Philosophical Explorations, Vol. 9, No. 1, March 2006

Routledge
Taylor & Francis Group

NEURO-COGNITIVE SYSTEMS INVOLVED IN MORALITY

James Blair, A. A. Marsh, E. Finger, K. S. Blair and **J. Luo**

In this paper, we will consider the neuro-cognitive systems involved in mediating morality. Five main claims will be made. First, that there are multiple, partially separable neuro-cognitive architectures that mediate specific aspects of morality: social convention, care-based morality, disgust-based morality and fairness/justice. Second, that all aspects of morality, including social convention, involve affect. Third, that the neural system particularly important for social convention, given its role in mediating anger and responding to angry expressions, is ventrolateral prefrontal cortex. Fourth, that the neural systems particularly important for care-based morality are the amygdala and medial orbital frontal cortex. Fifth, that while Theory of Mind is not a prerequisite for the development of affect-based 'automatic moral attitudes', it is critically involved in many aspects of moral reasoning.

Introduction

Understanding morality is of considerable intrinsic interest. Moreover, understanding morality is of direct relevance to clinical populations. Specific pathological conditions lead to breakdowns in 'morality', developmental psychopathy and 'acquired sociopathy'. By understanding morality, we may come to better understand these clinical conditions.

Until relatively recently, moral reasoning was considered by most psychologists to be an effortful, controllable rational process (Colby et al. 1983; Piaget 1932; Turiel 1983). However, more recently, models stressing the role of emotion have become prevalent (Blair 1995; Greene and Haidt 2002; Kagan and Lamb 1987; Moll, de Oliveirra-Souza, and Eslinger 2003; Nichols 2002).

A major claim of the current paper will be that there are a series of relatively separable neuro-cognitive systems that mediate particular types of reasoning. These various forms of reasoning are grouped together as 'moral reasoning'. The components of morality that these partially dissociable neuro-cognitive systems are thought to subserve are depicted in Figure 1. This paper will concentrate on neuro-cognitive systems involved in social convention and care-based morality.

Before an account of the neuro-cognitive systems involved in morality is developed, we will briefly consider Theory of Mind. While we do not believe Theory of Mind is necessary for the development of aversion to moral transgressions (cf. Blair 1995), we do believe that it plays a role in some aspects of moral reasoning.

ISSN 1386-9795 print/1741-5918 online/06/010013-15

DOI: 10.1080/13869790500492359

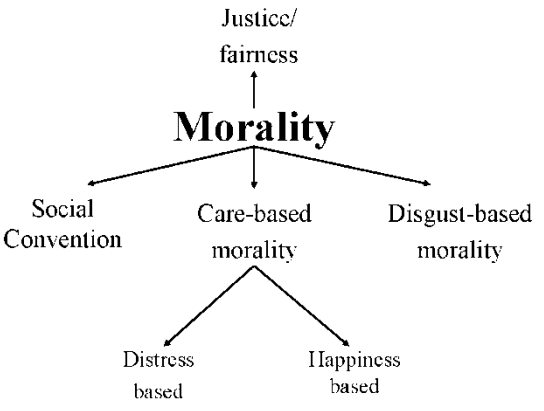

FIGURE 1
Dissociable components of morality

Theory of Mind

Theory of Mind refers to the ability to represent the mental states of others (Frith and Frith 2003; Leslie 1987; Premack and Woodruff 1978). The main neural regions implicated in Theory of Mind include medial frontal cortex, superior temporal sulcus and temporal pole (see Frith and Frith 2003).

There are three main positions on Theory of Mind: the theory-theory view (Gopnik and Meltzoff 1997; Wellman 1990), the modular view (Frith and Frith 2003; Leslie 1987) and the simulation view (e.g., Perner 1998). There is considerable debate regarding these three positions, a debate which we have no interest in entering into. However, we note that Leslie (1987) and the more recent simulation accounts (Gallese, Keysers, and Rizzolatti 2004) have attempted to detail a model of the computational architecture that is necessary for the representation of the mental states of others. In particular, the more recent simulation accounts have made reference to 'mirror neurons'; neurons which show activation when the participant is performing, or watching someone else perform, a specific action (Gallese, Keysers, and Rizzolatti 2004). The more general suggestion is that the observation of an action leads to the activation of parts of the same cortical neural network that is active during its execution. The observer understands the action because 'he knows its outcomes when he does it' (Gallese, Keysers, and Rizzolatti 2004, 393).

In short, recent accounts stress the importance of an integrated response to the interpretation of motor actions as a basis of many aspects of 'Theory of Mind'. We will argue below that this form of Theory of Mind is crucial for some aspects of moral reasoning.

Distinguishing Convention from Morality

Moral transgressions have been defined by their consequences for the rights and welfare of others and social conventional transgressions have been defined as violations of the behavioral uniformities that structure social interactions within social systems (Nucci 1981; Smetana 1981; Turiel 1983). Examples of moral transgression would be hitting another individual or damaging his property. Examples of conventional

transgressions would be talking in class or dressing in opposite gender clothing. Participants distinguish conventional and moral transgressions from the age of 39 months (Smetana 1981) and across cultures (Hollos, Leis, and Turiel 1986; Song, Smetana, and Kim 1987).

Conventional and moral transgressions are distinguished in three main ways. First, children and adults usually judge conventional transgressions as less serious than moral transgressions. For example, while all of the transgression situations, whether moral or conventional, are generally judged not *permissible*, conventional transgressions are more likely to be judged permissible than moral transgressions (Smetana 1985; Smetana and Braeges 1990).

Second, conventional transgressions are judged *differently* from (less modifiable than) moral transgressions. For example, conventional transgressions are judged more *rule contingent* than moral transgressions; i.e., individuals are more likely to state that conventional, rather than moral, transgressions are permissible in the absence of prohibiting rules (Nucci 1981; Smetana 1985; Smetana and Braeges 1990; Stoddart and Turiel 1985).

Third, when asked why it is wrong to talk in class or wear opposite gender clothes (conventional transgressions), participants will make reference to established rules that can be either explicit (that action is prohibited in this school) or implicit (that action is 'not the done thing'). In contrast, when asked why it is wrong to hit another or damage their property, participants are significantly more likely to make reference to the suffering of a victim (Turiel 1983).

Importantly, while children and adults usually judge conventional transgressions as less serious than moral transgressions, the distinction between conventional and moral transgressions cannot be reduced to one of seriousness. Participants do not judge all conventional transgressions as less serious than all moral transgressions. However, they always distinguish conventional and moral transgressions in their modifiablity judgments and justifications (see Stoddart and Turiel 1985; Turiel 1983).

It is important to note here that it is the absence of victims that distinguishes conventional from moral transgressions. If a participant believes that a transgression will not result in a victim, he/she will process that transgression as conventional. Smetana (1982) observed that whether an individual treats abortion as a moral transgression or conventional transgression is determined by whether he/she judges the act to involve a victim or not. In addition, Smetana (1985) found that unknown transgressions (specified by a nonsense word; i.e., X has done dool) were processed as moral or conventional according to the specified consequences of the act. Thus, 'X has done dool and made Y cry' would be processed as moral while 'X has done dool and the teacher told him off' would be processed as conventional.

Neuro-cognitive Systems Involved in the Processing of Conventional Transgressions

Kagan and Lamb (1987) argued that morality was distinguished from convention because morality was associated with emotional responding while convention was not. However, we would argue that there are emotional responses associated with conventional transgressions. The teacher experiencing a child continuously talking in the classroom is likely to experience anger. The classroom that has fallen into disorder likely reflects a teacher who lacks sufficient authority to instill respect in his/her pupils; the pupils have no expectations of the teacher's anger or do not find such expectations sufficiently aversive.

We suggest the existence of a system for Social Response Reversal (SRR) that is activated by aversive social cues (particularly, but not limited to, angry expressions) or expectations of such cues (as would be engendered by representations previously associated with such cues; i.e., representations of actions that make other individuals angry). This system is considered to (1) guide the individual away from committing conventional transgressions (particularly in the presence of higher-status individuals); and (2) orchestrate a response to witnessed conventional transgressions (particularly when these are committed by lower-status individuals) (Blair and Cipolotti 2000).

The model has clear ethological roots. The idea is that the SRR evolved as a system for the resolution of hierarchy interactions between conspecifics. Indeed, it has been suggested that the human angry expression evolved to mimic a high-status dominant face (Marsh, Adams, and Kleck 2005). Within-species aggression in most mammalian species is mediated by sub-cortical structures also involved in the basic response to threat. The suggestion is that the SRR is involved in the modulation of this aggressive response; increasing its probability under certain circumstances or decreasing its probability under others. As noted above, conventional transgressions are considered to be bad because of their disruption of the social order (Turiel 1983). Societal rules concerning conventional transgressions function to allow higher-status individuals to constrain the behavior of lower-status individuals. They may also, by their operation, serve to reduce within-species hierarchy conflict.

The activity of the SRR system is thought to be modulated by information on hierarchy and mental state (the latter provided by systems involved in Theory of Mind) (Berthoz et al. 2002). The form of modulation will be dependent on whether the individual is the perpetrator of (or considering being the perpetrator of), or is the witness to, the conventional transgression.

First, an individual is considering perpetrating a conventional transgression. In this situation, high-dominance potential witnesses should augment activity within the perpetrator's SRR engendered by expectations of the anger of these potential witnesses. This should lead to increased activation of alternative response options other than the conventional transgression about to be committed. The anger of high-dominance potential witnesses is to be particularly avoided. Low-dominance potential witnesses should not have this augmentation effect.

Secondly, an individual is a witness to the conventional transgression. In this situation, high-dominance perpetrators should suppress SSR activity, reducing anger in response to the perpetrator. In contrast, low-dominance perpetrators should augment SRR activity, increasing anger to the perpetrator and facilitating a reactive aggressive response to the perpetrator. Mental state information should also modulate SRR activity in witnesses. If the conventional transgression is recognized as intentionally committed, SRR activity and corresponding anger should be augmented, again facilitating a reactive aggressive response to the perpetrator. In contrast, if the conventional transgression is recognized as unintentional, SRR activity and corresponding anger should be reduced, decreasing the probability of a reactive aggressive response to the perpetrator.

The principal neural system particularly implicated in SRR is ventrolateral prefrontal cortex (Brodmann's Area 47). This region is particularly responsive to angry expressions, as well as other emotional expressions (Blair et al. 1999; Sprengelmeyer et al. 1998) and shows increased activity when the individual becomes angry (Dougherty et al. 1999). Importantly, this region shows activity when participants consider conventional

transgressions (Berthoz et al. 2002). Interestingly, this activity is equivalent whether the conventional transgression is intentional or unintentional (e.g., a person taking another person's seat on a bus versus a person accidentally falling into another person's space on a bus).

Given the role of intentionality in moderating reactions to conventional transgressions, the role of Theory of Mind in the processing of these transgressions is of interest. Autism is a severe developmental disorder described by the American Psychiatric Association's diagnostic and statistical manual (DSM-IV) as 'the presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activities and interests' (American Psychiatric Association 1994, 66). Individuals with autism show impairment in Theory of Mind (for reviews, see Baron-Cohen 1995; Hill and Frith 2003).

Children with autism pass the moral/conventional distinction described above (Blair 1996); they can recognize conventional transgressions set in a school-room. In contrast, individuals with autism show difficulty appropriately processing other types of conventional transgressions (Dewey 1991). For example, in one of the vignettes developed by Dewey (1991), an individual hears a baby he does not know cry in the middle of a park and investigates the diaper of the unknown baby in case it has hurt itself. Most people regard this as deeply inappropriate behavior and imagine the caregiver's anger if he/she returned to find an unknown stranger fiddling with his/her baby's diaper. Individuals with autism/Asperger's syndrome have difficulty with this task, probably because they fail to represent the caregiver's false belief that they might have harmful intent for the baby and therefore do not have any expectation of caregiver anger.

In short, at least two types of conventional transgressions can be considered: those where there is an explicit rule ('Do not talk during lessons') and those where the rule is less explicit (the rules for the touching of babies are context-specific and fluid). The argument here is that the SRR system is intact in individuals with autism/Asperger's syndrome, as evidenced in part by their apparently intact recognition of emotional expressions. There have been suggestions that patients with autism have difficulty recognizing the emotional expressions of others. However, the above only applies to studies where the groups have not been matched on mental age. When they are, children with autism have usually been found to be unimpaired in facial affect recognition (Adolphs, Sears, and Piven 2001). In short, SRR functioning, indexed by expression processing, appears intact. However, in situations where the recognition of appropriate social behavior (either as a witness or potential instigator) also requires the representation of mental states, individuals with autism have difficulty.

Patients with lesions of orbital/ventrolateral frontal cortex show SRR impairment. Such patients show difficulties with expression recognition (Blair and Cipolotti 2000; Hornak et al. 2003; Hornak, Rolls, and Wade 1996). Such lesions are a risk factor for engaging in inappropriate behaviors (Blair and Cipolotti 2000; Damasio 1994). In addition, patients with such lesions show difficulties processing conventional transgressions as indexed by performance on Dewey's (1991) social contexts task (Blair and Cipolotti 2000).

This work on emotional modulation of appropriate social behavior has interesting potential links with work on the social emotion of embarrassment. Leary (Leary, Landel, and Patton 1996) and others (Keltner and Buswell 1997) have suggested that embarrassment serves an important social function by signaling appeasement to others. When a person's untoward behavior threatens his/her standing in an important social group,

visible signs of embarrassment function as a non-verbal acknowledgement of shared social standards. Leary argues that embarrassment displays diffuse negative social evaluations and the likelihood of retaliation. The basic idea is that embarrassment serves to aid the restoration of relationships following social transgressions (Keltner and Buswell 1997). There is a good deal of empirical evidence to support this 'appeasement' or remedial function of embarrassment from studies of both humans and non-human primates (see, for a review, Keltner and Buswell 1997). For example, Semin and Manstead (1982) found that people reacted more positively to others following a social transgression if the transgressors were visibly embarrassed. In addition, Leary, Landel, and Patton (1996) presented evidence that people are actually motivated to convey embarrassment to others as a way of repairing their social image. Moreover, it should be noted that patients with orbitofrontal cortex lesions show difficulty in responding to the embarrassment of others (Beer et al. 2003).

If embarrassment does serve an important social function by signaling appeasement, the individual's perceived intention is likely to be crucial in determining whether they are expected to show embarrassment. If an individual intends to socially transgress, we might suspect that he/she will not display appeasement (i.e., embarrassment) afterwards. If the transgression is intentional, the transgressor is unlikely to be interested in the social relationship that has been broken. In contrast, if the violation of the social convention was unintended then we might expect clear displays of embarrassment; the individual will have realized that they have transgressed and wish to restore the social relationship. Recent work suggests that this is indeed the case (Berthoz et al. 2002).

In summary, we suggest that conventional transgressions activate a system for SRR. This system regulates goal-directed behavior in perpetrators (allowing the selection of alternative responses less likely to induce anger in others) as well as basic reactive aggressive impulses in witnesses. We believe that the activity of this system in both parties is modulated by information on hierarchy and mental state.

When processing conventional transgressions, for example in the context of the moral/conventional distinction task, we suggest that the participant judges them non-permissible because they are associated with expectations of anger. We suggest also that the participant's judgment of the seriousness of the transgression will be partly driven by the intensity of the expected anger. Of course, if the rule is removed (in the context of the modifiability judgments), then the participant should no longer expect others to be angry and therefore the action will no longer be judged as a transgression (talking in class is not a conventional transgression if the teacher says that you can talk in class).

Neuro-cognitive Mechanisms Involved in Processing Care-based Morality

Why are humans interested in care-based morality? Rationalist answers are strongly challenged by data from individuals with psychopathy. The classification of psychopathy identifies a relatively homogeneous pathology marked by pronounced emotional impairment (considerably reduced empathy and guilt) and behavioral disturbance (criminal activity and, frequently, violence) (Frick et al. 1994; Hare 1991). Such individuals show no general reasoning deficits (Blair 2004). However, they show pronounced difficulty with care-based morality both in their behavior (Frick et al. 1994; Hare 1991) and in their reasoning on tasks such as the moral/conventional distinction (e.g., Blair 1995).

The Violence Inhibition Mechanism (VIM) model of psychopathy and moral development was an early attempt to provide an account of why humans are interested in care-based morality (Blair 1995). The suggestion was that there was a system, VIM, that, when activated by distress cues (the sad and fearful expressions of others), resulted in increased autonomic activity, attention, and activation of the brainstem threat response system (usually resulting in freezing) (Blair 1995). VIM was thought to be activated whenever distress cues were displayed rather than being reliant upon contextual information about ongoing violence for activation.

Moral socialization, according to the model, occurred as a consequence of the pairing of distress cues, and consequently VIM activation, with representations of the act that caused the distress cues (Blair 1995). These representations of moral transgressions become triggers for the mechanism through their association with distress cues. The appropriately developing child thus initially finds the pain of others aversive and then, through socialization, the thoughts of acts that cause pain to others aversive. The idea was that individuals with psychopathy had disruption to this system such that representations of acts that cause harm to others do not become triggers for the VIM (Blair 1995).

Nichols (2002) provided a two-pronged critique of the VIM model: first, Nichols argued that the VIM model could not explain the processing involved in the moral/conventional distinction task. He argued that it did not provide an account of conventional reasoning and also did not explain how the concept of 'badness' could be understood. In the section above, we began to remedy this situation by describing a model of the neuro-cognitive systems involved in the processing of conventional transgressions. In the section below we will describe the neuro-cognitive systems involved in the processing of moral transgressions.

The second of Nichols' criticisms was that the VIM model did not provide an adequate account of judgments of wrong. The previous view of VIM suggested that actions/events paired with the distress of others would come to be regarded as aversive. However, as Nichols (2002) points out, the class of actions considered to be morally 'wrong' is only a subset of those that would be considered aversive on the basis of VIM activation. To use the example offered by Nichols (2002), natural disasters causing harm to people would be considered bad but not morally wrong. This issue will also be considered below.

The Integrated Emotion Systems Model

The Integrated Emotion Systems (IES) model was developed as an account of emotion (Blair 2004) but it also underpins our view on morality. There are two main components of this model that are particularly relevant here: an emotional learning system mediated by the amygdala and a system for 'decision making' on the basis of reinforcement expectations mediated by medial orbital frontal cortex.

The emotional learning system allows conditioned stimuli (CSs; i.e., representations of moral transgressions) to be associated with the unconditioned stimuli (US) of the victim's distress cues. At the anatomical level, this emotional learning system corresponds to the amygdala. Considerable work attests to the role of the amygdala in the formation of stimulus–reinforcement associations (Everitt et al. 2003; LeDoux 1998) such that it allows previously neutral objects to come to be valued as either good or bad. This emotional learning system functions very similarly to the earlier ideas on the functioning of the

VIM (though there are some notable differences). For one, it allows the individual to learn about both the 'goodness' and 'badness' of objects on the basis of moral socialization. In short, this system is involved in types of stimulus–reinforcement association formation other than simply 'moral transgression'–'distress cue' association (i.e., it will teach you that objects/actions are aversive because they are associated with the distress of others or because they are associated with pain).

Similar to the previous VIM model (Blair 1995), the major claim with respect to psychopathy is that individuals with psychopathy are impaired in the formation of stimulus–reinforcement associations (Blair 2004). It is argued that the expressions of fear and sadness serve as social unconditioned stimuli allowing conspecifics to teach the societal valence of objects and actions to the developing individual. Due to their impairment in the formation of aversive stimulus–reinforcement associations, individuals with psychopathy are less able to take advantage of this 'moral' social referencing and as a result are more difficult to socialize. In short, the position also allows the explanation of impairment in individuals with psychopathy in responsiveness to distress cues, fearful facial and vocal expression recognition, and the processing of the moral/conventional distinction (see Blair 2004). Moreover, work has shown that children with the emotional dysfunction associated with psychopathy are particularly difficult to socialize (Wootton et al. 1997).

The IES Model and Moral Reasoning

According to the IES model, the amygdala allows individuals to learn that specific actions/objects are either good or bad to conduct according to whether these actions/objects are associated with either the recipient's happiness or the victim's distress. Once the individual has learnt about a pro-social behavior/transgression, representation of the action will elicit an integrated emotional response that includes both the amygdala and medial orbital frontal cortex. We argue that this emotional response is effectively the individual's automatic 'moral attitude' to the representation. In line with this position, recent neuro-imaging studies of morality using different methodologies such as making moral decisions based on text descriptions of ethical dilemmas have all implicated both the amygdala and medial regions of orbital frontal cortex (see Luo et al. forthcoming).

We believe the amygdala plays a role in morality by allowing the association of representation of transgressions (interpersonal violence) with the aversive stimulus of the victim's fear/sadness (Blair 1995, 2001). We believe medial orbital frontal cortical activation is involved in decision making and response selection as a function of expected reinforcement information (Blair 2004). We believe medial orbital frontal cortex plays this role also in moral reasoning; it processes the expected reinforcement associated with the action (e.g., aversion engendered by the victim's distress or reward engendered by another's happiness). It uses this expected reinforcement information to determine avoidance or approach of the stimulus that elicited the reinforcement information. This will lead to the modulation of behavior, including verbal behavior. In short, we would propose that an individual's automatic moral attitude to an event involves an integrated neural response involving both the amygdala and medial orbital frontal cortex that is proportional to the emotive strength (due to previous learning) of the stimulus.

To return to the moral/conventional distinction, the suggestion is that the individual uses this 'automatic moral attitude' (i.e., the amygdala and medial orbital frontal cortical response to the stimulus) to influence his/her permissibility and modifiability judgments.

The individual does not regard a moral transgression as permissible partly because it is associated with aversive expectation information (due to its association with the victim's distress). Removing the rule does not alter this emotional response to the representation of the transgression; removing the rule does not alter the expectation of the aversiveness of the victim's distress. In other words, moral transgressions are still considered non-permissible by healthy developing individuals in their modifiability judgments. Finally, with respect to the individual's theories or justifications regarding why moral transgressions are wrong, considerable data demonstrate that stimuli that activate the amygdala lead to increased activity—through reciprocal feedback—of the representations of the stimuli that activated the amygdala (LeDoux 1998). In short, a representation of a transgression that activates the amygdala will be augmented and this representation will receive greater attention. Therefore, when the individual is engaged in a causal analysis of what caused the state of aversion that is the badness of the transgression, the representation perceived as the cause is more likely to be the representation of the transgression and particularly those aspects of the transgression most related to the amygdala activation, i.e., the distress of the victims. In short, individuals can develop a theory that moral transgressions are bad and prosocial behaviors are good because they hurt and help people, respectively.

The above section laid out a model of moral reasoning to provide a fuller account of performance on the moral/conventional distinction task answering, we believe, Nichols' (2002) first criticism. However, Nichols' second and more important criticism, that the VIM model did not provide an adequate account of judgments of wrong, remains unanswered. We will attempt to answer Nichols' second criticism in the section below where the nature of judgments of wrong and the role of Theory of Mind will be considered.

Theory of Mind and Judgments of Wrong

While Theory of Mind, the ability to represent the mental states of others (Frith and Frith 2003; Leslie 1987), was considered with respect to the processing of conventional transgressions, we have yet to consider it with respect to care-based morality. The account of moral reasoning presented in the previous section assumed that Theory of Mind played no role in the moral reasoning described. Such a position must therefore predict that individuals with Theory of Mind impairment will appropriately distinguish between moral and conventional transgressions. As noted above, individuals with autism show impairment in Theory of Mind (Baron-Cohen 1995; Hill and Frith 2003). In line with the suggestion that Theory of Mind is unnecessary for performance on the moral/conventional distinction test, individuals with autism show appropriate distinction of moral and conventional transgressions (Blair 1996).

However, this does not mean that Theory of Mind is irrelevant to moral reasoning. There is a considerable literature indicating the importance of information on the perpetrator's intent when assigning moral blame or praise that began with Piaget (Piaget 1932). The individual who intentionally swings a baseball bat into another individual's face has behaved far more 'wrongly' than the individual who unintentionally swings a baseball bat into another individual's face. In short, analogous to the suggestions above with respect to convention where Theory of Mind can impact upon the functioning of the SRR, Theory of Mind can influence the behavioral choices made by the systems involved in care-based moral reasoning. Intentional acts that harm others are responded to far more strongly than unintentional acts that harm others (cf. Zelazo, Helwig, and

Lau 1996). Moreover, in line with suggestions that Theory of Mind is necessary for the integration of intention information into moral reasoning, individuals with autism show reduced integration of such information into their moral reasoning (Steele, Joseph, and Tager-Flusberg 2003).

But what about judgments of wrong? Nichols suggests that moral judgment depends on two mechanisms: an affective mechanism that is activated by suffering in others and 'a Normative Theory prohibiting harming others' (2002, 226). This Normative Theory does not 'consist of a single simple rule. For instance, at least among adults, the Normative Theory allows that it is sometimes acceptable to harm a child for her long-term benefit' (2002, 226). Unfortunately, it is not completely clear what this Normative Theory does consist of. How precise do the specified conditions have to be? Is it acceptable to harm a child to improve her table manners? Or is it only acceptable to harm a child to prevent her engaging in life-threatening activities? Moreover, it remains unclear how this Normative Theory develops. Should there be individual differences in this Normative Theory? If so, why?

We do not believe it is necessary to propose the existence of a Normative Theory. We believe it is only necessary to consider the interaction of the neural systems involved in Theory of Mind with those engaged in the emotional response to the transgression situation. Actions that are 'wrong' rather than merely 'bad' are acts where there is intent to cause harm. The actions of an intentional agent that cause harm to others are 'wrong'. The actions of an unintentional agent (including natural disasters unless these are attributed to a divine intent) are 'bad'. As the level of victim distress increases, the act comes to be regarded as more 'wrong'/'bad' depending on the intention associated with the action. As it becomes clearer that the intent of the transgressor was to cause harm, the act becomes progressively more likely to be regarded as 'wrong' rather than 'bad'.

There are some situations that might appear contradictory with respect to the current framework; e.g., accidents due to drunk driving or the child punishment example used by Nichols. A drunk driver who backs into five people and kills them is likely to be regarded as 'wrong' rather than merely 'bad'. However, the driver clearly did not harm the five intentionally. We account for this situation with respect to the simulation view of Theory of Mind expounded above. If we represent the driver had the intent to become drunk, we, as part of the affective Theory of Mind process outlined above, generate valenced expectations of likely reinforcement associated with this action; a state of happy well-being with respect to the drunkenness but also, especially when primed by the presented story (and cultured prior expectations), an aversive expectation generated by the victims. In other words, when we calculate the drunk driver's internal mental state we calculate two valenced goals as a function of the automatic operation of the system: the appetitive reinforcement of the drunkenness and the aversive reinforcement of the victim's distress. Because these are expected outcomes of the behavior, they are considered the goals of the behavior. In short, the operation of the system implies that the drunk driver intended to take an action that could be expected to harm the victims and therefore should be considered 'wrong' rather than 'bad'.

With respect to the individual punishing the child 'for her own good', we again, according to the model, represent the punisher's internal state and represent two valenced goals: the aversive reinforcement of the child's distress as well as the appetitive reinforcement of the child's future well-being. In this situation, the judgment becomes not whether we regard the punisher as 'wrong' or 'bad' but 'wrong' or 'right'. According to the model,

this judgment is determined according to whether or not the aversive reinforcement of the child's distress outweighs the appetitive reinforcement of the child's future well-being.

This interaction of neural systems involved in moral reasoning and Theory of Mind is elegantly demonstrated in the work of Joshua Knobe. In Knobe's work, participants are given brief vignettes and then asked to determine whether particular behaviors within those vignettes were performed 'intentionally' (Knobe 2003). For example, two of the vignettes are:

(1) The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
They started the new program. Sure enough, the environment was harmed.

(2) The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
They started the new program. Sure enough, the environment was helped.

Most participants (85 per cent) consider that the chairman 'intentionally' harmed the environment (vignette 1) but few (23 per cent) consider that the chairman 'intentionally' helped the environment despite the identical vignette structure (Knobe 2003). While the interpretation of these data is debated, we would account for them by reference to the simulation view of Theory of Mind expounded above. If we represent the chairman had the intent to start the program, we, as part of the affective Theory of Mind process outlined above, generate valenced expectations of likely reinforcement associated with this action; reward with respect to the increased profit but punishment with respect to the harmed environment (in vignette 1). In other words, when we calculate the chairman's internal mental state we calculate two oppositely valenced expectations as a function of the automatic operation of the system: the appetitive reinforcement of the profit and the aversive reinforcement generated by the harmed environment. Because these are expected outcomes of the behavior (and are represented clearly because of their difference valences), they are considered the goals of the behavior. In short, the operation of the system implies that the chairman did intend to harm the environment; this was a valenced expectation attached to the goal. For the second vignette, when we calculate the first chairman's internal mental state we calculate two similarly valenced expectations; the appetitive reinforcement of the profit and the appetitive reinforcement generated by the helped environment. Because of the similar valence, the reinforcement expectation can be tagged simply to the expected profit; i.e., the chairman did not intend to help the environment. If this affect-based story is correct, individuals with psychopathy should show impairment on this task; in particular, they should not attribute negative intent to the chairman in vignette 1. We are currently testing this prediction.

In short, we believe that healthy individuals label behaviors as wrong if the action is intentional and generates aversion engendered by expectations, or the presence, of victims. Aversive consequences of actions will be considered intentional even if they

were not the individual's goal if they are expected consequences of the action (e.g., killings by drunk drivers). If the action can be expected to lead to future positive reinforcement even if currently aversion is being engendered by a victim, they may not be considered wrong (e.g., for some individuals, physical punishment).

General Conclusion

The goal of this paper was to specify the neuro-cognitive systems involved in mediating different aspects of morality. In this paper, the systems thought to be involved in social convention and distress care-based morality were considered. However, as depicted in Figure 1, we consider that there are at least two additional, partially separable (at the neuro-cognitive level) 'moralities'. These are justice/fairness and disgust-based morality. Relatively little work has considered justice/fairness, certainly from a neuro-cognitive perspective. However, this relative paucity is likely to be rapidly addressed.

More work has considered disgust-based morality (Haidt 2001; Nichols 2002). Disgusted expressions, like fearful, sad and happy expressions, are reinforcers. Usually, they provide information about foods (Rozin, Haidt, and McCauley 1993). In particular, they allow the rapid transmission of taste aversions; the observer is warned not to approach the food that the expresser is displaying the disgust reaction to. Disgusted expressions have been shown to engage the insula and putamen (Phillips et al. 1997; Sprengelmeyer et al. 1998) and patients with damage to the insula present with selective impairment for the recognition of disgusted expressions (Calder et al. 2000). While disgusted expressions frequently convey information about foods, they are also used to convey distaste at another individual's actions. In short, we can develop a disgust-based morality; the emotional force behind the proscribed actions is not anger, as is the case for social conventions, or sadness/fear, as is the case for care-based morality, but disgust. Given disgust-based learning recruits regions currently not thought to be dysfunctional in psychopathy, we believe that disgust-based morality may be intact in individuals with psychopathy. Work is under way to test this hypothesis.

We described here systems involved in social convention and care-based morality. We suggested both gain their power by being 'built upon' basic emotional responses. We suggested that both of these forms of morality are mediated by at least partially separable neural systems. Finally, we suggested that while Theory of Mind is not necessary to learn about social conventions or care-based morality, it is necessarily involved in much moral (conventional and care-based) reasoning. Theory of Mind, for example, is a developmental prerequisite for understanding that something is not merely bad, but morally wrong.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the NIH: NIMH.

REFERENCES

ADOLPHS, R., L. sears, and J. PIVEN. 2001. Abnormal processing of social information from faces in autism. *Journal of Cognitive Neuroscience* 13 (2): 232–40.

AMERICAN PSYCHIATRIC ASSOCIATION. 1994. *Diagnostic and statistical manual of mental disorders*. 4th ed. (DSM-IV). Washington, D.C.: American Psychiatric Association.

BARON-COHEN, S. 1995. *Mindblindedness: An essay on autism and theory of mind*. Cambridge, Mass.: MIT Press.

BEER, J. S., E. A. HEERAY, D. KELTNER, D. SCABINI, and R. T. KNIGHT. 2003. The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology* 85: 594–604.

BERTHOZ, S., J. ARMONY, R. J. R. BLAIR, and R. DOLAN. 2002. Neural correlates of violation of social norms and embarrassment. *Brain* 125 (8): 1696–708.

BLAIR, R. J. R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1–29.

——. 1996. Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders* 26: 571–79.

——. 2004. The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain and Cognition* 55 (1): 198–208.

BLAIR, R. J. R., and L. CIPOLOTTI. 2000. Impaired social response reversal: A case of 'acquired sociopathy'. *Brain* 123: 1122–41.

BLAIR, R. J. R., J. S. MORRIS, C. D. FRITH, D. I. PERRETT, and R. DOLAN. 1999. Dissociable neural responses to facial expressions of sadness and anger. *Brain* 122: 883–93.

CALDER, A. J., J. KEANE, F. MANES, N. ANTOUN, and A. W. YOUNG. 2000. Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience* 3: 1077–78.

COLBY, A., L. KOHLBERG, J. GIBBS, and M. LIEBERMAN. 1983. A longitudinal study of moral judgement. *Monographs of the Society for Research in Child Development* 48: 124.

DAMASIO, A. R. 1994. *Descartes' error: Emotion, rationality and the human brain*. New York: Putnam (Grosset Books).

DEWEY, M. 1991. Living with Asperger's syndrome. In *Autism and Asperger's syndrome*, edited by U. FRITH. Cambridge: Cambridge University Press.

DOUGHERTY, D. D., L. M. SHIN, N. M. ALPERT, R. K. PITMAN, S. P. ORR, M. LASKO, et al. 1999. Anger in healthy men: A pet study using script-driven imagery. *Biological Psychiatry* 46 (4): 466–72.

EVERITT, B. J., R. N. CARDINAL, J. A. PARKINSON, and T. W. ROBBINS. 2003. Appetitive behavior: Impact of amygdala-dependent mechanisms of emotional learning. *Annual New York Academy of Sciences* 985: 233–50.

FRICK, P. J., B. S. O'BRIEN, J. M. WOOTTON, and K. MCBURNETT. 1994. Psychopathy and conduct problems in children. *Journal of Abnormal Psychology* 103: 700–7.

FRITH, U., and C. D. FRITH. 2003. Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences* 358: 459–73.

GALLESE, V., C. KEYSERS, and G. RIZZOLATTI. 2004. A unifying view of the basis of social cognition. *Trends in Cognitive Science* 8: 396–403.

GOPNIK, A., and A. N. MELTZOFF. 1997. *Words, thoughts, and theories*. Cambridge, Mass.: MIT Press.

GREENE, J., and J. HAIDT. 2002. How (and where) does moral judgment work? *Trends in Cognitive Science* 6: 517–23.

HAIDT, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108 (4): 814–34.

HARE, R. D. 1991. *The hare psychopathy checklist—revised*. Toronto: Multi-Health Systems.

HILL, E. L., and U. FRITH. 2003. Understanding autism: Insights from mind and brain. *Philosophical Transactions of the Royal Society B* 358: 281–89.

HOLLOS, M., P. LEIS, and E. TURIEL. 1986. Social reasoning children and adolescents. *Journal of Cross Cultural Psychology* 17: 352–74.

HORNAK, J., J. BRAMHAM, E. T. ROLLS, R. G. O. MORRIS, J. DOHERTY, et al. 2003. Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain* 126: 1691–712.

HORNAK, J., E. T. ROLLS, and D. WADE. 1996. Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal damage. *Neuropsychologia* 34: 247–61.

KAGAN, J., and S. LAMB. 1987. *The emergence of morality in young children*. Chicago: University of Chicago Press.

KELTNER, D., and B. N. BUSWELL. 1997. Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin* 122 (3): 250–70.

KNOBE, J. 2003. Intentional action and side effects in ordinary language. *Analysis* 63: 190–93.

LEARY, M. R., J. L. LANDEL, and K. M. PATTON. 1996. The motivated expression of embarrassment following a self-presentational predicament. *Journal of Personality* 64: 619–37.

LEDOUX, J. 1998. *The emotional brain*. New York: Weidenfeld and Nicolson.

LESLIE, A. M. 1987. Pretense and representation: The origins of 'theory of mind'. *Psychological Review* 94: 412–26.

LUO, Q., M. NAKIC, T. WHEATLEY, R. RICHELL, A. MARTIN, and R. J. R. BLAIR. Forthcoming. The neural basis of implicit moral attitude—an IAT study using event-related FMRI. *NeuroImage*.

MARSH, A. A., R. B. ADAMS, and R. E. KLECK. 2005. Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychology Bulletin* 31: 1–14.

MOLL, J., R. DE OLIVEIRRA-SOUZA, and P. J. ESLINGER. 2003. Morals and the human brain: A working model. *Neuroreport* 14: 299–305.

NICHOLS, S. 2002. Norms with feeling: Towards a psychological account of moral judgment. *Cognition* 84 (2): 221–36.

NUCCI, L. 1981. Conceptions of personal issues: A domain distinct from moral or societal concepts. *Child Development* 52: 114–21.

PERNER, J. 1998. Simulation as explication of prediction-implicit knowledge about the mind. In *Theories of theories of mind*, edited by P. Carruthers and P. K. Smith. Cambridge: Cambridge University Press.

PHILLIPS, M. L., A. W. YOUNG, C. SENIOR, M. BRAMMER, C. ANDREWS, A. J. CALDER, et al. 1997. A specified neural substrate for perceiving facial expressions of disgust. *Nature* 389: 495–98.

PIAGET, J. 1932. *The moral development of the child*. London: Routledge and Kegan Paul.

PREMACK, D., and G. WOODRUFF. 1978. Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences* 1 (4): 515–26.

ROZIN, P., J. HAIDT, and C. R. MCCAULEY. 1993. Disgust. In *Handbook of emotions*, edited by M. LEWIS and J. M. HAVILAND. New York: The Guilford Press.

SEMIN, G. R., and A. S. MANSTEAD 1982. The social implications of embarrassment displays and restitution behaviour. *European Journal of Social Psychology* 12 (4): 367–77.

SMETANA, J. G. 1981. Preschool children's conceptions of moral and social rules. *Child Development* 52: 1333–36.

———. 1982. *Concepts of self and morality: Women's reasoning about abortion*. New York: Praeger.

———. 1985. Preschool children's conceptions of transgressions: The effects of varying moral and conventional domain-related attributes. *Developmental Psychology* 21: 18–29.

SMETANA, J. G., and J. L. BRAEGES. 1990. The development of toddlers' moral and conventional judgments. *Merrill-Palmer Quarterly* 36: 329–46.

SONG, M., J. G. SMETANA, and S. Y. KIM. 1987. Korean children's conceptions of moral and conventional transgressions. *Developmental Psychology* 23: 577–82.

SPRENGELMEYER, R., M. RAUSCH, U. T. EYSEL, and H. PRZUNTEK. 1998. Neural structures associated with the recognition of facial expressions of basic emotions. *Proceedings of the Royal Society of London, Series B* 265: 1927–31.

STEELE, S., R. M. JOSEPH, and H. TAGER-FLUSBERG. 2003. Brief report: Developmental change in theory of mind abilities in children with autism. *Journal of Autism and Developmental Disorders* 33: 461–67.

STODDART, T., and E. TURIEL. 1985. Children's concepts of cross-gender activities. *Child Development* 56: 1241–52.

TURIEL, E. 1983. *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

WELLMAN, H. M. 1990. *The child's theory of mind*. Cambridge, Mass.: Bradford Books, MIT Press.

WOOTTON, J. M., P. J. FRICK, K. K. SHELTON, and P. SILVERTHORN. 1997. Ineffective parenting and childhood conduct problems: The moderating role of callous-unemotional traits. *Journal of Consulting and Clinical Psychology* 65: 292–300.

ZELAZO, P. D., C. C. HELWIG, and A. LAU. 1996. Intention, act, and outcome in behavioural prediction and moral judgment. *Child Development* 67: 2478–92.

James Blair (author to whom correspondence should be addressed), Chief, Unit on Affective Cognitive Neuroscience, Mood and Anxiety Disorders Program, National Institute of Mental Health, 15K North Drive, Room 206, MSC 2670, Bethesda, MD 20892-2670, USA. E-mail: blairj@intra.nimh.nih.gov

Abigail Marsh, Mood and Anxiety Disorders Program, National Institute of Mental Health, National Institute of Health, Department of Health and Human Services, Bethesda, MD 20897, USA. E-mail: marsha@mail.nih.gov

Elizabeth Finger, Mood and Anxiety Disorders Program, National Institute of Mental Health, National Institute of Health, Department of Health and Human Services, Bethesda, MD 20897, USA. E-mail: fingere@intra.nimh.nih.gov

Karina Blair, Mood and Anxiety Disorders Program, National Institute of Mental Health, National Institute of Health, Department of Health and Human Services, Bethesda, MD 20897, USA. E-mail: peschark@mail.nih.gov

Qian Luo, Mood and Anxiety Disorders Program, National Institute of Mental Health, National Institute of Health, Department of Health and Human Services, Bethesda, MD 20897, USA. E-mail: luoj@mail.nih.gov