Moral Intuition: Its Neural Substrates and Normative Significance

James Woodward
Division of the Humanities and Social Sciences, 101-40

Corresponding Author: John Allman
Division of Biology, 216-76

California Institute of Technology
1200 E California Blvd.
Pasadena, CA 91125, USA

Abstract

Philosophers use the phrase "moral intuition" to describe the appearance in consciousness of moral judgments or assessments without any awareness of having gone through a conscious reasoning process that produces this assessment. This paper investigates the neural substrates of moral intuition. We propose that moral intuitions are part of a larger set of social intuitions that guide us through complex, highly uncertain and rapidly changing social interactions. Such intuitions are shaped by learning. The neural substrates for moral intuition include fronto-insular, cingulate, and orbito-frontal cortices and associated subcortical structure such as the septum, basil ganglia and amygdala. Understanding the role of these structures undercuts many philosophical doctrines concerning the status of moral intuitions, but vindicates the claim that they can sometimes play a legitimate role in moral decision-making.

Keywords: von Economo neurons, insula, moral intuition

1.

By “moral intuition” contemporary philosophers mean moral assessments, judgments, or responses to behavior in actual or hypothetical scenarios, where these responses typically occur quickly or automatically and carry with them a strong feeling of authority or appropriateness but where one need not be (and often is not) aware of any conscious reasoning process that leads to this assessment. Intuition, in this sense, is meant to contrast with moral judgments that are reached on the basis of some extended process of deliberate or explicit reasoning. Consider the well-known trolley problem (Foot, 1978; Thomson, 1976) which consists of the following pair of examples: In the first, a runaway trolley is headed toward a group of five people and will kill them unless diverted. You are able to flip a switch which will divert the trolley onto a siding, but if you do so one person on the siding will be killed. In the second version, again there is a runaway trolley

which will kill five people unless stopped. In this case, however, the only way to stop the trolley is to push a fat man into its path. This will kill the fat man but save the other five.

Most people, at least in western societies and perhaps more generally, have the immediate and strongly felt moral intuition that re-directing the trolley onto the track on which one person is standing in order to save the five in the trolley's path is morally permissible. Most people also have the intuition that pushing the fat man into the path of the trolley in the second example is not morally permissible. However, notoriously, most people have little insight into or conscious awareness of the process that has generated these responses and have great difficulty articulating more general reasons or principles which would "explain" or justify these particular responses.

As a second illustration, consider the following example, originally due to Williams (1973). You are an explorer in a remote jungle in South America. You come upon a village and the local military official, Pedro, who tells you that the Indian inhabitants of the village have been engaging in anti-government activity. Pedro says that he intends to execute ten (arbitrarily selected) inhabitants as a reprisal measure, but adds that if you will shoot one of these yourself, he will spare the other nine. If you refuse he will immediately kill all ten. Many people regard this a deep moral dilemma—they have the strongly felt intuition that it would be very wrong to kill the one, but also feel the weight of the consideration that ten will die if they fail to act.

A broadly similar understanding of moral intuition can be found among psychologists and neurobiologists, although with the important difference that psychologists are far more likely than philosophers to think of affect and emotion as playing an important role. (see below). Jonathan Haidt (2001) describes moral intuition as "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." He 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, no psychological problems resulting from the incest and so on. When subjects are asked to explain or justify their judgments they appeal initially to possible harms/bad consequences (possible creation of a child with birth defects, etc.) and then when reminded that the case is one in which it is stipulated that these harms will not be present, they retreat to saying that the action just seems wrong, although they cannot explain why—a response that Haidt describes as "moral dumbfounding". Haidt takes such examples to illustrate the independence of moral intuition from processes of deliberate, explicit reasoning.

Although no one doubts that, as an empirical matter, we have such intuitive moral responses, there is a great deal of disagreement about their nature, about the processes that underlie or generate them, and about their legitimate role, if any, in moral and political argument. Many moral philosophers think that comparison with intuition provides at least a prima-facie standard for evaluating more general moral principles or theories—that is, it is a consideration in favor of a moral theory if it generates judgments

that agree with widely accepted intuitions, and a consideration that counts against it if it yields judgments that contravene accepted intuitions. Thus it counts against a moral theory (it is "contrary to intuition") if it tells us that it is permissible to push the fat man in front of the trolley, and it is a point in favor of a moral theory if it yields the opposite judgment. Among those taking this view, there are a range of different attitudes regarding the stringency of this requirement. Some writers hold that agreement with intuition is the only standard for assessing a moral theory and that in principle at least any disagreement between theory and intuition is grounds for rejecting the former. On this view, the deliverances of intuition are accepted at face value, and the task of moral theory is simply to describe or systematize these, but not to override or replace them. Sometimes this view takes the form of the more specific suggestion that intuitions play something like the role of "observation" in science—just as the task of scientific theorizing (it is claimed) is to explain what we observe, so the task of moral theory is to explain or justify our intuitions. Other writers suggest that although agreement with intuition is one consideration in assessing moral theories, there are other considerations as well. For example, one popular view, associated with John Rawls (1971), is that intuitions about particular cases and judgments about more general principles should be mutually adjusted in the light of each other in a process of "reflective equilibrium". In this view, individual intuitions are not sacrosanct; they may be rejected based on considerations of overall coherence with other intuitions and general principles, although intuition still remains as an important constraint on moral theorizing.

Still other writers (Bentham 1789; Unger, 1996) take a much more negative and dismissive view of the role of moral intuition. They suggest that such intuitions are often (or usually or even always) the product of "prejudice," "self-serving bias," or arbitrary contingencies or idiosyncrasies of enculturation or upbringing and that even single subjects will often have inconsistent intuitions. They conclude that appeals to intuition should play at best a very limited role in guiding moral argument and decision-making, and that considerations of consistency and theoretical coherence may justifiably lead us to reject or override many or most moral intuitions. (As we note below, this stance is rarely followed consistently.) In this spirit, Peter Singer (1974/2002, p. 47) suggests that

> we should take seriously the assumption that all of the particular moral judgments we intuitively make are likely to derive from discarded religious systems, from warped views of sex and bodily functions, or from customs necessary for survival of the group in social and economic circumstances that now lie in the distant past, in which case it would be best to forget all about our particular moral judgments.

Just as there is no agreement about the proper role of appeals to intuition within moral argument, there is also (and relatedly) no consensus about the nature and character of intuition itself. In addition to those who favor a bias or prejudice view of intuition, there are, among those who take a less dismissive view of its status, those who advocate a rationalist picture in which moral intuition is like insight into logical or mathematical truths or into the status of propositions that are in some suitably broad sense *"a priori."* On this view, the truths revealed by intuition are thought to be self-evident and rationally compelling in just the same way that mathematical truths are; it is exactly their status as *a*

priori truths that makes them seem so obvious and irresistible. Others (e.g. McDowell (1985), and McGrath (2004)), guided by the analogy with observation, think of moral intuition as like, or perhaps even literally, an instance of ordinary visual perception and as sometimes yielding moral knowledge for just this reason--according to McGrath (2004), "If Jim knows that the children acted wrongly in setting the cat on fire, then…he has this piece of knowledge because he perceives that the children acted wrongly in setting the cat on fire" (p. 227). Still others mix perceptual and reason-based analogies as in Roger Crisp's (approving) characterization of Sidgwick's view of intuition as "not gut feeling but a belief which after careful observation presents itself as a dictate of reason" (2002, pp. 70- 71).

A common thread in much of the discussion of intuition by moral philosophers is the denial that emotional processing plays any very important role in moral intuition or at least in the kind of intuition that has legitimate probative force in moral argument. For example, Frances Kamm (1993) recommends the following method in ethics which draws on the idea that moral intuition can deliver *a priori* truths that are independent of contingent facts about human emotional responses:

> [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 goes on to say:

> 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 or the chemical structure of water. Intuitions are appropriate to ethics because ours is an a priori, not an empirical investigation. (1993, p. 8)

Other writers hold, unlike Kamm, that emotions are likely to be involved in many cases of moral intuition, but they tend to focus on the ways in which emotions detract from the normative or epistemic credentials of our intuitions. For example, in a recent discussion of moral intuition, Sinnott-Armstrong (2006, p. 203) suggests that the involvement of emotion in moral intuition often restricts the range of considerations to which subjects respond: "Emotions stop subjects from considering the many factors in these examples. If this interpretation is correct, then many pervasive and fundamental moral beliefs result from emotions that cloud judgment." While Sinnott- Armstrong also allows for the possibility that the involvement of emotion can sometimes enhance judgment, his focus tends to be on the potentially distorting effects of emotion. We will argue below that under the right circumstances the involvement of the emotions in moral intuition can

enlarge (rather than restrict) the range of morally relevant considerations that subjects take into account and in this way lead to morally superior judgments and decisions.

Many but not all philosophers writing on moral intuition also associate intuition with very specific metaphysical doctrines. For example, Stratton-Lake in the introduction to his recently edited volume *Ethical Intutionism* (2002) claims that this philosophical position (and presumably also those who regard intuition as a source of moral knowledge or information) are committed to *cognitivism* (that the beliefs which are the outputs of intuition are the sorts of things that can be true or false) as well as *realism* and *non-naturalism* (roughly that there are "objective" facts about rightness and wrongfulness that are "out there" in the world and that make moral beliefs true or false but which also at the same time have a special metaphysical status which makes them not identical with ordinary, garden-variety "natural" facts of the sort that might be studied by science).

Philosophical views about intuition also vary considerably as to what sorts of things are possible or trustworthy objects of intuition. Some writers who appeal to moral intuition focus largely or entirely on intuitions about particular cases--for example, particular episodes in which we see that some animal is mistreated and have the immediate intuition that this particular action is wrong. Other writers hold that we also have intuitions about more general and abstract moral principles or considerations, as in Sidgwick's well-known contention that among the "ethical axioms—intuitive propositions of real clearness and certainty" is the "self-evident principle that the good of any one individual is of no more importance from the point of view (if I may say so) of the Universe than the good of any other" (Sidgwick, 1907). Indeed, some writers (Singer, Unger) argue that intuitions about general principles are *more* trustworthy or deserve to be taken more seriously than those concerning particular cases, because the latter are more likely to be subject to various biases.

Regardless of one's views about the normative status of moral intuition we may inquire, in a naturalistic vein, about the psychological and neural systems that are associated with or subserve such intuition, about how these relate to the systems associated with other sorts of psychological and reasoning processes, and about how such systems contribute to moral and other kinds of decision-making. 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 its legitimate role in moral argument. Both sets of questions will be pursued in this essay. As we will see, a more empirically adequate understanding of moral intuition should lead us to reject a number of widely accepted views (in both philosophy and psychology) about its nature and about its normative significance.

2.

We first summarize the general picture of moral intuition that will emerge from our discussion and then turn to details. We think of moral intuition as belonging to the general category of *social cognition*, where this has to do with the information processing involved in navigating the social world: Predicting and understanding the behavior and

mental states of others, recognizing behavior and intentions that are potentially beneficial or harmful to oneself or those one cares about and responding appropriately, anticipating and recognizing how others are likely to respond to one's own choices, and so on. The neural areas activated when subjects have "moral intuitions" in response to moral decision tasks (Moll et al., 2002) seem to be areas involved in social cognition more generally, or at least in those aspects of social cognition that involve relatively fast, automatic and affect-laden processing. These areas, 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, Baron-Cohen et al., 1999). The first three of these structures are also involved in the detection and monitoring of visceral, bodily sensations including in particular 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 sort of fast processing seems necessary for the successful real time prediction of others' behavior in socially complex situations and for generating suitable responses to such behavior. Typically such processing involves the integration of a large number of disparate social cues and considerations into a coordinated response, where the complexity and high dimensionality of these cues as well as the need for quick responses may make it difficult to employ explicit and self-conscious cost/benefit calculations or other explicit rule-based strategies. Indeed, as we shall see, there is evidence that subjects who attempt to employ strategies based exclusively on rule based deliberation often end up making normatively worse choices than subjects who employ strategies that allow a greater role for "intuition" and the emotional processing associated with it.

We will argue that understanding the structures and processes subserving moral intuition should lead us to reject many ideas about moral intuition that are common in the philosophical literature. First, an accurate empirical picture of moral intuition strongly suggests that there is no specialized or dedicated faculty devoted just to moral intuition or moral cognition. Instead, our capacity for moral intuition (and moral cognition and decision making) largely derive from and are structured by our more general capacities for social cognition. There is thus no reason to suppose that what underlies our capacity for moral intuition is a capacity to detect or reason about non-natural or metaphysically mysterious properties of some kind–instead what we are responding to are features of our social world (including the behavior and mental states of others) and how these affect what we care about.

Second, and contrary to the inclination of many philosophers to dismiss or downplay the role of emotion and affect in moral intuition, there is, as already remarked, considerable empirical evidence that the neural areas involved in paradigmatic cases of moral intuition are also centrally involved in emotional processing and that manipulation of emotional processing affects the content of one's moral intuitions. Moreover, much of this emotional processing is unconscious, so that subjects are often unable to tell when

(or which) emotional processes have played a role in generating their intuitions, and thus are unable to follow the advice of Kamm and others to discount intuitions in which emotion has played a role. Of course, one may simply stipulate that by "moral intuition" one means something that does not involve emotion, but such stipulation seems arbitrary and unmotivated, given the character of the neural systems that are activated in what are ordinarily thought of as cases involving the exercise of moral intuition. Put slightly differently, if one decides that by definition, "intuition" cannot involve emotional processing, then many cases which philosophers like Kamm think of as involving appeals to intuition may turn out to involve no such thing because of the involvement of affective processes in generating the responses in question. There might be a point to such a stipulation if, from a normative point of view, intuitions that do not involve emotion or affect were, in addition to being common, somehow more reliable or likely to lead to better moral judgment and decision, but there are strong reasons to doubt that this is the case. In fact, there is considerable empirical evidence that subjects with damage to the areas involved in emotional processing (and in moral intuition) make decisions that both 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). Good moral decision-making seems to require the involvement of emotional processing and affect, which is not to deny that it also involves processes that look more purely cognitive. We thus deny the very widespread view that the involvement of emotion in moral intuition and decision-making, usually leads to decisions that are normatively inferior (in comparison with decisions that are based more purely on reason). While we agree with Sinnott-Armstrong's 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 of the right sort of emotion can have the opposite effect (Sinnott-Armstrong, 2006).

The role of emotional processing in the generation of paradigmatic cases of moral intuition also leads us to reject the common assimilation in the philosophical literature of moral intuition to visual perception or to insight into logical, mathematical or other sorts of *a priori* truths. If the visual perception idea was correct, one would expect subjects with intact visual processing but damage to areas involved in emotional processing to have the same intuitions as normal subjects–as we shall see, the empirical evidence tells strongly against this. Similarly, if moral intuition is a special case of insight into *a priori* truths like those found in logic and mathematics, one would expect subjects with damage to emotional processing areas but intact areas that are known to be involved in logical or mathematical reasoning to have unaffected intuitions—again, this is not what is found. Other considerations support a similar conclusion: even among normal subjects the deliverances/assessments of moral intuition are simply not as widely shared as the judgments that result from ordinary visual perception and are sensitive to the impact of culture and experience in ways that visual perception is not. Similarly for logical/mathematical insight–among those with the requisite training, the judgments resulting from such insight are far more universal than the deliverances of moral intuition, and the procedures for checking whether such insight is correct are very different from those involved in assessing the correctness of moral intuition, if indeed there are procedures of the latter sort at all.

Third, we reject the common view (shared, for example, by writers like Greene and Haidt, whose idea are discussed in more detail below) that the processes that underlie the generation of moral intuitions are typically or always relatively "primitive" (Greene, 2004, p. 389), relatively hard-wired or innate, fixed by our evolutionary history and not subject to subsequent modification by experience, and that these structures function only as relatively coarse-grained "alarm signals" that do not do "sophisticated information processing". (Greene, 2004; forthcoming) Our contrary view is that emotional processing and the structures that underlie 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 in the form of explicit rules. This implicit learning can be quite flexible and context sensitive–intelligent rather than stupid. Indeed, there is evidence that under the right circumstances the structures underlying moral intuition and emotional processing associated with them are often better at integrating complex multi-dimensional environmental clues that are relevant to good decision-making than our efforts at explicit conscious deliberation.

Of course, as we have already noted, for moral philosophers, the central question about moral intuition is its significance, if any, for the assessment of various moral claims. Is the fact that we have strong intuitive reactions, either in favor or against some action, policy or proposed principle relevant to how we ought to morally assess these, and if so, why? Philosophers often respond to these questions by constructing blanket defenses of or condemnations of moral intuition. We will not follow either of these courses here. We think that better questions to ask are these: Do intuitive moral responses sometimes contain information that we can recognize as relevant to good moral decision-making and if so, what is this information? Under what circumstances is such information likely to be present or absent? Does "intuition" sometimes play some functional role in moral decision-making that is not or could not be played by other sorts of psychological processes? What do we lose if ignore such responses or try to replace them entirely with some alternative? We will try to sketch answers to these questions below.

Before turning to details, however, several additional remarks by way of orientation are in order. As the brief sketch of Stratton-Lake's views above reminds us, philosophers often frame issues about the role of moral intuition in terms of categories like "truth", "knowledge", and "justification". They ask questions like the following: Are most (or even some) of our moral intuitions or the judgments/beliefs associated with them "true"? Do intuitions yield moral "knowledge" or "justified" moral belief? If the moral beliefs associated with intuitions are sometimes "true", then what are the "truth makers" for such beliefs–facts about non-natural properties, such as the "rightness" of various actions, as Stratton-Lake supposes? In what follows we will try as best we can to avoid such questions, for several reasons. First, they are highly controversial and it is unclear how to resolve them. Second, we don't think that we need to take a stand on them for the purposes of this essay. In particular, we don't think that the claims we make about the processes that underlie moral intuitions or even the claims that we make about its normative significance require that we answer such questions. We thus think, in opposition to Stratton-Lake, that discussion of the role of intuition in moral judgment

does not require commitment to any particular metaphysical doctrine–either cognitivism, realism, non-naturalism, or their denial.

Consider an example that figures in our discussion in section 4. A fireman is in a building in which there is a fire and suddenly has the "intuition " that the situation is highly dangerous and that he and his men should get out immediately. As it turns out, this intuition is "correct" (trustworthy, reliable, better than the contrary intuition that he should stay–choose your favorite approval word) in a straightforward common sense way: The floor on which they were standing was about to collapse because there was a hidden fire on the floor beneath them, into which they would fall. The fireman's intuition would have been "incorrect" or "misguided" if, for example, there was no hidden fire below and he and his crew were in no danger. We hold that one can make such claims about correctness/ incorrectness and justify them without committing to any particular metaphysical/epistemological view about whether the fireman's intuition amounts to knowledge, or about whether it has a non-natural truth-maker and so on. The facts/considerations to which we can appeal to assess the correctness of the fireman's intuition are facts of a perfectly straightforward "natural" sort–the hidden fire, the effects on the firemen if the floor had collapsed, and so on.

We think that a similar contention is true for moral intuition–while assessments of certain moral intuitions as more reliable than or normatively superior to others is a matter that is admittedly often far more controversial than our assessment of the fireman's intuition, it does not help to resolve such controversies to become enmeshed in metaphysical arguments about whether moral beliefs can be true or false and so on. To the extent that there is some way of deciding or arguing about whether some moral intuitions about a proposed course of action are worth taking more seriously than others (see below), we can address such issues without taking a stand on the metaphysical issues described above. To the extent that there are no such procedures for assessing moral intuition, excursions into metaphysics will also be pointless.

A second issue that deserves brief comment concerns the nature of our claims about the role of emotion in moral judgment. A great deal of recent philosophical discussion of this subject has focused on the semantics or pragmatics of *moral language*—when one makes a moral judgment, is the meaning or content of the moral judgment or its pragmatic function captured by construing it as having to do in some way with the expression or endorsement of certain emotions? Issues of this sort are explored in recent expressivist or sentimentalist accounts of moral judgment such as those due to Gibbard (1990) and Blackburn (1998). We take no stand about such philosophy of language projects, which we see as sharply distinct from our own. Our claims have to do with the causal role of emotional processing and associated neural structures in moral intuition, the effects on moral intuition when such processing is absent, and so on. We don't think that anything in particular follows from these claims for contentions about the linguistic meaning, content, or function of moral judgments, including those associated with intuition. Specifically, we don't think that our emphasis on the causal role of emotional processing in moral intuition commits us (at least in any straightforward and direct way) to an expressivist or emotivist account of the meaning of moral terms.

We conclude this section with some general remarks by way of orientation for the non-philosophical reader on the contrast between two different kinds of moral theories: *consequentialist* and *deontological*. The contrast itself is entirely orthodox and will figure figure importantly in our discussion below. However, the significance we assign to the contrast and in particular our distinction between that which we call *parametric* consequentialism and *strategic* consequentialism and our suggestion that the latter sometimes yields judgments that are closer to deontological theories than the former is much more controversial, although important to our overall argument.

We may think of consequentialism as the doctrine that the rightness or wrongfulness of actions depends entirely on the goodness or badness of the consequences that the action produces. This characterization is of course highly non-specific since it leaves open both what counts as a consequence and how the goodness or badness of consequences should be assessed. (Classical utilitarianism of the sort represented by Bentham (1789) and among contemporary philosophers, Peter Singer, is a more specific version of consequentialism with the goodness of consequences being assessed in terms of their overall utility.) In practice, consequentialists have characterized their view in opposition to deontolgical approaches. These deontological approaches come in a variety of different versions but tend to share one or more of the following anti-consequentialist commitments: (a) the manner or way in which an action leads to an outcome or the structure of the intention with which the action is produced has moral significance independently of the goodness or badness of the outcome itself[1], (b) certain actions are morally prohibited, because, for example, they violate people's rights, or are unjust or unfair even if they produce the best consequences overall, (c) actions can be wrong because they treat people as mere means to the production of good consequences or because they fail to treat people with appropriate respect or dignity. (Kant (1785) is the paradigmatic deontologist; contemporary deontologists include philosophers like Kamm (1993) and Thomas Nagel (1972).) As an illustration of (a), and perhaps (c), many deontologists will judge that there is a crucial moral difference between the two versions of the trolley problem above, despite the fact that in both versions the number of lives at stake under the available options is the same. Specifically, it may be claimed that pushing the fat man in front of the trolley involves using him as a mere means for saving the lives of the five and is morally wrong for this reason. By contrast, in the first version of the trolley problem, although diverting the trolley onto the siding results in the death of the one, it does not involve using him as a means (his death is a mere "side-effect") and hence is morally permissible[2]. On the other hand, consequentialists (or at least parametric consequentialists--see below) tend to see no fundamental difference between the two

[1] More specifically, many deontologists claim that it matters morally whether an outcome results from an action or an omission and whether the outcome is an end at which the actor aims or a means to that end, or whether it is instead a mere side-effect of these.
[2] We use this merely to illustrate one characteristic deontological treatment of the trolley problem. It is well-known (cf. Thomson, 1976) that an account along these lines is not, even from a deontological perspective, normatively adequate to deal with all of various permutations of the problem imagined by philosophers.

versions of the trolley problem and in both cases recommend the action that saves the greatest number of lives–that is flipping the switch in the first version and pushing the fat man in the second. Deontological theories thus yield judgments that are in agreement with most people's intuitions about both versions of the problem while consequentialist theories yield judgments that are in agreement with intuition about the switching version but not with the fat man version. Consequentialists are thus inclined to dismiss our intuitions about the fat man version of the problem as simply mistaken.

As an illustration of (b) and perhaps (a), many deontologists will think that it would be wrong to kill the one to prevent the murder of the ten in the Explorer example or at least that this example presents a very serious moral dilemma, in which it is far from obvious what to do. The reason why this action would be wrong is that it would violate the one's right to life and even though it obviously would be desirable to save the ten (if there were some other way of accomplishing this) it matters morally how this good consequence is brought about—it is impermissible to bring it about by murdering the one. If the explorer refuses to kill the one, with the result that Pedro kills the ten, the explorer does not intend to kill the ten and hence is not responsible for the deaths. It is Pedro who kills them, not the explorer. By way of contrast, consequentialists have tended to suppose that if the facts are as stipulated in the example, it is obvious that the explorer should kill the one, since this will save the greatest number of lives.

So far we have simply been describing how self-styled consequentialists and deontologists have tended to react to these examples. One might well wonder, however, whether consequentialism *per se* yields the judgments that are identified as "consequentialist" in the above examples, and relatedly, whether the opposition between consequentialist and deontological approaches is necessarily as stark as portrayed above. In fact, it seems clear that the judgments identified as consequentialist are reached by restricting the relevant consequences (in part by stipulating away other relevant considerations) just to a comparison of the number of lives saved (in the examples as described) under the various courses of action available to the decision maker. Thus in the Explorer case, those consequentialists who think it obvious that you should shoot the one often simply assume or stipulate that the example has various other features that remove many other considerations that would be present in real life cases of this sort, and which would be relevant from both a consequentialist and a deontological perspective. For example, it is typically assumed that the explorer somehow knows for certain that Pedro will not renege on his end of the deal, killing the other nine after you kill the one, that killing the one will not create an incentive for Pedro and others like him to make similar threats in the future, and so on. Thus the consequentialist decision-maker does not have to take these possibilities into consideration, as he would need to in the real world.

It is a natural thought that if one were to instead work with richer examples with the features that would be present in realistic, real-life cases (e.g. uncertainty about what Pedro will do), it would be far less obvious that a sophisticated consequentialism (which is both sensitive to uncertainty and attempts to take into account both immediate and more indirect or long run consequences) would recommend cooperating with Pedro or

that it would necessarily yield judgments that are starkly different from more "deontological" theories in other cases.

We will not try to seriously argue for this claim here (doing so would require another paper, if not a book) but we do wish to suggest the following distinction between two different forms of consequentialsm which will inform the remainder of our discussion. *Strategic* or sophisticated consequentialists are sensitive to the dynamic, interactive or strategic aspects of moral decision making and to the uncertainties that result from these. Strategic consequentialists recognize that when they make moral decisions they are typically embedded in an ongoing interaction with other actors who will respond in complex ways that are not easy to predict, depending on the decision-maker's choices, and that these responses will in turn present the original decision-maker with additional decisions and so on–in other words, that they are involved in a complex repeated game of some kind. Strategic consequentialists thus tend to be sensitive to the incentives that their choices create, to the informational limitations and asymmetries they face, and to the opportunities for misrepresentation these create, and also to considerations having to do with motives and intentions, since these are highly relevant to predicting how others will behave. They also recognize that one's present choices may affect one's future behavior (by, for example, creating habits or tastes or by sensitizing or desensitizing one to various outcomes). In addition, they recognize that those with whom they are dealing may not be consequentialists, and may be heavily influenced by non-consequentialst considerations in deciding how to respond the decision-maker's original choices. John Stuart Mill's moral and political thought (Mill, 1859) has many elements characteristic of strategic consequentialism, and the same holds for Peter Railton (2003) and Alan Gibbard (1990), among contemporary philosophers.

By contrast, parametric consequentialists tend to think of the behavior of others as fixed or parametric, independently of their choices, so that they don't need to worry about long run interaction effects, incentives, and so on. In part because of this, they think of the decision problems they face as having a relatively simple structure about which they are likely to have adequate information. While the methodology that naturally goes along with strategic consequentialism is some (empirically adequate) version of game theory, the methodology that guides the parametric consequentialist is closer to classical, one person decision theory in which the decision-maker maximizes value, assuming a fixed environment. Among contemporary moral philosophers, Peter Singer (1993) is one of the clearest exemplars of this sort of approach, as is Peter Unger (1996), to the extent that his views are consequentialist.

The relevance of this distinction to moral philosophy is that many of the normative recommendations that are regarded as characteristically consequentialist in that literature seem to follow from parametric versions of consequentialism but do not obviously follow from more strategic versions, which at least in some circumstances arguably yield judgments that are closer to traditional deontological approaches. (One indication of this is that the normative recommendations advanced by strategic consequentialists like Mill, Gibbard and Railton often look closer to what deontologists recommend than do the recommendations of parametric utilitarians like Singer.) This in

turn has important consequences for the origins of moral intuitions and deontological intuitions in particular. As we have seen, many philosophers (including many if not most deontologists) have supposed that moral intuition delivers *a priori* truths like those of logic and mathematics, hence truths that are independent of and do not have their source in experience. This idea fits badly with many empirical observations, including observations about the role of brain areas involved in social/emotional processing in moral intuition and the fact that, as we shall see, these areas seem differentially involved in characteristically deontological intuitions.

Our remarks about the relationship between strategic versions of consequentialism and traditional deontolgical theories suggests an alternative source for deontolgical intuitions, one that fits better with these empirical observations. On this alternative picture, moral intuitions in general, including deontological intuitions, do at least sometimes reflect the operation of experienced-based learning mechanisms, including those that rely on emotional processing, rather than insights into a *priori* truths. In particular we think that these intuitions sometimes incorporate information of the sort that is the distinctive focus of more strategic versions of consequentialism—that is, information about the mental states of others affected by our actions, how they are likely to respond to our choices, and so on. For reasons that will be explained below, it may be difficult to make all of this information fully explicit or to employ all of it in self-conscious deliberation–hence decision making that is influenced by intuition, including deontological intuitions, may yield normatively superior outcomes to decision procedures that attempt to completely eschew reliance on intuition.

The remainder of this essay is organized as follows. Section 3 advances some proposals about the role of the orbito-frontal, insular and anterior cingulate cortices in social and moral intuition. Sections 4 and 5 relate our views to other recent empirical work on moral intuition and social intuition and cognition. Sections 6 and 7 then explore some issues regarding the normative status of moral intuitions in the light of our discussion.

3.

In this section we develop a neurobiological theory of intuition. However, first we want to provide a definition of intuition and contrast this with deliberation. Intuition is a form of cognition in which many variables are rapidly evaluated in parallel and compressed into a single dimension. This compression facilitates fast decision-making. In contrast with deliberative cognition, we typically are not aware of the logical steps or assumptions underlying this process although intuition is based on experience-dependent probabilistic models. Instead we feel the intuitive process as visceral sensations (gut feelings). Intuition involves the rapid comparison of a current transaction with previously experienced similar events and a visceral assessment of the probability of a favorable or unfavorable outcome for the current transaction. Deliberation is much slower and typically involves the serial processes of inductive and/or deductive reasoning. Both intuition and deliberation are logic driven, but forms of logic differ, with intuition being based on probabilistic inference and deliberation on conscious, usually verbally mediated,

reasoning. Intuition is plastic; it is *not* instinct, although instinctive feelings may contribute to it. Emotional value judgments contribute to both intuition and deliberation. Focused attention and the exclusion of other information are required for deliberation, but not for intuition. We tend to rely on intuition in complex situations involving many variables and a high degree of uncertainty and which demand immediate decisions, which are characteristic features of many social interactions. We tend to rely on deliberation in situations which lend themselves to explicit step-by-step verbally mediated reasoning where a rapid response is not required. Many social interactions occur too rapidly, are too complex, and involve too much uncertainty to permit the exercise of deliberative thought.

We propose that moral intuitions are part of the larger set of social intuitions that guide us through complex, highly uncertain and rapidly changing social interactions. Our moral intuitions, like social intuitions generally, tend to become more finely differentiated as we gain experience in life. The neurobiological substrate for these intuitions includes the insular, cingulate, and orbito-frontal cortices and associated subcortical structures such as basal ganglia and amygdala, all of which have been implicated by many functional imaging and brain lesion studies (Adolphs, 2006; Allman et al., 2005; Baron-Cohen et al., 1999; Berthoz et al., 2002; Damasio, 1994; de Quervain et al., 2004; Rilling et al., 2002; Sanfey et al., 2003; Shin et al., 2000; Singer et al., 2004a,b; Zald and Kim, 2001). Brain lesion studies have also revealed a very interesting feature of the neurobiological development of moral intuition. Damage to orbito-frontal cortex during the first few years of life has a profound impact on adult moral intuition and judgment (Anderson et al., 1999), which stands in contrast with early damage to speech and motor cortex, which is well compensated for during later development (Finger et al., 2000). This is not to imply that the system does not continue to differentiate through later stages of development, but that certain crucial circuits must be operative at early stages for the later stages to occur. We will return to the effects of early orbito-frontal lesions on moral intuitions in section 6.

We propose that a prime input to the neural circuitry for moral intuition is insular cortex. In all mammals, the insular cortex contains a representation of the motor and sensory systems involved in the ingestion and digestion of food (Rolls, 2005; Small et al., 1999). It is thus responsible for the regulation of food intake; the ingestion of nutritious food and the rejection of toxins. In primates, and especially in humans, there is an additional set of discrete inputs arising from the body that signal sharp pain, dull pain, coolness, warmth, itching and sensual touch (Craig, 2003) which endows primates with a much more highly differentiated cortical system for bodily awareness, with all the potential for neuronal plasticity and learning that are characteristic of cortical circuits. These highly differentiated inputs convey important elements of interpersonal contact and reflect the evolutionary development in primates, and especially in humans, of enhanced capacities for registering the awareness, individual identity, and memory of that social contact.

The regulation of food intake is expressed in the primordial opposed emotions of lust and disgust, the consumption of the nutritious and the spitting out or vomiting of the toxic. Disgust means literally “bad taste,” and facial expressions of disgust powerfully

activate anterior insular cortex (Phillips et al., 1997), demonstrating the social component to this insular processing. We propose that the neural substrate for lust-disgust served as the evolutionary template for substrates for the complex social emotions that tend to occur in polar opposites such as love-hate, gratitude-resentment, self-confidence-embarrassment, trust-distrust, empathy-contempt, approval-disdain, pride-humiliation, truthfulness-deception, and atonement-guilt. The first of each of these pairs generally favors the formation of social bonds and the second tends to disrupt bonds. They thus range from the poles of prosocial to antisocial. Many of these complex social emotions are known to activate fronto-insular (FI) and anterior cingulate cortex (ACC). The social emotions for which this has been demonstrated thus far include lust (Karama et al., 2002), love (Bartels and Zeki, 2000), resentment (Sanfey et al., 2003), embarrassment (Berthoz et al., 2002), trust (Singer et al., 2004a), empathy (Singer et al., 2004b), deception (Spence et al., 2001), and guilt (Shin et al., 2000), and it is likely that the others will be found to activate FI and ACC. We propose that the circuitry in insular cortex that originally processed lust-disgust served as a template for the evolution of the circuitry responsible for polar social emotions in FI and ACC. This is consistent with evidence that primitive mammals were solitary and that complex social behaviors are specializations within specific taxa of mammals, such as primates or cetaceans (Martin, 1990). Just as the insula has the capacity to integrate a large array of complex gustatory experience into visceral feelings leading to a decision to consume or regurgitate, so fronto-insular cortex integrates a vast array of implicit social experiences into social intuitions leading to the enhancement or withdrawal from social contact. By this theory, it is no accident that our language is full of visceral metaphors for social interactions, because they reflect the underlying neural processes. Interactions with specific individuals are “delicious” or “nauseating”, and these visceral feelings have powerful moral dimensions.

Just as our capacity for the interpretation and appreciation of foods becomes more differentiated with experience and maturity, so does our capacity to differentiate more complex social emotions. Primary flavors such as sweetness and saltiness appeal strongly to children, whereas adults favor more complex flavors that involve the sour and the bitter, such as those imparted by the processes of fermentation and the slow aging of tannins in wine. Children tend to prefer tastes elicited by compounds with simple molecular structures (salt, sugar and fat) while adults often prefer tastes elicited by much more complex compounds. Indeed the characterization of these complex compounds challenges modern analytical chemistry. Similarly, children’s moral intuitions involve strong black and white “with me or against me” feelings, while mature adults will experience a far greater range and subtlety of moral intuitions that more closely match reality. Adult subjects with lesions involving FI experience a narrower range of social emotions than do normal subjects (Zygourakis et al, 2006).

FI and ACC are active when subjects make decisions under a high degree of uncertainty (Critchley et al., 2001). These areas are involved in the subjective experience of pain in oneself and empathy for the pain experienced by a loved one (Singer et al., 2004b), which are powerfully magnified by uncertainty. They are also active in situations involving social uncertainty and pain such as the experience of guilt and embarrassment (Shin et al., 2000; Berthoz et al., 2002). Humor, which activates FI and ACC in

proportion to subjective ratings of funniness (Watson et al., 2006), may serve as a way to recalibrate intuitive judgments in changing social situations, thus resolving uncertainty, relieving tension, engendering trust, and promoting social bonding. The experience of humor has a similar ontogeny to the appreciation of complexity in flavors or moral ambiguity. Just as children love sweets, and possess simple black and white moral intuitions, so they tend to enjoy slap-stick cartoons, while mature adults have the capacity to enjoy more richly nuanced forms of humor that often involve the appreciation of irony.

In this context it is interesting that FI and ACC contain a class of large bipolar cells, the von Economo neurons (VENs), that are found in humans and great apes but not in other primates (Allman et al., 2001; Allman et al., 2005). Thus the VENs are a recent evolutionary development that emerged since the divergence of hominoids from other primates. The VENs develop late in ontogeny as well as phylogeny. They first appear in small numbers in the 35th week of gestation and at birth only about 15% of the mature number are present. This postnatal increment in VEN population may arise by differentiation from some pre-existing cell type or by migration from a potentially proliferative zone in the ventricles. The VENs are more numerous in the right hemisphere than in the left, which is probably related to the right hemispheric specialization for the social emotions (Allman et al., 2005). The VENs are selectively destroyed in fronto-temporal dementia, which is characterized by difficulties is moral intuition, self-awareness, appetite control, and bizarre humor (Seeley et al, 2006 and William Selley, personal communication). The VENs are also greatly reduced in number in a genetic condition, agenesis of the corpus callosum, which is characterized by abnormalities in social cognition and humor (Kaufman et al., 2006).

VEN functions are revealed by immuno-cytochemical staining with antibodies to neurotransmitter receptors. The VENs are strongly labeled with antibodies to the dopamine D3 receptor (Allman et al., 2005), which may signal the expectation of reward under uncertainty (Fiorillo et al., 2003). FI and ACC activity is coupled to situations in which the subject sustains a gambling loss (punishment) and then switches to a different behavioral strategy (O'Doherty et al., 2003), implying that in normal subjects these areas are involved in adaptive decision-making and cognitive flexibility. FI is also activated in gambling tasks when the subjects anticipate that their luck is about to change, which is a form of intuition (Elliott et al., 2000).

The serotonin 2b receptor is also strongly expressed on the VENs (Allman et al., 2005), and this receptor is rarely expressed elsewhere in the central nervous system (Baumgarten and Göthert, 1997). However, the serotonin 2b receptor is also strongly expressed in the human stomach and intestines where it promotes contractions of the smooth muscles responsible for peristalsis (Borman et al., 2002). Serotonin may serve as an antagonistic signal to dopamine, with serotonin signaling punishment and dopamine signaling reward. The activation of the serotonin 2b receptor on VENs might be related to the capacity of the activity in the stomach and intestines to signal impending danger or punishment (literally "gut feelings") and thus might be an opponent to the dopamine D3 signal of reward expectation. The outcome of these opponent processes could be an

evaluation by the VEN of the relative likelihood of punishment versus reward and could contribute to "gut level" or intuitive decision-making in a given behavioral context.

ACC and FI are known to have an important role in interoception, or the conscious awareness of visceral activity (Craig, 2004; Critchley et al., 2004). In his theory of "somatic states", Damasio (1994) proposed that this monitoring of sensations arising from the gut is crucial to adaptive decision-making. The presence of a serotonin receptor on the VENs that is otherwise rare in the brain, but common in the viscera, suggests an interesting extension of the concept that these areas are monitoring activity in the gut. Perhaps the expression of the serotonin 2b receptor on the VENs represents a transposition of this function from the gut into the brain, with these circuits emulating how the gut would respond but over a faster time course. This circuitry would enable the individual to react more quickly to threatening circumstances than if that individual depended solely on monitoring sensations arising from the gut. Also, the emulation of gut activity in the cortex would permit a greater degree of plasticity and learning to occur than would be the case in the mesenteric nervous system.

Far from being evolutionarily primitive, the neurobiological system involved in moral intuition possesses at least three recently evolved components. First, there are the highly differentiated inputs to the insula in primates and especially in humans that subserve the sensations of pain, coolness, warmth and sensual touch which are important elements in differentiating individual identity in social contact and in self-awareness. Second, there is the emergence of a novel circuit component, the VENs in apes and their great elaboration in humans in FI and ACC, cortical areas which are strongly implicated in the social emotions and humor. Third, there is the expansion in apes and especially in humans of anterior orbito-frontal cortex, area 10, which has been specifically implicated in moral decision-making in emotionally charged situations (Greene et al., 2001). Damage to anterior orbito-frontal cortex in the first few years of life has a devastating impact on the capacity for moral intuition in adulthood. We believe that these recently evolved systems are part of an adaptive complex supporting the increased complexity of hominoid and especially human social networks. We hypothesize that the VENs and associated circuitry enable us to reduce complex social and cultural dimensions of decision-making into intuition, thus facilitating the rapid execution of decisions.

4.

Within this framework of the distinction between intuition and deliberation, it is natural to contrast moral (and other forms of social) *intuition* with moral *reasoning*, understood as involving conscious inference, deliberation, or theorizing, associating the former with the operation of the automatic system, and the latter with the deliberative system. Jonathan Haidt adopts this view of matters in the characterization of intuition quoted in section 1. A very similar view of moral intuition is adopted by Joshua Greene, in remarks quoted below.

We think that the general contrast between automatic and deliberative processing and the association of moral intuition with the former is very plausible. This association

is supported by, among other things, the work of Haidt and of Greene described below which emphasizes the quick, automatic character of much moral intuition and the difficulties that people have in providing reasoned explanations for these intuitions. However, in our view, some other very common claims in the literature about the characteristics of the two systems and their role in judgment and decision-making are far less defensible and have led to some misguided conclusions about moral intuition. Caricaturing only slightly, many researchers[3] seem to adopt the following view: While conceding that the automatic system (and emotional processing in general) may involve "useful heuristics" that in the right circumstances produce results that are in conformity with generally accepted normative standards of rational, prudent or adaptive behavior, the literature tends to emphasize those circumstances in which the system produces results that are sub-optimal or contrary to such standards--that is, results that are biased or rationally indefensible. The automatic system is thus seen as employing "error prone" or "maladaptive" procedures. It is also commonly suggested that the reasoning system, when operative, tends to produce more satisfactory results or results more in keeping with normative standards of rational judgment. Thus susceptibility to various fallacies in probabilistic inference (e.g. assigning higher probability to a conjunction than to each of its conjuncts, as in the well-known Linda is a feminist bank teller example (Kahneman and Tversky, 1972)) is taken to be a consequence of the operation of the intuitive system, which may or may not be corrected by more rational processing. A similar picture is taken to apply to the moral realm: moral intuitions are produced by automatic processing that is error prone and hence in need of correction by rational deliberation.

It is also commonly assumed not just that there is a distinction between the styles of processing employed by the two systems, but that the two systems are entirely non-overlapping and temporally distinct in the their operations, and in particular that the operation of the automatic system usually temporally precedes any operation of the deliberative system and that once the latter becomes operative (if it does) , the automatic system is no longer operative and, moreover, no longer needs to be operative for normatively good outcomes. Instead, once the deliberative system becomes operative, it is able to intervene and correct whatever errors have been produced by the automatic system. Errors or unsatisfactory outcomes thus occur when the deliberative system fails to come into play or fails to adequately oversee the outputs of the automatic system, and the remedy for such errors is to encourage more active involvement by the reasoning system. Often this line of thought is accompanied by a sort of slide from the very plausible idea that the intuitive system sometimes makes mistakes that can be corrected by the deliberative system to the far less plausible idea that the deliberative system would make better decisions if had no or little input from intuitive system–a line of thought we find in several of the writers discussed below.

Often this picture is accompanied by the further idea that the automatic system is more primitive, both phylogenetically and developmentally, than the deliberative system,

[3] In addition to Greene, non-philosophers who adopt something like this parametric view in connection with moral and political decision-making include Baron (1994) and Sunstein (2005).

more rigid (and modular) in its operation, and less modifiable by learning or experience. For example, in a recent paper, Greene, et al. (2004) contrast what they call "personal" moral judgments, which "are largely driven by social-emotional responses" with other sorts of judgments, which they call "impersonal" and which (they claim) are more driven by "cognitive" processes. Personal moral violations have a "ME HURT YOU" structure where the HURT component involves "primitive" kinds of violations–assault rather than insider trading (Greene et al., 2004, p. 389). The responses triggered by such personal violations are "pre-potent" and are heavily affect-laden–they have the characteristics associated with moral intuition in accounts like Haidt's. The authors suggest that we share these responses with other primates and that they are thus relatively old in evolutionary terms.

These authors also claim that often at least these "intuitive" responses issue in deontological or non-consequentialist/non-utilitarian intuitions or judgments–e.g., the judgment that it would be wrong to push the fat man in front of the trolley. By contrast, they claim that brain areas associated with "abstract reasoning and cognitive control" (dorso-lateral prefrontal cortex and parietal areas, according to the authors) are involved in more impersonal moral judgments and these "cognitive" processes have a preferred behavioral outcome, namely that of favoring utilitarian moral judgments. (We note for future reference that when the authors speak of utilitarian judgments, they often mean judgments that would be endorsed by parametric rather than strategic versions of that theory). According to the authors, these brain areas "house some of our species' most recently evolved neural features and cognitive abilities." When a subject judges that utilitarian/consequentialist considerations justify a personal moral violation that is contrary to intuition (pushing the fat man in front of the trolley) these more cognitive processes compete with and succeed in suppressing our more emotional (deontological) responses–this is reflected, for example, in the longer reaction times of subjects that make utilitarian judgments. Given this association of utilitarian/consequentialist judgment with structures that are seen as sophisticated and uniquely human, and deontological judgment with structures that are older and more primitive, it is not surprising that at least one of the authors (Greene) makes it clear elsewhere (Greene, in press) that he takes utilitarianism to be normatively superior to deontological moral theories. In this respect, Greene's view is the analogue, within the moral realm, of a two systems view of non-moral judgment and decision–making, in which the role of the more cognitive deliberative system is to suppress and correct the normatively inferior responses of the automatic system.

In contrast to the views just described, our position is that good prudential and moral decision making requires the integrated deployment of both the automatic and deliberative systems (and cognition and emotion) working together and mutually supporting one another. There is considerable empirical evidence (e.g. Damasio, 1994) that subjects with damage to ventro-medial prefrontal areas (including the medial orbito-frontal cortex) and/or insula, who are not able to make use of emotional processing and the signals or "intuition" this generates, often make decisions that are much "worse" by any reasonable standard, in terms of their impact on the subjects themselves and those around them, than those who do not have these deficits. This provides strong prima-facie

reason to think that such processes also play an important role in moral decision-making and to doubt that the normatively better decisions are always associated with greater involvement of the "deliberative" system and reduced reliance on intuition and emotional processing. There is also empirical evidence, described below, suggesting that normal subjects who attempt to rely entirely on conscious deliberation make normatively worse decisions on non-moral problems of high complexity than subjects who rely more on unconscious processing.

We will return below (Section 6) to Greene's association of deontological responses with emotional processing and the "intuitive system," but to the extent that this association is accepted, it also seems to us to provide reason to think that characteristically deontological intuitions should not be automatically dismissed in making moral decisions–there may be information in these intuitions that is not readily accessible to more deliberative, self-consciously analytical forms of decision-making. We also emphasize that it is simply factually incorrect to think that the systems and neural structures involved in moral intuition and social emotional processing have been retained in unaltered form from other primates–instead, as remarked above, these systems have undergone very substantial changes in humans and support forms of social cognition and emotional processing in humans that are not present in other primates.

5.

So far we have been using words like "intuition" and "social intuition" without any very detailed discussion of what these notions involve or of how such intuitions are acquired. We think of intuition in general (including social and moral intuition) as the result of implicit learning involving various neural structures. It is characteristic of such learning that it is achieved on the basis of probabilistic cues that are present in a series of individual trials. Subjects produce a response of some sort (either judgment or behavior) and then receive feedback from the environment about the correctness or appropriateness of this response. "Correctness" here typically has to do at least in part with the subject's desires and interests--the subject learns fruits of this color taste good, but fruits of that color do not, that others with this trait are friendly, those without it are hostile, etc. When learning is implicit, subjects who receive feedback may be able to achieve more correct responses over time but without being aware of which cues they are responding to in achieving improved responses, and without being conscious of any explicit rule which guides those responses.

A well-known non-social illustration is provided by an experiment of Lewicki et al. (1987)[4]. The experimental task was to determine as quickly as possible in which

[4] This experiment as well as the experiment from Lewicki (1986) are discussed in Lieberman (2000) who also uses them to illustrate the connection between intuition and implicit learning. Lieberman also discusses a number of other experiments illustrating intuitive social cognition.

quadrant of a screen a target stimulus appeared. Trials were presented in blocks of seven– in the first six the target was presented by itself; on the seventh the target was presented in the presence of distractors that made it difficult to identify. Unknown to the subjects, there was a subtle relationship between the locations of the target on the first six trials and the location on the seventh. Subjects received eleven hours of practice and over the course of this the speed with which they were able to identify the location of the target on the seventh trial improved. Lewicki et al. were able to show, on the basis of other experimental manipulations, that this improvement was due in part to subjects having learned the relationship between the first six and seventh trials, and was not just due to greater familiarity with the task. However, subjects were unaware of this relationship--it was learned only implicitly.

A dramatic real-life example of implicit learning and reliance on intuition is described in Klein (1998). A lieutenant fireman with a great deal of experience and his crew are trying to put out what is apparently a small fire in a kitchen. They stand in the living room spraying water on the fire but this has less impact than expected. Then, in Klein's words,

> The lieutenant starts to feel as though something is not right. He doesn't have any clues; he just doesn't feel right about being in the house, so he orders his men out of the building—a perfectly standard building with nothing out of the ordinary.
>
> As soon as they leave the building the floor on which they had been standing collapses [because of a much larger hidden fire in the basement below]. (Klein, p. 32)

When questioned subsequently, the lieutenant had little insight into what prompted his decision, attributing it to a "sixth sense". The lieutenant reported that at the time of the fire he did not consciously entertain the thought that the house had a basement and that the basement contained a much larger fire. However, under further questioning, it became apparent that there were specific cues to which he had responded—the fire did not react as expected to the water, the living room was much hotter than would be expected for a fire of that size, etc.—all of which prompted the reaction that he did not know what was going on but that "something was not right", which in turn led to the decision to evacuate. Again, we see broadly the same pattern as in Lewicki's experiment—implicit learning on the basis of past experience which leads to a normatively appropriate "intuition" but without extensive deliberative reasoning and indeed with little awareness of the processes that generate the intuition or the cues on which it is based.

A well known example of compromised implicit learning and intuition is provided by the behavior of patients with orbitofrontal damage on the Iowa Gambling Task (cf. Bechara et al., 1994). In this task, subjects choose cards from one of several decks and win or lose money depending on the card drawn. Some of these are "bad" decks--they contain some cards with high rewards but the overall pay-off from these decks is negative. Other "good" decks have overall positive payoffs. Normal subjects learn to differentiate the good from bad decks and to draw from the former fairly quickly.

Moreover, they do so before they are able to provide an explicit rationale or justification for their selections. Measurement of their galvanic skin responses shows aversion to the bad decks well in advance of any conscious decision to avoid them. In some cases, subjects report having "gut feelings" that the bad decks are to be avoided. By contrast, OFC patients perseverate with the bad decks, in some cases even after they become consciously aware that they are losing money in drawing from them. They also fail to exhibit the galvanic skin responses of normal individuals. They thus appear not to show the implicit, intuition-based learning which is characteristic of normal subjects and as a consequence make sub-optimal decisions.

Many experiments have demonstrated that implicit learning is important in the social domain–indeed, it is thought that an enormous amount of real life social learning is implicit. In another experiment, Lewicki (1986) presented information to subjects about personality traits of people portrayed in photographs. Unknown to subjects there was a correlation between these traits and hair length in the photographs. Subjects were then asked to guess about the personality traits of people represented in new photographs. They did so by extrapolating the hair length/trait correlation, even though they remained unaware of this correlation and when asked to justify their trait ascriptions, pointed to the eyes rather than the hair length of those portrayed in the photographs.

Although this particular example may seem relatively trivial, there is a great deal of evidence for similar implicit learning in more complex social interactions. For example, in experimental studies of behavior in complex strategic environments (markets, auctions, bargaining games, etc.) with repeated interactions a generic result is this: subjects change their behavior over time in such a way that, according to some relevant criterion, their performance improves–for example, they make choices that result in their earning more money. In this sense, they learn (implicitly) more normatively correct or appropriate patterns of behavior. However, their verbal accounts of why their behavior is successful and even of the cues that they take to be guiding their behavior are often rather confused and bear little relation to the real reasons their behavior is successful.

When philosophers talk about "moral intuition", they typically have in mind responses that take the form of judgments (at least potentially verbalizable) about some example or episode which may be directly experienced, but more commonly is merely described in some verbal scenario. It is important to bear in mind however that social intuition generally and moral intuition in particular will often be linked not just to judgment but to non-verbal action or behavioral response in real life social interactions–indeed much of the significance and value of moral and social intuition rests on the contributions that it makes to such behavior. It is the capacities associated with social intuition that allow us to respond quickly and appropriately to the behavior of others, to correctly predict their behavior, to detect their intentions and other mental states and their reactions to our behavior and to adjust accordingly. In real life cases, it is typically intuitive social processing that allows us to detect from someone's facial expression or "body language" that they are annoyed or afraid, that they are likely to co-operate or betray us, and to adjust our behavior accordingly. The standard view is that more

deliberate reasoning processes are too slow and perhaps too unconstrained in other ways (see below) to do this effectively in real time–thus (some) reliance on intuitive processing in the social domain is essential for successful social interaction. This is illustrated by those with autism spectrum disorders, who are severely deficient in social intuition and must rely entirely on more deliberative and calculative reasoning to guide their social behavior, with the result that this behavior is often defective and inappropriate (Allman et al, 2005). To think of social intuition in this way is to think of it as involving processing and acquiring information (learning) about the social world, just as the fireman in Klein's example acquires information relevant to the prediction of a hidden fire, but without the phenomenological features philosophers associate with learning, such as explicit formulation of alternative hypotheses or conscious systematic weighing of evidence. If, as we have suggested, moral intuition is often best thought of as a form of such social intuition, it is far from obvious that we should try to entirely avoid reliance on it in moral judgment.

6.

We turn now to the very difficult question of what our discussion implies about the normative significance of intuitive moral responses. Is the fact that some action, policy, or principle comports with or conflicts with "moral intuition" relevant to the moral assessment of that action, and if so, why?

When posed in this very general way, we doubt that the question admits of any illuminating answer[5]. Reasons to discount the moral significance of at least some "intuitions" are easy to come by: there is considerable variation in people's intuitive moral responses to similar actions both across cultures and within cultures, with different people having "inconsistent" intuitions. Even a single person's intuitions may seem inconsistent and unstable over time. Many moral intuitions are based on mistaken empirical beliefs or have normatively unattractive sources–racial and gender biases and so on. Moreover, we have suggested that the most obvious routes to assimilating moral intuition to processes or experiences that are often veridical or truth-promoting (like visual perception or the experience of self evident steps in mathematical proof) rest on mistaken empirical assumptions about the processes that generate such intuitions. There is also experimental evidence that intuition in some situations can be biased by recent experience in such as way as to result in poorly adaptive decision-making (Kovalchik and Allman, 2006).

[5] Commenting on the general issue of whether the quality of decision-making is improved by adopting more intuitive or more deliberative strategies, Lieberman (2000) writes, "the unexciting answer proposed here is, it depends" (p. 110). We agree. However, we also note that given the willingness of a number of writers to claim that more deliberative strategies will always or usually lead to normatively superior decisions, it is by no means trivial to claim that there are many circumstances in which reliance on intuition or deliberation and intuition together can enhance the quality of decision-making.

For all of these reasons, we are not inclined to try to construct a wholesale defense of appeals to moral intuition. Instead, we will confine ourselves to some suggestions about the kinds of information that may, under the right circumstances, be conveyed by moral intuition, and its relevance to moral decision-making.

Our first observation in this connection is suggested by many of the examples of intuitive judgment described above. At least in the non-moral realm, it is entirely possible for an intuitive judgment to be normatively correct (reliable, trustworthy) even though the considerations that show the judgment to be correct are not fully known or recognized by or transparent to the subject at the time of the judgment. Thus, assuming that the normatively correct decision in the Iowa Gambling experiment is to maximize expected income, normal subjects begin to make normatively correct choices and to avoid the "bad" deck well before they can recognize or articulate in words what makes the choice of this deck inferior to the choice of the good deck, and while some subjects are able subsequently to recognize or explain why the choice of one deck is superior to the other, by no means all subjects who make normatively correct choices are able to do this. A similar pattern is present in Klein's example of the fire-fighter who gives the order to evacuate–the intuition that immediate evacuation is appropriate comes first and the cues on which the intuition is based (and which rationalize it or show it to be reasonable) are not immediately accessible to the fire fighter–not part of the content of his intuition. It was only later, and then with the help of others, that it was possible to reconstruct the factors that led to this intuitive judgment.

These observations have several immediate consequences that are relevant to philosophical debates about the role of moral intuition. First, it is a mistake to suppose, as many utilitarian critics of appeals to moral intuition do, that if subjects are unable to provide a systematic justification or "rational reconstruction" of the underlying basis for their intuitions, those intuitions must be unreliable, or not worth taking seriously. We should expect that in the typical case, both for moral and other kinds of intuitions, subjects will be unable to provide such reconstructions and have at best limited insight into the processes that produce their intuitions. This is consistent with those intuitions containing useful information and being normatively defensible.

Second, there are good reasons to be skeptical of the common tendency among moral philosophers to make heavy use of supposed intuitions about highly unrealistic examples and/or examples with which people have little if any experience. We include in this category examples that are frankly science fiction-like, or physically or biologically impossible. (Judith Thomson's (1971) case of spores that become attached to furniture and grow into people, Michael Tooley's (1972) example of pills that turn kittens into human babies, and so on). We also include examples in which features that are usually or ordinarily present (because of facts about human psychology, human social and political behavior, and what people can reasonably know) are stipulated to be absent or very different from how they typically are. Cases in point include versions of the Explorer example in which, it is stipulated that you know for certain that Pedro will carry out his threat to the kill the ten if you do not kill the one, that he will not harm the ten (ever) if you do kill the one, that Pedro will never make such threats again, that no one will ever

find out if you do kill the one, that killing the one will have no additional ill effects on you and so on. Also in this category are standard ticking bomb examples involving torture: terrorists have put a bomb in place that will kill thousands soon unless disarmed; you have captured one of the terrorists who you know knows where the bomb is but he refuses to divulge this information; you know for certain that if you torture him he will reveal this information, and that if you do not torture him, he will not reveal the information; it is also certain that the torture being contemplated will only be used in cases of this very sort and never more widely.

Philosophers appeal to such examples for a variety of reasons–because they are looking for cases that discriminate among competing moral theories and sometimes the only cases that will do this are unrealistic (a fact that is of itself of considerable interest) or because they hope to "isolate" which features of examples make a difference for our judgments by mentally removing other supposedly confounding features or by equalizing such features across pairs of examples.

One reason for being skeptical of appeals to intuitions about such unrealistic examples is simply that, as we have suggested, the deliverances of moral intuition are most worth taking seriously when we have repeated experience with feedback that results in implicit learning. If we are presented instead with examples with which we have little or no such experience–either because they are impossible or highly unlikely or atypical -- then it is unclear why we should take our intuitive responses seriously or what they show. This is perhaps obvious enough in the science fiction examples but we think it holds as well for examples that make assumptions that we have good reason to believe are highly unrealistic. Thus, if, as we would claim, the historical record indicates that once the use of torture is legitimized in extreme situations, it is virtually always "abused" in the sense of being inflicted in situations that are not "extreme" and on people who do not have information of imminent momentous crimes that cannot be obtained in other ways, then it is highly problematic to think that anything of moral importance can be learned by considering our reactions to imaginary cases in which none of these features are present. To the extent that our intuitions about torture, or about whether you should succumb to the threats of people like Pedro have been shaped by any process of learning with feedback, what they have been shaped by (and are sensitive to) is actual, real life experiences in which the stipulated-away features have usually or always been present. Moreover, it is implausible that people have introspective access which allows them to identify isolatable features of situations to isolatable features to which their intuitions are responding. Instead, to the extent that a learning with feedback process is operative, it is likely that intuitions are often shaped in a very holistic way by experiences with real-life situations in all of their embedded complexity, with people often having very limited access to the specific factors which shape particular features of these overall reactions.

In the non-moral cases described above, we have urged that there is a straightforward un-mysterious naturalistic explanation (involving repeated experience with feedback resulting in implicit learning) for how subjects come to make intuitive judgments that are normatively correct. There is no need to posit special faculties (over and above ordinary sense perception and emotional responses) that put us in touch with

non-natural properties, or that yield insights into *a priori* truths. The question we now want to explore is whether a similar story might be told about moral intuition, at least under the right conditions. Suppose that one is trying to decide whether it would be morally right to perform some possible course of action A. Suppose (we hope uncontroversially) that one is more likely to arrive at a morally defensible decision/assessment of A if this assessment reflects the operation of some process that exhibits the right sort of sensitivity and responsiveness to facts having to do with how oneself and others will be affected by A, how others are likely to respond to A, how this will affect all those concerned and so on -- call this the morally relevant information. (We make no restrictive assumptions about what this information may involve—it may include the kinds of information that both (parametric and strategic) utilitarians and deontologists think are relevant to moral decisions). Suppose one decides by consulting one's intuitive reactions to A. One thing this might involve, for example, is imagining or simulating doing A and seeing what it feels like, what one's emotional reaction to having performed this imagined action is or alternatively what it would feel like to be on the receiving end of A. If one's intuitive reactions to A track the morally relevant information in the right way, then these reactions may provide us with morally good advice about what to do in much the same way as the intuitive responses of the normal subjects in Damasio's card experiment help to guide them in making prudentially good decisions. In effect, one would be using oneself (and more specifically one's intuitive moral reactions) as a sort of instrument (a "moral thermometer") that yields recommendations about what to judge or do, just as the subjects in Damasio's experiment use their visceral reactions to the two decks of cards to guide their choice[6]. Someone following this procedure might take the strong reaction of disgust and outrage he or she felt upon seeing the photographs of prisoner abuse at Abu Ghraib as a prima-facie indication of (an emotional signal of) the wrongfulness of the treatment of prisoners that occurred there.

This suggestion raises an obvious issue that needs to be addressed. In the non-moral cases described above, it is relatively straightforward what the subjects' intuitive reactions are tracking or responding to when they are correct or appropriate, and the normative standards involved in judgments of appropriateness are uncontroversial. The fire captain's intuitions track the presence of an unusually dangerous fire, and the intuitions of the normal subjects in Damasio's experiments track something like their expected monetary rewards from the two decks. We think that the intuitions in question are normatively appropriate because it is uncontroversial that these are the features that good judgment and decision-making should track in these cases. By contrast, whatever the features may be tracked by our intuitive reactions to, say, different versions of the trolley problem or the Explorer dilemma, it may seem inevitable that the moral

[6] To the extent that one relies in this way on one's moral intuitions in guiding moral judgment, we might think of this as a kind of intuition-based moral reliablism, but without (in our formulation) carrying with it the realist commitments of standard versions of reliabalism. That is, the idea is that one's reactions track considerations that are morally relevant, but we don't necessarily have to cash out the notion of morally relevant considerations terms of the detection of moral facts (see below).

significance (if any) of these features is going to be highly controversial. Here, in other words, we apparently run up against the difficulty that while there is a widely shared consensus about what the fireman should be responding to, there is much less consensus in many cases involving moral intuition.

One way of out this dilemma is to ask what would be left out if we were to eschew all appeals to (or reliance on) moral intuition in moral judgment and decision-making. Suppose that considerations that virtually all moral theories agree are relevant to good moral decision-making are likely, as a matter of empirical fact, to be neglected if we follow such a decision procedure and that we have empirical evidence that those who are unable to make use of moral intuition tend to make decisions that virtually all widely accepted moral theories agree are morally defective. Then we would have, as it were, generic reasons to think that moral intuitions sometimes track considerations of moral importance and that we would be ill-advised to entirely neglect them, even though it would remain an open question (to be settled on some different, presumably case by case basis) what normative significance any particular moral intuition has.

The general picture we advocate of the contribution of moral intuition to moral judgment and decision-making is the following: in many real life cases (and of course especially in those that have the feel of a moral dilemma in which conflicting considerations seem to be present), the consequences of our actions for all of those affected will be extremely difficult to evaluate analytically – that is, by undertaking to construct an explicit list of all of the various ways that ourselves and others will be affected and then deciding what to do by employing some explicit rule that tells us how to combine or synthesize this information into a single judgment. This is so for a variety of different reasons. As suggested in section 2, a very important aspect of good moral decision-making involves successfully anticipating how others will react or respond to our choices, how we in turn may be led to respond and so on. In other words, moral decision-making has a *strategic* or *dynamic* aspect; typically we must think of ourselves as choosing not in a parametric environment in which the behavior of others is fixed and independent of our choices but rather in an environment in which the behavior of others will be changed in various complex ways by our choices. Predicting via explicit analytic calculation what will happen in such environments if we were to make one choice rather than another is notoriously very difficult. Moreover, both the behavior of others and the moral significance of our choices for them will of course depend on their beliefs, preferences, emotions, and values, including the beliefs and emotions they come to hold as a consequence of our choices—a choice which seems otherwise justifiable to us may be seen as insulting or demeaning by some of those affected, and a good moral decision-maker will need to recognize when this is the case, and take this consideration, along with many others, into account in deciding what to do. Thus many moral decisions are made in situations in which a large number of different considerations are relevant (they are high dimensional problems) and under conditions of very imperfect information in which there is a high level of uncertainty about the full impact and likely consequences of our actions.

One role of the social emotions and of moral intuition is to help overcome the limitations of purely analytical or rule-based decision-making procedures such as cost-benefit analysis. The problem with trying to make moral decisions on such a purely analytical basis is that it is very likely we will leave out (or fail to pay sufficient attention to or to be motivated by) considerations that are important, even from a cost-benefit perspective. This is in part because it seems to be true, as an empirical fact about human beings, that the number of different dimensions or different kinds of considerations that we are able to fully take into account in explicit conscious rule guided decision-making is fairly small – as we note below, empirical studies show that when presented with high dimensional decision problems that require the integration of many different kinds of information, those who attempt to decide entirely by conscious rule based deliberation often ignore all but a few dimensions of the decision problem. For example, in the moral realm we may fail to fully anticipate how our actions will be perceived by others (e.g. that they will be perceived as insulting or humiliating), how others will respond to our actions, how we and others may be affected by these responses and so on. In addition, defects in empathetic identification or sympathy may have the consequence that we are insufficiently motivated to take into account the impact of our actions on others.

As a concrete illustration, consider the apparent approval by high U.S. officials of interrogation procedures that involved parading male Muslim prisoners naked and in sexually suggestive situations in public places and in the presence of women, the use of religious insults, and threats involving dogs—procedures that anyone with even a superficial knowledge of Arab culture would recognize as highly degrading, humiliating, and offensive. Consider also the contention by some American commentators that these procedures were tantamount to "fraternity pranks" and thus not morally objectionable. Assume for the sake of argument that these procedures were prompted at least in part by a desire to obtain militarily useful information from the prisoners. Whatever the value of this objective, it seems apparent that the decision-makers and commentators failed to appreciate the full impact or significance of these procedures as experienced by the prisoners or the response of the rest of the world when these activities became public, the damage from which to U. S. interests almost certainly outweighed whatever benefits were obtained from the interrogations. In other words, they omitted or failed to see the significance of considerations that were unquestionably both normatively and prudentially relevant to their decision–making.

One thing that engagement of the social emotions and moral intuition (and the visceral processing that accompanies them) can contribute in cases like this is to enlarge the scope of the considerations that are taken into account in decision making so that relevant factors of the sort described above are included as well. One way this can be accomplished is through the role of such emotions in simulating the mental states of others. A large body of evidence suggests that 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—that is, in representing the mental states of others, we activate the emotional areas and processing in ourselves that are

involved in the those mental states when experienced by others[7] (Damasio, 1994). By further simulating how we would behave in the presence of this mental state in ourselves, we may also be able to predict successfully how others will behave, given that they have this mental state. In turn, the use of this simulation process has the important additional consequence that it has at least some tendency to have motivational force in our own behavior, since the simulation works by our actually undergoing aspects of the processing underlying the mental state we are detecting. Thus when we recognize that, say, another person has been humiliated by some activity by simulating this emotion in ourselves, this both helps us to predict how that person is likely to behave in response to that activity but also alters our own motivational set—both by directing our attention to the fact of the humiliation and making it more salient than it otherwise would have been and perhaps also by encouraging us to react negatively to it (regarding it as hurtful and disgusting). The "moral intuitions" many of us had in response to the humiliating and degrading photographs of prisoners at Abu Ghraib involved, we suppose, something like this process.

On the other hand, if someone does not employ this sort of simulation heuristic very readily—either because of damage to brain areas involved in such processing or because they have learned to disregard the intuitive emotional responses that normally result from (or are engaged in) such processing, then the considerations we have just described—the recognition of the humiliation (or at least its full depth), and appreciation of its likely effect on the behavior of others is more likely to be missed. Instead, attention may be focused almost exclusively on other aspects/dimensions of the decision problem, such as the desirability of obtaining information. Although we have no way of knowing for sure, we surmise that this is what happened when interrogation techniques like those employed at Abu Ghraib were approved.

This argument depends on the assumption that attempting to engage in purely analytic modes of decision-making (fully explicit rule-based deliberation) and neglecting to engage more intuitive modes of processing can lead to the neglect of relevant considerations and hence to normatively inferior decisions. There are both theoretical and empirical considerations that support this claim. The theoretical considerations have to do with the fact that conscious deliberation is restricted in the number of dimensions or attributes of a problem to which it can pay attention or take into account. Those who attempt to rely exclusively on this mode of decision-making thus often neglect or fail to take into account important dimensions of the problem that they face. By contrast, there is evidence that more intuitive, unconscious modes of processing are better at integrating "large amounts of information…into an evaluative summary judgment" and can lead to superior judgments and decisions for this reason (Djiksterhuis et al., 2006).

There is also empirical evidence that supports these contentions. In a recent experiment, Djiksterhuis et al. (2006) presented subjects with a choice among different models of automobile. In one condition (the simple condition) these were characterized

[7] The role of "mental simulation" in intuitive decision-making in both social and non-social contexts is also emphasized by Klein.

by a small number of favorable or unfavorable attributes, in different combinations (e.g. one car would have 3 favorable and one unfavorable attribute, another the opposite profile). In another condition, the cars were characterized by a large number of such attributes. Subjects were then either (a) asked to think about the cars for four minutes before choosing a favorite car (conscious thought condition) or (b) were distracted for four minutes by another task that required their attention and then asked to choose. (unconscious thought condition). In this second condition (b), subjects made normatively good choices (as measured by the number of favorable attributes the chosen car had) for both cars with simple and complex attributes, with no difference between these two conditions. In the conscious condition, performance on the simple attribute task was about the same as performance in the unconscious thought task, but performance on the complex attribute task was markedly inferior. Similar results were found on other choice tasks.

In another, more ecologically realistic, study shoppers who made complex choices (e.g. furniture) were compared with those who made simpler choices in real life settings like department stores. Immediately after making their purchases, shoppers were queried about whether they had engaged in extensive conscious thought about their purchases prior to making them and then three weeks later were asked how satisfied they were with their purchases. For complex choices, conscious thinkers reported less satisfaction than those who reported engaging in less conscious deliberation.

Similar results have been reported in other experiments–for example, Wilson et al. (1993, see also Wilson, 2002) report that students offered their choice of posters for room decoration and who are encouraged to engage in prior deliberation report less satisfaction with their choices in comparison with those who chose without much deliberation, and that when subjects in romantic relationships are asked to provide analytical reasons concerning how their relationships are going, these predict the future of the relationship less well than those who are asked just to report feelings about the relationship.

These results suggest that unconscious processing, and the intuition or emotions to which it leads, can sometimes lead to better decisions than conscious deliberation, at least when the decisions involved are “personal” or “prudential”. It is important to appreciate, however, as Wilson emphasizes, that this is not to say that having more information rather than less makes for worse decisions, or that it is better to be uninformed, inexperienced, or naïve about the subject matter of one’s decisions. Instead, in Wilson’s words, what the evidence seems to support is the conclusion that

> We should gather as much information as possible to allow our adaptive unconscious to make a stable, informed evaluation rather than an ill-informed one. Most of us would agree that it would not be wise to marry the first person we are attracted to. If we spend a lot of time with someone and get to know him or her very well, and still have a very positive gut feeling, that is a good sign. (2002, p. 171)

We suggest that a similar conclusion holds for moral decision making. When what we call moral intuition is functioning in a normatively appropriate way it will reflect the operation of what Wilson calls the adaptive unconscious on a range of relevant considerations and experiences, issuing in a similar sort of gut feeling about the appropriateness or inappropriateness of some course of action. At least sometimes such intuitions will lead to judgments/decisions that are superior to those arrived at on the basis of more deliberative and rule-based decision strategies.

7.

We noted above the evidence from Greene et al. (2004) that emotional areas are more active in at least some cases in which subjects make deontological as opposed to consequentialist (that is, parametric consequentialist) moral judgments in response to hypothetical scenarios. Partial support for this interpretation is also provided by a forthcoming study by Koenigs et al., that found that patients with damage to ventro-medial prefrontal cortex (and hence impaired emotional processing) made more utilitiarian judgments than normals in response to the same scenarios, although only on so-called "hard" dilemmas which are not resolvable by appeal to some simple generally accepted rule and which are assumed to require considerable emotional processing. As remarked above, Greene et al. interpret their results as showing that deontological judgments reflect the presence of distorting emotional factors which bias or interfere with the normatively superior judgments that would be recommended by more purely consequentialist considerations. We think, however, the results from ventro-medial prefrontal cortex (VMPC) patients might reasonably lead one to wonder whether emotional processing and the deontological intuitions with which they are apparently associated always detract from the goodness of moral decision-making.

We want to propose an alternative interpretation of this data in the light of the considerations described above. We begin by reminding the reader of two of the distinctive features of deontological moral theories and judgments as opposed to consequentialist theories (or at least parametric versions of those theories). First, deontological theories are generally understood to attach an independent weight to the structure of the intentions and motives with which agents act, in addition to the consequences those actions produce. Thus deontologists characteristically claim that it often matters morally whether an outcome results from an action or an omission, whether the outcome is intended as an end or means or instead occurs as a "side effect" of one's action, whether the outcome occurs as a result of one's own action or instead as a result of the actions of others, and so on. Second, deontologists (at least of the Kantian variety) seem (at least in their own view) to attach a greater or different weight to considerations having to do with dignity, respect, and not "using people as mere means" than many consequentialists.

Our proposal is that one reason why we (sometimes) find greater involvement of emotional processing when subjects make "deontological" judgments is simply that (a) emotional processing is more likely to be involved when subjects attend to facts having to do with motives and intentions (because such attention requires simulation employing

emotional processing, at least when these are present in situations involving complex, high dimensional choices) and (b) sensitivity to considerations having to do with dignity and respect requires empathy, and this too requires extensive involvement of neural areas like frontal insular cortex that engage in emotional processing. By contrast merely counting up the number of deaths associated with each of two options and choosing the option that produces the fewer deaths will often be a more mechanical rule-based matter and may not require much emotional processing. The qualification concerning high dimensional choices is present under (a) because we want to explicitly allow for the possibility that in some simple cases the deontological option also may be chosen on the basis of some simple, consciously accessible rule. For example, perhaps some subjects may generate "deontological" responses by employing a rule that says e.g., "it is always worse to produce a result via an action rather than by allowing it to occur". This may occur even in the absence of emotional processing. We would expect emotional processing to be particularly likely to be involved when the choice is complex and high dimensional, where there is no consciously accessible rule indicating what to do, and where emotional processing can play the integrating and synthesizing role described above. In other words, the involvement of social emotions in tracking intentions, motives, etc. seems to become crucial when the tasks the subject faces are complex, and there is no consciously accessible rule to guide the subject's behavior. This would explain the Koenigs et al. result that VMPC patients with compromised emotional processing generate deontological judgments like those of normals on relatively simple cases involving action/omission choices but diverge from normals in producing more "utilitarian," less "deontological" judgments in cases like the fat man version of the trolley problem in which there appears to be no rule that is consciously accessible to most subjects that generates the deontological response. Hauser (2006) shows that normal subjects are able to articulate rules rationalizing deontological judgments in some simple cases, such as those involving the act/omission distinction but are unable to articulate such rules in more complex cases in which they make deontolgical judgments.

Thus while it initially may seem puzzling and paradoxical that deontological intuitions involve extensive emotional processing, given that prominent deontologists like Kant have held that the source of such intuitions was to be found in "reason" rather than the emotions, when one considers what deontological judgments are sensitive to, at least in complex cases, the involvement of the emotions is just what one would expect.

If this suggestion is correct, then the characteristic moral intuitions of deontologists about outcomes that are intended vs. those that are mere side-effects, about using people as mere means, about treating them with dignity and respect and so on are at least tracking something. There is no reason to think of these intuitions as mere emotional noise, or as due entirely to outmoded religious dogma. But why suppose that what is tracked is morally significant or important.

This is a very large question, but a sketch of a possible answer can be found in our remarks above. Sensitivity to facts about people's intentions matters to good moral decision-making because these facts are relevant in all sorts of different ways to what will happen when we and others choose various courses of action, to how others are likely to

respond to our choices, and more generally because such information is highly relevant to an appreciation of the strategic structure of human interaction. Consequentialist decision makers who neglect such considerations (parametric consequentialists, as we called them) are likely to end up with results that even by their own lights are morally undesirable. A similar story can be told about decision makers who neglect considerations having to do with dignity and respect, as our brief treatment of Abu Ghraib illustrates.

To provide a further illustration that puts some concrete detail on this abstract claim about the connection between deontological-looking intuitions and situations with a complex informational structure having to do with motives, intentions, and dynamic interactions among people, let us return to the Explorer example. Recall the abstract structure: A threatens to produce some morally very bad outcome (e.g. to kill ten people) unless B does something that also involves a bad outcome (e.g. killing one person, C) but an outcome that is judged to be less bad than the outcome that A is threatening. In such cases, many deontologists will have the strong moral intuition that B should not kill C, even if the result is that A kills ten. At least some consequentialists will instead judge that B should kill C and will dismiss the contrary deontological intuition as misguided (perhaps on the grounds that it is a naïve, emotionally mediated overgeneralization of a salutary reluctance to kill in most ordinary situations to a case in which such killing is morally indicated.)

It seems to us, however, that there is a plausible case to be made that it is instead this consequentialist judgment that rests on a very naïve and incomplete analysis that neglects the relevance of the sorts of considerations having to do with intentions and strategic structure to which we have been drawing attention. To begin with, in most realistic circumstances (recall that we have argued that our intuitions insofar as they are worth taking seriously are shaped by these), the recommendation that B should (morally ought to) kill C, will amount to advocacy of a system of rules or practices and associated incentives[8] that allow A and others of similar mind (the As) to put Bs under a moral obligation to do what such As wish them to do, even when this involves the Bs acting in ways that (in the absence of the threat) all would agree to be wrong. Incentives are thus created for As to attempt to achieve their purposes by threatening to perform morally

[8] We recognize that some consequentialists will contend that we are not entitled to invoke rules and practices here. They will argue that the correct consequentialist recommendation is that B should kill the one but that this recommendation should be one-off, applying to this case only and concealed from others so that (supposedly) there is no issue of advocacy or creation of a practice. What B should do is one thing; what consequentialists should publicly advocate another. While we lack the space for detailed discussion, we think this argument will not work if we confine ourselves to realistic circumstances: among other things, if B follows this advice, A and the nine who survive will certainly know of his behavior and it is unclear why they will refrain from making this information public. More generally, there are all of the problems attendant on the consequentialist decision-maker/advice giver making different recommendations to different people/groups, the likelihood of this being eventually discovered, the likely consequences of this discovery and so on.

objectionable actions and for Bs to succumb to such threats. Thus by imposing appropriate threats As can put Bs under obligations not just to kill but to rob banks and deliver the money to them, to release As' confederates from prison and so on. By contrast, under a moral arrangement of the sort embodied in a more deontological set of rules, As will not be able to put Bs under obligations to do what would otherwise be morally objectionable by threatening something worse.

It is true, of course, as parametric consequentialists will emphasize, that if we have a system of rules according to which B is under no obligation to kill C, and B consequently refuses to do so, A may follow through on his threat to kill the ten. Note, however, that (again in realistic circumstances) B has no strong grounds for predicting that A will follow through on his threat if and only if B does not kill C. This is so for several reasons. First, given A's behavior, it is likely that his intentions and purposes, whatever they may be, give him reasons to want the involvement of B in the killing of C. (If not, why doesn't A simply kill C himself?) Insisting that B must not kill C, regardless of A's threat certainly blocks *these* intentions of A. Moreover, if these are A's only murderous intentions, and B refuses to co-operate, A has no motive for carrying through on his threat to kill the ten[9]. If on the other hand, A has other murderous desires/intentions as well – he's all too happy to kill the ten or various other victims himself, or has independent reasons for wanting them dead, then we have no particular grounds for accepting A's representations that he will not kill the ten as long as long as B kills C. We also have good reason to worry that if B complies, A will impose similar or worse threats in the future – e. g. B kills C, thereby saving the ten for the moment, but A makes new threats involving the ten next week. We don't claim, of course, that these considerations show for certain that B will not be successful in preserving the ten by complying with A's wishes, but merely that A's future behavior is highly uncertain and that B and others should take this fact into account in his calculations.

These observations barely scratch the surface (we have ignored, for example, the likely reaction of C and his family and friends to B's choices[10]), but they do perhaps begin to suggest the enormous complexity of the considerations to which an adequate

---

[9] Perhaps the situation is this: A wants C dead but is not in a position to kill him (C is well–guarded, etc.) B is in a position to kill C. Furthermore B is a parametric consequentialist and A recognizes that he can obligate B to kill C by threatening to kill the ten, thus achieving A's desire that C dies. If, on the other hand, B is not a parametric consequentialist and refuses to kill C, then unless A has some independent reason for wanting the ten dead, he has no motive for carrying through on his threat toward them.

[10] If C and his friends and relatives are all thorough-going consequenitialsts, then insofar as B's killing C is the optimal outcome from a consequentialist point of view, they all (including C) will enthusiastically welcome it. But in the real world, with more realistic assumptions about human motivation, C and his friends are unlikely to see things this way. If they have the opportunity, they will resist B, violently if necessary. If B succeeds in killing C, C's friends are likely to seek vengeance or justice against B, being unimpressed by the argument that B did the optimal thing from a consequentialist perspective. All of this should figure in B's calculations.

moral analysis of examples of this sort should be sensitive, and the extent to which they centrally have to do with understanding the intentions of the actors, the strategic structure of their interactions, how others are likely to respond to their choices, and so on. Again, we emphasize how consequentialist treatments of such examples that urge us to discount our "deontological" intuitions and instead recommend that B should kill C tend to ignore or downplay these considerations, focusing just on the ten who are threatened if B refuses vs. the one who will die if B acts. We suggest that this restriction in focus is not an accident—it is plausible that our deontological intuitions are tracking or responding to these additional considerations and that those who do not feel the pull of these intuitions will also be those who tend to neglect these factors in their moral assessment. Although we lack the space for discussion, we think that a parallel story might be told about many of the other examples involving competing deontological and consequentialist intuitions in the philosophical literature.

We recognize that the response of many deontologists will be that this suggestion has too much of a consequentialist feel to it to capture the true source and character of their intuitions. It will be said, for example, that the deontological prohibitions on B's killing C in the above scenario focus just on the intrinsic wrongness of B's killing C and are experienced as absolutist or unconditional in character; hence they cannot have anything to do with contingent facts about how, e.g., A is likely to behave under a moral scheme that permits him to place B under an obligation to kill C, etc. Our response to this is to reemphasize the observation made above that in cases of non-moral intuition it seems to be true as an empirical matter that people have relatively limited insight into the source and character of their intuitions—their phenomenology often reveals little about their provenance, either about what causes them or about what if any deeper normative justification they may have. We suspect that a similar conclusion is true for many moral intuitions: people may be right to have the intuition that it is wrong for B to kill C, but this intuition by itself may disclose little about why it is held – either about the factors that cause it or why it is a good or justifiable intuition to have.

We conclude this section with a final example, which we also take to illustrate the general point that it is a realistic and not just abstract possibility that the ability to empathize and to experience complex social emotions may enlarge the range of considerations to which the decision maker is sensitive (rather than, as some would have it, merely "clouding" judgment) and that this in turn may lead to greater sensitivity to "deontological" intuitions. (Here the relevant considerations are not primarily others' intentions, but rather have to do with factors like loyalty and friendship.) The example is an interview with a patient with developmental frontal damage – hence impaired emotional processing from an early age (from a study by Corinna Zygourakis, Ralph Adolphs and John Allman). The subject is asked to view a clip from a documentary film which describes the efforts of Hungarian Jews to survive during the last days of world war II when they are being rounded up by the Nazis and sent to concentration camps. In this particular clip, one of the survivors describes a long march to a concentration camp. He and two other boys promised to stick together, no matter what happened to them on this difficult march. One of the boys, however, was injured and began limping. A German soldier noticed the limping boy and shot him, while his friends were too scared to stand

up for him.

The patient was asked to rank the top three emotions that the person in the film clip feels (with 1 being the strongest emotion, 2 being the second strongest emotion, and 3 being the third strongest emotion). She could choose from a list of: anger, disgust, embarrassment, empathy, fear, guilt, happiness, pain, sadness, shame, surprise. The patient listed pain, anger, and sadness but not guilt or shame (which normal subjects rank very highly in their list of top three emotions). The patient clearly understood what was going on in the film since she correctly responded to all multiple-choice questions about objective aspects of the film. However, when asked how she relates to the characters, she responded that she "couldn't imagine being in that situation".

The patient was also asked the following question: "Did the person (actor) do the right or wrong thing in the situation depicted in this film clip? Circle a number from -5 to 5, where -5 is the most morally reprehensible/morally unacceptable thing the person could have done, and 5 is the most virtuous/morally acceptable thing the person could have done". The patient responded by circling a 2. When asked to explain why she answered this way, she said, "I would have done the same. Either three people would die, or just one."

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

The 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 would usually be regarded as) a “utilitarian” direction. Even if, in the final analysis, we agree with the patient’s judgment about this particular case, it seems overwhelmingly likely that her judgment about other cases in which empathetic identification and appreciation of the emotions experienced by others is morally important is likely to be defective[11].

[11] There is an important additional point to be made about the role of intuition and emotional processing in moral decision-making that has to do with the significance of genuine uncertainty in such decisions. Decision procedures like explicit cost benefit analysis require that we assign definite probabilities to the various possible outcomes of the options facing the decision-maker. That is, they provide no way to incorporate genuine uncertainty in the sense of absence of knowledge of these probabilities (or even

8.

We have been arguing that there may be a prima-facie case for taking our intuitive responses seriously as moral guides when certain conditions are met. One of the most important of these is that the conditions for some form of learning with corrective feedback be present (or some other process that plays a similar role) with this learning tracking considerations that are agreed to be morally relevant. This may take a number of different forms. The most straightforward possibility is that the subject himself has previous direct personal experience of the action in question (or whatever is the object of moral assessment), either as a doer of the action, or as the person acted upon, or as someone who has direct experience of what it is like for others who engage in the action or are on its receiving end. In the case of torture, this would include those who themselves have been tortured or have witnessed torture and its effects. It is on the basis of these considerations that we should take e.g. the reaction of John McCain or Jacobo Timerman to interrogation techniques that involve torture more seriously than the reaction of Dick Cheney or Rush Limbaugh.

The idea of an intuition-forming process that provides for an opportunity of learning with feedback from experience can be broadened in at least two ways. First, even in the absence of direct personal experience with the situation we are assessing our intuitive reactions may be influenced in various ways by the experience of others in a way that provides prima-facie support for taking those reactions seriously. (This may be the result of deliberate and self-conscious adoption of those reactions or it may simply reflect the fact that those reactions have been assimilated into the general culture in such a way that others are encouraged to have them.) Thus even in the absence of direct experience with torture, we may acquire strong intuitive reactions from those who have had such experience – either our contemporaries or those with such experiences in the past. For example, the strong revulsion of many political thinkers and actors in the seventeenth and eighteenth centuries, including the U. S. founders, who lived in a time when various forms of torture and cruel and degrading treatment were common have become to varying degrees part of the political and moral culture of many liberal democracies and influence contemporary moral intuition. The reactions of these historical figures reflect their lived experience with practices involving torture, as they are actually

---

absence of knowledge of the state space of possible outcomes) into our decision-making. However, as urged above, uncertainty in this sense is a pervasive feature of social and moral decision-making. Explicit cost benefit methodologies may thus encourage overconfidence about our ability to predict or calculate expectations about what will happen and may lead us to attach insignificant weight to the possibility that there are entirely unknown dangers associated with our action and to fail to take "worst case scenarios" sufficiently seriously, on the grounds that they are thought to be very unlikely. By contrast, several neural systems involved in emotional processing (FI and ACC) respond strongly to uncertainty (Critchley et al, 2001). The involvement of such systems in decision-making can thus help correct for overconfidence and encourage cautious behavior in the face of genuine uncertainty.

employed in real life contexts, and deserve to be taken seriously for this reason. A similar point holds for those who have lived under or have serious knowledge of contemporary regimes that employ torture.

A closely related point is that by gathering information about the behavior of others in morally charged situations, and what happens when they make various choices in real life situations, we may put ourselves in a situation in which our intuitive reactions are being shaped by a learning with feedback process with desirable characteristics. For example, one may learn about actual historical cases in which people faced choices about whether to co-operate with evildoers in the hope of preventing them from doing even worse things (or in the hope of at least ameliorating the bad consequences of their behavior). Or one may learn about what happens when governments either give in or, alternatively, refuse to do so when terrorists take hostages and threaten to kill them unless the governments do their bidding.

Yet another way in which a learning with feedback story about the shaping of intuitions may apply even in the absence of direct personal experience is through analogizing of the situation we are assessing to one with which we do have experience. Few Americans have direct experience with torture, but virtually all of us have experience with the intentional infliction of pain and humiliation. People can (and presumably often do) use their experience with actions falling in these more general categories to influence their reactions to torture. (There is, however, the obvious danger that reactions generated in this way may miss much of what is most distinctive and important about the unfamiliar particular case before us – torture is in important respects very unlike other episodes involving the infliction of pain and humiliation.) Similarly, very few people have experience with scenarios in which someone threatens others with death in order to get a second party to commit murder. However, many of us have had experience with situations in which someone threatens to do something bad to others unless we do something we regard as wrong or ill-advised and have learned, through our own experience or from the experience of others, that it is disastrous to give in to such threats in these circumstances.

9.

We conclude with a summary of our main claims, with an emphasis on those that make distinctive empirical predictions.

1) The neural structures and capacities that underlie many prototypical cases of moral intuition are also involved in the processing of social emotions and fast social cognition more generally. These are at least somewhat distinct from the structures that are activated in logical or mathematical reasoning and also from those activated in visual perception. To the extent this is so, the ideas that moral intuition is relevantly like insight into *a priori* logical or mathematical truths or like visual perception (either in terms of the mechanisms on which it relies or the conditions under which it is likely to be reliable) are misguided.

2) Because of 1), subjects with damage to brain areas known to be involved in the processing of social emotions will have different moral intuitions in some cases than normal subjects or will perhaps lack such intuitions. Thus, for example, the moral intuitions of prefrontal patients and autistics will be different from those without these deficits. Manipulation of emotional processing whether by behavioral or pharmacological means will also affect the intuitions of normal subjects. Subjects who fall within a range of normal functioning but exhibit more or less than average sensitivity or susceptibility to social emotions (e.g. those with mild Asperger's syndrome) will also exhibit distinctive patterns of moral intuition.

3) It is an empirical question whether human beings make better moral decisions by avoiding use of moral intuition and emotional processing and instead deciding entirely on the basis of deliberate, conscious reasoning strategies. The analogous question has already been investigated in cases involving prudential or personal decisions (e.g. consumer purchases) and there is evidence that the involvement of intuition and unconscious emotional processing produces better decisions.

4) Moral decision makers who do not employ normal emotional processing in their decisions (either because they are unable to do so because of neurological abnormalities, or for other reasons) will tend to neglect certain relevant dimensions of moral decision making—especially those having to do with information about intentions, motives, and likely reactions on the part of those affected. Such subjects will think about moral decisions parametrically rather than being sensitive to their strategic structure, neglecting considerations having to do with how others will respond to initial choices, long run incentive effects that alter the behavior of others, and so on. In this respect their judgments and decisions will look more "utilitarian", in the parametric sense of utilitarian described above. Thus we should expect more utilitarian judgments on the part of subjects with autism spectrum disorders, fronto-temporal dementia, as well as patients with damage to ventro-medial prefrontal cortex.

5) We would expect the following abilities/behavior to co-vary together: (a) Good skills at mind reading/social cognition, including the ability to accurately ascribe mental states to others and to predict their behavior, (b) susceptibility to social emotions like guilt, embarrassment, gratitude, trust, pride/pleasure in the achievement and good fortune of others and awareness of when one is experiencing such emotions, (c) tendency toward empathetic identification with others as indicated on standard empathy measures, (d) tendency to be sensitive in moral judgment to certain of the considerations (intentions, motives, respect) emphasized in deontological moral theories, (e) because of (d) superior moral decision making as judged by both deontological theories and sophisticated (non-parametric) versions of consequentialism.

Acknowledgements: We thank Ralph Adolphs and Steve Quartz for their insightful comments on earlier drafts of this paper, and Corinna Zygourakis and Ralph Adolphs for providing the data from the developmental orbito-frontal patient described in Section 7. We also thank Walter Sinnott-Armstrong for helpful correspondence regarding his own

views. This work was supported in part by a grant from the James S. McDonnell Foundation.

References

Adolphs, R., 2006. How do we know the minds of others? Domain-specificity, simulation, and enactive social cognition. *Brain Res.* 1079, 25-35.

Allman, J.M., Watson, K.K., Tetreault, N.A., Hakeem, A.Y., 2005. Intuition and autism: a possible role for Von Economo neurons. *Trends Cogn. Sci.* 9, 367-373.

Allman, J.M., Hakeem, A., Erwin, J.M., Nimchinsky, E., Hof, P., 2001. Anterior cingulate cortex: The evolution of an interface between emotion and cognition. *Ann. N.Y. Acad. Sci.* 935, 107-117.

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., Damasio, A.R., 1999. Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience* 2, 1032-1037.

Baron, J. 1994. Non- consequentialist decisions. *Behavioral and Brain Sciences* 17, 1-42.

Baron-Cohen, S., Ring, H., Wheelright, S., Bullmore, E., Brammer, L., Simmons, A., Willilams, S., 1999. Social intellilgence in the normal and autistic brain: an fMRI study. *European Journal of Neuroscience* 11, 1891-1989.

Bartels, A., Zeki, S., 2000. The neural basis of romantic love. *Neuroreport* 11,, 3829-3834.

Baumgarten, H.G., Göthert, M., 1997. (Eds.), *Serotonergic Neurons and 5-HT Receptors in the CNS*. Springer-Verlag.

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.

Bentham, J., 1789. *An Introduction to the Principles of Morals and Legislation.* Latest edition: Adamant Media Corporation, 2005.

Berthoz, S., Armony, J.L., Blair, R.J.R., Dolan, R.J., 2002. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain* 125, 1696-1708.

Blackburn, S., 1998. *Ruling Passions.* Oxford University Press, Oxford.

Borman, R.A., Tilford, N.S., Harmer, D.W., Day, N., Ellis, E.S., Sheldrick, R.L., Carey,

J., Coleman, R.A., Baxter, G.S., 2002. 5-HT receptors play a key role in mediating the excitatory effects of 5-HT in human colon in vitro. *Br. J. Pharmacol.* 135, 1144-1151.

Brown-Séquard, C., 1874. Dual character of the brain. *Smithsonian Misc. Collect.* 15, 1-21.

Craig, A.D., 2004. Human feelings: why are some more aware than others? *Trends in Cognitive Sciences* 8, 239-241.

Craig, A.D., 2003. Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology* 13, 500-505.

Crisp, R., 2002. Sidgwick and intuitionism. In: Stratton-Lake, P., (Ed.), *Ethical Intuitionism: Re-evaluations.* Oxford University Press, Oxford, pp. 56-75.

Critchley, H.D., Wiens, S., Rotshtein,, P., Öhman, A., Dolan, R.J., 2004. Neural systems supporting interoceptive awareness. *Nat. Neurosci.* 7, 189-195.

Critchley, H.D., Mathias, C.J., Dolan, R.J., 2001. Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron* 29, 537-545.

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

de Quervain, D.J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. *Science* 305, 1254-1258.

Dijksterhuis, A., Bos, M.W., Nordgren, L.F., van Baaren, R.B., 2006. On making the right choice: The deliberation-without-attention effect. *Science* 311, 1005–1007.

Elliott, R., Dolan, R.J., Frith, C.D., 2000. Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cereb. Cortex* 10, 308-317.

Finger, S., Beyer, T., Koehler, P.J., 2000. Dr. Otto Soltmann (1876) on development of the motor cortex and recovery after its removal in infancy. *Brain Res. Bull.* 53, 133-140.

Fiorillo, C.D., Tobler, P.N., Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* 299, 1898-1902.

Foot, P. 1978. The problem of abortion and the doctrine of the double effect. In: *Virtues and Vices*, Basil Blackwell, Oxford.

Gibbard, A., 1990. *Wise Choices, Apt Feelings.* Oxford University Press, Oxford.

Greene, J. D. (in press). The secret joke of Kant's soul. In: Sinnot-Armstrong, W. (Ed.), *Moral Psychology, Vol. 3: The Neuroscience of Morality*, MIT Press, Cambridge, MA.

Greene, J.D., Nystrom, L E., Engell, A.D., Darley, J.M., 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400.

Greene, J., Haidt, J., 2002. How (and where) does moral judgment work? *Trends Cogn. Sci.* 6, 517–523.

Greene, J.D., Sommerville, R., Nystrom, L.E., Darley, J.M., Cohen, J.D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108.

Haidt, J., 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychol. Rev.* 108, 814-834.

Hauser, M.D., Cushman, F.A., Young, L., Kang-Xing Jin, R., Mikhail, J. (Forthcoming) A dissociation between moral judgments and justifications. *Mind Lang.*

Kahneman, D., Tversky, A., 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychol.* 3, 430-454.

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

Kant, I., 1785. *Groundwork of the Metaphysics of Morals.* translated by Paton, H.J., Harper TorchBooks.

Karama, S., Lecours, A.R., Leroux, J.M., Bourgouin, P., Geaudoin, G., Joubert, S., Beauregard, M., 2002. Areas of brain activation in males and females during viewing of erotic film excerpts. *Hum. Brain Mapp.* 16, 1-13.

Kaufman, J, Paul, L, Manaye, K, Korenberg, J, Hof, P, and Allman, J. (2006)VonEconomo neurons are selectively vulnerable in agenesis of the corpus callosum, under submission.

Klein, G., 1998. *Sources of Power.* MIT Press, Cambridge, MA.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Hauser, M.D., Cushman, F.A., Damasio, T. (Submitted) Damage to ventromedial prefrontal cortex results in extreme utilitarian moral judgments.

Kovalchik, S., and Allman, J. (2006) Impaired reversal learning in normal individuals on

a modified Iowa Gambling Task: An emotional decision-making study of healthy young and elderly individuals. Cognition & Emotion. In press.

Lieberman, M., 2000. Intuition: A social cognitive neuroscience approach. *Psychol. Bull.* 126, 109-137.

Lewicki, P., Czyzewska, M., Hoffman, H., 1987. Unconscious acquisition of complex procedural knowledge. *J. Exp. Psychol. Learn.* 13, 523-530.

Lewicki, P., 1986. *Nonconscious social information processing*. Academic Press, New York.

Martin, R.D., 1990. *Primate Origins and Evolution*. Princeton University Press.

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

McGrath, S., 2004. Moral knowledge. *Philos. Pers.* 18, 209-228.

Mill, J.S., 1859. On Liberty. In: Robson, J.M. (Ed.), *Collected Works of John Stuart Mill, Vol. 18*. University of Toronto Press, Toronto, pp. 213-310.

Moll, J., de Oliveira-Souza, R., Eslinger, P.J., Bramati, I.E., Mourao-Miranda, J., Andreiuolo, P.A., Pessoa, L., 2002. The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *J. Neurosci.* 22, 2730-2736.

Moll, J., Eslinger, P.J., Oliveira-Souza, R., 2001. Frontopolar and anterior temporal cortex activation in a moral judgment task. *Arquivos de Neuro-Psiquiatria* 59, 657–664.

Nagel, T., 1972. War and massacre. *Philos. Public Aff.* 1, 123-144.

O'Doherty, J., Critchley, H., Deichmann, R., Dolan, R.J., 2003. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J. Neurosci.* 23, 7931-7939.

Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C.R., Gray, J.A., David, A.S., 1997. A specific neural substrate for perceiving facial expressions of disgust. *Nature* 389, 495-498.

Preston, S., de Waal, F., 2002. Empathy: Its ultimate and proximate bases. *Behav. Brain Sci.* 25, 1-72.

Railton. P., 2003. *Facts and Values: Essays Toward a Morality of Consequence*.

Cambridge University Press, Cambridge.

Rawls, J., 1971. *A Theory of Justice*. The Belknap Press of Harvard University Press.

Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., Kilts, C., 2002. A neural basis for social cooperation. *Neuron* 35, 395-405.

Rolls, E.T., 2005. Taste, olfactory, and food texture processing in the brain, and the control of food intake. *Physiol. Behav*. 85, 45-56.

Sanfey, A.G., Rilling,, J.K., Aronson, J.A., Nystrom L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755-1758.

Seeley, W, DeArmond, S, Carlin, D, Macedo, M, Bush, C, Miller, B, and Allman, J. 2006. Early frontotemporal dementia targets neurons unique to apes and humans. under submission.

Shin, L.M., Dougherty, D.D., Orr, S.P., Pitman, R.K., Lasko, M., Macklin, M.L., Alpert, N.M., Fischman, A.J., Rauch, S.L., 2000. Activation of anterior paralimbic structures during guilt-related script-driven imagery. *Soc. Bio. Psychol*. 48, 43-50.

Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., Frith, C.D., 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.J., Frith, C.D., 2004b. Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157-1162.

Singer, P., 1974/2002. Sidgwick and reflective equilibrium. *The Monist* 58. Reprinted in Kulse, H. (Ed.), 2002. *Unsanctifying Human Life*. Blackwell Publishing, Oxford, pp 27-50. Page references are to reprinted version.

Sidgwick, H., 1907. *The Methods of Ethics, 7th Edition*. Macmillan, London.

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

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

Small, D.M., Zald, D.H., Jones-Gotman, M., Zatorre, R.J., Pardo, J.V., Frey, S., Petrides, M., 1999. Human cortical gustatory areas: A review of functional neuroimaging data. *Neuroreport* 10, 7-14.

Spence, S., Farrow, T., Herford, A., Wilkinson, I., Zheng, Y., Woodruff, P., 2001. Behavioral and functional anatomical correlates of deception in humans. *Neuroreport* 12, 2849-2853.

Stratton-Lake, P., 2002. (Ed.), *Ethical Intuitionism: Re-evaluations*. Oxford University Press, Oxford.

Sunstein, C. 2005. Moral heuristics. *Behav. Brain Sci.* 28, 531-542.

Thomson, J.J., 1976. Killing, Letting Die, and the Trolley Problem. 59 *Monist* 59, 204-217.

Thomson, J.J., 1971. A defense of abortion. *Philos. Public Aff.* 1, 47-66.

Tooley, M., 1972. Abortion and infanticide. *Philos. Public Aff.* 2, 37-65.

Unger, P., 1996. *Living High and Letting Die*. Oxford University Press, New York.

Watson, K.K., Matthews, B.J., Allman, J.M., 2006. Brain activation during sight gags and language-dependent humor. *Cereb. Cortex* Advance Access online: doi:10.1093/cercor/bhj149.

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.

Wilson, T.D., Lisle, D.J., Schooler, J.W., Hodges, S.D., Klaaren, K.J., LaFleur, S.J., 1993. Introspecting about reasons can reduce post-choice satisfaction. *Pers. Soc. Psychol. B.* 19, 331-339.

Zald, D.H., Kim, S.W., 2001. The orbitofrontal cortex. In: Salloway, S.P., Malloy, P.F., Duffy, J.D. (Eds.), *The Frontal Lobes and Neuropsychiatric Illness*. American Psychiatric Publishing, Inc., Washington, D.C.

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.