmes, Poline, Price, & Frith, 1996). The SPM Anatomy Toolbox (Eickhoff et al., 2005) was used to guide anatomical and probabilistic cytoarchitectonic localization of the resulting clusters.

In addition to these whole-brain analyses we performed several Region of Interest (ROI) analyses based on ROIs taken from previous studies. These included a mentalizing-related region in the dorsal medial prefrontal cortex (obtained using a well-established mentalizing localizer task (Saxe & Kanwisher, 2003)) that was found in a previous study by Wrayt et al. (2012) to predict later prosocial behavior (MNI coordinates: [8 62 26]). We also included a second dorsal medial prefrontal region from the same study that predicted upcoming altruistic decisions during a donation game ([22 64 14]). Using the MarsBaR toolbox (Brett, Anton, Valabregue, & Poline, 2002) we created 9-mm spheres around these coordinates (since they would overlap at a size of 10 mm). For correlation analyses we extracted the mean beta values for several contrasts of interest from these ROIs using the Region of Interest Extraction (REX) toolbox (Whitfield-Gabrieli, 2009).
3. Results

3.1. Validity of experimental paradigms and manipulation

To confirm the validity of our paradigms and the effectiveness of our similarity manipulation, we first performed several general analyses. First, we assessed whether our similarity induction procedure indeed resulted in the similar target persons being regarded by participants as more similar to themselves than the dissimilar persons. In the Mentalizing Task participants chose the same statement as for themselves more often for the similar person (i.e., in 72% of the trials on average) than for the dissimilar person (in 43% of the trials; Wilcoxon signed-rank test; \( Z = 4.077; p < 0.001 \)). In addition, the similar target person was rated higher on all measures of affiliation and interpersonal closeness (\( F_{31} = 69.56; p < 0.001 \)), including similarity (\( T = 9.01; p < 0.001 \)), likability (\( T = 3.80; p = 0.001 \)), connectedness (\( T = 7.756; p < 0.001 \)), familiarity (\( T = 5.26; p < 0.001 \)) and cooperativeness (\( T = 2.39; p = 0.023 \)). Furthermore, feelings of connectedness with the similar person were predicted by the extent to which participants attributed the mental states that applied to themselves also to this person in 43% of the trials; Wilcoxon signed-rank test; \( Z = 4.077; p < 0.001 \)). Response times during mentalizing with the similar and dissimilar persons did not differ (paired \( t \) test; \( M_{\text{SIMILAR}} = 7.93 \) s; \( M_{\text{DISSIMILAR}} = 7.77 \) s; \( T = 0.880; p = 0.386 \)).

Assessing the neural areas involved in mentalizing across conditions in the Mentalizing Task, by testing for the contrast [Mentalizing > baseline], revealed an extensive network of areas typically involved in mentalizing (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014) including large portions of the medial prefrontal cortex, precuneus, and parts of the middle temporal gyrus and temporal pole. The Empathy Task, too, engaged the reliably established empathy for pain network consisting of bilateral anterior insula (AI) and anterior midcingulate cortex (aMCC) (Lamm et al., 2011), as well as numerous other regions (Table S2) when we tested for activation related to [Painful > Non-Painful stimulation] across both Target Persons. This indicated that the Empathy Task was effective in evoking empathic responses, which was further confirmed by the fact that activation for this contrast in the left AI (ROI obtained from Lamm et al. (2011)) was positively correlated with trait empathic concern, a sub scale of the IRI (\( r_s = 0.44; p = 0.012 \)), and that empathy ratings (i.e. unpleasant affect experienced when others were in pain) across conditions were correlated with trait empathy (\( r_s = 0.57; p < 0.001 \)). Overall empathy for pain ratings also correlated with prosocial behavior, as indexed by the number of shocks taken over from the other participants (\( r_s = 0.47; p = 0.006 \)). This suggests that across conditions, behavioral measures of empathy were predictive of prosocial behavior. Taken together these validity checks suggest that our experimental manipulations induced the expected neural and behavioral responses.

3.2. Main effects of similarity on mentalizing, empathy and prosocial behavior

3.2.1. Effects of similarity on brain activation during mentalizing

We tested for differential effects of mentalizing with dissimilar and similar persons on neural activity by contrasting activation during the mentalizing decision phase involving the DISSIMILAR > SIMILAR target person. This revealed bilateral activation covering the pars triangularis of the inferior frontal gyrus (Brodmann area (BA) 45; see Fig. 2A, Table 1), corresponding to the midventrolateral prefrontal cortex (mid-VLPFC) (Ridderinkhof, van den Wildenberg, Segalowitz, & Carter, 2004). On the left side, the cluster extended caudally into the pars opercularis (BA 44); on the right side it extended dorsally into the middle frontal gyrus. Testing for the reverse contrast by comparing mentalizing with SIMILAR > DISSIMILAR others did not yield any significant activation clusters.

To more closely examine the relation between this activation pattern and our similarity manipulation we extracted the mean beta values per participant from these clusters. We found that the size of the effect in the left VLPFC cluster, i.e. of the differential activation related to mentalizing with the dissimilar versus the similar person, was correlated with the extent of differentiation between the target persons in ratings of similarity (\( r_s = 0.39; p = 0.038 \)).

![A Mentalizing with Dissimilar > Similar person](image)

![B Medial dmPFC ROI](image)

![C Lateral dmPFC ROI](image)

Fig. 2. (A) Increased activity in bilateral ventrolateral prefrontal cortex (VLPFC) obtained by contrasting mentalizing with the DISSIMILAR > SIMILAR target person. (B) Illustration of the more medial spherical dorsomedial prefrontal (dmPFC) ROI (MNI coordinates: 8 62 26). (C) Illustration of the more lateral spherical right dorsomedial prefrontal (dmPFC) ROI (MNI coordinates: 22 64 14); both ROIs were obtained from Waytz et al. (2012).

### Table 1

MNI stereotactic coordinates of the local maxima of the activation clusters resulting from the contrast of mentalizing with the DISSIMILAR > SIMILAR Target Person in the Mentalizing Task.

<table>
<thead>
<tr>
<th>Area</th>
<th>Hemisphere</th>
<th>Peak MNI-coordinates</th>
<th>Cluster size (voxels)</th>
<th>( T ) value</th>
<th>( p ) value (corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentalizing with Dissimilar &gt; Similar person</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferior frontal gyrus (pars triangularis)</td>
<td>Left</td>
<td>(-36) 30 18</td>
<td>296</td>
<td>5.52</td>
<td>0.001</td>
</tr>
<tr>
<td>Inferior frontal gyrus (pars triangularis)</td>
<td>Right</td>
<td>38 30 14</td>
<td>539</td>
<td>5.45</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
This confirmed that response in this region was sensitive to subjectively experienced differences in similarity between the target persons.

3.2.2. Effects of similarity on empathy

Testing for effects of similarity on empathy ratings, using a 2 x 3 repeated measures ANOVA with factors Target Person (SIMILAR, DISSIMILAR) and Stimulation Type (PAINFUL, NON-PAINFUL) revealed a main effect of Target Person (F(2,67) = 7.26; p = 0.011), with similar persons evoking higher empathy ratings than dissimilar persons. Post hoc paired t tests showed that this effect was driven by non-painful trials, i.e. observing non-painful stimulations involving the similar person was experienced as more unpleasant than those involving dissimilar persons (T = 2.53; p = 0.017); no difference was found for painful stimulations (T = 1.09; p = 0.286). However, in the debriefing form we asked participants whether empathizing felt equally easy with both co-players, and found that of those participants who indicated that they found it easier to empathize with one co-player than with the other (37%), 92% indicated that it was easier with the similar co-player, indicating that overall, empathic responses to the pain of similar persons might not have been more intense, yet introspectively seemed to have been more easily provoked.

We assessed the effects of similarity on brain activity during the Empathy Task, by contrasting, within our 2 x 3 repeated measures ANOVA with factors Target Person (SELF, SIMILAR, DISSIMILAR) and Stimulation Type (PAINFUL, NON-PAINFUL), stimulation epochs involving the similar versus the dissimilar person. We tested for SIMILAR > DISSIMILAR effects (1) in PAINFUL stimulation trials, (2) NON-PAINFUL trials, and (3) collapsed over PAINFUL and NON-PAINFUL trials; none of these analyses revealed any significant activation clusters. Applying the reverse contrasts, i.e. DISSIMILAR > SIMILAR, did not yield any effects either. ROI analyses involving spheres in bilateral AI and aMCC obtained from Lamm et al. (2011) did also not result in significant effects. Thus, across participants, we did not observe any systematic effects of similarity on activation strength during empathy trials.

3.2.3. Effects of similarity on prosocial behavior

There was no main effect of similarity on prosocial behavior, as assessed with a paired t test comparing the number of shocks taken over from the similar versus the dissimilar person in the Prosocial Task (T = 0.23; p = 0.821). Similar to the findings of the Empathy Task, however, responses obtained with the debriefing form somewhat contradicted this result: of those participants who indicated that they tended to favor one co-player over the other in this task (53%), 82% expressed a preference for the similar co-player. This suggests that a tendency to favor the similar target person over the dissimilar one might have been present, but that its behavioral manifestation was overridden.

In line with the absence of behavioral differentiation, testing for effects of similarity on brain activity during the decision phase in which shocks had to be distributed between the participants and a presumed co-player, by contrasting SIMILAR with DISSIMILAR trials using a paired t test, did not reveal any effects; neither did the reverse comparison (DISSIMILAR > SIMILAR).

3.3. Associations between recruitment of prefrontal areas during mentalizing and empathy and prosocial behavior

3.3.1. Links between dmPFC activation during mentalizing and empathy and prosocial behavior

In order to address our main research question and examine associations between the involvement of dorsomedial prefrontal areas in mentalizing on the one hand, and measures of affiliation, empathy and prosocial behavior on the other hand, we performed a number of correlation analyses (see Table 2). For these analyses, we extracted mentalizing-related brain activation in two ROIs in dmPFC obtained from Waytz et al. (2012; see Methods section for more details): a medial one (MNI coordinates 8 62 26; see Fig. 2B) and a somewhat more lateral one (MNI coordinates 22 64 14; see Fig. 2C). Both of these areas were part of the network related to the overall mentalizing > baseline contrast, confirming their involvement in mentalizing processes. For each participant, we extracted the mean beta values in these ROIs for the contrast Mentalizing > baseline, as an index of overall involvement of the area in mentalizing. Using this procedure, we found that response in the more lateral dmPFC region during mentalizing, irrespective of Target Person, predicted prosociality, as indexed by the overall number of painful shocks taken over by the participant from her presumed co-players (r = 0.421; p = 0.020), thus directly replicating Waytz et al.’s (2012) finding. We did not find such a pattern for the more medial dmPFC ROI (r = 0.29; p = 0.123).

Although across participants we did not find significant differential activation in response to the two target persons in these ROIs, the more lateral region showed higher activation for the dissimilar than the similar person (mean contrast estimates Mentalizing DISSIMILAR > SIMILAR in units of SE: 1.692; p = 0.101). The more medial area did not show such a pattern (mean contrast estimates/SE: 0.049; p = 0.961). Yet, in both areas we found a trend positive correlations between differential activation related to mentalizing with the dissimilar versus the similar person, and the extent of differentiation between the target persons in ratings of subjective similarity (r = 0.32; p = 0.088 for the more lateral dmPFC region; r = 0.32; p = 0.090 for the medial dmPFC region). This indicated that both areas showed some subthreshold sensitivity to our manipulation.

We then explored further associations between the extent of differentiation in these regions (obtained by extracting the mean beta values for the contrast Mentalizing DISSIMILAR > SIMILAR) and other measures indicative of [bias in] empathy or prosocial behavior. First of all, differential response in both the more medial and the more lateral dmPFC area did not correlate with differentiation between similar versus dissimilar persons in empathy ratings, nor with the extent to which the similar person was favored over the dissimilar one in prosocial behavior.

We then assessed how differential (DISSIMILAR > SIMILAR) response in dmPFC during mentalizing was related to differential (SIMILAR > DISSIMILAR) brain response to the two target persons during empathy for pain. To this end, we created ROIs in the anterior middlingulate cortex (aMCC) and left and right anterior insula (AI), using the activation clusters for the contrast PAINFUL > NON-PAINFUL across target persons in the Empathy Task (see Table S2). From these three ROIs, we extracted the mean activation in response to painful stimulation of the SIMILAR > DISSIMILAR target person. This revealed two trend negative correlations between response in the more medial dmPFC region and differential response in left (r = 0.32; p = 0.087) and right anterior insula (r = 0.3; p = 0.091). No such correlations were found for the more lateral dmPFC region (all p > 0.203). Thus, stronger recruitment of the medial dmPFC during mentalizing with the dissimilar as compared to the similar person seems associated with a reduction in empathic bias for the similar person’s pain on the level of brain responses.

In the more lateral dmPFC, the extent of differential (DISSIMILAR > SIMILAR) response during mentalizing was furthermore positively correlated with trait empathic concern (r = 0.54; p = 0.002) and overall empathy for pain ratings (r = 0.44; p = 0.015). Thus, persons high in empathic concern tended to more strongly recruit this part of dmPFC when mentalizing with dissimilar, as compared...
with similar, persons. Persons high in empathic concern also tended to show a stronger behavioral differentiation between the two persons in the mentalizing task, as reflected in a stronger tendency to assign mental states that applied to themselves to similar persons and to assign opposing mental states to dissimilar persons ($r_s = 0.37; p = 0.047$).

### 3.3.2. Links between VLPFC activation during mentalizing and empathy

In addition to assessing the role of the dmPFC, we correlated the strength of the differential VLPFC activation found in the whole-brain analysis for the contrast DISSIMILAR > SIMILAR with the amount of differentiation between the similar and the dissimilar person in the Empathy Task and the Prosocial Task. This revealed a negative correlation between left VLPFC activation and the difference in empathy for pain ratings involving the similar versus the dissimilar person ($r_s = -0.42; p = 0.021$). To examine what drove this effect, we median split the sample according to the size of the differential (i.e., DISSIMILAR > SIMILAR) activation in the VLPFC and found that participants with strong differentiation showed increased activation of VLPFC for the dissimilar person (mean parameter estimates DISSIMILAR/SIMILAR: 0.963/0.543), rather than participants with low differentiation showing reduced activation for the similar person (mean parameter estimates DISSIMILAR/SIMILAR: 0.530/0.487). Thus, stronger recruitment of this area during mentalizing with the dissimilar person was associated with a reduction in biased empathy for the similar person later on. We did not find any correlation between differential VLPFC activation and differentiation in prosocial behavior towards the similar versus dissimilar person.
3.3.3. Links between prefrontal activation during mentalizing and self-centered pain processing

Instigated by the observation that participants who showed increased differential recruitment of lateral dmPFC during mentalizing with the dissimilar person tended to be higher in trait empathic concern, we wanted to explore how participants’ mentalizing response covaried with their self-centered, rather than other-related stress responses. To this end, we assessed correlations between mentalizing responses in lateral dmPFC and VLPFC and a more state-like measure of participants’ self-related distress responses: their responses while they underwent painful stimulation themselves. In order to isolate the affective (rather than the sensory) component of this pain response, we created ROIs that captured the overlap between self-related pain responses and empathy for pain responses, by testing for the conjunction between the contrast PAINFUL > NON-PAINFUL across target persons in the Empathy Task (see Table S2) and the contrast SELF PAINFUL > NON-PAINFUL. This resulted in activation clusters in the left insula (MNI coordinates [−32 22 2]; size 442 voxels), the right insula ([34 24 2]; size 271 voxels) and in aMCC ([10 10 58]; size 483 voxels), all of which are part of the “pain matrix”. From these ROIs we extracted mean activation in response to SELF PAINFUL > NON-PAINFUL stimulation. Activation in the left anterior insula ROI showed a trend positive correlation with behavioral painfulness ratings (rs = 0.34; p = 0.054), confirming the adequacy of this ROI for assessing self-related distress experienced during first-hand pain.

Correlation analyses then showed that response in all three pain matrix areas was negatively correlated with involvement of the more medial dmPFC area in mentalizing (aMCC; rs = −0.40; p = 0.029; left AI: rs = −0.44; p = 0.014; right AI: rs = −0.64; p < 0.001). In other words, participants who recruited this dmPFC area more strongly during mentalizing tended to show a less intensive self-related affective neural response to pain. We did not find any significant correlations with overall or differential mentalizing responses in the more lateral dmPFC area. Finally, we also found a negative correlation between the size of the neural response to self-experienced pain in the right AI and the extent overall mentalizing-related activity in the left VLPFC cluster that was found in the whole brain analysis (rs = −0.42; p = 0.022). Thus, individuals with a less intense self-related pain response showed a more pronounced overall mentalizing response in the medial dmPFC area and in the left VLPFC.

4. Discussion

In the present study we examined the effects of mentalizing with similar and dissimilar others on behavioral and brain measures of empathy and prosocial behavior. We were specifically interested in assessing which cognitive processes mediate the facilitative effects of mentalizing and associated dorsomedial prefrontal activity on prosocial behavior known from previous studies. To this end, we presented participants with a mentalizing task in which they had to estimate the mental states of a similar and a dissimilar fictitious other person. We then assessed participants’ behavioral and neural responses during an empathy task, in which these persons received painful hand stimulation, and in a prosocial task, in which participants had to decide on how to distribute a series of further painful shocks between themselves and each person. In addition to predicting main effects of target person similarity on empathy and prosocial behavior, we reasoned that two cognitive processes potentially driving the effects of mentalizing, self-projection and other-enhancement/self-inhibition, would be associated with different patterns of effects. That is, we expected self-projection processes to be reflected in greater mentalizing-related dmPFC involvement for similar than dissimilar persons, with the extent of this differential activity predicting the extent to which the similar person would be favored in empathy, prosocial behavior and other affiliative measures. Effortful other-enhancement and self-inhibition, in contrast, was expected to be associated with more pronounced dmPFC responses for dissimilar than similar persons, with the extent of this differential activity countering a bias towards the similar person in empathy, prosocial behavior and affiliation.

4.1. Main effects of similarity on behavioral and brain responses

Overall, our tasks were effective in evoking genuine mentalizing and empathic processes, as was confirmed by several manipulation checks on brain and behavioral data. The similarity manipulation did not induce a clear-cut bias in empathy or prosocial behavior, though. Overall, similar persons evoked higher self-reported empathy, although this effect was driven by non-painful trials. In the prosocial task the two persons were treated equally. Yet, the similar person received markedly higher ratings on all affiliation and interpersonal closeness scales. Besides, explicit statements during debriefing indicated that participants tended to experience a bias towards the similar person in both empathizing and prosocial behavior but seemed to have overridden behavioral manifestation of this tendency. Thus, although overall, participants were reluctant to overtly treat their presumed co-players differently, similar persons tended to be evaluated more positively.

Mentalizing with the dissimilar person did not yield any differential activation in dmPFC areas but evoked robust responses in bilateral mid-ventrolateral prefrontal cortex (mid-VLPFC). This area is known to be recruited under conditions of response ambiguity in which a response has to be selected from several alternatives (Levy & Wagner, 2011) and is thought to be involved in inhibiting prepotent but inappropriate responses (Ridderinkhof et al., 2004). This is in line with the notion that having to mentalize with the dissimilar person creates more response uncertainty: while the mental states of the similar person seem fairly predictable (i.e., one can rely on one’s own mental states to predict them), those of the dissimilar person are harder to predict, resulting in both more response uncertainty and a stronger need to suppress the initial urge to select the mental state that applies more to oneself. Notably, recent studies have provided more direct evidence for a specific and critical role for this area in the ability to resist interference from one’s own perspective when reasoning about other people’s mental states (Hartwright et al., 2012; Samson et al., 2005, 2015; see also Samson et al., 2005; Schurz & Tholen, 2016). Thus, intensified VLPFC response during mentalizing with the dissimilar person may reflect efforts to actively generate an accurate response under uncertain conditions (i.e., when the own mental states are not suitable as a template), thereby inhibiting the own dominant response. This interpretation is further supported by the correlation of differential VLPFC response with the extent of subjective differentiation between the similar and dissimilar person in ratings of similarity.

4.2. Predictive relations between mentalizing and later tasks

To examine predictive links between prefrontal involvement in mentalizing and later behavioral and brain measures of empathy and prosocial behavior, we performed several correlation analyses. These analyses indicated that differential recruitment of prefrontal areas within and outside of the neural network typically involved in mentalizing is associated with later variations in empathy and the intensity of self-centered stress responses.

More specifically, mentalizing-related responses in an ROI in lateral dmPFC, found in a previous study to be predictive of
altruism, predicted overall prosocial behavior also in the present study, further confirming the genuineness of this link. This lateral dmPFC area and the left VLPFC area found in our main imaging analysis tended to be more active during mentalizing with the dissimilar person. In addition, in all three areas, increased differential response during mentalizing with the dissimilar person was associated with increased differentiation in similarity ratings between the target persons. Thus, all areas showed some sensitivity to (subjectively experienced) similarity of the target persons, with stronger neural responses for dissimilar than similar persons.

In VLPFC, the extent of the differential response during mentalizing seemed associated with a diminishment in favoring of the similar person in the empathy task, as suggested by negative correlations with differentiation in empathy. A similar pattern was found for the medial dmPFC area: in this area, increased differential response during mentalizing with the dissimilar persons predicted a reduction of empathic bias for the similar persons in the left and right anterior insula, which are part of the empathy for pain network.

In the lateral dmPFC, differential response during mentalizing was furthermore associated with higher self-reported empathy for pain during the Empathy Task and with trait empathic concern, that is, with a disposition towards engaging in other-oriented emotions. In line with this, a negative association between mentalizing-related prefrontal responses and self-centered orientation was also suggested by a negative correlation between overall recruitment of medial dmPFC and left VLPFC during mentalizing and the extent of distress during one’s own shocks, as indexed by BOLD response in pain matrix areas. The link between reduced self-centered pain responses and self-inhibition abilities will have to be confirmed by future research though, to disentangle inhibition processes from individual differences in pain sensitivity.

4.3. Implications of findings for hypotheses

Although different prefrontal areas involved in mentalizing seem to modulate subsequent empathic and prosocial responses in different ways, the overall pattern of findings provides some novel insights into the cognitive process that might drive the known facilitative effects of mentalizing on prosociality. More specifically, our findings seem to suggest that this process reflects the extent of effortful enhancement of other perspectives, rather than the extent of self-projection. That is, rather than engaging in simulation processes in which the self is projected into the other’s situation, generating a prediction of the other’s perspective, thereby inhibiting the more dominant own perspective, seems to be the mechanism in mentalizing that enhances subsequent prosocial orientation.

This view is supported by several findings (see Table 2). First, involvement of one of these areas, the lateral dmPFC, in mentalizing is predictive of prosociality. Second, response in this area, as well as in the left VLPFC and to some extent the medial dmPFC, tends to be more pronounced for dissimilar than similar persons. As mentioned before, although VLPFC was not predictive of prosociality, this region has been associated with the inhibition of immediate, self-related responses so as to enhance an alternative perspective. And, third, in medial dmPFC and left VLPFC the extent of increased mentalizing-related activity for the dissimilar seems to counteract subsequent empathic bias towards the similar person.

Engaging in this other-enhancement process thus seems to help people to overcome a self-other distinction that would otherwise impede prosocial responses. The propensity to engage in this process furthermore seems more pronounced in people high in dispositional empathic concern and in overall subjective empathic responses during the Empathy Task. More generally, it seems to be associated with a better ability to disengage from self-centered responses, which may result in a general shift in processing mode from (effortless) self-centered to (effortful) other-oriented processing, as suggested by the finding that it is related to less intense responses to self-related stressors.

4.4. Relevance of our findings in light of other studies

With this study we elaborated on previous studies which separately showed that (1) mentalizing facilitates empathy and prosocial behavior (e.g. Majdandžić et al., 2012; Oswald, 1996); (2) activity in dmPFC areas during mentalizing is predictive of its prosocial effects (Waytz et al., 2012) and (3) mentalizing with similar and dissimilar others differentially recruits dmPFC areas (Tamir & Mitchell, 2010). Rather than merely confirming these links, we aimed to clarify the functional role of dmPFC and other prefrontal areas in mediating them, that is, assess how mentalizing facilitates prosocial behavior and empathy. Using a similarity manipulation, we sought to disentangle the contributions from two potential mechanisms: self-projective simulation and other-enhancement requiring self-inhibition. The use of a within-subjects design with three consecutive tasks involving the same two fictitious target person allowed us to assess associations between differential mentalizing, empathic, and prosocial responses across tasks.

The similarity manipulation evoked only moderate biases in empathy across participants, and no bias in prosocial behavior. This might have been due to robust inequity aversion tendencies in many participants, counteracting inclinations to favor one target person. However, comparing patterns of differentiation across tasks within participants allowed for a more fine-grained image, in which several prefrontal areas classically involved in mentalizing (dmPFC) and in response control under uncertain or conflicting conditions (VLPFC) show patterns of correlations that suggest a decisive role for self-inhibition and other-enhancement rather than pure self-projection processes. Notably, these areas and the broader networks of which they are part of most likely do not contribute in the same way to mentalizing and its facilitative effects on prosocial behavior.

More specifically, activity in the more lateral dmPFC region is predictive of general prosociality, but also shows some subthreshold sensitivity to similarity (i.e. stronger responses during mentalizing with the dissimilar person) which is more pronounced in participants high in trait empathic concern. This area was shown in a previous study to predict an upcoming altruistic decision in which one’s own interests are overridden to enhance others’ welfare (Waytz et al., 2012) suggesting that in our study too it might be involved in enhancing a perspective different from one’s own, thereby suppressing the own current perspective.

Activity in the more medial dmPFC area, a “classical” mentalizing region, is not predictive of the extent of overall prosocial behavior, but shows some sensitivity to our similarity manipulation in the form of increased recruitment for dissimilar persons, which seems to counteract a later neural bias in empathy in favor of the similar person. Mentalizing-related response in this area is also strongly negatively related with aversive responses to self-related distress. As suggested by previous studies, the medial dmPFC seems to be involved in generating representations based on uncertain, internally retrieved information (Mitchell, 2009) while disengaging from more immediate, external inputs.

The mid-VLPFC falls outside the typical “mentalizing network” (Frith & Frith, 1999; but see Schurz & Tholen, 2016); however, it is strongly sensitive to our similarity manipulation and seems to be associated with a downregulation of empathic bias and self-related aversive responses. As mentioned above, increased recruitment of this area might denote enhanced effort to generate an
appropriate response (in this case, an estimation of someone’s mental states) under uncertain conditions. This seems also reflected in a better ability to disengage from immediate, prepotent responses, such as pain responses.

Overall, the combined recruitment of these prefrontal areas during mentalizing might thus contribute to a salient representation of the other person as a full-blown individual with mental states, along with a disengagement from the own current perspective. This might pave the way for further behaviors requiring other-oriented processing, such as empathy and prosocial behavior.

Our study is in line with previous evidence that medial prefrontal cortex recruitment during socio-cognitive thought is correlated with mentalizing accuracy (Zaki, Weber, Bolger, & Ochsner, 2009), trait empathy (Wagner, Kelley, & Heatherton, 2011) and prosocial behaviors (Masten, Morelli, & Eisenberger, 2011). Importantly, it complements previous findings that engaging in mentalizing promotes later altruism (Waytz et al., 2012) by pinpointing a more specific mechanism that might drive these effects. It also extends notions that mentalizing might involve an anchoring-and-adjustment process (Epley & Gilovich, 2001; Epley et al., 2004) that requires increased engagement in the case of dissimilar others (Tamir & Mitchell, 2010), by showing that the flexible recruitment of such processes may shape future social behavior. Although medial prefrontal areas are broadly involved in cognition, which is clearly dismissed by the term “mentalizing region”, our study is in line with the general notion that this region is specialized in generating representations of not readily observable entities based on internally retrieved information (Mitchell, 2009), and that higher levels of prediction uncertainty increase processing demands in this area (Köster-Hale & Saxe, 2013).

4.5. Limitations and interpretational issues

It should be noted that our similarity manipulation exerted only weak effects on behavioral and brain indices of empathic and prosocial responses. That is, we did not find a main effect of similarity on prosocial preference, and observed only partial effects on empathy measures. This may have several reasons. First, our similarity manipulation was deliberately subtle: despite their different preferences and opinions, overall both target persons may have seemed quite similar to the participants. We used such a subtle manipulation since we wanted to exclude a priori effects of liking (due to e.g. similar versus different political views) in order to have a cleaner measure of the mere effects of engaging in mentalizing on empathy and prosocial behavior. Yet, these differences may have been too delicate to produce robust differences in neural processing during empathy and prosocial behavior. More specifically, although previous research has shown that activity in the empathy for pain network (i.e., aMCC and anterior insula) can be modulated by the social relation with the target person, these studies tended to use more extreme social manipulations that might have resulted in a complete blocking of empathic responses (e.g. Singer et al., 2006). The empathy for pain network may be less sensitive to more subtle differences in social relation with the target persons. Second, the Empathy Task was designed to induce “top-down” rather than “bottom-up” empathy responses. That is, rather than being evoked by emotion contagion (e.g. facial expressions) or visual displays of physical harm, empathic responses had to arise from internally generated representations based on abstract cues (the lightning symbol). This was done to eliminate contributions from factors such as facial appearance, eye contact, pain expression, and bottom-up vicarious responses to visual stimuli in our tasks, to test for specific effects of mentalizing on empathy in a maximally stringent manner. It may however have resulted in decreased sensitivity to the rather subtle target person manipulation. Third, the design of the Prosocial Task allowed participants to treat the target persons equally. Therefore, the notoriously strong tendency for people to avoid inequity may have overridden behavioral manifestation of covert preferences for the similar target person in our participants. To evoke more robust effects, future studies should thus rely on stronger similarity manipulations, and should aim to induce stronger overall empathic responses. In addition, preferences in prosocial behavior might be assessed more effectively using a forced choice paradigm in which participants are required to favor one target person over the other (Majdandžić et al., 2016).

Although we thus did not find robust main effects of similarity, calculating differential responses to the two target persons as a proxy for socio-affective bias, and relating the size of these responses to differential neural responses during mentalizing allowed us to make inferences on the cognitive process underlying the prosocial effects of considering others’ mental states. It should be noted, however, that these inferences are based on indirect evidence. That is, they are based on the central assumption that the relative involvement of self-projection processes is stronger during mentalizing with similar persons, while processing demands for self-inhibition are stronger during mentalizing with dissimilar persons. While we consider this a valid assumption, future research should confirm our conclusions by using more direct measures to tap into these processes. For instance, localization tasks could be used to pinpoint neural signatures of self-inhibition versus self-projection processes a priori. Alternatively, explicit instructions or task constraints could be used as a further manipulation to vary the extent of self-projection, other-enhancement, and self-inhibition, after which ensuing modulations of empathy and prosocial behavior could be assessed.

Importantly, with the present account, which emphasizes the importance of self-inhibition and other-enhancement in facilitating prosocial behaviors, we do not intend to preclude a role for self-projection processes in mentalizing. Indeed, it is difficult to account for the widely observed facilitative effects of similarity on the speed and accuracy of mental state attribution (Epley and Gilovich, 2001; Niedenthal et al., 2000) without presume some form of self-based simulation. As has been suggested by others, mentalizing likely involves a dual-stage process involving both self-projection and subsequent self-inhibition (Epley et al., 2004). Yet, we suggest that, of these component processes, the one that seems particularly critical for the facilitative effects of mentalizing on prosocial orientation seems to be the enhancement the other’s perspective at the expense of one’s own perspective. This will be easier to accomplish if the other is more similar to oneself, since it requires less adjustment of or disengagement from the self template. However, that does not imply that the prosocial effects of mentalizing are a direct result of the extent of this initial self-projection. Rather, our findings suggest that it seems to be the subsequent effortful enhancement-inhibition processes that are key to the vividly represented “mental” other deserving to be treated well.

Nonetheless, it might be argued that our pattern of findings could in principle also reflect differences in the extent of self-projection. This seems not plausible for several reasons. First, if more pronounced self-projection would be reflected in more dmPFC activation with respect to baseline, we would expect activation predicting prosociality to be stronger for similar than dissimilar persons. This is not the case. If on the other hand more self-projection would be reflected in more efficient, and hence less strong activation with respect to baseline, we would expect the activation (although stronger for dissimilar than for similar persons) to be negatively correlated with overall prosocial behavior. This is also not consistent with the present findings; nor with those of Waytz et al. (2012).