The Neurobiology of Moral Cognition: Relation to Theory of Mind, Empathy, and Mind-Wandering

9

Danilo Bzdok, Dominik Groß, and Simon B. Eickhoff

Contents

Abstract

A sense of morality forms the fabric of human societies. There is an ongoing debate whether the cognitive and emotional sources of moral decisions might be closely related to theory of mind, an abstract–cognitive capacity, and empathy, an automatic–affective capacity. That is, moral decisions are believed to imply representation of other individuals' thoughts and emotional states, respectively. Moreover, it has been noticed that neural activation patterns during moral cognition are very similar to the brain areas engaged during mind-wandering, i.e., neural correlates of an endogenously controlled state in the absence of a specific mental task.

Investigation of the neural substrates underlying moral cognition was greatly facilitated by the advent of neuroimaging techniques. This growing number of observation on brain activation patterns during the aforementioned tasks now

D. Bzdok • S.B. Eickhoff (✉)
Institut für Neurowissenschaften und Medizin (INM-1), Jülich, Germany
e-mail: danilo.bzdok@rwth-aachen.de; s.eickhoff@fz-juelich.de

D. Groß
Medical School, RWTH Aachen University, Aachen, Germany
e-mail: dgross@ukaachen.de

provides rich substrates for a quantitative integration of the current literature. Such large-scale integration, identifying brain areas consistently engaged by moral, social, empathic, and unconstrained cognition, then provides a quantitative basis for the comparison of their neuronal implementation. This chapter thus quantitatively assesses and reviews the neurobiological relationship between the moral network and the neural networks subserving theory of mind, empathy, and unconstrained cognition.

In conclusion, the neural network subserving moral decisions probably reflects functional integration of distributed heterogeneous networks, is dissociable into cognitive and affective components, as well as highly similar to the brain's default activity pattern.

Introduction

Moral behavior has classically been thought to be based on rational (i.e., rather conscious, controlled, and effortful) thinking. Rational explanations assumed that moral behavior arises from a conscious weighing of different rules, norms, and situational factors. In contrast, the role of emotion and intuition in moral thinking (thought to represent an unconscious, automatic, and effortless way) has been less often emphasized (Haidt 2001). Emotional explanations emphasized the influence of intuitive, subconscious emotional states that are rapidly evoked by a given situation. Taken together, *abstract–inferential* and *automatic–emotional* processing have been implicated and contrasted in philosophical, psychological, and biological accounts of moral behavior.

The association of psychological categories, such as decision-making or emotional influences thereon, with brain activity in the underlying neural networks has been greatly promoted by the development of functional neuroimaging. Positron emission tomography (PET) and the noninvasive functional magnetic resonance imaging (fMRI) allow the in vivo investigation of functional specialization in the human brain. Based on local changes in cerebral blood flow and glucose or oxygen metabolism, these techniques allow inference on regional increases in neural activation during the performance of specific tasks. Often, the neural correlates of a given task (reflecting a mental process of interest, e.g., moral decision-making) are isolated by subtraction of the activation measured during a closely related task (a control task, such as semantic or abstract decisions) that is supposed to carry the same confounds (e.g., reading) but not to evoke the mental process of interest. Over the last two decades, functional neuroimaging has then provided a wealth of information on the cerebral localization of various psychological tasks, including moral decision-making.

Notions of rationality and emotionality also serve as explanations in contemporary imaging research on the neural correlates underlying moral decisions (moral cognition). Joshua Greene (in the USA) and Jorge Moll (in Brazil) can probably be considered the protagonists in the ensuing debate. Results from fMRI studies by Greene and colleagues (Greene et al. 2001; 2004) were consistently interpreted as revealing a neuroanatomical dissociation between emotional responses and

subsequent explicit cognitive modulations in moral cognition. However, fMRI findings by Moll and colleagues (Moll et al. 2005a, 2006; Moll and Schulkin 2009) were interpreted as revealing various different psychological processes without any specific neural correlates, including group-oriented (i.e., pro-social) and self-oriented (i.e., egoistic) affective drives, in moral cognition.

It is important to note that the rational and emotional facets of moral cognition are, by theoretical arguments and empirical research, closely related to two other aspects of social interaction: theory of mind (ToM) and empathy. ToM refers to the ability to contemplate other's thoughts, desires, intentions, and behavioral dispositions by *abstract inference* (Frith and Frith 2003; Premack and Woodruff 1978). Evidently, moral decisions are influenced by whether or not an agent's action is perceived as intentional or accidental, which crucially relies on mental state inference, i.e., ToM. Consistently, behavioral data from subjects with high-functioning autism, known for impoverished ToM abilities, suggested an involvement of ToM in moral judgments, given that these individuals relied less on the agent's intentions and more on action outcomes (Moran et al. 2011). Empathy, on the other hand, refers to *intuitively* adopting somebody's *emotional state* while maintaining the self–other distinction (Decety and Jackson 2004; Singer and Lamm 2009). More specifically, empathy can be subdivided into (partially intertwined) emotional and cognitive components (Shamay-Tsoory et al. 2009). Phylogenetically and ontogenetically earlier "emotional empathy" is closely related to emotion contagion and simulation systems, while later developing, more advanced "cognitive empathy" is more related to perspective-taking and imagination systems. In particular, empathy is different from and tends to precede sympathy, which does not necessarily result in identical affect in the observer, but for instance in pity or compassionate love. In moral decisions, experiencing empathy was shown to alleviate harmful actions towards others (Eisenberger 2000). Conversely, deficient empathy skills are a clinical hallmark of psychopathic subjects and are believed to contribute to their morally inappropriate behavior (Hare 2003). Taken together, the aforementioned debate on the contribution of cognitive and emotional factors to moral decision-making may be reframed as the question whether the neural correlates of moral decisions are closer related to those of ToM or empathy – or whether there is a distinct moral module serving moral cognition.

The Neural Architecture of Moral Cognition

In the present analysis we tried to avoid the pitfalls of descriptive verbal summaries of neuroimaging results, which are inevitably subjective and hence potentially biased. Such critical verbal analyses tend to focus on a limited number of preselected aspects and tend to be biased by the authors' own adherence to a specific research area. In contrast to classical review articles, coordinate-based meta-analysis (CBMA) is hypothesis-free, data-driven, and, hence, objective by algorithmically weighing all results equally. As the CBMA method is not skewed by subjectivity, it precludes overinterpretation of expected, easily interpretable

findings and neglect of unexpected, barely reconcilable findings in neuroimaging research. CBMA might therefore help to point out consistent, yet frequently ignored findings (Eickhoff and Bzdok 2012).

Rather than a critical verbal discussion, we therefore recently performed a quantitative CBMA of the neuroscientific literature on moral cognition using the activation likelihood estimation (ALE) algorithm (Eickhoff et al. 2009). This approach represents a powerful approach to gain a synoptic and, in particular, unbiased view of distributed neuroimaging findings. In this context, ALE addresses the following key question: Where in the brain is the convergence of the included experiments' activation foci higher than expected by chance? ALE thus offers a unique opportunity to quantitatively measure concordance between neuroimaging results without the implicit assumptions of neuroanatomical terminologies, which are at times inconsistently employed. It also allows relating different neural processes to each other by performing meta-analyses on different neuropsychological phenomena. This integration and synthesis of neuroimaging data thus permits statistically defensible inference on the neural basis of psychological processes across a large number of experimental implementations and subjects samples.

The presented summary of neuroimaging studies on moral cognition included all those experiments that required participants to make (covert or overt) appropriateness judgments on actions of one individual towards others. In these studies, participants evaluated mainly textual, sometimes pictorial social scenarios with moral violations or dilemmas. While this approach has by far dominated the neuroscientific investigation of moral decisions, it should be noted, however, that it largely equates to assessing the neural correlates of identifying and judging moral violations. In contrast, they are less focused on "rightful action," i.e., the implementation of moral thoughts and knowledge in one's own behavior. Nevertheless, the analysis presented in Fig. 9.1 represents the current state of neuroimaging evidence for moral cognition. The obtained pattern of converged brain activation is in very good agreement with descriptive reviews of fMRI studies on moral cognition (J. Greene and Haidt 2002; Moll et al. 2005b). In the following, we will discuss the presumed functional implications of the individual brain areas that resulted as significant loci of convergence.

The *medial prefrontal cortex* (mPFC) is a highly associative brain region implicated in a range of complex tasks, such as action monitoring, free thought, autobiographical memory recall, and the perception of others. In fact, consistent activity in the mPFC during moral cognition was found all along its dorsoventral axis (simply put: from the upper to the lower parts of the middle front side of the brain), including the dorsomedial prefrontal cortex (dmPFC), frontopolar cortex (FP), and ventromedial PFC (vmPFC). From a conceptual perspective, the more dorsal versus more ventral parts of the mPFC are discussed to relate to cognitive versus affective processes, controlled versus automatic processes, explicit versus implicit social cognition, goal versus outcome pathways, as well as other-focus versus self-focus. Direct evidence for such fundamental distinction is however still limited. It is noteworthy that mPFC damage early in life can leave intellectual abilities intact while leading to hindered acquisition of social conventions and

Fig. 9.1 Meta-analysis results on moral cognition. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to moral cognition (67 neuroimaging experiments). Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *AM* amygdala, *dmPFC* dorsomedial prefrontal cortex, *FP* frontal pole, *mPFC* medial prefrontal cortex, *PCC* posterior cingulate cortex, *prec* precuneus, *TP* temporal pole, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (cf. Bzdok et al. 2012)
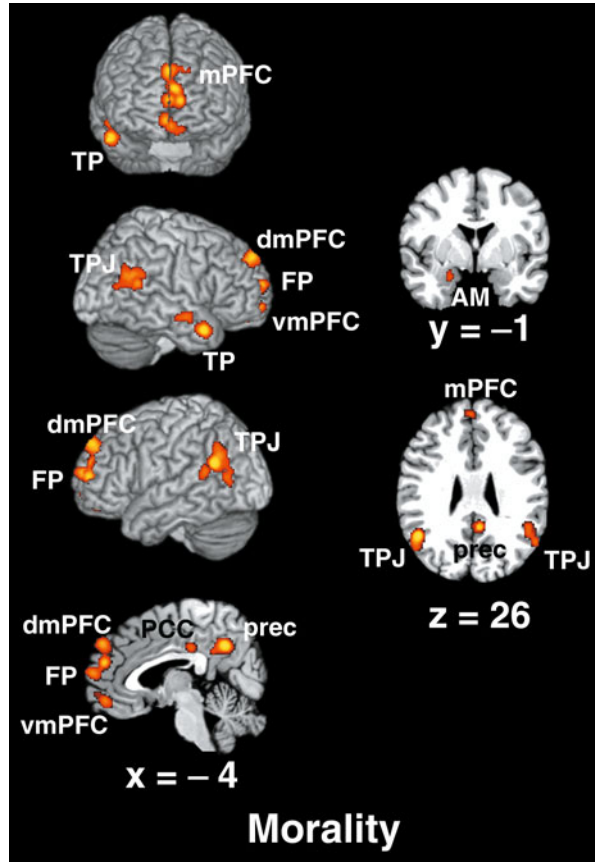
moral rules (Anderson et al. 1999). Early lesioned patients (much more than adult-onset prefrontal patients) display immoral behaviors, such as stealing, physical violence, and absence of remorse in the context of impaired moral reasoning (Moll et al. 2003). In short, a child's moral development can be disrupted by early mPFC damage.

The dmPFC has axonal connections with the temporo-parietal junction (TPJ; especially connected to superior frontal gyrus) and the precuneus (especially connected BA8/9), which have both likewise been implicated in the meta-analysis of moral cognition. The *temporo-parietal junction* is a supramodal association area whose heterogeneous functional profile seems to range from attentional reallocation, filtering irrelevant stimuli, and prediction generation over processing embodied self and predicting others' action intentions to agency. Paralleling its functional diversity, the TPJ literature offers various neuroanatomical labels for this area, including the terms angular gyrus, inferior parietal lobule, posterior superior temporal sulcus, supramarginal gyrus, BA 39, PGa/PGp, as well as "pli courbe." Consequently, interpretation of the inconsistently named TPJ can be challenging

given increased metabolic activity across disparate psychological tasks, stimulus domains, and experimental modalities.

The *precuneus* is another highly integrative area, which is believed to generate internally directed thoughts in form of self-referential visuospatial imagery (Cavanna and Trimble 2006). Consistently, the precuneus appears to mediate covert reallocation of spatial attention, that is, spatial cognition in the absence of physical (e.g., eye) movements (Gitelman et al. 1999), which led to its informal nickname "mind's eye." This proposed domain-spanning role might potentially explain its various domain-specific functional involvements, such as in visual rotation, deductive reasoning, autobiographical memory retrieval, and mental navigation in space.

The *posterior cingulate cortex* (PCC) is adjacent to but distinct from the precuneus by its connections to the limbic system and thus close relation to emotion processing (Margulies et al. 2009). This area is most frequently proposed to be important for the modality-independent retrieval of autobiographical memories and their integration with current emotional states (Maddock 1999).

As another affect-related brain region, the *amygdala* (AM) is believed to automatically extract biological significance from the environment and to shape appropriate behavioral responses (Bzdok et al. 2012; Sander et al. 2003). This functional concept covers its involvement in classical conditioning, emotion regulation, social cognition, reward processing, and memory formation (Adolphs 2010; Bzdok et al. 2012; LeDoux 2000). Considering that the amygdala is probably the brain area most unequivocally linked to emotion processing and given its heightened activity across the meta-analyzed neuroimaging studies, the reverse inference on an involvement of emotional brain systems in moral cognition seems justified.

Finally, the *temporal pole* (TP; here liberally referring to the entire anterior temporal lobe/BA38) was repeatedly proposed to store verbal and nonverbal semantic knowledge, in particular, context-independent social semantic knowledge, including values and concepts of social events (cf. Olson et al. 2007; Ross and Olson 2010; Zahn et al. 2007). Examples of such conceptual social knowledge would be the meaning and ramifications of "deceitfulness" or how people dress appropriately according to given situations. In line with this, neurological lesion of the TP entails social semantic deficits, such as failing to name human actions or to recognize the name, voice, handwriting, odors, or face of familiar people. Moreover, they may result in behavioral and personality disturbances, ranging from compulsively eating flower decorations on tables to general apathy to other's distress (cf. Gorno-Tempini et al. 2004).

It can be concluded that moral cognition is neurally implemented by brain areas that tend to be highly interconnected, not specific for any single psychological task, and not dependent on a specific sensory modality but rather multimodal and "associative." Those brain areas are moreover implicated in complex psychological processes, including social cognition, autobiographical memory retrieval, mental imagery, and reallocation of attention to internal information, all of which might contribute to the final "psychological outcome," i.e., moral judgment.

Two brain areas that have also been repeatedly discussed to subserve moral cognition, however, were not revealed by the meta-analysis. First, the posterior

superior temporal sulcus (pSTS) is involved in audiovisual integration and processing biologically significant movement (Hein and Knight 2008) and was reported to increase neural activity during moral cognition in several papers. The lack of convergence in the pSTS may readily be explained by inconsistent neuro-anatomical labeling. As a rule of thumb, activation in the surroundings of this cortical area was often interpreted as "pSTS" in the morality literature and as "TPJ" in the ToM literature. The convergent activation in the meta-analysis on moral cognition is, however, neuroanatomically corresponding to the TPJ. That is, observed recruitment of the TPJ during moral tasks was perhaps recurrently mislabeled as pSTS, which might have confused discussion of the TPJ and pSTS in previous neuroimaging studies on moral cognition.

Second, the delineated "moral network" did not include the dorsolateral pre-frontal cortex (dlPFC), conventionally interpreted as important for attention shifting and decision-making, although brain activity in this area was discussed in several original papers and reviews in moral neuroscience. The dlPFC was specif-ically proposed to reflect the engagement of abstract reasoning processes and cognitive control in moral cognition (Greene et al. 2004). Heightened dlPFC activity was thus argued to promote utilitarian responses by overriding prepotent socio-emotional behavioral tendencies. The absence of *consistent* metabolic increase in the dlPFC during moral decisions might be parsimoniously explained by selective recruitment. That is, this brain region might be recruited by the "core" moral network as an auxiliary functional module depending on the specific cogni-tive set imposed by specific moral decisions. Rather than being part of the "core" network, the dlPFC might have been observed to increase activity in difficult personal moral judgments and approving personal moral violations because of those decisions' increased cognitive demand. The same principle of context-dependent recruitment of supplementary areas might hold true for other brain regions that were associated with moral cognition repeatedly but not consistently, including but not restricted to the anterior insula, anterior cingulate cortex, and lateral orbitofrontal cortex. Also for these regions, inconsistent neuroanatomical labeling cannot be excluded as a confounding factor in the previous literature on moral cognition.

The Neurobiological Relationship Between Moral Cognition and Theory of Mind

Performing another meta-analysis on brain activity evoked by theory of mind then allowed elucidating the correspondence of the neural substrates consistently engaged by this task and moral cognition, two psychologically related mental processes (cf. introduction). We included those neuroimaging studies into the meta-analysis of theory of mind that required participants to adopt an intentional stance towards others, that is, predict their thoughts, intentions, and future actions. These studies mostly presented cartoons and short narratives that necessitated understanding the beliefs of the acting characters. The results of the meta-analysis

Fig. 9.2 Meta-analysis results on theory of mind. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to theory of mind (68 neuroimaging experiments). Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *dmPFC* dorsomedial prefrontal cortex, *FP* frontal pole, *IFG* inferior frontal gyrus, *MTG* middle temporal gyrus, *mPFC* medial prefrontal cortex, *prec* precuneus, *TP* temporal pole, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (cf. Bzdok et al. 2012)
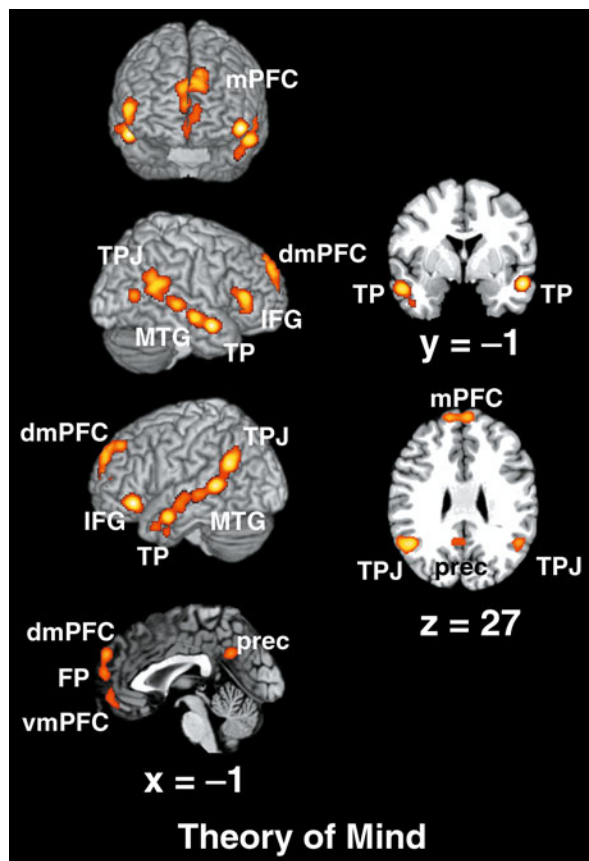
of ToM (Fig. 9.2) are consistent with earlier meta-analysis of such neuroimaging studies (Spreng et al. 2009). Conceptually, the convergence across studies on ToM resulted in an *abstract–inferential social–cognitive network* implicated in the recognition and processing of others' *mental states*.

Brain activity patterns during moral cognition and ToM overlapped in the bilateral vmPFC, FP, dmPFC, and TPJ, as well as the right TP (Fig. 9.3). This extensive convergence indicates that moral cognition and ToM engage a highly similar neural network. The homologous neural implementation, in turn, entices to speculate about a close relationship between these two psychological processes and the experimental tasks to probe these. The interest in the neurobiological relationships between moral cognition and ToM recently gained momentum, which entailed publication of a small number of targeted neuroimaging studies.

A seminal fMRI study investigated the interaction of a protagonist's initial intention and subsequent action outcome by explicit moral judgments of short written stories (Young et al. 2007). The bilateral TPJ, dmPFC, and precuneus

Fig. 9.3 Relationship between the moral network and the neural network underlying theory of mind. Overlapping activation patterns between the meta-analysis on moral cognition and the difference analysis between ToM and empathy. Coordinates in MNI space (cf. Bzdok et al. 2012)

showed significant signal effects for the interaction of negative versus neutral beliefs versus outcomes. The right TPJ showed the biggest signal increase in attempted (intention) but failed (outcome) harm, that is, when nothing bad actually happened despite what the protagonist planned. Given that moral cognition and mental state attribution were probably part of the participants' cognitive set in all experimental conditions, right TPJ activity appeared to reflect special emphasis on the agent's thoughts when weighing various contextual features against each other. Consistently, transient disruption of right TPJ activity using repetitive transcranial magnetic stimulation was found not to impair moral judgments per se (Young et al. 2010). Rather, this manipulation reduced the influence of the protagonist's belief without completely eliminating it as an input to judgment formation. It is important to note, however, that further evidence from neuroimaging (Mitchell 2008) and lesion (Apperly et al. 2007) studies questioned the *specificity* of the TPJ for belief processing. Nevertheless, right RTPJ activity appears to be, comparing to other relevant areas, particularly related to processing mind states in explicit moral cognition.

The ensuing notion that the TPJ may represent a crucial link between moral cognition and ToM was confirmed by another fMRI study that set out to detail encoding and integration of intentions in the context of moral judgments (Young and Saxe 2008). While "encoding" consists in merely creating a representation of the protagonist's belief, "integration" then consists in flexibly weighing the moral judgment depending on the interaction of intention and outcome. The bilateral TPJ and precuneus were related to both encoding the protagonist's belief and integrating it with other relevant contextual features. In fact, brain activity in these regions did not differ according to belief content, in particular, its valence (negative vs. neutral). In contrast, the dmPFC was related to processing belief valence during the integration phase. The authors thus proposed that the TPJ and precuneus mainly process beliefs, while the dmPFC mainly processes morally relevant aspects of the presented stories in constructing a coherent moral judgment. Analogous to the TPJ, it is important to note that dmPFC activity might not be *specific* to belief processing as it was for instance also linked to language coherence (Ferstl and von Cramon 2002). In addition to belief, outcome, and valence, metabolic responses in the dmPFC and bilateral TPJ during moral judgments were observed to vary according to the previously experienced fairness of the judged agent (Kliemann et al. 2008). This suggests that the dmPFC and bilateral TPJ might also integrate available memory of an agent's personality traits in explicit moral judgments. Another fMRI study showed selective metabolic increase in the dmPFC, right TPJ, and precuneus in response to morally relevant facts in short stories without explicit mental state information or an explicit necessity for moral judgments (Young and Saxe 2009). This finding can be taken to argue that brain regions typically related to ToM might be implicated not only in *explicit or controlled* but also *implicit or automatic* moral cognition.

In line with the meta-analytic overlap, the reviewed fMRI studies suggest that moral cognition might involve reconstructing personality attributes and intentions

of agents as well as their integration with action outcomes and other relevant contextual features when reasoning about morally significant social scenarios. It should, however, not be underestimated that experimental similarities (i.e., employed stimuli and paradigm) between studies on moral cognition and ToM might have contributed to the observed neural homology. In other words, the congruency may appear larger than it actually is due to shared low-level features (textual descriptions, cartoon, requirement to make a judgment).

The Neurobiological Relationship Between Moral Cognition and Empathy

In contrast to ToM, the correspondence between moral cognition and empathy has hardly been investigated in targeted neuroimaging research despite their relatedness on psychological grounds (cf. introduction). The correspondence between the neural substrates reported in those largely separate lines of research was therefore juxtaposed by individual meta-analysis on either topic (Figs. 9.1 and 9.4). We included those neuroimaging studies into the meta-analysis of empathy that aimed at eliciting the conscious and isomorphic (i.e., happiness in other induces happiness in oneself) experience of somebody else's affective state. Put differently, in these studies participants were supposed to "feel into" and thus know what another person was *feeling* (rather than *thinking*, which would be related to ToM). These studies employed mostly visual, sometimes textual, or auditory stimuli that conveyed affect-laden social situations which participants watched passively or evaluated on various dimensions. Conceptually, the convergence across studies on empathy resulted in an *automatic–emotional social–cognitive network* implicated in vicariously mapping others' *affective states*.

Please note that a meta-analytic distinction between emotional and cognitive empathy cannot and should not be done at this point. It cannot be done because there are currently not enough available neuroimaging studies on cognitive empathy. It should not be done because assuming a clear-cut neurobiological dissociation between emotional and cognitive empathy would constitute a fairly strong a priori hypothesis about how psychological constructs map on brain organization.

Brain activity related to both moral cognition and empathy converged significantly in an area of the dmPFC (Fig. 9.5). An fMRI study identified a similar area as highly selective for processing guilt (Wagner et al. 2011), an emotion closely related to moral and social norm violation. More specifically, guilt was proposed to promote interpersonal relationships by immediately providing actual or anticipated emotional feedback for the acceptability of actions (Tangney et al. 2007). Moreover, the dmPFC has consistently been related to the (possibly interwoven) reflection of own and simulation of others' mind states (Bzdok et al. 2013). One might therefore cautiously conclude that convergence in this highly associative cortical area might reflect complex representational social–emotional processing. Additionally, the individual meta-analyses revealed the left AM in moral cognition

Fig. 9.4 Meta-analysis
results on empathy. Whole-
brain renderings as well as
sagittal, coronal, and axial
slices depicting the significant
results of the ALE meta-
analyses of eligible
neuroimaging experiments
(published until 2010) related
to empathy (112
neuroimaging experiments).
Coordinates in MNI space.
All results were significant at
a cluster-forming threshold of
$p < .05$. *ACC* anterior
cingulate cortex, *AI* anterior
insula, *AM* amygdala, *BS*
brainstem, *dmPFC*
dorsomedial prefrontal
cortex, *IFG* inferior frontal
gyrus, *MTG* middle temporal
gyrus, *PCC* posterior
cingulate cortex, *SMA*
supplementary motor area,
TPJ temporo-parietal junction
(cf. Bzdok et al. 2012)
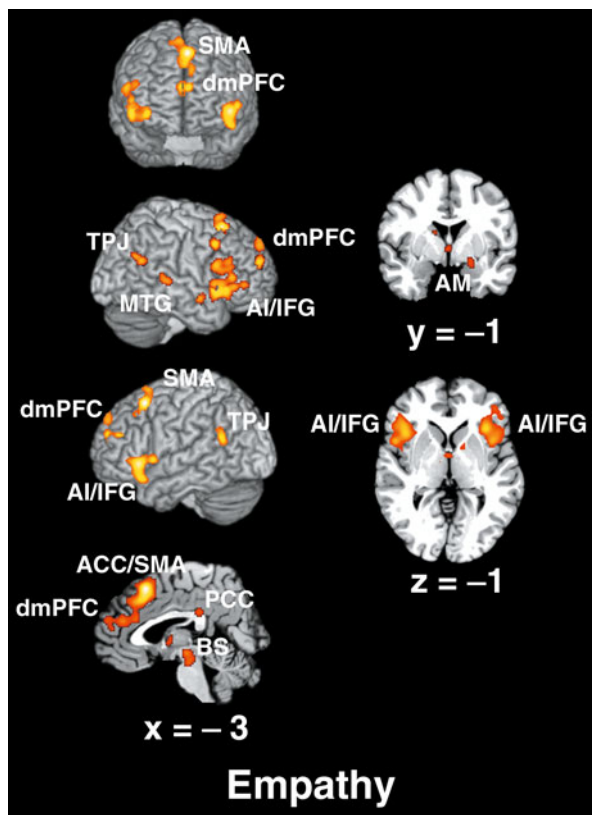
and the right AM in empathy, that is, the same area in contralateral hemispheres.
A role of the AM in moral cognition is supported by correlation between its neural
activity and the participants' level of self-reported emotional arousal when
presented with visual stimuli of people being harmed (Decety et al. 2012). More
specifically, it is known that AM activity typically increases in the left hemisphere
in controlled, elaborate social-cognitive processes and in the right hemisphere in
automatic, basic emotional processes (Markowitsch 1998; Phelps et al. 2001). This
lateralization pattern potentially explains the consistent engagement of the left AM
in moral cognition (more controlled/elaborate) and right AM in empathy (more
automatic/basic). Furthermore, activity in the PCC was found in adjacent, yet
nonoverlapping, locations during moral cognition and empathy. Neural activity in
this brain area was observed in hearing and recalling affective autobiographical
episodes, dealing with coherent social scenarios, viewing familiar faces, as well as
emotional planning. The PCC was thus repeatedly proposed to integrate retrieval of
past experiences and ongoing emotion processing, which is potentially shared by
moral cognition and empathy.

To sum up the meta-analytic evidence, some aspects of affective processing are
probably shared by moral cognition and empathy, as the respective meta-analyses

Fig. 9.5 Relationship
between the moral network
and the neural network
underlying empathy. *Bottom
panel*: overlapping activation
patterns between the
meta-analysis on moral
cognition and the difference
analysis between empathy
and ToM. *Top panel*: sagittal
and coronal slices of
juxtaposed results from the
meta-analyses on moral
cognition and empathy to
highlight similar convergence
in the posterior cingulate
cortex and amygdala.
Coordinates in MNI space
(cf. Bzdok et al. 2012)
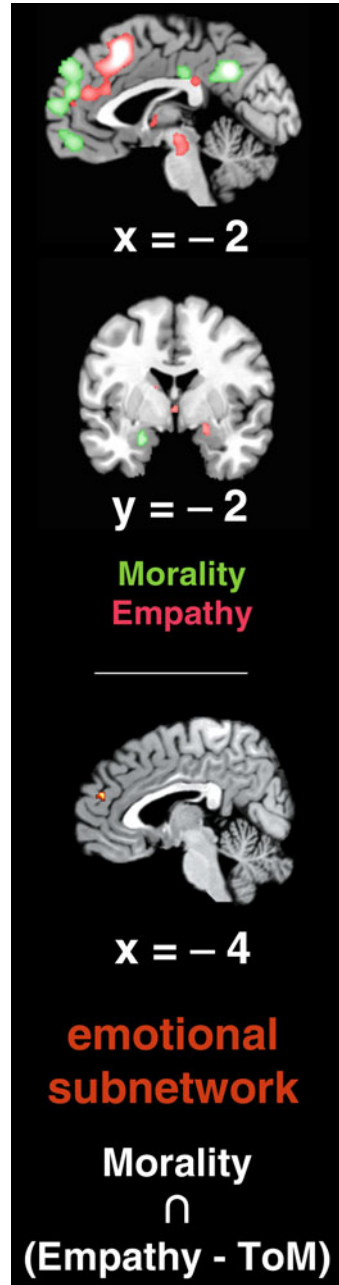
revealed convergence in the dmPFC (direct overlap), AM (oppositely lateralized), and PCC (closely adjacent clusters). A single direct overlap in the dmPFC might further suggest representational socio-emotional processing to be common to moral cognition and empathy. More generally, it is interesting to note that the neural correlates of moral cognition are more closely related to the neural signature of ToM than to that of empathy. It is an important question whether this is an epiphenomenon of methodological idiosyncrasies or illustrates brain network dynamics in everyday life.

The Neurobiological Relationship Between Moral Cognition and Mind-Wandering

It is becoming increasingly clear that brain areas pertaining to social cognition are topographically highly similar to brain areas that increase activity in the idling, unconstrained mind, the so-called "default mode," and decrease activity during stimulus-driven, goal-directed tasks (Shulman et al. 1997; Spreng et al. 2009). More specifically, brain areas underlying unconstrained cognition were consistently associated with a number of complex, introspective mental tasks, including contemplating mind states, self-focused reflection, mental navigation of the body in space, autobiographical memory recall, and, more generally, envisioning situations detached from reality. Performing separate meta-analyses on moral and unconstrained cognition hence allowed elucidating the correspondence between the neural substrates consistently engaged by these two mental states. We included those neuroimaging experiments from the BrainMap database (Laird et al. 2011) into the meta-analysis of unconstrained cognition whose metadata indicated them to provide coordinates of brain deactivation (Fig. 9.6). Brain activity patterns during moral reasoning and unconstrained cognition overlapped in the vmPFC, dmPFC, precuneus, and bilateral TPJ (Fig. 9.7). Consequently, those brain areas consistently implicated in moral cognition indeed lower their activity during stimulus-driven, goal-directed cognition. More broadly, the observed similarities of the neural networks underlying moral and unconstrained cognition favor a possible relationship between the physiological baseline of the human brain and a psychological baseline implicated in constant social cognition.

What is the common denominator of moral and unconstrained cognition? It was speculated that the human brain might have evolved to, by default, predict environmental events using mental imagery. In particular, autobiographical memory supplies building blocks of social semantic knowledge. Isolated conceptual scripts may then be reassembled to enable forecasting future events (Tulving 1983). Constructing detached probabilistic social scenes could thus influence perception and behavior by estimating saliency and action outcomes (Boyer 2008; Schilbach et al. 2008). Ultimately, the tonically active default mode network might be adapted to gathering sensory information for the probabilistic mapping of the external world in order to optimize the organism's behavioral response. That is,

Fig. 9.6 Meta-analysis results on mind-wandering. Whole-brain renderings as well as sagittal, coronal, and axial slices depicting the significant results of the ALE meta-analyses of eligible neuroimaging experiments (published until 2010) related to mind-wandering (533 neuroimaging experiments), i.e., brain activity in the absence of a specific task. Coordinates in MNI space. All results were significant at a cluster-forming threshold of $p < .05$. *dmPFC* dorsomedial prefrontal cortex, *HC* hippocampus, *FP* frontal pole, *MFG* middle frontal gyrus, *mPFC* medial prefrontal cortex, *prec* precuneus, *SMG* supramarginal gyrus, *TPJ* temporo-parietal junction, *vmPFC* ventromedial prefrontal cortex (Schilbach et al. 2012)
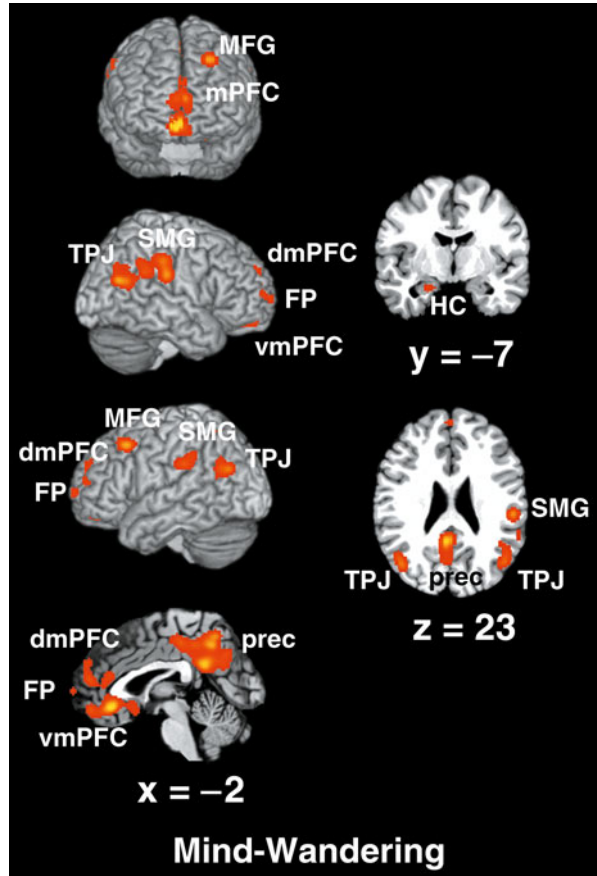

pondering over morally relevant social situations and simply letting the mind float might both imply contemplation of hypothetical social scenes that guide actual behavior.

Conclusion and Future Directions

The conjunction of the above quantitative and qualitative reviews suggests that moral cognition might rely in large extent on neural systems related to social cognition. The group of social processes subserving moral cognition appears to include contemplating mind states, emotion processing, and internally directed cognitions, such as self-referential thought, mental imagery, and perhaps even prediction generation. Additionally, the neural correlates of moral cognition were dissociated into more rational and more emotional subsystems by reference to

Fig. 9.7 Relationship
between the moral network
and the neural network
underlying mind-wandering.
Overlapping activation
patterns between the
meta-analyses on moral
cognition and
mind-wandering, i.e., brain
activity in the absence of
a specific task. Coordinates in
MNI space (cf. Bzdok et al.
2012; Schilbach et al. 2012)
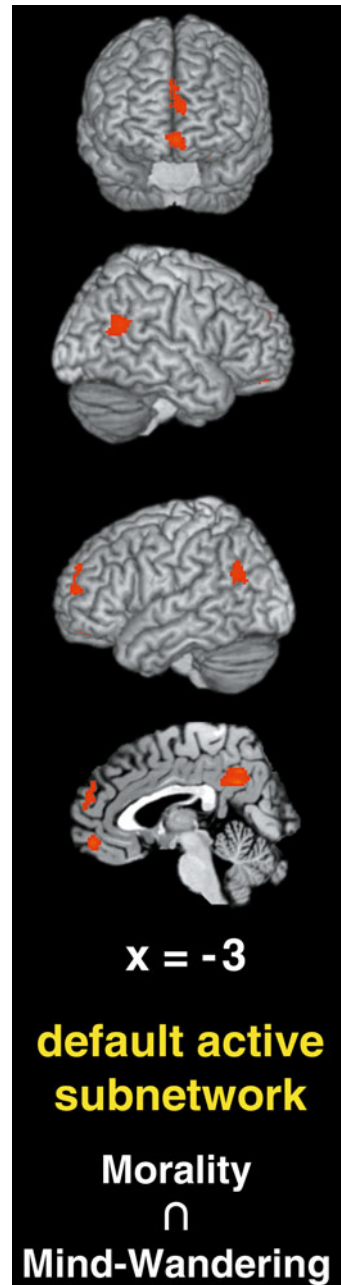
a socio-cognitive framework (ToM cognition) and a socio-affective framework (empathy), respectively. The neural network underlying moral cognition might thus be considered "domain-global" as it reflects functional integration of distributed brain networks. Put differently, no part of the brain is uniquely devoted to moral cognition but this capacity is very likely deployed across several heterogeneous functional domains. This contention concurs with the observation that there is also no single brain region specific for the development of antisocial/psychopathic behavior deficient in moral cognition (Raine and Yang 2006).

Shifting from the neural level to psychological concepts, the question "what brain areas support moral judgment?" might already be ill-posed, or at least inaccurate, because moral cognition is unlikely to be a unified psychological entity. That is, distinct (sets of) cognitive components probably support distinct classes of moral judgments. Even more fundamentally, "morality" as such might actually not be localizable in the brain at all given that it constitutes a complex cultural concept, that is, a phenomenon of human cultural evolution. Considering from a phenomenological, psychological, and philosophical point of view, there might even be nothing unique to the notion of "morality" itself. What is unique to moral judgments comparing to, for example, judgments of complex social scenarios? In short, we cannot measure "morality" itself in the human brain. Instead, we can measure brain activity of individuals lying in a neuroimaging scanner while thinking about moral issues. For these reasons, it might actually be naive to search for something like a "distinct moral module" in the first place. Furthermore, moral psychology and moral neuroscience mainly concentrated on moral decision-making, rather than the very manifestation of morality – moral behavior (see Moll et al. 2006 for an exception).

From a methodological perspective, it is unclear to what extent existing neuroimaging studies on moral cognition suffer from this potentially limited ecological validity. That is, the used experimental tasks might only partially involve the neural processes that navigate real-life moral behavior. In particular, complicated dilemmas borrowed from moral philosophy were often employed as stimulus material. This is epitomized by the "trolley dilemma" that prompts a decision between either letting five people die or actively causing the death of another single person to save the life of those five people. A tendency for artificial moral scenarios, on top of the experimental constraints of neuroimaging environments, could have entailed a systematic overestimation of cognitive versus emotional processes. The observed bigger correspondence of the moral neural network with that of ToM, rather than empathy, might thus be epiphenomenal of established experimental features.

Moreover, neuroimaging results were often discussed by qualitative comparison between studies that differ in various crucial aspects, including stimulus material (text vs. pictures/movies vs. numbers in neuroeconomic games), the participants' perspective (engaged second-person vs. observant third-person perspective), control conditions (morally relevant vs. morally irrelevant, high-level vs. low-level), or continuity (single-shot judgments vs. multi-trial paradigms). It is conceivable that the neural instantiation of moral judgments and comparisons between those are highly susceptible to such experimental differences, especially in light of the integrative character of moral cognition.

From a neuroethical perspective, it can be said that the neuroscience of moral decision-making may be able to contribute to ethics by providing descriptive results. But one must be wary of overextending the logic of neuroimaging with regard to morality. Although moral decision-making correlates with the activation of specific regions in the brain, this does not necessarily mean that moral judgment can be reduced to this activation. A central problem is thus that an attempt is made under scientific conditions to establish which brain areas are particularly active during moral cognition without it being possible to define exactly what moral cognition as such is and without the existence of an objectifiable moral theory (Gazzaniga 2007).

Some critics even raise much more fundamental objections to imaging studies of moral cognition. They pose the rhetorical question: How can I hope at all to discover "facts" which prove "values"? All moralities and all moral decisions are in fact based on values and norms, and these cannot, or at least cannot necessarily be reduced or ascribed for their part to (neurophysiological) facts but are intrinsically subjective and hence may only be accessible on phenomenological accounts. A second question is equally important: What is the consequence if I regard and acknowledge specific activities of neuronal tissue as the basis for moral evaluations and decisions? If values and norms can be described as sequences of cellular processes, does this not remove the basis for morality? (Vogelsang 2008).

Neil Levy (2007) attempted to summarize this "challenge from neuroscience to morality" in four consecutive sentences:

1. Our moral theories, as well as our first-order judgments and principles are all based, more or less directly upon our moral intuitions.
2. These theories, judgments and principles are justified only insofar as our intuitions track genuinely moral features of the world.
3. But our moral intuitions are the product of cognitive mechanisms which evolved under non-moral selection pressures and therefore cannot be taken to track moral features of the world; hence
4. Our moral theories, judgments and principles are unjustified.

It emerges from the above that the study and objectification of processes involving moral decisions and judgments poses fundamental problems. The ability to image brain activity is much greater than the ability to draw clear conclusions with regard to questions of morality from it.

Critics also object that it is not clear to them why knowledge about neuronal processes should help us at all when it comes to morality and moral decisions. They point out that human morality – beyond all modern empirical methods of access – has always been a main topic of philosophical debate and that this debate will continue to be needed in the future (Brukamp and Groß 2010; Groß 2010). Particularly when normative conclusions are to be made on the basis of these empirical findings, the problem of the "ought" fallacy arises, as a (moral) "ought" cannot simply be derived from a (neuronal) "is." This also makes it necessary to raise the question of the scientific and normative rules for dealing with scientific studies on moral cognition. What standards must be met by the scientists who carry

out and publish studies of this kind? Is it enough for them to be experts in neuroscience esp. neuroimaging or should (additional) expertise in the field of morality and (neuro)ethics be required?

Irrespective of the abovementioned questions and concerns, detailing the neurobiological nature of moral cognition is a goal worth pursuing. Importantly however, moral neuroscience should strive for an explanation of the understanding of morality which underlies it, for realistic moral scenarios and for a more rigorous across-study discussion (cf. Knutson et al. 2010). These suggestions might help to minimize the risk of investigating "in vitro moral cognition."

Cross-References

▶ Beyond Dual-Processes: The Interplay of Reason and Emotion in Moral Judgment
▶ Mental Causation
▶ Moral Cognition: Introduction
▶ Moral Intuition in Philosophy and Psychology
▶ The Half-Life of the Moral Dilemma Task: A Case Study in Experimental (Neuro-) Philosophy

References

Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences, 1191*(1), 42–61. doi:10.1111/j.1749-6632.2010.05445.x. NYAS5445 [pii].

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience, 2*(11), 1032–1037. doi:10.1038/14833.

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W. L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal, "reality-unknown" false belief and false photograph tasks. *Cognition, 103*(2), 300–321. doi:10.1016/j.cognition.2006.04.012. S0010-0277(06) 00082-5 [pii].

Boyer, P. (2008). Evolutionary economics of mental time travel? *Trends in Cognitive Science, 12*(6), 219–224. doi:10.1016/j.tics.2008.03.003. S1364-6613(08)00093-4 [pii].

Brukamp, K., & Groß, D. (2010). Neuroenhancement – A controversial topic in contemporary medical ethics. In P. A. Clark (Ed.), *Contemporary issues in bioethics* (pp. 39–50). [Ebook]. www.intechopen.com/books/contemporary-issues-in-bioethics/neuroenhancement-a-controversial-topic-in-medical-ethics. Rijeka

Bzdok, D., Langner, R., Hoffstaedter, F., Turetsky, B. I., Zilles, K., & Eickhoff, S. B. (2012a). The modular neuroarchitecture of social judgments on faces. *Cerebral Cortex, 22*(4), 951–961. doi:10.1093/cercor/bhr166. bhr166 [pii].

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012b). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function, 217*(4), 783–796. doi:10.1007/s00429-012-0380-y.

Bzdok, D., Laird, A., Zilles, K., Fox, P. T., & Eickhoff, S. (2012, in press). An investigation of the structural, connectional and functional sub-specialization in the human amygdala. *Human Brain Mapping*. doi:10.1002/hbm.22138

Bzdok, D., Langner, R., Schilbach, L., Engemann, D. A., Laird, A. R., Fox, P. T., Eickhoff, S. B. (2013). Segregation of the human medial prefrontal cortex in social cognition. *Front Hum Neurosci, 29*(7), 232. doi: 10.3389/fnhum.2013.00232.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(Pt 3), 564–583. doi:10.1093/brain/awl004. awl004 [pii].

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 71–100. doi:10.1177/1534582304267187. 3/2/71 [pii].

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex, 22*(1), 209–220. doi:10.1093/cercor/bhr111. bhr111 [pii].

Eickhoff, S. B., & Bzdok, D. (2012). Meta-analyses in basic and clinical neuroscience: State of the art and perspective. In S. Ulmer & O. Jansen (Eds.), *fMRI – Basics and clinical applications* (2nd ed.). Heidelberg: Springer.

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping, 30*(9), 2907–2926. doi:10.1002/hbm.20718.

Eisenberger, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology, 51*, 665–697.

Ferstl, E. C., & von Cramon, D. Y. (2002). What does the frontomedian cortex contribute to language processing: Coherence or theory of mind? *NeuroImage, 11*, 1599–1612.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358*(1431), 459–473. doi:10.1098/rstb.2002.1218.

Gazzaniga, M. S. (2007). *Wann ist der Mensch ein Mensch? Antworten der Neurowissenschaft auf ethische*. Fragen: Patmos.

Gitelman, D. R., Nobre, A. C., Parrish, T. B., LaBar, K. S., Kim, Y. H., Meyer, J. R., & Mesulam, M. (1999). A large-scale distributed network for covert spatial attention: Further anatomical delineation based on stringent behavioural and cognitive controls. *Brain, 122*(Pt 6), 1093–1106.

Gorno-Tempini, M. L., Rankin, K. P., Woolley, J. D., Rosen, H. J., Phengrasamy, L., & Miller, B. L. (2004). Cognitive and behavioral profile in a case of right anterior temporal lobe neurodegeneration. *Cortex, 40*(4–5), 631–644.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science, 6*(12), 517–523. S1364661302020119 [pii].

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. doi:10.1126/science.1062872. 293/5537/2105 [pii].

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400. doi:10.1016/j.neuron.2004.09.027. S0896627304006348 [pii].

Groß, D. (2010). Traditional vs. Modern neuroenhancement. Notes from a medico-ethical and societal perspective. In H. Fangerau & T. Trapp (Eds.), *Implanted mind (= Science Studies)* (pp. 137–157). Transcript: Bielefeld.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. doi:10.1037//0033-295x.108.4.814.

Hare, R. D. (2003). *The Hare psychopathy checklist – Revised* (2nd ed.). Toronto: Multi-Health Systems.

Hein, G., & Knight, R. T. (2008). Superior temporal sulcus – It's my area: Or is it? *Journal of Cognitive Neuroscience, 20*(12), 2125–2136. doi:10.1162/jocn.2008.20148.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949–2957. doi:10.1016/j.neuropsychologia.2008.06.010. S0028-3932(08)00261-3 [pii].

Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., & Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *Social Cognitive and Affective Neuroscience, 5*(4), 378–384. doi:10.1093/scan/nsq005. nsq005 [pii].

Laird, A. R., Eickhoff, S. B., Fox, P. M., Uecker, A. M., Ray, K. L., Saenz, J. J., Jr., & Fox, P. T. (2011). The brainmap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Research Notes, 4*(1), 349. doi:10.1186/1756-0500-4-349. 1756-0500-4-349 [pii].

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184. doi:10.1146/annurev.neuro.23.1.155.

Levy, N. (2007). *Neuroethics. Challenges for the 21st century*. Cambridge: Cambridge University Press.

Maddock, R. J. (1999). The retrosplenial cortex and emotion: New insights from functional neuroimaging of the human brain. *Trends in Neurosciences, 22*(7), 310–316. S0166223698013745 [pii].

Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., & Petrides, M. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences of the United States of America, 106*(47), 20069–20074. doi:10.1073/pnas.0905314106. 0905314106 [pii].

Markowitsch, H. J. (1998). Differential contribution of right and left amygdala to affective information processing. *Behavioural Neurology, 11*(4), 233–244.

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex, 18*(2), 262–271. doi:10.1093/cercor/bhm051. bhm051 [pii].

Moll, J., & Schulkin, J. (2009). Social attachment and aversion in human moral cognition. *Neuroscience and Biobehavioral Reviews, 33*(3), 456–465. doi:10.1016/j.neubiorev.2008.12.001. S0149-7634(08)00204-2 [pii].

Moll, J., de Oliveira-Souza, R., & Eslinger, P. J. (2003). Morals and the human brain: A working model. *Neuroreport, 14*(3), 299–305. doi:10.1097/01.wnr.0000057866.05120.28.

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignacio, F. A., Bramati, I. E., Caparelli-Daquer, E. M., & Eslinger, P. J. (2005a). The moral affiliations of disgust: A functional MRI study. *Cognitive and Behavioral Neurology, 18*(1), 68–78. 00146965-200503000-00008 [pii].

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005b). Opinion: The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*(10), 799–809. doi:10.1038/nrn1768. nrn1768 [pii].

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America, 103*(42), 15623–15628. doi:10.1073/pnas.0604475103. 0604475103 [pii].

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 2688–2692. doi:10.1073/pnas.1011734108. 1011734108 [pii].

Olson, I. R., Ploaker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: A review of findings on social and emotional processing. *Brain, 130*, 1718–1731. doi:10.1093/Brain/Awm052.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience, 4*(4), 437–441. doi:10.1038/86110. 86110 [pii].

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *The Behavioral and Brain Sciences, 1*(4), 515–526.

Raine, A., & Yang, Y. (2006). Neural foundations to moral reasoning and antisocial behavior. *Social Cognitive and Affective Neuroscience, 1*(3), 203–213. doi:10.1093/scan/nsl033.

Ross, L. A., & Olson, I. R. (2010). Social cognition and the anterior temporal lobes. *NeuroImage, 49*(4), 3452–3462. doi:10.1016/j.neuroimage.2009.11.012. S1053-8119(09)01196-3 [pii].

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*(4), 303–316.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457–467. doi:10.1016/j. concog.2008.03.013. S1053-8100(08)00037-8 [pii].

Schilbach, L., Bzdok, D., Timmermans, B., Fox, P. T., Laird, A. R., Vogeley, K., & Eickhoff, S. B. (2012). Introspective minds: Using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. *PloS One, 7*(2), e30920. doi:10.1371/journal.pone.0030920. PONE-D-11-20368 [pii].

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain, 132*, 617–627. doi:10.1093/Brain/Awn279.

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., & Petersen, S. E. (1997). Common blood flow changes across visual tasks. 2. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience, 9*(5), 648–663.

Singer, T., & Lamm, C. (2009). The social neuroscience of empathy. *Annals of the New York Academy of Sciences, 1156*, 81–96. doi:10.1111/j.1749-6632.2009.04418.x. NYAS04418 [pii].

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489–510. doi:10.1162/jocn.2008.21029.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon.

Vogelsang, F. (2008). Aufgaben und Perspektiven einer künftigen Neuroethik. In F. Vogelsang & C. Hoppe (Eds.), *Ohne Hirn ist alles nichts. Impulse für eine Neuroethik* (pp. 11–22). Neukirchener.

Wagner, U., N'Diaye, K., Ethofer, T., & Vuilleumier, P. (2011). Guilt-specific processing in the prefrontal cortex. *Cerebral Cortex, 21*(11), 2461–2470. doi:10.1093/cercor/bhr016. bhr016 [pii].

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*(4), 1912–1920. doi:10.1016/j.neuroimage.2008.01.057. S1053-8119(08)00087-6 [pii].

Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience, 21*(7), 1396–1405. doi:10.1162/jocn.2009.21137.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240. doi:10.1073/pnas.0701408104.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*(15), 6753–6758. doi:10.1073/pnas.0914826107. 0914826107 [pii].

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6430–6435. doi:10.1073/pnas.0607061104. 0607061104 [pii].