Philosophical Issues, 18, Interdisciplinary Core Philosophy, 2008

WHAT ARE MORAL INTUITIONS AND WHY SHOULD WE CARE ABOUT THEM? A NEUROBIOLOGICAL PERSPECTIVE*

John Allman and Jim Woodward
California Institute of Technology

1.

Although appeals to "moral intuition" are ubiquitous in contemporary moral philosophy, there is little agreement about either the nature of moral intuition itself or its legitimate role in moral reasoning. Some philosophers suggest that in morality, intuition is nothing more than a repository of misinformation, and biases that lack any rational justification (Singer 1974). Others advocate a methodology in which agreement with intuition is close to a necessary condition for the acceptability of a moral theory (Kamm 2007).

This paper explores a number of issues concerning the role of intuition in moral argument. Our strategy is to use what is known empirically about the neural and psychological structures underlying intuitive judgments, both in morality and more generally. We will argue that these empirical results help to constrain the legitimate uses of appeals to intuition

Philosophers and others have meant many different things by "moral intuition" and as a result any characterization of this notion must be somewhat stipulative.[1] Nonetheless, we think the following description from the psychologist Jonathan Haidt captures a number of its central features:

> [moral intuition] is the sudden appearance in consciousness of moral judgment, including affective valence (good–bad, like–dislike) without any conscious awareness of having gone through steps of search, weighing evidence or inferring a conclusion. (2001)

Haidt gives as an example the immediate judgment most people have that brother-sister incest is wrong, even in a case in which the most obvious

forms of harm are stipulated to be absent—the pair are consenting adults, there is no possibility of pregnancy, and so on. When subjects are asked to justify their judgment, they appeal initially to possible harms/bad consequences (creation of a child with birth defects, etc.) and then, when reminded that these harms are absent, retreat to saying the action just seems wrong, although they cannot explain why. Haidt describes this as "moral dumbfounding", a notion to which we return below.

Haidt's example illustrates a number of features that characterize moral intuition. The natural contrast is with the results of deliberate, explicit reasoning from a previously accepted moral theory or set of rules. Intuition tends to be fast and relatively automatic, while reasoning is typically slower and involves deliberate effort and conscious awareness of a train of thought leading to a conclusion. As Haidt says, the deliverances of intuition simply appear in consciousness, typically without the subject being consciously aware of the processes lead to the intuition. While people may, with varying degrees of success, offer theories or rules that justify or rationally reconstruct their intuitive judgments, intuitions seem to precede any theoretical reconstruction that follows.

A second illustration which we introduce for future reference is provided by the well-known trolley problem. A run-away trolley is headed toward 5 people and will kill them unless diverted by a switch, in which case it will kill 1 person. Most people judge it permissible to flip the switch. On the other hand, most people judge it impermissible to push a large man in front of the trolley, killing him but stopping the trolley and saving the five. Almost no one is able to provide a reasoned justification for these judgments; instead, they present themselves as immediate reactions whose source or basis is not readily consciously accessible. (Hauser, Young, Cushman 2008)

Among philosophers who take a sympathetic view of moral intuition, two analogies (sometimes mixed together) predominate. Some writers adopt a *rationalist* picture: moral intuition, at least when "veridical", is like insight into logical, mathematical or other "*a priori"* truths: like recognizing that the angles of a Euclidean triangle "must" sum to 180 degrees. On this view, truths revealed by moral intuition are self-evident, rationally compelling, and independent of any empirical presuppositions in the same way (it is supposed) mathematical truths are. These truths are discovered through general reasoning abilities, rather than our capacities for emotional response or more specialized capacities for social cognition

Frances Kamm (1993) endorses something like this conception of moral intuition in describing her preferred moral methodology:

> [one] begins with responses [that is, "intuitions"] to particular cases—either detailed practical cases or hypothetical cases with just enough detail for hypothetical purposes. [One then tries] to construct more general principles from these data. . .

She continues:

> The responses to cases with which I am concerned are not emotional responses but judgments about the permissibility or impermissibility of certain acts. . ..
>
> These judgments are not guaranteed to be correct [but] if they are, they should fall into the realm of *a priori* truths. They are not like racist judgments that one race is superior to another. The reason is that the racist is claiming to have "intuitions" about empirical matters and this is as inappropriate as having intuitions about the number of the planets. . . Intuitions are appropriate to ethics because ours is an a priori, not an empirical investigation. (1993, p. 8)

A second, quite different analogy compares moral intuition to ordinary sense (typically visual) perception (cf. McDowell 1985). In the paradigmatic case, one has an intuitive moral response to an actual experience—e.g., one sees someone being tortured and has the intuition that this is wrong. The suggestion is that this intuitive assessment is not just *prompted* by perception, but is relevantly *like* perception—it is direct, immediate, and relatively automatic in the way perception is and also can be veridical, if the subject is in the right position or has the right sensitivities. The common idea that moral intuitions stand to moral theory in something like the way that observations in science stand to scientific theorizing also draws on the idea that moral intuition is relevantly like perception

Another issue running through discussions of moral intuition concerns the role of emotional processing. As we understand her, Kamm advocates the following combination of views: (1) if emotion plays a role in one's intuitions ("responses"), these are not legitimate data for moral theorizing. (Presumably because emotions detract from rationalistic credentials of intuitions) (2) Fortunately, it is possible to have intuitions that are not infected with emotion. (3) These are the ones on which we should rely. Interestingly, many critics of the use of intuition in moral argument agree with Kamm that, when present, emotion often distorts moral judgment, but also think this infection unavoidable, and hence that intuitions are unreliable. Thus the critics accept (1) but reject (2). We argue below that many paradigmatic examples of moral intuition involve neural areas that are associated with emotional processing, and that this processing plays a causal role in generating intuitions. However, contrary to (1), in the right circumstances involvement of emotion can lead to normatively superior judgment and decision-making.

Regardless of one's views about the normative status of moral intuition, it seems uncontroversial that human beings have such intuitions and rely on them in judgment and decision-making. Thus one may inquire, in a naturalistic vein, about the psychological and neural systems that underlie such intuitions and about how these relate to the systems associated with other sorts of psychological and reasoning processes. For example, to what extent do systems associated with moral intuition overlap with those involved

in mathematical reasoning or visual perception, and what is the role of emotional processing in such systems? What happens to people's moral judgment and behavior when these systems are compromised? These are important empirical questions in their own right, but one might also hope that a better understanding of the sources and character of moral intuition will help to clarify whether and when it has a legitimate role in moral argument

2.

We begin by summarizing the general picture of moral intuition we will defend. We think of moral intuition as belonging to the general category of *social cognition.* Social cognition has to do with information processing involved in navigating the human social world: predicting the behavior and mental states of others, evaluating these (as potentially beneficial or harmful to oneself or those one cares about), and responding appropriately. We follow a great deal of recent theorizing by thinking of social cognition in terms of a *dual process* model. Such models suggest that social cognition involves two kinds of processing: slow deliberate serial reasoning in which we self-consciously weigh the evidence for alternative hypotheses and, in contrast, processing that is relatively fast and automatic, where recruitment of emotion may play an important role. It is social cognition involving this second sort of processing that we specifically associate with moral intuition. Our reason is that the neural areas activated when subjects have "moral intuitions", at least when these involve responses to complex multi-faceted moral decision tasks characteristic of moral dilemmas, seem to be just the areas active in aspects of social cognition involving automatic, affect-laden processing. Furthermore, damage to these areas affects moral intuition. The areas in question, which include orbito-frontal, insular, and anterior cingulate cortices and the amygdala, are involved in the processing of various complex social emotions such as guilt, embarrassment, resentment resulting from unfair treatment, and in the recognition of emotions in others (Shin et al. 2000; Berthoz et al. 2002; Singer et al. 2004a; Sanfey et al. 2003). The first three structures are also involved in the detection and monitoring of visceral, bodily sensations, including those associated with food ingestion and expulsion and with introspective awareness of one's own feelings (Craig 2004; Critchley et al. 2004). They are also involved in empathy (Singer et al. 2004b), and in making decisions under conditions of social uncertainty regarding the behavior of others (Sanfey et al. 2003; Singer et al. 2004a). This suggests a connection between intuitive social cognition and awareness of one's own feelings, perhaps because we use the latter via some simulation process to predict and understand the behavior of others. (See below)

Although the outputs of intuitive social cognition may be conscious (taking the form of judgments that, e.g., one trusts someone, or has just embarrassed them), the processing leading to these judgments often is not. Fast processing is necessary for successful real time prediction of others' behavior in socially complex and highly interactive situations and for generating suitable responses to such behavior. Often such processing involves the unconscious integration of many disparate social cues and considerations into a coordinated response Both this complexity and high dimensionality and the need for quick responses mean that subjects cannot just rely on conscious deliberation and calculation. Emotional processing can play a role in facilitating quick appropriate responses. More generally, the involvement of emotions can improve the quality of moral decision-making via a variety of routes: by focusing attention, enhancing empathetic identification, and (via simulation) aiding in the acquisition of information about their beliefs, intentions, and motives.

Understanding the structures and processes involved in moral intuition should lead us to reject many common ideas about this notion. First, there is no specialized or dedicated faculty devoted just to moral intuition or, for that matter, moral cognition. Instead, our capacity for moral intuition largely derives from our more general capacities for social cognition. When we have moral intuitions, we are not detecting *sui generis* "non-natural" properties—instead we are responding to features of our social world, as well as features of the natural world affecting what we care about. Second, as noted, there is considerable empirical evidence that the neural areas involved in many paradigmatic cases of moral intuition are also centrally involved in emotional processing. Moreover, much of this emotional processing is either (i) unconscious and/or (ii) such that subjects are not able to tell whether it influences their responses. As a consequence, subjects are often to follow advice to discount intuitions in which emotion has played a role.

The case for rejecting intuitions involving emotional processing presumably rests on the assumption that intuitions not involving emotion or affect are likely to lead to better moral judgment and decision-making. However, there are reasons to doubt this is always or even usually the case. Subjects with damage to areas involved in emotional processing (and in moral intuition) but intact processing in other areas make decisions that in terms of their effects on self and others are "bad" by the standards of virtually all widely accepted criteria for prudential and moral decision-making (Damasio 1994). The empirical evidence thus suggests that good moral decision-making often requires the involvement of emotional processing and affect, which is not to deny that it also involves processes that look more purely cognitive. While we agree with Sinnott-Armstrong's (2006) claim that the involvement of certain kinds of emotion in moral judgment can lead to the neglect of morally relevant considerations, and a narrowing of moral focus, we also think that the involvement emotions under the right condition can have the opposite

effect, broadening the range of considerations to which the agent responds. Moreover, what ultimately matters morally is not just judgment, but choice and action. In addition to its potentially beneficial contribution to moral judgment, the involvement of emotion in moral decision-making can provide a motivation for action that may otherwise be lacking.

The role of social/emotional processing in the generation of moral intuition also leads us to reject the common assimilation of moral intuition to visual perception or to insight into *a priori* truths. While it is true that some areas involved in moral intuition and judgment (such as the amygdala) are also closely linked to structures involved in visual processing, and while emotion certainly influences visual attention, many areas involved in moral intuition are not part of the "visual" system, however expansively construed. Moreover, it is not true that subjects with intact visual processing but damage to areas involved in social cognition and emotional processing (such as insula and VMPFC) make the same intuitive moral judgments as normal subjects, and still less is it true that such subjects exhibit the same behavior. In addition, subjects with intact emotional processing and social cognition but impaired visual processing seem to exhibit no differences in moral intuition from normal subjects with unimpaired visual processing. Similarly, if moral intuition is a special case of insight into *a priori* truths, one would expect subjects with damage to emotional processing areas but intact areas that are known to be involved in logical or mathematical reasoning (e.g. intraparietal sulcus) to have unaffected intuitions—again, this is not what is found.

We suggested above that it is illuminating to connect moral intuition with so-called automatic processing within a dual processing framework. Nonetheless, we reject certain claims about moral intuition sometimes associated with that framework. According to such claims, the automatic processes underlying social cognition are not just different from deliberative processes but are also "primitive", unsophisticated, retained in relatively unaltered form from our primate ancestors (Greene 2004, p. 389), and relatively unmodifiable by experience. In addition, these processes are seen as error-prone and as requiring correction via the intervention of more deliberative processes which have all of the virtues that automatic process lack—deliberative processes are flexible, sophisticated, and distinctively human. It is thus concluded that judgment tends to be normatively superior to the extent it is more fully controlled by the deliberative system. Our contrary view is that emotional processing and the structures underlying moral intuition can be heavily influenced by learning and experience, although the learning in question is often implicit and subjects often have difficulty formulating what is learned as explicit rules. This implicit learning can be quite flexible and can involve highly complex information processing. Moreover, many of the structures identified above as underlying moral intuition and social cognition have *not* been retained in unaltered form from other primates—instead, these structures have undergone very substantial changes in humans and support

forms of social cognition and emotional processing in humans not present in other primates.[2] Rather than its being the case that judgment and decision-making are always improved by relegating them entirely to the control of the deliberative system and removing any influence of automatic processing, we think the empirical evidence suggests that normatively good decision-making in both the prudential and moral realms requires the integrated deployment of both the automatic and deliberative systems (and cognition and emotion).

3.

The normative issue that most interests moral philosophers is the question of whether and how intuitive reactions are relevant to moral assessment. The temptation is to respond by constructing either a general defense or an equally general condemnation of moral intuition. We will not follow either course here. We think better questions to ask are these: Do intuitive moral responses sometimes contain information recognizable as relevant to good moral decision-making? If so, what is this information? Does "intuition" sometimes play a functional role in moral decision-making not played by other psychological processes? What, if anything, would we lose if we ignored such intuitive responses?

Let us begin with some non-moral examples in which intuition plays or fails to play a role in decision-making. Klein (2003) describes an incident involving two NICU nurses, one expert, the other a relative novice. The novice is taking care of a premature infant who seems (to her) stable and in no danger, although she notes that the baby seems somewhat lethargic, and that her temperature has fallen, but is still within the normal range. Then the experienced nurse notices that 'something about the baby "just looked funny" (Klein p. 16), that the baby seems "mottled", and that her belly is "rounded". After questioning the novice nurse, the experienced nurse tells the doctor the baby is in "big trouble"—as indeed she is, since she is in the early stage of a sepsis infection which likely would have been fatal if not treated immediately.

What is going on in this example? The experienced nurse has learned from past experience that certain cues are diagnostic of an infant in serious medical difficulty (For example, the rounded stomach indicates the baby was not digesting her food). This learning process involves the nurse being exposed to relevant cues (e.g., having to do with the physical appearance and behavior of the baby) in a series of cases, formulating a judgment (e.g., baby X is ill, baby Y is not ill) on the basis of these cues, and then receiving feedback about this response, typically taking the form of results of independent tests establishing whether the baby is ill, hence whether the nurse's initial judgment was correct. Over time, with the right experience and feedback, the nurse's

judgment improves—she goes from being a novice to an expert. However, at least to a large extent, the nurse's learning is likely to be implicit in the sense that she may not be fully aware of the cues on which she relies or able to explicitly formulate some rule which guides or reproduces her judgments and which she can explain to others. As Klein notes, another distinguishing feature of the expert is that she is able to integrate cues in a holistic, interactive fashion—the novice nurse recognized a number of the individual cues but failed to put them together to reach a correct diagnosis.

Examples of such implicit learning have been widely reported in the psychological literature (e.g., Lewicki et al. 1987) There is strong evidence that such learning occurs in the social domain and in connection with tasks that are emotionally (or morally or prudentially) important to the learner (Lieberman 2000).

One of the best known examples of normatively significant implicit learning is provided by the Iowa Gambling Task (Bechara 1994). Subjects choose among card decks and win or lose money depending on which cards are drawn. Some ("bad") decks have overall negative pay-off, although they contain some cards with high positive rewards. Other ("good") decks have a net positive pay-off, but contain some cards with negative pay-offs. The pay-off maximizing strategy is to draw from the good decks and avoid the bad. Normal subjects learn this fairly quickly, well before they are able to formulate explicit reasons for doing so. Indeed, measurement of their galvanic skin responses shows aversion to bad decks develops before any conscious decision to avoid them, with some subjects reporting an initial "gut feeling" that they should avoid certain decks. By contrast, subjects with damage to orbitofrontal cortex show quite different behavior, continuing to draw from bad decks long after normal subjects abandon them, and indeed, in some cases even after they become consciously aware that they are losing money from these decks. OFC patients also fail to show the galvanic skin responses of normals. According to Damaiso's "somatic marker" hypothesis, normal subjects receive, as a result of implicit learning, emotional signals that influence them to avoid bad decks before consciously formulating reasons for doing so. In contrast, OFC patients don't access the same emotional signals and hence fail to exhibit the implicit, emotion-based learning characteristic of normal subjects. In this case, at least, this seems to contribute to normatively inferior choices.

Might a parallel story be told about moral intuition? Let us first sketch how such a story might go and then consider its plausibility and implications. Suppose one is trying to decide whether it would be morally right to perform some action A, such as pushing someone in front of a trolley. Suppose imagining doing A (or where this is more appropriate, imagining someone else doing A) generates the intuitive reaction that this would be wrong. This negative intuitive reaction will often include a strongly aversive emotional signal. An agent who is guided in part by this signal will be in this respect

like the normal subjects in the Iowa Gambling Task or the skilled NICU nurse. For example, someone might take the strong reaction of disgust and outrage she felt upon seeing the photographs of prisoner abuse at Abu Ghraib as a prima-facie indication (an emotional signal) of the wrongfulness of such treatment. Similarly, for the intuitive reaction that it is wrong to push someone in front of a trolley.

4.

The obvious questions raised by this picture are these: why suppose that such reactions ever track anything of moral significance? If they sometimes do track morally significant information, what is this information? In pursuing these issues we immediately encounter the following difficulty: Different moral theories propose different criteria for normative assessment and may take different information into account in arriving at such assessments. If we have to establish which moral theory is correct before we can answer questions about the normative significance of intuition, we are unlikely to make progress on the latter issue. We think that, fortunately, there is a way out of this difficulty: there are some generic features of good moral decision-making about which there ought to be general agreement, regardless of the specific moral theory to which one subscribes. It is sensitivity or responsiveness to these generic features that we believe is sometimes enhanced by involvement of emotional processing and use of "moral intuition".

Before turning to details, however, a caveat is in order. We take it as uncontroversial that moral intuition sometimes does reflect the influence of misinformation, bias, and so on, just as critics allege. This follows from our contention that moral intuition reflects the influence of learning from experience and that when it has some claim to moral significance, this be will because that it tracks certain factors and not others. There will be many cases in which normatively relevant factors have not played a role in shaping people's intuitions. When this is so, there will be no reason to assign any normative significance to them. Our interest is simply in making the case that moral intuition can sometimes reflect information of moral significance; whether it actually does in any particular case will depend on the details of that case.

The first feature to which we draw attention is the highly complex, interactive character of many situations in which we make moral decisions, including those which present themselves as moral dilemmas. Such situations are interactive in the sense that those affected by one's choices are not passive; instead the moral decision-maker's choices (and such other factors as the intentions with which she acts) prompts those affected to *react* by choosing one response rather than another, these responses in turn lead

others (including the initial decision-maker) to additional responses and so on. Thus moral decision-makers often face situations with a *strategic* structure, in which they need to consider how their choices will influence the behavior of those with whom they interact, rather than *parametric* decision problems in which they can assume the behavior of others is fixed, independently of the decision maker's choices. On any plausible moral theory, accurately anticipating the consequences of such patterns of choice and responses will be important in good moral decision-making. Unfortunately, examples in moral philosophy often edit out such features.

As an illustration consider Williams' (1973) example of the explorer, Jim, who must choose between killing one person in which case the local captain will release nineteen others, or doing nothing in which case the captain will kill all twenty. In Williams' presentation and subsequent discussion, it is stipulated that Jim (and the reader) know for certain how others will behave in response to Jim's choices: the captain will keep his word if Jim kills the one; if Jim refuses the captain will carry out his threat. It is also apparently assumed (although not made explicit) the captain will not return to kill more villagers the following week, and so on. We are also given no information (presumably because this is assumed to be irrelevant) about what the captain's motivations are in wishing to involve Jim in the killing.

In real life, by contrast, it will be highly uncertain about how the other actors in this situation will behave and it will need to consider the possibility that the captain will kill the other nineteen if Jim kills the one. (Or should we believe although willing to murder twenty people, the captain can be counted on to keep his promises?) Also at least potentially relevant to Jim's decision are the captain's larger purposes. Does the captain wish to blackmail Jim or draw him into some larger pattern of collaboration or use him to facilitate the achievement of additional morally bad ends, as is often the case in real life examples in which people face similar choices? In real life, Jim's choice should be sensitive to these (and many other) considerations.

Good moral decision-making also requires anticipating not only obvious consequences (e.g., damage to life and limb) but also more subtle psychological consequences (e.g., whether our actions will be perceived as insulting or humiliating.) Both recognizing these non-obvious costs and benefits and correctly anticipating the behavior of others requires information about their mental states: their intentions, beliefs, desires, emotions, and so on. In real-life cases (as opposed to the hypothetical scenarios constructed by many moral philosophers), getting such information and using it effectively will be a very complex problem requiring the integration of many disparate considerations in the face of substantial uncertainty.

Our suggestion is that intuition and emotional processing, when they are the product of the right sort of implicit learning, can reflect complex assessments of the mental states of others and their likely behavior in interactive situations, as well as the likely consequences of such behavior,

appropriately corrected for uncertainty. A large body of evidence suggests we often detect and represent the mental states of others (including their beliefs, preferences, intentions, and emotions) by simulating these via our own emotional processing—by activating the emotional areas and processing in ourselves involved in the those mental states when experienced by others (Damasio 1994). By further simulating how we would behave in the presence of these mental states, we may also predict how others will behave. Simulation (and the empathetic identification it can facilitate) also helps decision-makers to recognize non- obvious psychological costs/benefits of their actions. In turn, the use of simulation has the important additional consequence that it often has some tendency to have motivational force in the simulator's own behavior. Thus when we recognize that another person has been humiliated by simulating this emotion, this also alters our own motivational set—both by directing attention to the humiliation and by encouraging us to react negatively to it.

Moreover, both empirical evidence and theoretical considerations suggest that sometimes assessments based on intuition and emotional processing are not only faster than assessments based on conscious deliberation, but also more accurate and normatively superior (when assessed by uncontroversial criteria.) One reason is that conscious deliberation often operates as "bottleneck", since the amount of information that can be consciously processed and integrated is relatively limited. Conscious deliberators presented with high-dimensional problems tend to focus on just a few dimensions, neglecting other dimensions even when they have (or can readily obtain) information about these, resulting in normatively inferior decisions. Unconscious processing, including processing involving emotions, need not be limited in this way. Thus experimental evidence (Djiksterhuis et al. 2006) shows that consumers provided with information about many relevant attributes of goods, and who are experimentally induced to consciously deliberate show less retrospective satisfaction with their choices than those provided with similar information but induced not to consciously deliberate. Similar results are reported by Wilson (2002) for a variety of different decisions, including those involving romantic relationships.

As an application to moral decision-making, consider the use by the U.S. government of interrogation techniques in Iraq that included, in addition to ordinary physical torture, sexual humiliation religious insults, and threats involving dogs—experiences regarded as highly degrading and offensive within Arab culture and which were imposed for just this reason. Consider also the contention by some American commentators that these procedures were just "fraternity pranks" and thus not morally objectionable. Assume, for the sake of argument, that use of these techniques was prompted in part by a desire to obtain militarily useful information and by the assessment that the value of this information outweighed any associated costs. In retrospect, it seems apparent the decision-makers failed to fully appreciate the impact

of these procedures on the prisoners or the costs to U.S. interests when these activities became public. The U.S. decision-makers also failed to anticipate that once authorized, such techniques would not be used in a limited number of controlled, preauthorized situations (as they claimed they intended) but instead would be used more widely in many episodes of gratuitous sadism without any purpose linked to the obtaining of information, even though, as a historical matter, this usually seems to happen when torture is authorized. More generally, it seems uncontroversial that the decision-makers failed to properly take account of a number of considerations that were both normatively and prudentially relevant, even though information that should have led them to recognize their relevance was readily available. Engagement of moral intuition (and accompanying emotional processing) helps in such cases to enlarge the scope of the considerations taken into account—e.g., the decision maker who is guided in part by such emotional responses will more likely to recognize and be influenced by the humiliating nature of the interrogation techniques employed. These additional considerations need not be fully taken into account as part of a process of explicit, conscious deliberation; instead they may operate in ways not fully or readily consciously accessible and which reflect previous implicit learning, the upshot being overall emotional signals of revulsion toward use of the interrogation techniques just described—a "moral intuition" that the treatment of the prisoners was wrong.

This model suggests that moral judges and decision-makers who do not employ their capacities for emotional processing and the intuitive responses that result from these—either because of neural damage or they have not learned to do so—will also have some tendency to neglect or be insensitive to certain aspects or dimensions of the moral problems they encounter. For example, they may be insensitive to what we called the strategic aspects of such problems: failing to correctly anticipate how others will perceive or react to their behavior or to register their full range of emotions.

An illustration is provided by a patient with developmental frontal damage—hence impaired emotional processing from an early age (Zygourakis, Adolphs and Allman 2006). The patient viewed a documentary film clip from a concentration camp survivor. The survivor and two other boys promised one another to stick together, no matter what. When one of the boys began limping, he was shot by German soldiers, while his friends were too scared to protest.

When asked to rank the three strongest emotions that the narrator feels, the patient listed pain, anger, and sadness but not guilt or shame (which normal subjects rank in their top three emotions). When asked what characters felt, she responded that she "couldn't imagine being in that situation". When asked "Did the person (actor) do the right or wrong thing in the situation depicted in this film clip?" with -5 representing the most morally unacceptable and 5 the most morally acceptable, the patient responded with

a 2, explaining, "I would have done the same [as the narrator]. Either three people would die, or just one."

Whether or not one thinks that the boys made a morally correct (or at least defensible) choice in abandoning their friend, it seems uncontroversial that, unlike the early orbito-frontal-damaged patient, most people do not think the choice was a straightforward one, with the only relevant consideration being number of lives saved. Instead, for most people, considerations having to do with loyalty, solidarity, friendship, and the promise to stick together will also be recognized as relevant; this is why normal subjects rank "shame" and "guilt" as among the emotions that characters in the film are likely to feel.

The OFC patient is clearly sensitive to a narrower range of morally relevant considerations than normal subjects, and this is the result of deficiencies in her ability to empathize and to experience complex social emotions like guilt and shame. This in turn leads her to see the dilemma the boys face as one-dimensional, with the only relevant consideration being number of lives saved, thus biasing her judgment in (what is standardly regarded as) a "consequentialist" direction. Even if we agree with the patient's judgment about this case, it seems likely her judgment will be defective about other cases involving empathetic identification and appreciation of the emotions experienced by others.

We emphasize that in saying that good moral decision-making should take account of the mental states of others and their likely responses to our actions and that intuitive automatic processing can facilitate this, we do *not* mean to assume that some version of consequentialism is correct. Consequentialist theories assume that (1) ultimately *only* consequences are morally relevant and, furthermore, that (2) there is some procedure that permits the summation or aggregation of consequences of different kinds affecting different people into a single index capturing everything morally relevant. Our contention that (at least some of) the generic considerations relevant to good moral decision-making include information about the mental states of others and their likely behavior does not commit us to either (1) or (2) above. Indeed, we think that our view of moral intuition fits naturally with some of the characteristic themes of deontological moral theorizing—for example, the importance that deontologists place on the structure of intentions and their concern with notions like respect. We discuss this immediately below.

5.

In a well-known series of studies, Greene et al. (2001, 2004) found more activation in so-called emotional areas in subjects who make (what the authors describe as) "deontological" choices in comparison with subjects

who make "utilitarian" choices in certain moral dilemmas. For example, this pattern holds for subjects judging it wrong to push the large man into the trolley in comparison with those who judge this is permissible. One apparently natural interpretation, favored by Greene, is that subjects making deontological choices have strongly negative emotional responses to such actions as direct intentional killing, especially by physical contact, but have no corresponding affect-laden responses to causing death in more indirect ways, as by flipping a switch. Without wishing to dispute that this captures part of what happens on when subject make deontological choices, we want to suggest an alternative (but not inconsistent) interpretation which relates these results to other recent empirical findings.

The classification of choices in dilemmas like the trolley problem as deontological versus utilitarian/consequentialist is completely standard in philosophical discussion and has been taken over by neuroscientists who use such scenarios. Nonetheless, the classification is potentially misleading. The choices standardly labeled as utilitarian/consequentialist (such as pushing someone in front of the trolley, killing one person to prevent someone else from killing ten) are not consequences of utilitarianism per se, but rather consequences of a particular version of utilitarianism which is parametric rather than strategic. We lack space to argue for this in detail, but our basic thought is that in a version of consequentialism which pays attention to motives and intentions, and the interactive character of moral decisions, it is far from obvious that choices like pushing someone in front of the trolley or killing one to prevent someone else killing twenty are morally optimal[3]. However, in what follows, we will not assume this claim is correct. We introduce it only to motivate the further idea that in addition to the distinction between consequentialism and deontology, there is another important distinction in moral theorizing: the distinction between theories and analyses that take account of strategic factors like motives, intentions and non-obvious mental states and those that do not. Familiar deontological theories build in reference to such considerations, while consequentialist theories may or may not take them into account, although if they will do so, it will be only in an instrumental way. Another potentially important distinction is between approaches or styles of moral decision-making that appeal to relatively simple rules that (it might appear) can be applied in a fairly mechanical way (e.g., minimize the number of deaths) and those approaches that take account of more complex constellations of considerations in ways that may not be readily captured in simple rules that are accessible to the decision-maker.

With this in mind, consider some additional empirical results. Hauser et al. (2007) found that subjects were readily able to produce a rule to justify their choice in some moral dilemmas (e.g., when the choice was between action and inaction) but not in others (e.g., when the choice involved using people as a means, as with pushing someone in front of

a trolley). In an imaging study, Borg, Hynes, van Horn, Grafton, and Sinnott-Armstrong (2006) found activation in "emotional" areas primarily in dilemmas associated with using people as means but *not* in dilemmas in which the choice is between action and inaction. Finally, Koenigs et al. (2007) found that patients with damage to VMPFC (and hence impaired emotional processing) made more "utilitarian" judgments than normals (which is what one would expect from Greene's results) but only on so-called "hard" dilemmas which are not resolvable by appeal to consciously accessible rules and which are thought on the basis of reaction time data to be difficult and conflict-inducing. (The large man version of the trolley problem is an example) On other, simpler moral dilemmas, VMPFC patients were just as "deontological" as normals.

We suggest that a natural interpretation of these results is that when subjects make a moral judgment or decision that is dictated by the application of a relatively simple rule, little emotional processing need be involved. This is so whether the rule in question is utilitarian-looking ("Minimize the number of deaths") or deontological-looking ("It is worse to produce a bad result via an action rather than an inaction"). Emotional processing tends to be employed by normals when they face (what they regard as) moral problems of some complexity, and where they have no simple rule available to dictate their decisions—this is particularly likely to be true in cases in which complex assessments of the intentions, motives and emotional states of others are required. In such cases, people tend to rely on emotional signals to guide their choices—they have "intuitions" which they are unable to explain or justify by simple rules. They use such signals to guide them to judgments that they do not reach on the basis of conscious calculation. It is their inability to employ such processing that accounts for the distinctive pattern of judgments in VMPC patients found by Koenigs et al, with abnormal judgments appearing only for hard dilemmas.[4]

On this interpretation, the heightened activation in "emotional" areas when subjects make some "deontological" (as opposed to "consequentialist") choices occurs (at least in part) for the following reasons. It is characteristic of deontological theories that they require taking account of such factors as the structure of intentions (whether some outcome is intended as an end or means, or alternatively is a mere side effect, etc.) and that they also attach considerable moral importance to considerations like dignity and respect. Attempting to take such considerations into account in complex cases requires considerable emotional processing, for the reasons outlined above. We thus conjecture that a self-styled utilitarian who, consistently with his utilitarian commitments, takes such considerations into account to the extent that they contribute instrumentally to the maximizations of overall consequences in a complex problem might well exhibit similar activation in emotional areas. If this is correct, there may not be anything distinctively "deontological" about the use of emotional processing.

6.

What follows from these ideas about the legitimate role of intuition in moral argument? First, intuitive responses are more likely to track morally relevant information if they are consequences of processes of learning in which those responses are shaped by experience with real-life events resembling those being evaluated. In some cases this experience will be direct and personal, as is the case for moral intuitions about torture that have been shaped by being tortured or close acquaintance with the victims of torture. These grounds provide a strong prima-facie case for taking the moral intuitions of John McCain and Jacabo Timmerman about torture more seriously than those of Dick Cheney. In other cases, the relevant experience may be more indirect; it may come, for example, from living in a society in which certain practices are permitted and observing the consequences of these. The strong opposition of the founding figures of the U. S. to torture was not based, in most cases, on direct personal experience, but these figures lived in or knew about political cultures in which such mistreatment was widely employed and their intuitions reflected their information about torture as it actually occurred in real life and are worth taking seriously for this reason. Similarly, for the intuitions of people who have lived under regimes in which arbitrary arrests and imprisonment are common. In still other cases, the relevant learning will be vicarious—based, for example, on accounts provided by others of historical or contemporary episodes. For example, those who wonder what happens (and what sorts of intuitions are appropriate) when murder is permitted on the grounds that (it is claimed to be) necessary to save larger numbers from other threats may consult Mandelstam (1970). In yet other cases, people may avail themselves of imaginative identification based on analogy—although few in our society will have direct experience with torture, everyone will have some experience with physical pain and humiliation.

A second, related issue concerns the common philosophical practice of appealing to intuitions about highly unrealistic or even impossible hypothetical examples—examples in which implicit learning is not possible because exemplars of the relevant sort are very rare or non-existent. Here the rationale for the use of intuition outlined above is unavailable. We include under this heading examples that are frankly science-fiction-like such as Judith Thomson's (1971) case of spores that settle in furniture and grow into people. More controversially, we also include examples that, although not physically or biologically impossible, stipulate away features that would ordinarily be present (because of facts about human psychology, social organization, or what people can know). As an illustration, consider a standard "ticking bomb" hypothetical scenario in which a captured terrorist is known to have information about the location of a bomb that unless disarmed will soon kill many people. It is stipulated that the terrorist will reveal the location if and

only if he is tortured, and also that if permitted, torture will only be used on this occasion (and perhaps others exactly like it) and will never be used in other, different situations in which would be unjustified. What should we make of people's intuitions (whatever they might be) about whether torture is justified in such a case? It is arguable that experience with real life cases shows that torture tends not to be a very reliable means for eliciting truth, is rarely if ever the only means for doing so, and if permitted at all tends to be used indiscriminately in cases in which virtually everyone would agree that its use is unjustified. Suppose, for the sake of argument, that these empirical claims are correct. To the extent that a moral judge's intuitions about torture have been shaped by experience, they will reflect the operation of such considerations, rather than any experience with cases like the ticking bomb scenario.

In consequence, when a moral judge is asked to consult her intuitions about ticking bomb cases, the results are unlikely to be illuminating. One possibility is the judge imports reactions formed in response to realistic cases involving torture into her reaction to the hypothetical case—i.e., she may not "correct" for the distinguishing features of the hypothetical case. (Recall that this is essentially what subjects do in Haidt's example involving brother/sister incest) If so, invoking the hypothetical case accomplishes nothing: one might as well stick with realistic examples. Another possibility is that the distinguishing features of the hypothetical case influence the judge's intuition. If so, it is unclear that this intuition reflects anything of moral relevance, since it is not the product of the sort of learning, which (on the account we have presented) sometimes results in intuitions serving as a source of morally relevant information. Relatedly, it is unclear what would count as a normatively appropriate modification in the judge's intuition in order to accommodate the new, hypothetical situation. Moreover, since the processes generating the original intuitions regarding realistic cases of torture are to a substantial extent "automatic" and not consciously accessible, it is also unclear what operations the moral judge should perform to appropriately take account of the distinguishing features of the new scenario or how the judge (or the rest of us) can recognize whether this has been done. In practice, many moral philosophers seem to assume that if the judge is consciously aware of the distinguishing features of the new situation, this ensures that she will be able to appropriately adjust her intuitive response. If anything like our account of moral intuition is correct, we see no basis for this assumption.

To further explore some of these issues consider some remarks of Kamm's about two examples of Singer's. Singer (1993) notes we have the intuitions that it is very wrong not to save a drowning child at the cost of ruining a $500 suit, but that it is not wrong not to save a child from starvation overseas at the cost of $500. He attempts to use the first intuition to motivate a principle underming the second. Kamm (2007) objects, in part on methodological grounds. She holds that in comparing intuitions about

different examples, we must employ "perfectly equalized cases". In particular, ". . . in order to see whether one variable (near/far, kill/let die) makes a moral difference we must compare two cases that differ only with respect to this variable, holding the contextual factors constant" (p. 347). Singer's two cases differ along many dimensions besides the difference in spatial proximity—for example, in pond case, only you are in a position to help while this will not usually be true for starving child cases. To deal with this, Kamm proposes that we should instead compare intuitions about cases differing only along the single dimension of physical proximity. (Kamm suggests this is analogous to a controlled experiment in which only the putative cause is varied across the treatment and control groups.) Thus she compares her intuitions about two cases: In *Near Many* she is near a pond from which she may rescue many children by spending money. Other possible rescuers are present, but fail to rescue. *Far Many* is exactly similar except that children are drowning far away (but she can still save them by spending money). Kamm's intuitions are that she has a stronger obligation to rescue in *Near Many* than in *Far Many*, and similarly for other variants comparing near and far rescues. She takes this to suggest that physical proximity itself is relevant to obligations to rescue.

Let us put aside the issue of whether Kamm's intuitions are "correct", whatever that might mean and focus on her methodological claims. Kamm is correct that Singer's examples differ along many dimensions, but we think it dubious that her method of equalization adequately addresses whatever problems this creates. The difficulty is that merely *telling* oneself (or some other moral judge) to consider cases equalized along all but one dimension, does not ensure that one will not import reactions formed by real life cases, which typically differ along *many* dimensions. Such spill-over from real life cases is particularly likely if, as we have argued, people have limited conscious access to the factors influencing their intuitions, since in this case, they may be unable to detect such influences. If so, people's reactions to imaginary (including equalized) cases will not be influenced by just those cases, considered in themselves, but will by an uncontrolled mixture of other factors as well—thus these reactions will not be relevantly like controlled experiments.

If intuition does not furnish direct access (whether of a quasi-perceptual or rationalistic sort) into a realm of moral truth, what positive or constructive role can it play in moral argument? One of our themes has been that it is plausible to think of it as functioning in moral cases in broadly the same way it functions in non-moral examples (the NICU nurse, the participants in the Iowa gambling task). Following this model, we might think of moral intuition as functioning in part like an alarm bell—a "signal" that grabs our attention and communicates an overall evaluation of a situation and perhaps recommends a course of action. This signal may track or reflect morally relevant information, but of course it also may not—as Singer says,

it may merely reflect bias and misinformation. Thus, if there are doubts about whether an intuition reflects anything of moral importance, these cannot be resolved just by scrutinizing the intuition itself. But what the intuition can do is to direct attention to various aspects of the situation to which the intuition is a response that deserve moral scrutiny. Thus someone who has an aversive gut reaction to certain decks in the Iowa gambling task would do well to direct attention to those decks and explore whether they have features that make avoiding them a normatively appropriate response. If such features are not immediately apparent, it will often be a good strategy to look harder, rather than discarding the intuition as misguided.

We might think of moral intuitions about Abu Grahib similarly: as alarms or signals communicating the assessment that something is wrong and immediate attention/remediation is required. Again we emphasize that the mere occurrence of this signal does not show that it is appropriate. But it should raise our level of suspicion and lead us to explore the possibility that wrong-making features may be present in the situation. In the most straightforward case, we will find independent support (that is, independent of the occurrence of the intuition itself) for the presence (or absence) of such features, but the features themselves may be non-obvious and identifiable only if we immerse ourselves in the empirical details of the situation and exercise our capacities for emotional engagement.

A natural reaction is that if intuitions play the role just described, they must be dispensable "in principle". When making a moral judgment, why not put aside our intuitions, and focus instead on whatever the intuition tracks, and directly assess its moral significance, using whatever moral theory/analysis we think appropriate? The problem with this suggestion should be clear from our discussion above: when humans face complex, multi-dimensional moral problems, they have, as a matter of empirical fact, a strong tendency to fail to adequately take into account some relevant dimensions of the problem, and hence to make normatively inferior decisions. The sorts of moral theories constructed by philosophers, which typically focus on just a few dimensions of assessment, encourage (or at least do not counteract) this tendency. Assigning weight to intuition can help to avoid this danger. In effect, intuition can suggest the presence of "outside" or surprising information (from the view of the decision-maker) that is morally relevant but whose role may not be adequately recognized in the moral theory or system of belief the decision-maker employs.

For these reasons, we favor a methodological use of intuition that moves in the opposite direction from the methodology employed by many contemporary moral philosophers. We think that philosophical discussion should focus on rich, multi-dimensional examples in all their real-life complexity—examples taken from historical, psychological, or social scientific investigations, or where appropriate, from imaginative literature and film. We should try to understand the constellations of features that tend as a

matter of empirical fact to be present in such examples. If the topic is torture, we should try to understand how and in what circumstances and with what motivations and results torture has been used in real-life situations. If the topic is collaboration with evil doers (If some threatens to kill ten people unless I kill one, should I do so?), we should look at actual real life examples in which people collaborate or refuse to do so and ask what tends to happen in such cases, how the collaborators themselves and those with whom they collaborate behave, and so on. The reactions or intuitions on which we should focus should be those of people who have had experience with such real life examples and we should try to understand the empirical features of the examples that shape their intuitions. We should resist the temptation to regard features that, as an empirical mater, tend to be present in realistic examples as inessential aspects we should control for or strip away. Thus if torture, if authorized at all, tends to be used widely in circumstances in which it is unwarranted, we should consider examples having this feature, not hypothetical examples from which this feature is eliminated. When we consider a question like "what are our obligations to needy people in distant countries", rather than focusing on examples involving drowning children (Singer), machines that will rescue large numbers of children if we put in money (Kamm), and people who will bleed on the interiors of cars if taken to the hospital, we should look to detailed studies and overall judgments by developmental economists, anthropologists and others with knowledge and experience relevant to such questions, recognizing that these involve many complex considerations about the effectiveness of aid, the incentives it creates both for the recipients themselves and relevant third parties like local governments. What is morally valuable in the intuitions of knowledgeable, experienced people will reflect the influence of such considerations.

Notes

* Many thanks to Walter Sinnott-Armstrong, Joshua Knobe, and Liane Young for helpful comments. A companion essay to this paper (Woodward and Allman, 2007) provides more neurobiological detail and philosophical background.

1. In particular we will focus only on intuitions understood as responses to particular cases, and not judgments about general principles, such as "intuitions" about the correctness of utility maximization or the categorical imperative as fundamental moral principles. We conjecture that the psychological processes underlying the latter are rather different from those underlying the former, that areas involved in logico-mathematical reasoning are more likely to be active in the latter, and there is less shaping by experience.
2. See e.g. Seeley et al. (2006), Allman et al. (2005).
3. But recall our discussion of realistic versions of Williams' example in which (we suppose) it is not clear Jim should kill the one.
4. A somewhat similar interpretation is suggested in Koenigs et al. (2007).

References

Allman, J., Watson, K., Tetreault, N. and Hakeem, A. 2005. "Intuition and Autism: A Possible Role for Von Economo Neurons." *Trends in Cognitive Science*, 9(8):367–373

Bechara, A., Damasio, A.R., Damasio, H., Anderson, S.W., 1994. "Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex." *Cognition*, 50: 7–15.

Berthoz, S., Armony, J., Blair, R., Dolan, R., 2002. "An fMRI Study of Intentional and Unintentional (embarrassing) Violations of Social Norms." *Brain*, 125, 1696–1708.

Borg, J., Hynes, C., van Horn, J. Grafton, S. and Sinnott-Armstrong, W. 2006. "Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation." *Journal of Cognitive Neuroscience*, 18, 5 803–817.

Craig, A.D., 2004. "Human Feelings: Why Are Some More Aware Than Others?" *Trends in Cognitive Sciences*, 8, 239–241.

Critchley, H. Wiens, S., Rotshtein, P., Öhman, A., Dolan, R. 2004. "Neural Systems Supporting Interoceptive Awareness." *Nature Neuroscience*, 7, 189–195.

Damasio, A., 1994. *Descartes Error*. Norton, Boston.

Dijksterhuis, A., Bos, M., Nordgren, L., van Baaren, R. 2006. "On Making the Right Choice: The deliberation-without-attention effect." *Science*, 311, 1005–1007

Greene, J., Sommerville, R., Nystrom, L., Darley, J., Cohen, J. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science*, 293, 2105–2108.

Greene, J., Nystrom, L., Engell, A., Darley, J. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron*, 44, 389–400.

Haidt, J., 2001. "The Emotional Dog and Its Rational Tail: A social intuitionist approach to moral judgment." *Psychological Review*, 108, 814–834.

Hauser, M., Cushman, F., Young L., Kang-Xing J. Mikhail, J. Forthcoming. "A Dissociation between Moral Judgments and Justifications." *Mind and Language*, 22: 1–21.

Hauser, M., Young, L. and Cushman, F. 2008. "Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions." in W. Sinnott-Armstrong, ed. *Moral Psychology, Volume 2: The Cognitive Science of Morality*. MIT Press, Cambridge, MA.

Kamm, F., 1993. *Morality, Mortality, Volume I: Death and Whom to Save From It*. Oxford University Press, New York.

Kamm, F., 2007. *Intricate Ethics*. Oxford University Press, New York.

Klein, G., 2003. *The Power of Intuition*. Currency Books, New York.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Hauser, M.D., Cushman, F.A., Damasio, T. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments." *Nature*, 446, 908–11.

Lewicki, P., Czyzewska, M., Hoffman, H., 1987. "Unconscious Acquisition of Complexprocedural Knowledge." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 523–530.

Lieberman, M., 2000. "Intuition: A Social Cognitive Neuroscience Approach." *Psychological Bulletin*, 126, 109–137.

Mandelstam, N. 1970. *Hope Against Hope*. New York: Random House.

McDowell, J., 1985. "Values and Secondary Qualities." In T. Honderich, (ed.), *Morality and Objectivity*. Routledge and Kegan Paul, London, pp. 110–129.

Sanfey, A., Rilling, J., Aronson, J., Nystrom L., Cohen, J., 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science*, 300, 1755–1758

Seeley, W., Carlin, D., Allman, J., Macedo, M., Bush, C., Miller, B., DeArmond, S., 2006. "Early Fronto-Temporal Dementia Targets Neurons Unique to Apes and Humans." *Annals of Neurology*, 60: 660–667.

Singer, P., 1974. "Sidgwick and Reflective Equilibrium." *The Monist*, 58: 490–517.

Singer, P., 1993. *Practical Ethics*. Cambridge University Press, Cambridge.

Singer, T., Kiebel, S., Winston, J., Dolan, R., Frith, C., 2004a. "Brain Responses to the Acquired Moral Status of Faces." *Neuron*, 41, 653–662.

Singer, T., Seymour B., O'Doherty, J., Kaube, H., Dolan, R., Frith, C., 2004b. "Empathy for pain involves the affective but not sensory components of pain." *Science*, 303, 1157–1162.

Sinnott-Armstrong, W., 2006. *Moral Scepticisms*. Oxford University Press, New York.

Shin, L., Dougherty, D., Orr, S., Pitman, R., Lasko, M., Macklin, M., Alpert, N., Fischman, A., Rauch, S., 2000. "Activation of Anterior Paralimbic Structures During Guilt-Related Script-Driven Imagery." *Biological Psychiatry*, 48, 43–50.

Thomson, J.J., 1971. "A Defense of Abortion." *Philosophy and Public Affairs*, 1, 47–66.

Williams, B., 1973. *Utiliarianism: For and Against*, Cambridge University Press, Cambridge

Wilson, T., 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Harvard University Press, Cambridge, MA.

Woodward, J. and Allman, J. 2007. "Moral Intuition: Its Neural Substrates and Normative Significance." *Journal of Physiology-Paris* 101: 179–202.

Zygourakis, C., Adolphs, R., Tranel, D., and Allman, J. 2006. "The Role of the Frontoinsular Cortex in Social Cognition: FI Lesions Impair Ability to Detect Shame, Guilt, Embarrassment, and Empathy, under submission."