**Emotion Versus Cognition in Moral Decision-Making: A Dubious Dichotomy** [1]

## I. Introduction

My goal in this essay is to explore some issues having to do with the contrast between "emotion" and "cognition" and the ways in which these figure in moral judgment and decision-making. I begin by sketching a view which I call the *Rationalist Dichotomy* (RD) position. This assumes a sharp dividing line between human "cognitive" (rational, conceptual) capacities and capacities labeled as "emotional" or affective and valorizes the former at the expense of the latter. I will then suggest that current understanding of how the brain works and of the functions of neural areas commonly described as "emotional" undermines the RD position and instead suggests an alternative picture, which I will call the *Integrative Non-Dichotomy* (IND) view. According to the IND view, emotion and cognition are not sharply distinct and emotional processing, properly understood, plays (and ought to play, on virtually any plausible normative theory, including utilitarianism) a central role in moral judgment. The implications of the IND view for the interpretation of various experimental results regarding moral decision-making and for "rationalist" projects in moral philosophy more generally will then be explored.

*The Rationalist Dichotomy position*. The RD view has been very influential, both historically and in recent theorizing, and both among those adopting naturalistic approaches to moral judgment (e.g. , Greene et al., 2004) and among those adopting less naturalistic and more aprioristic treatments (e.g., Parfit, 2011). The RD view sometimes takes a "scientific" form but more commonly (among moral philosophers) takes a form which assumes the superiority of reason but is non-committal about neurobiological and evolutionary details. In its "scientific" form RD is commonly framed in terms of the assumption that our reasoning abilities are "modern" in evolutionary terms (perhaps unique to human beings), highly sophisticated in terms of the information processing they carry out, and flexible in terms of their ability to respond optimally to a wide range of circumstances. By contrast, "emotion" is claimed to be produced by structures that are ancient in evolutionary terms (the product of the "reptilian" part of the triune brain, as Paul MacLean (1990) notoriously claimed), "primitive" in terms of the information processing they can accomplish, and inflexible and stereotyped in operation, often encoding responses that are genetically fixed and relatively unmodifiable. On this view, emotions are sometimes heuristically useful as "alarm bells" that alert us to events in our environment that impinge on our welfare, as when we have a fear response to a snake. However, precisely because they are primitive, inflexible and insensitive to many relevant considerations, emotions are likely to lead us astray when we need to make complex or sophisticated decisions. Although as noted above, the positions of many contemporary moral philosophers differ from the views just described in making no specific claims about the neural structures involved in cognitive or emotional processing, they nonetheless hold to the basic commitments of the RD position—in particular, its

[1] Thanks to Ralph Adolphs, Josh Greene, Antonio Rangel, and Kyle Stanford for helpful discussions.

positive assessment of the involvement of reasoning in moral decision-making and its negative assessment of the effects of emotion. Strikingly, these RD commitments are shared both by many philosophers sympathetic to utilitarianism (Greene, Singer, Parfit) and by many philosophers favoring alternatives to utilitarianism ( Kamm, 1993, many contemporary neo- Kantians, and, on many interpretations, Kant himself). These commitments provide one of the main motivations for the adoption of some variety of moral rationalism, according to which the source of moral requirements is found in "reason", conceived as something distinct from and independent of "emotion".

The neural structures generally described as involved in "emotional" processing include ventro-medial prefrontal cortex, orbital frontal cortex (which I will generally classify together as VMPFC/OFC), anterior cingulate cortex (ACC), insula, amygdala, and other structures like ventral striatum involved in reward processing. If we are to assess RD claims about the generally negative impact of emotion on moral judgment and decision-making, we need to understand what these areas *do* in the brain. What, if anything, do these structures contribute to decision –making, moral and otherwise? To what features of an organism's environment and/or the organism itself (for example, other neural structures) are these structures responsive? If one were to somehow remove their contribution to decision-making, replacing them with contributions commonly described as purely cognitive or "reason-based", what would be the likely upshot? Would this lead to normatively better decisions by some plausible standard? Is it even possible for neurotypical subjects to systematically do this?

*The Integrative Non-Dichotomist position.* In contrast to the RD view, I will argue for a very different account of the contribution to judgment and decision-making of the "emotional" structures mentioned above. On the IND account, which I believe is better supported by current brain research, these structures (and especially structures like OFC/VMPFC) engage in complex, sophisticated information processing and computation. What they compute are *values* or *rewards* associated with distal stimuli and with actions directed to those stimuli. The structures under discussion are highly flexible and capable of sophisticated forms of learning, particularly in social contexts, which are often informationally very complex. Contrary to what is sometimes thought, these structures have *not* been retained in relatively unmodified form from our non- human ancestors. Instead, they have continued to change under the distinctive selection pressures to which human beings have exposed, and as a consequence are importantly different, both anatomically and functionally, from homologous structures in non-human primates. They are used in information-processing tasks that in some respects are importantly different from those for which they are used in other animals. The result is that humans have emotional responses that differ in important respects from those of other animals, including other primates.

An illustrative example is provided by the human insula. In non-human mammals (e.g. rats), this structure is involved, among other things, in assessment of taste and food intake that is potentially harmful ( generating literal "disgust" reactions), as well as the monitoring of interior bodily states. In human beings, this structure is involved in (it has been co-opted or re-used for) a wide variety of other tasks, including empathetic identification, decisions regarding charitable donations, affective response to pain in self and others, reactions to perceived unfairness in economic interactions, and assessments of risk. Needless to say, not all of these activities are engaged in by rats – and when humans

engage in these activities, it is not by making use of a rat insula overlaid by more sophisticated "purely cognitive" control structures. The distinctive functions of the human insula are supported by distinct anatomical structures (cf. Allman et al., 2010)

The alternative picture of the role of neural structures involved in "emotional" processing I will defend has a number of consequences that should be of interest to moral philosophers and philosophers of mind/psychology. First and most fundamentally, it suggests that we should be skeptical of the idea that reason and emotion are sharply distinct and mutually exclusive categories. Second, (and relatedly) we should skeptical of attempts to ground morality purely in "reason", conceived as something distinct from "emotion". Third, there is no reason to suppose that *in general* the quality of judgment and decision- making (in either moral or non-moral domains) would be improved to the extent that emotional processing plays no role in such decisions. This is not to deny, of course, that involvement of particular emotions in some decisions can detract from the goodness of those decisions. For example, decisions made in intense rage are often not good decisions, either morally or prudentially. However, this observation obviously does not imply that judgment and decision-making would be improved if completely uninfluenced by the emotion-processing structures mentioned above. My view is that the general question: "does the involvement of emotion enhance or undermine the quality of decision-making?" rests on (or strongly suggests) a mistaken empirical presupposition -- that evaluative judgment and decision-making among neurotypical subjects could be typically carried out without the involvement of emotional processing, the only question being whether this would be a good or bad thing. Since, on my view, judgment and decision-making, both in the moral domain and elsewhere, usually involves emotional processing (see below for the qualification intended by "usually"), I believe that the real issue is not *whether* emotion should be involved but rather *how* it should be involved. Thus, insofar as questions about the impact of emotion on decision-making have determinate answers, a better approach is to ask more specific and nuanced questions: e.g., for a particular kind of decision, made under such and such conditions, does the involvement of this particular sort of emotion improve or detract from judgment?

Before turning to details of my argument, however, several additional clarificatory comments are in order. First, the claims that follow are intended as *empirical* and *causal* claims about the operation of certain neural structures and their influence on moral judgment and decision-making, rather than as semantic or conceptual claims about, e.g., how emotion enters into the "meaning" of moral judgments (as when it is claimed that moral judgment "essentially involves" the expression of emotion). The sorts of claims I will be defending are claims like (1.1) "insula activation causally influences judgments about the appropriateness of contributing to charity" rather than claims like (1.2) "when someone says stealing is wrong what is meant is that speaker has a strong negative emotional response (or a certain pattern of insula activation in response) to stealing".

Second, in what follows, I sometimes claim that certain decisions are normatively superior to others, either prudentially or normatively. Since my intention in this paper is not to defend any particular global normative theory but rather simply to explore how various neural structures influence moral decision-making, some explanation of the bases for these claims about normative superiority is in order. In an attempt to minimize reliance on question-begging normative assumptions, I have followed a strategy of

appealing to standards that are regarded as uncontroversial by most people and are endorsed by many different normative approaches. For example, VMPFC patients systematically make choices with effects like the following: they result in large financial losses, unemployment, and alienation of friends, family and co-workers. Identifying such decisions as prudentially defective does require normative assumptions, but these assumptions are not usually taken to be controversial. Similarly, subjects who donate money to charity in experiments described below show greater activation in neural areas identified as involving "emotional" processing, in comparison with non-donors. Common sense moral judgment as well as most moral theories, whether utilitarian or deontological, regard donations to charity as sometimes morally praise-worthy or at least not morally inferior to decisions not to donate. I assume such standards in what follows rather than arguing for them[2].

I should also emphasize that these assessments of decisions as normatively superior/inferior involves appeal to standards that are *independent* of any valuation of the neural processes leading to these decisions. In other words, I do not assume that decisions involving some particular level of cognitive and/or emotional processing are, for that reason alone, inferior or superior to other decisions. In my view, normatively inferior decisions can result both from processes in which so-called cognitive factors dominate and those in so-called emotional factors dominate, although it is also true (and consistent with this) that damage to areas like VMPFC/OFC tends to lead to consistently inferior prudential decisions.

## 2. The Basic Picture

I first sketch the basic picture I will defend and then turn to details. I see the structures mentioned above as involved in emotional processing (VMPFC/OFC and so on) as all having the function of processing information about positive and negative reinforcers, or, as they are also described in the neurobiological literature, "rewards" and "punishers"[3]. The current understanding is that these structures *compute* (literally -- see below) *values* (for the organism) associated with reinforcers and actions undertaken to provide reinforcers. *Primary reinforcers* are reinforcers which are such that it does not have to be learned that they are rewarding, because, e. g., this is specified genetically. Thus sweet-tasting stimuli are for most people automatically experienced as rewarding. Primary reinforcers include stimuli producing specific sensory states but also very likely include stimuli associated with more general and abstract rewards. For example, many

[2] In addition to the considerations described in this paragraph, other possible bases for non-question-begging normative judgments to which I sometimes appeal below include consistency (e.g., choices violating formal constraints like transitivity are regarded as normatively defective ) and whether a decision fails to take account of considerations that are regarded as relevant by virtually all competing normative theories. The general problem of how to identify non-question begging constraints for the normative assessment of moral judgments in the absence of any generally accepted moral theory is an important and difficult one which deserves more attention from philosophers.

[3] The following remarks are heavily influenced by Rolls, 2005, as well as by the other papers cited below.

human beings experience certain kinds of cooperative or mutually beneficial social interactions as intrinsically rewarding and social isolation or rejection as punishing. In many cases, this is probably something that does not need to be learned. This preference for pro-social interactions emerges very early in human development, prior to the acquisition of language and sophisticated reasoning abilities, and seems to be one of many features distinguishing us from other primates, such as chimpanzees, who do not seem to have the sorts of affective processing that lead them to value cooperative interactions in the way that humans often do[4]. The view taken here is thus that the value humans attach to cooperation and pro-social interactions is not just a matter of our being smarter or more rational than chimpanzees or better calculators of long-term self-interest, but also reflects a difference in the emotions we experience and what we care about.

Organisms like ourselves who are able to act flexibly require a great deal more than just a genetic specification of primary reinforcers. In particular, organisms capable of behavioral flexibility must be able to *learn* about *secondary reinforcers*—stimuli or objects that are predictive of the arrival of primary reinforcers, but often only imperfectly or probabilistically. For example, assuming that sweetness signals the presence of a substance that is biologically valuable, an organism must learn that certain foodstuffs will taste sweet and that others are foul-tasting. This involves a process by which the reward value that attaches to the primary reinforcer also comes to be associated with stimuli that are predictive of it, so that reward circuitry also becomes activated when, say, ripe blueberries are present because these are predictive of sweetness. In fact, the *same* reward circuitry and processing that are activated by primary reinforcers are also activated by secondary reinforcers or the expectation that they will occur. In human beings, reward learning and processing is sufficiently flexible that all sorts of stimuli may acquire the status of secondary reinforcers through learning. For example, monetary reward, although of no intrinsic biological significance, activates the same reward processing circuitry that is involved in responding to more "biological" stimuli like pleasing tastes. A similar point holds for many socially and morally relevant stimuli. Thus choices to give to charity, as well as decisions to behave cooperatively in prisoner's dilemmas activate the same neural circuitry as is active when subject experience rewards connected to sensory gratification. One important consequence is that there appears to be nothing in the brain corresponding to a "moral module" and no neural structures specifically devoted only to moral judgment or the processing of moral values.

As mentioned above, in many ecologically realistic circumstances secondary reinforcers are associated only probabilistically with the arrival of primary reinforcers. This is particularly true of the social environment in which, for example, we may have (at best) probabilistic information about how others are likely to behave toward us. For this reason, the structures involved in reward processing might be expected to be—and in fact are—highly sensitive to probabilistic information relevant to the arrival of reinforcers. In particular, there is evidence that structures like insula and VMPFC/OFC integrate information about reward value with relevant probabilistic information –

[4] For details, see, e.g., Hamann, et al. (2011) and Decety et al. (2011). As argued below, this is one reason why the rationalist project of deriving moral requirements humans find compelling from "reason" alone, independently of features of specifically human emotional processing, seems unlikely to succeed.

indeed, these structures seem to compute features such as the expected value of reward or reward variance (Quartz, 2009). Structures like VMPFC/OFC are also involved in rapidly adjusting estimates of expected reward value as the probability distributions governing rewards change, as in reversal learning, when previously rewarding outcomes become punishing and vice –versa. Again this role is particularly important in social contexts—for example, when your previous friend suddenly becomes your enemy or vice –versa. In addition, some neural structures involved in emotional processing and valuation such as the amygdala are sensitive to (among many other things) the presence of *uncertainty* as opposed to *risk.* In situations involving risk one does not know the outcomes of one's choices with certainty but does know the probability distribution governing those outcomes. By contrast, in situations involving uncertainty one does not know the probability distribution governing the outcomes resulting from one's actions, as would be the case if one were required to bet on a coin toss without knowing anything about the bias of the coin. Human beings (and presumably other animals as well) treat risk very differently from uncertainty, as evidenced by such phenomena as the Ellsberg paradox. In particular, they tend to be very averse to uncertainty—averse in the sense that they try to avoid choices involving uncertainty if possible, and, when these are unavoidable, often choose very conservatively in such cases, employing worse-case or maximin, rather than expected utility reasoning. These features of human emotional response involving risk and uncertainty figure importantly in moral and political theorizing: compare Rawls' characterization of his original position, which conceives of the parties as choosing under conditions of uncertainty, with Harsanyi's treatment, which treats this situation as involving only risk. Because the Rawlsian parties choose under conditions of uncertainty, it seems (given our aversion to uncertainty) intuitive or psychologically natural to suppose, as Rawls claims, that they will employ conservative, maximin reasoning. By contrast, because as Harsanyi conceives of matters, his parties face a situation of risk, his claim that they will employ expected utility reasoning in that situation and thus be led to some form of utilitarianism also seems plausible. In effect, Rawls' treatment trades on the difference between the neural processing underlying risk and uncertainty and the way that the latter affects evaluation-- this is one of many cases in which features of human neural processing involving "emotion" affect which moral principles we find appealing or intuitive. The sensitivity of many of the structures involved in emotional processing to complex and changing probability information is one reason why it is appropriate to think of these as involved in sophisticated and flexible information processing, rather than myopic, stereotyped responses.

One important way in which the values associated with secondary reinforcers are learned involves *reinforcement learning*. Put abstractly, a system involved in such learning produces signals—call them P -- that predict the arrival of some rewarding stimulus S' (sweet taste) on the basis of some other stimulus S (the sight of ripe berries). The system detects whether S' in fact arrives (and is rewarding) and then compares this with what was predicted by P. The discrepancy between these two (between actual and predicted reward) yields a reward prediction error. This error is combined with the original prediction P to yield a new updated prediction P' of the reward value expected on the basis of S which is then again compared with the actual reward. The process is iterated until the prediction error approaches zero. As I note below, reward prediction error signals of this sort (with the right normative characteristics) have been found in a

number of the structures involved in emotional processing including insula, amygdala, and VMPFC/OFC, which is one reason why it seems appropriate to think of these structures as involved in learning.

As noted above, many commentators (particularly those influenced by RD views) think of emotions as associated with relatively fixed and stereotyped responses to environmental conditions that are unaffected by learning and experience. It should be obvious from my discussion above that this is a misconception: one of the advantages of having "emotional" structures that compute and represent values is that this greatly enhances behavioral flexibility and normatively appropriate modification of behavior as a result of learning. Having the capacity to compute value assessments which reflect input from a variety of sources and to then use this to guide judgment and action allows organisms to move beyond fixed input/output patterns of response[5].

All of this has been very abstract and the reader may well wonder what it has to do with moral judgment and decision-making. More detailed examples will be discussed below, but to illustrate what I take the connection to be, consider the following generic example, which is a composite loosely based on a study of reports of autobiographical memories from ordinary subjects assembled by Jessica Escobedo at Caltech[6]. These include reports by subjects of things that they had done that they regarded as wrong or discreditable in their lives. The reports fell into a number of different categories, but one not uncommon set conformed to the following pattern: The subject S, often an adolescent, chose some course of action A, either not thinking at all about its impact on a second person or being willing to discount this impact because S anticipated other benefits from doing A or because S underestimated its cost for the second person. For example, in one report, the subject, call her Sally, avoided going to the fair with her best friend so that she could go with another group of girls who were more popular. Sally then became aware that her best friend felt hurt and distressed as a consequence of this action and this in turn led Sally to feel distressed, guilty and ashamed —that is, to have aversive feelings about having done A that outweighed whatever benefits obtained from A. Sally reported that these feelings were far worse than she expected. She described herself as having been "taught a lesson" by this experience and resolved not to do anything similar in the future.

This is an example of one-shot learning, rather than the more gradual learning that might take place in the presence of a probabilistic relationship and a less strongly aversive reinforcer but in other respects it seems to conform to the pattern described above. Initially, Sally predicts certain benefits and certain costs from choosing A, but she is wrong about these: the actual consequences of A are more aversive and less rewarding than predicted and this leads her to avoid A-like actions in the future. In at least this sense, Sally made a mistake or fell into an error in her choice of A—an error about how rewarding/punishing A would be to her—and she learned in the sense that she altered her valuation of A- like actions so that this valuation better reflected the rewarding and punishing features of the action for her. Put slightly differently, Sally learned in the sense that she acquired new and more accurate information about the reward value (for her) of A. It is these notions of "information", "learning", "mistake", "error" and so on

---

[5] Cf. Rolls, 2005.

[6] For some details about this data base, see Escobedo and Adolphs, 2010.

that I have in mind when I describe structures like VMPFC/ OFC as engaged in learning about reward value and involved in error correction.

More systematic studies of moral learning show episodes of the sort just described are common (although they are certainly not the only form of moral learning—see below). In his research on young children, Martin Hoffman (2001) describes a process, which he calls "induction", which works in roughly the following way: A child engages in what an adult would regard as a moral transgression (Mary hits Harry, causing him to cry). The adult then draws Mary's attention to Harry's reaction and says something like "How would you like it if Harry did that to you?" , thus (in many cases) creating or strengthening ("inducing") an empathetic response in Mary and an averse reaction to what she has done. Hoffman claims, as an empirical matter, that induction is much more effective than other measures (such as punishing Mary without further explanation) in altering behavior and in getting children to voluntarily conform to moral requirements. James Blair (2005) in his studies of psychopaths claims that normal human adults and children possess what he calls a violence inhibition mechanism which, at least in many circumstances, produces an aversive response when they cause harm to others; this leads them to learn to avoid harm-causing activities. This mechanism is impaired in psychopaths with the result that they do not learn to avoid behavior that harms others.

I recognize that many philosophers will wish to resist the suggestion that Sally (or the children described by Hoffman) have undergone "moral learning" or the like. They may wish to claim, for example, that Sally has learned only that (2.1) actions of kind A are aversive or punishing for her; and this is very different from (2.2) learning that A was morally wrong. I fully agree that (2.1) is different from (2.2) and nothing in my discussion, either here or subsequently, is meant to imply otherwise. As I see the example, Sally learns (2.2) and part of the causal explanation for why she learns (2.2) is the occurrence of (2.1) and the processing underlying it. The claim that the process described above causally influences Sally's judgment that she behaved wrongly does not require that we commit ourselves to the idea that her moral judgment involves "nothing but" her having an emotionally aversive reaction when she learns of the impact of her behavior on her friend. Fortunately for our purposes, as we shall see below, these causal claims by themselves suggest important consequences for moral psychology and moral theory—we don't need to supplement them with unnecessary "nothing but" claims.

The moral learning present in Sally's case represents just one possibility, although a particularly important one. Human beings engage in many other forms of moral learning – they can learn not just from experiencing the consequences of their own actions, as Sally does, but also from moral instruction (often in the form of stories and illustrations) in which they learn to have adverse or favorable reactions to various sort of behavior without actually engaging in that behavior. But these cases too seem to involve the same sorts of emotional processing and reward structures.

### 3. The Role of Orbital Frontal and Ventromedial Pre-frontal Cortex

As noted above, many different neural structures are involved in emotional processing and the computation of reward value. Among such structures, the OFC/VMPFC plays a particularly important role. Following many others (e.g. Rolls, 2005), I take OFC/VMPFC to contain a final, common, integrated value signal that is

computed via some variant of the reward learning process described above from inputs provided by other neural structures, where these depend on the nature of the decision problem faced. Some of these inputs come from structures like the amygdala and insula that are often described as involved in "emotional" processing, including processing of information having to do with the emotional states of others (so-called "hot" theory of mind). Other inputs come from structures like the right temporo-parietal junction (RTPJ) and posterior superior temporal cortex (pSTC), often regarded as involved in more "cognitive" or "cooler" theory of mind tasks such as the ascription of beliefs and intentions to others . For example, in a charitable giving task described below, willingness to give is correlated with activity in OFC/VMPFC and this signal is in turn modulated by input from other areas, including both pSTC and insula. The former is thought to be involved in processing information about the desires and intentions of recipients and the latter is involved in empathetic identification. In addition, input that modulates value signals in OFC/VMPFC can also come from what are often taken to be paradigmatic "cognitive control" structures like dorso-lateral prefrontal cortex (DLPFC). (Structures taken by Greene and others to be fundamentally different from "emotional" structures). Although there is some uncertainty about whether there is an independent "value signal" originating in DLPFC (or whether instead DLPFC merely modulates signals in VMPFC/OFC), the weight of opinion at present seems to be that *all* rewarding and punishing stimuli, as well as their combined or integrated significance are eventually represented in some way in OFC/VMPFC. It is thus assumed that these structures have the function of combining reward signals concerning quite different stimuli into a "common currency" that guides comparison and choice among different alternatives, as when an animal must choose between continuing to exploit a food source and a mating opportunity. Such a common currency is thought to be required if decision-makers are to choose in a minimally consistent way, avoiding, e.g., intransitive preferences. (Damage to VMPFC/OFC produces, among many other pathologies, violations of transitivity—Schoenbaum et al, 2011.) Moreover, as one would expect from the characteristics needed if such a signal is to guide all-things-considered judgment and decision in situations in which probabilities attach to outcomes, it does not merely represent that some stimulus or outcome is valued more or less than another but instead represents something more like reward magnitude—in other words the signal has a "cardinal-like", rather than merely ordinal, character, which makes possible the representation of quantities like expected reward.

As is well –known from cases like that of Phineas Gage or EVR (Anderson et al., 1999), OFC/VMPFC patients suffer from many pathologies in choice and decision-making, apparent both in laboratory contexts and real life. The best characterization of these incapacities are a matter of on-going disagreement but they include choices that are highly imprudent, failure to learn from mistakes, problems in emotional regulation, and failure to conform to ordinary norms of social behavior. A recent review (Schoenbaum et al. 2011) takes the fundamental impairment to be a failure to learn about the valuations of

expected outcomes [7] , which is very much in line with the account I have advocated, but in any event such patients are profoundly disabled and require institutionalization.

Several other features of OFC/VMPFC are worth emphasizing. First, not only are there extensive projections forward from structures like the amygdala and insula to OFC/VMPFC, there are also backwards projections from OFC/VMPFC to these structures. These allow OFC/VMPFC to influence and modulate these structures, rather than just serving as receivers to which they broadcast. Acquisition/ learning of normatively appropriate responses in structures like the amydala is thought to depend on appropriate modulatory input from structures like OFC/VMPFC—for example, according to one prominent theory (Blair, 2005), psychopathy is associated with deficiencies in VMPFC/ amygdala connections, leading to failure to develop aversive responses to harm inflicted on others. In general, during the process of normal maturation, from childhood to adulthood, functional connectivity between VMPFC/OFC and other structures like amygdala and insula increases (Decety, 2011).

Given the evidence just reported about the role of a common set of structures involved in the representation of the values of stimuli as food and money, a very natural expectation is that these same structures should also be involved in moral evaluation and the representation of moral value. This is exactly what is reported in a number of recent papers. In an fMRI experiment investigating brain regions involved in charitable giving (Hare et al. 2010), subjects were given an amount of money which they could donate to charities. Subjects were imaged in both a forced giving task in which they were required to make donations and a "free choice" task in which they made their own decisions about donations. In the latter task, but not the former, a value signal was present in VMPFC/OFC which correlated with amount donated, suggesting that this signal computed the value to the subject of making a donation. This signal in VMPFC/OFC was found in turn to be modulated by areas known to be involved in social cognition, in this case anterior insula and posterior superior temporal cortex. pSTC is thought to be involved in shifting attention to another's perspective and to contain an updating signal in contexts in which the optimal choice involves detecting the intentions of others (Behrens et al, 2008, Hampton et al. 20008) . Both regions have been found in other studies to be active in tasks involving altruistic giving. The general picture is thus again one in which a value signal computed in the "emotional area" OFC reflects input from other areas, and in a way that seems normatively appropriate, both in the sense that charitable giving itself is sometimes normatively appropriate and in the sense that it seems normatively appropriate that areas involved in empathy and social cognition and perspective taking should influence decisions about charitable donations. This is because appropriate charitable giving requires awareness of which choices will benefit others and these areas process such information.

Given these empirical results should we think of VMPFC/OFC (or for that matter, the insula and amygdala) as "cognitive" or as "emotional" areas? In some important respects, much of the activity in these areas seems to be cognition-like or representational, or at least to be involved in "information processing". First, these areas

[7] These authors explicitly deny (and provide evidence against the claim) that the primary role of OFC is to inhibit inappropriate responses, as suggested in the Shenhav and Greene paper discussed below, or to flexibly encode associative information.

are involved in calculation, computation, and learning and these are activities which are often thought of as "cognitive". In addition, signals in VMPFC/OFC, as well as in structures like insula are representational in the sense of representing quantities like expected reward and reward variance. Moreover, valuations associated with his area are attached to representations of objects and situations computed in occipital and temporal areas that are relatively cognitively complex and in some cases rather abstract – for example, to a subject's representation of "snake" (evaluated as dangerous) or "offer received in an ultimatum game" (evaluated as unfair and exploitive)**.** Notions of error and misrepresentation also often seem applicable to activities in these areas, sometimes in very straightforward ways. For example, areas like insula can misrepresent experiences or evaluations of other subjects affected by one's actions, in the sense of failing to accurately represent pain experienced by others. More generally, these areas can also misrepresent in the sense of failing to accurately represent the bearing or significance of various environmental stimuli for the subject or whether they should be regarded as sources of concern. For example, a harmless snake may be evaluated as dangerous and so on. Finally, as emphasized above, structures like VMPFC/OFC combine inputs from many different sources, including those usually regarded as cognitive such as the pSTC which also may make it seem appropriate to think of the VMPFC/PFC as "cognitive".

On the other hand, if one associates "reasoning" or "cognition" with deliberate effortful inference involving the conscious manipulation of proposition-like mental structures of the sort carried out in logic and mathematics, including the use of complex combinations of these strcutures, then because the information processing carried out in structures like VMPFC typically does not involve this sort of conscious inference, labels like "cognition" and "reasoning" seem misleading. In addition, the processing in VMPFC appears to be dedicated and specialized in the sense that it is quantities like expected reward and so on that are computed—there is no suggestion that this machinery might be used for other sorts of abstract reasoning. This assessment is reinforced when one considers that there is independent evidence that logical and mathematical reasoning is generally carried out in areas like dorso-lateral prefrontal cortex (DLPFC) that are distinct from the areas standardly associated with emotional processing. Characterization of areas like VMPFC as involving "reasoning" will also be misleading if one also associates reasoning with the absence of affect or if one thinks of one of the roles of reasoning to be the suppression of emotion or its replacement with other sorts of "purely rational" motivations.

These considerations lead me to suggest that invoking a dichotomy between cognition and emotion (= non-cognitive) and then arguing about which side of this dichotomy the activity of structures like VMPFC/OFC should be placed is probably not a fruitful way of proceeding. If by "emotional" we mean structures that are affective, involved in evaluation (appraisal of what ought to be the case) and in motivation, then structures like VMPCF/OFC are indeed involved in "emotional" processing, but this does not mean that they are non-representational, impervious to influence from learning or by cognitive structures, or passive in the sense of not involved in modulation and control of other neural structures. Processing in such structures seem to involve an intermixture of elements we assocate with information-processing and affect.

Whether or not this skepticism about the emotion/cognition dichotomy is correct, it definitely seems wrong to think of the role of the insula in the charitable donation

experiment as merely a "distractor" or alarm bell which, unless suitably controlled, would undermine the superior decisions that would otherwise be made by purely "cognitive" processing. Judged by both deontological and utilitarian standards, the normatively superior decisions in this experiment, will at least sometimes involve donating to charity rather than keeping all of the money for oneself. Moreover, willingness to do this, as we have seen, is correlated with and apparently causally influenced by level of insula activation. Thus, assuming the above standard for what is normatively appropriate in charitable giving, subjects whose choices are *not* influenced by their "emotional" insulas make normatively inferior decisions, in comparison with those who are so influenced[8].

### 4. Greene on Deontological versus Utilitarian Decision-Making

With this as background, I want now to turn to some very interesting and influential papers by Joshua Greene on the neural structures involved in moral judgment[9]. In Greene et al., 2004, subjects were presented with a series of moral dilemmas designed to elicit choices that were either more "utilitarian" or more "deontological". Greene et al. found greater activation in VMPFC among deontological (as opposed to utilitarian ) decision makers and more activation in DLPFC among utilitarian as opposed to deontological decision-makers. On the basis of their identification of VMPFC as an "emotional" area and DLPFC as "cognitive", Greene et al. concluded that deontological decisions were differentially associated with emotional processing and utilitarian decisions with more cognitive processing. In a similar vein, reports from several sources (e.g. Koenings et al., 2007) that patients with lesions to VMPFC make more utilitarian decisions, were taken to show that utilitarian decisions

---

[8] Of course, it might be responded that such subjects are irrational (or "non-rational"), since their choices are influenced by "emotion", and that even in the absence of any influence from emotional areas like insula, subjects who are "rational" would be led just through the exercise of their reason to donate large amounts to charity ( perhaps on the grounds that reason tells me that it would be "irrational" to give any particular preference to myself over any other sentient creature in the universe or some such.) If this is understood as an empirical claim and "rational" subjects are understood as those showing a high level of activity in cognitive areas like DLPFC, then this response predicts that subjects with such high levels of cognitive activity and no activity in emotional areas like insula will make large charitable donations—a claim for which, as far as I know, there is no evidence. If the claim is understood as an apriori or conceptual one (in effect building a willingness to donate into the definition of "rationality") then one faces the problem that subjects with a high level of activity in what are thought to be "cognitve" areas still may not be rational according to this definition.

[9] Although I express some skepticism below about several of Greene's claims, I want to say explicitly that, probably more than any other researcher, he has played a central role in developing and driving forward the whole area of investigation of the neural structures underlying moral judgment. Disagreement with him is a measure of the extent to which his ideas have established the agenda and framework for discussion in this area.

rely less or not at all on emotional processing, again on the basis of the claim that the VMPFC is an emotional area. Following the RD position outlined in Section 1, Greene et al. also hold that greater involvement of "cognitive" processing generally leads to normatively superior decisions. They thus conclude that these results about neural processing provide support for the superiority of utilitarianism as a normative doctrine, a claim which has also been endorsed by Singer (2005).

A more recent paper by Shenhav and Greene (2010) complicates these claims in an interesting way. These authors presented subjects with hypothetical choices involving rescuing groups of people of different sizes with different probabilities of success. They found what they describe as:

> a BOLD signal in VMPFC/MOFC correlated with the "expected moral value" of decision options -- the interaction between the value of various outcomes and probability. This is consistent with the hypothesis that this region supports the integration of positive and negative reward signals into a more abstract representation of value, a kind of decision ''currency''. (p. 671)

They add, "our results suggest that an individual's sensitivity to lives/saved lost in the context of moral judgment is in part determined by the same mechanisms that determine the individual's sensitivity to the probability of losses and to overall reward in the context of economic decision-making" (673). Thus, their results fit very nicely with the ideas described in earlier sections of this paper, according to which the OFC/VMPFC contains an overall, integrated signal which represents moral, as well as other kinds of value, and which takes account of information about probabilities.

As so far described, the results from this paper say nothing about distinctively utilitarian moral decision –making. In fact, Shenhav and Greene found a value signal in OFC/VMPFC among decision-makers who have what they consider to be a utilitarian orientation but they also found such a signal among more deontological decision-makers. This is exactly what one should expect if, as argued above, OFC/VMPFC is involved in the representation of value for *all* those who make judgments or decisions, whether utilitarian or deontological. Of course, this value representation will differ in detail in utilitarian and non-utilitarian decision-makers, since these subjects do, after all, make different decisions, with this difference presumably reflecting differences in input from other neural structures, but both the results from Shenhav and Greene and those described in earlier sections suggest that this will not be a matter of the deontologically inclined making use of an entirely different system or mechanism of value representation and computation from those who make utilitarian judgments. This has an important consequence: If VMPFC/OFC has a role in *all* decision-making, then it appears that one can't argue for the moral superiority of utilitarianism over deontology merely on the grounds that the "emotional" VMPFC/OFC is involved only in the latter.

Is there nonetheless some basis for retaining the idea that utilitarian decision-making has a distinctive neural signature? As I understand them, Shenhav and Greene remain committed to the idea that that utilitarian decision-making disproportionately involves "cognitive" processing, although they now identify the OFC as involved in such processing (whereas previously Greene regarded it as an "emotional" area) . They write:

> We found that increased activity in bilateral lateral OFC … was associated with more frequent endorsement of utilitarian trade-offs. According to Greene et

> al.'s dual-process theory of moral judgment…, utilitarian judgments are driven primarily by controlled cognitive processes, which may compete with countervailing emotional responses. These results are broadly consistent with this dual- process theory, given the implication of lateral OFC in reducing the influence of emotional distracters on judgments (regulating pain and negative emotions, (favoring delayed rewards over more immediate ones), inhibiting socially inappropriate behaviors), and more generally controlling the influence of emotional responses that interfere with the pursuit of more distal goals (673).

Noting that Greene had previously suggested that utilitarian judgment was associated with increased activity in DLPFC (a "cognitive" area) rather than OFC, Shenhav and Greene go on to suggest that DLPFC may be associated with effortful cognitive control of emotions, while OFC may be involved in more "implicit modulation of affective representations", adding that this is "consistent with the aforementioned literature implicating the lateral OFC in performing a gating or weighing function as opposed to the overriding of a prepotent response" (674).

Let me make what I hope are some constructive suggestions about this, which may or may not accord with Greene's current views. First, the change in classification of OFC (or OFC/VMPFC) from "emotional" to "cognitive" seems to me support the misgivings expressed earlier about the clarity and usefulness of this distinction. Second, phrases like "inhibition", "reducing the influence of emotional distractors" , and "modulation" suggest a range of different pictures concerning the relation between the VMPFC/OFC and the input which it regulates. It is important to be clear about which of these pictures is most appropriate. It seems fairly clear both from Shenhav and Greene's paper and from the other results reported in section 3 that it is misguided to think of the role of the VMPFC/ OFC as one of entirely removing or suppressing the influence of "emotional" or affective factors on decision-making, allowing only purely cognitive or rational factors to be operative. Assuming one thinks of structures like the amygdala and insula as engaged in emotional processing, then a picture according to which information from these sources is integrated or synthesized with from information from other sources by the VMPFC/OFC when judgment and decision-making are functioning in normatively appropriate ways seems more correct than a model in which emotional influences are excised from decision-making, either by VMPFC/OFC or other structures. This is not at all to deny that judgment and decision in which VMPFC/OFC exercises top-down control or regulation are often normatively superior to judgments and decisions in which these structures are not playing this role—if nothing else, this is suggested by the poor quality of decision-making in subjects in which these structures are damaged. But it does suggest that this normative superiority, when present, is not just a matter of "cognition" inhibiting "emotion". Rather than thinking of the activity of VMPFC/ OFC as "cognitive" simply on the grounds that it involves regulation of emotion, it seems to me more natural (echoing the arguments of section 3) to think of these structures as neither exclusively cognitive or exclusively emotional, but rather as combining elements of both.

I can expand on this point and also on the role of effortful "cognitive control" involving structures like DLPFC in decision-making by comparing Shenhav and Greene's experiment to two very interesting papers by Hare et al. (2009, 2011) on

dietary choice. In the experiments reported in Hare et al. 2009, self-described dieters were scanned while asked to rate food stuffs for health and tastiness separately. For each subject, a reference item was selected that was neutral for both taste and health and subjects were then presented with choices between various food items and these neutral items. Subjects who were self-controllers (that is, subjects who chose healthy items) were found to make their decisions on the basis of both health and taste and exhibited activity in VMPFC that reflected the integrated influence of both health and taste considerations, as expressed in the earlier ratings. By contrast, signals in VMPFC of non-self controllers reflected the influence only of taste evaluations. The crucial difference was that in the self-control group, VMPFC activity was modulated by DLPFC, generally thought of as a more cognitive area , which is consistent with a large body of independent evidence implicating the DLPFC in "cognitive control". At least in these experiments, however, there was no evidence for a "value signal" within DLPFC; instead DLPFC influenced or modulated evaluation of food items, by influencing the value signal within VMPFC.

The second paper confirmed and extended this picture, showing that when individuals made healthy choices and exercised self-control, the value signal in VMPFC was modulated by DLPFC in such a way that greater value was attached to healthier foods in comparison with tasty foods.

Suppose, for the sake of argument, that, as Greene claims, increased activity in areas associated with "effortful" cognitive or executive control like DLPFC (in addition to "gating" activity on the part of VMPFC/OFC) is characteristic of utilitarian decision-makers. Focusing for the moment just on the role of DLPFC, if what goes on among such "utilitarian" subjects is like what goes on among subjects who exercise dietary self-control, perhaps what one should expect is the following: in both deontological and utilitarian subjects, a value signal associated with choice and judgment will be present in VMPFC /OFC, which is just what Shenhav and Greene found. However, (continuing this line of thought) for utilitarian (as opposed to deontological) subjects, there will be more activity in DLPFC and perhaps other cognitive areas which will operate so as to modulate this value signal. Put differently, if one wants to continue to think of VMPFC/OFC as an "emotional" area, then the difference between deontological and utilitarian subjects will not be, as some of the language in Greene's earlier papers perhaps suggested, that the former rely solely on input from emotional areas and the latter solely on input from cognitive areas, but rather that both rely on valuation signals in VMPFC in judgment and decision-making, but, in the case of utilitarian decision makers, in contrast to deontological decision makers, this emotional input will in addition be more influenced by processing in areas associated with cognition and executive control. I will add that if, as I have already suggested, the "emotional" and "cognitive" labels are not terribly helpful in this context, it might be even better to put matters in terms of a contrast between valuation that includes but is not limited to areas associated with executive control and working memory versus valuation not involving those areas, or involving them to a lesser degree (but perhaps involving other areas to a greater degree—see below). Going further one might also suggest, in contrast to a standard "dual process" picture in which there are two "systems", asymmetrically related, in which one has the role of inhibiting or correcting mistakes made by the other, it may be more appropriate to think in terms of on- going reciprocal interaction and feedback, involving a number of

"systems" rather than just two. Consistently with all this, one might still retain the idea that greater involvement of systems associated with cognitive control, like DLPFC and whatever else one might want to include in this category, typically leads to normatively superior decisions and that in virtue of such involvement, these decisions tend to be "utilitarian", thus yielding an argument for utilitarianism as a normative theory along broadly the same lines that Greene has endorsed before.

I put this forward as a way of framing or reframing Greene's views –- as a friendly amendment (or perhaps a description of what he now thinks). I want now, however, to raise some questions about the resulting picture: First, note that given the overall argument above, it is not clear why one should expect to see areas involved in cognitive control *only* when there is "utilitarian" choice. An alternative, and in many ways more natural, assumption is that higher levels of "cognitive control" are required whenever a subject successfully suppresses and acts against a strong initial adverse reaction to some course of action, whether this action is recommended by utilitarianism of not[10]. As an extreme (but real) case, consider those fanatical Nazis who described themselves as feeling empathy for their victims and revulsion at killing them but who self-consciously set out to suppress these feelings in favor of what they took to be their duty to conform to Nazi doctrine, acting out of what some commentators have described as a sort of perverted or distorted Kantianism that prescribed duty's for sake. One might conjecture that these Nazis had rather high levels activity in DLPFC and other cognitive control areas as they struggled against their feelings of humanity but their choices were not "utilitarian" either in the sense that they conformed to what utilitarianism understood as a normative doctrine requires or in the sense that they were attempting to decide on the basis of utilitarian considerations.

A less extreme case, conceivably subject to experimental investigation, might involve someone who is committed to conforming to some requirement regulating diet or religious observance and is tempted, for either self interested or non-religious other regarding reasons not to follow this requirement (e.g. perhaps helping someone in need would involve violating this requirement). If Rangel's diet experiment is any guide, successful resistance to the temptation to violate the requirement would involve activity in DLPFC and other areas associated with cognitive control, including perhaps VMPFC/OFC, even though successful resistance may not (and often will not) be recommended by utilitarianism, or based on explicit utilitarian calculation, or, from the point of view of many of us, normatively appealing.

One possible response concedes the possibility just described but contends it is not the ecologically most common or usual situation. That is, it might be suggested that most often, although admittedly not always, when people exert strong executive/cognitive control they are reasoning in a utilitarian ways and conversely when people engage in

[10] Another way of putting this point is that the dilemmas employed by Greene are those in which the "utilitarian" choices require actions which for most people require suppression of some strong adverse reaction to the course chosen. The issue us whether neural activity observed in connection with such choices reflect this fact, rather than distinctively utilitarian content of the choices. For some similar concerns, see Kahane (2012).

utilitarian reasoning, they tend to make use of areas involved in such control. Because such involvement tends to improve the normative quality of decision- making, decisions based on utilitarian reasoning may be expected to be normatively superior.

This leads to another set of issues. As is perhaps obvious, I have been using "utilitarian decision-making" in a way that is ambiguous: it might mean (4.1) "decision coinciding with the recommendations of correctly applied utilitarian principles" or it might mean (4.2) "decision made via deliberate explicit utilitarian calculation that attempts to take account of all benefits and costs". The argument associating involvement of VMPFC/OFC and DLPFC with "utilitarian decisions" seems most naturally understood as an argument that associates these areas with effortful, explicit utilitarian deliberation about costs and benefits— with decisions that are utilitarian in sense (4.2) above. It is an old idea that if one wishes to produce outcomes that are best in the sense of conforming to what is recommended by utilitarian moral theory (decisions that are utilitarian in sense 4.1 above) the best strategy in many cases may not be to try to engage in explicit utilitarian calculation (sense 4.2). It may be that Greene is assuming the opposite—that judgments involving explicit utilitarian calculation are likely to produce superior outcomes by utilitarian standards. Perhaps this is right, but it is not obviously right, even if one is a utilitarian.

There is another point to be made about the notion of "utilitarian judgment" in sense 4.1 above: The judgments described as utilitarian in the empirical literature on moral decision-making reflect a very particular (and arguably controversial) conception of what utilitarianism requires. Roughly speaking, this conception involves a commitment to what I have elsewhere called "parametric" as opposed to "strategic" utilitarianism (Woodward and Allman, 2007). The difference between these two conceptions may be illustrated by Williams' example in which Jim, an explorer in the jungle, is told by Pedro that he will shoot ten villagers unless Jim shoots one (in my version of the story I stipulate that this one is distinct from the ten.) One sort of utilitarian analysis (the parametric sort) takes this to a very simple decision problem. There are two possibilities: (i) One person will die if Jim shoots, (ii) ten if he does not, the consequences under (i) are better than the consequences under (ii), therefore utilitarianism recommends Jim should shoot. With this understanding of utilitarianism, a subject who judges that Jim should not shoot is regarded as making a "deontological" judgment.

I described this utilitarian analysis as "parametric" because it takes the decision problem faced by Jim to have a very simple transparent structure characterized by a few fixed and stable parameters that, moreover, Jim knows for certain: it is assumed, for example, that Jim can take Pedro at his word, so that the relevant probabilities are all either zero or one, and that the only relevant considerations are the number of lives saved under each course of action. In effect, the parametric utilitarian treats the dilemma Jim faces as a simple arithmetic problem. A more "strategic" utilitarian analysis would take into account considerations such as the following: What is the probability that Pedro will act as he claims, rather than, say, killing the ten after Jim kills the one? Is there even a well-defined, knowable probability here or is situation faced by Jim characterized by extreme uncertainty, in the sense described in section 2, with the result that there is no basis for any expected utility calculation of the sort utilitarians advocate performing? What protections do the villagers have against Pedro after Jim kills the one and leaves,

with Pedro remaining? More generally, what are Pedro's plans and intentions in making this offer? Does he see some advantage in involving Jim in the killing? Does he intend some form of blackmail? Discrediting of Jim so that he will not be a credible witness? Would Jim be more effective in saving lives if he were to refuse to participate in the killing and go on to publicize Pedro's behavior to the outside world? What other indirect effects might follow from Jim's participation (or not) in the killing?

When these additional considerations are taken into account, it is no longer so obvious that the action recommended by "utilitarianism" is for Jim to the kill the one, rather than refusing, as at least some versions of deontology would require. In any case, I take it that it is the choice conforming to the recommendations of the strategic rather than the parametric analysis that is required by utilitarianism, since a consistent utilitarian must take into account all available information about the expected effects of his choices, as a properly peformed strategic analysis will do.

Similar points can be made about many of the other standard "utilitarianism versus deontology" dilemmas in the philosophical literature—pushing the big guy in front of the trolley, using the organs of one patient to save the lives of five others and so on. In of these cases, a strategic form of utilitarianism might recommend the same course of action as standard versions of "deontology".

Several consequences follow from this observation. First, it is not obvious that judgments regarding such dilemmas are measures of whether subjects are "utilitarian" as opposed to "deontological", rather than measures of whether subjects are parametric as opposed to strategic utilitarians. Second, consider the information that strategic as opposed to parametric utilitarians consider. This includes (4.3) theory of mind information about the intentions, beliefs, desires and plans of the various actors in a situation, (4.4) related to these, dynamic or strategic considerations about how the situation may evolve over time (how Pedro will respond to Jim's choice, what choices will be available to Jim in the light of that response), and (4.5) information about probabilities and the presence of uncertainty. As noted above, so-called emotional areas are heavily involved in processing all of these kinds of information. Since, as argued above, utilitarians *ought* to take such information into account, this provides further reasons for thinking that, given a commitment to utilitarianism, the involvement of such areas in moral judgment and decision-making can sometimes lead to normatively superior results. It also provides additional reasons to be skeptical of the claim that it is a mark of utilitarian decision-making per se that it does not involve such areas or involves them only minimally[11].

[11] It seems to me that a similar conclusion also follows for most plausible non-utilitarian moral theories. This is because most plausible moral theories will require decision-makers to be guided by accurate information about how their actions affect others, how others are likely to respond to one's choices and so on. (Of course different theories may disagree about how such information is relevant.) As long as it is acknowledged that such information is often morally relevant in some way, it will be normatively better to take such information into account rather than ignoring it. This in turn requires the involvement of neural areas involved in processing such information, including "emotional" areas.

## 5. Moral Rationalism

There is more that might be said about this line of thought but rather than succumbing to the temptation for further exploration, I want to step back and use some of the experimental results I have been describing to raise some more general questions about the role of "reason" and "emotion" in moral judgment and decision-making. As I noted in the introduction, a very striking feature of great deal of contemporary moral philosophy (at least to an outsider like me) is its strongly rationalist flavor. For the purposes of this paper, "moral rationalism" can be taken to be the conjunction of two claims: (5.1) moral claims are the sorts of claims that can be true or false, in (as adherents of this doctrine say), "a mind-independent way" and (5.2) true moral claims can be "grasped" or recognized as such just through the operation of "reason"—this recognition does not require "emotional" processing. In this respect, the recognition of such truths is, according to moral rationalists, very much like the recognition of mathematical truth. Just as (it is supposed) emotion or affect has no role (either causally in learning or in justification) to play in judgments about mathematical truths, similarly for moral judgment

Many contemporary deontolgists and many contemporary utilitarians seem committed to moral rationalism: both groups favor analogies between, on the one hand, moral judgment and the processes leading to moral judgment and, on the other, judgments about mathematical truth and the processes underlying the recognition of such truths. Such analogies occur throughout Parfit, 2011 and in the work of deontologists like Kamm, who writes as follows about intuitive responses to hypothetical moral dilemmas:

> 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

---

Let me add that some readers may find it tempting to take this consideration to vindicate the suggestion that greater involvement of "cognition" leads to better moral decisions on the grounds that taking into account more rather than less information automatically amounts to a greater involvement of cognition. I regard this as not so much wrong as unilluminating: on this suggestion, all of the emotional processing areas discussed in this essay are doing "cognitive" processing simply in virtue of doing information-processing. "Cognition" has been *defined* in such a way that it no longer contrasts with emotion and indeed so that pretty much anything the brain does is "cognitive". The view vindicated is thus not the RD view or the views expressed in Greene, 2004.

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

I will have nothing directly to say about the rationalist claim (5.1) concerning the truth-aptness of moral claims, but I want to suggest that that the rationalist claim (5.2) about the processes involved in moral judgment does not fit very well with the empirical facts described in this essay. Think of the Hare et al. imaging study of dieters. The picture of valuation and decision-making that emerges from this and other studies described above might loosely be described as having both Kantian and Humean aspects. Following Kant, it appears there is an aspect of the self (involving the DLPFC and other "cognitive" structures) associated with our reasoning abilities that can stand back from, assess, and attempt to influence more immediate desires (e.g., for tasty food). We can have such desires for immediate sensory reward and yet fail to endorse them or act on them. We can also use structures like the DLPFC to alter our evaluation of immediate rewards relative to more distant goals like health. Contrary to Hume's famous remark that "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them", if we associate "reason" with the DLPFC, it looks as though reason is *not* confined to merely assessing various strategies for realizing goals given to us by a preference or valuational structure that is fixed and independent of our reasoning abilities. Instead, reason can influence or modulate that structure. Yet at the same time— here the underlying picture looks more Humean— the way in which "reason" exerts this influence is by modulating an affect-laden value signal that reflects input from other sources (including "emotional" areas) outside of DLPFC. (If it is part of Humeanism that something affective or connotative must be present for evaluative judgment or decision or action to occur, and the presence of a value signal in VMPFC/OFC corresponds to this affective element, then this strand of Humeanism seems vindicated.) As noted above, at least in the experiments described in Hare et al. 2009, 2011, the DLPFC does not have its own value signal which somehow supplants or replaces the signal in VMPFC, so that DLPFC generates, as it were, purely reason – based valuations. Instead DLPFC modulates evaluative signals that are influenced by other, "emotional" input as well. Moreover even if, as some claim, DLPFC generates its own value signal, it remains the case that this signal is apparently integrated with value signals from other neural structures that are known to be involved in "emotional" processing.

Suppose we understand the notion of "reason" narrowly, as having to do with the sorts of abilities involved in logical or mathematical reasoning and "moral rationalism" to be the view that considerations supplied by reason, so understood, can by themselves generate moral requirements. Then, as far as the experimental results described above go, they do not seem to provide support for a picture according to which claim (5.2) (above) in moral rationalism describes the process by which people typically come to hold the moral evaluations that they do. That is, when people make moral judgments or hold moral values, these are, as an empirical matter, not typically generated by or grounded in just their reasoning capacities, acting, so to speak, on their own. Instead, other structures which are affective or emotional also play central causal roles in moral judgment and valuing, just as they do in other sorts of valuation. As far as their casual genesis goes, in normal cases, involving intact brains, our values result from the

interaction and integration of the output of these structures with the influence of other structures that we tend to think of as more cognitive or reason-based.

One possible response to this is to concede these causal claims, but to hold that they do not in any way undermine moral rationalism. Recall that, as formulated above, moral rationalism, holds only that moral truths "can be" grasped as such by reason, not that this is the way they are typically recognized. Thus (it might be argued), the acceptability of moral rationalism turns on whether moral requirements, correct moral judgments and evaluations, and so on, are deriv*able* from considerations supplied by reason, understood as above. It may be true, as an empirical matter, that most people are not caused to make the moral judgments they make solely as result of their contemplation of considerations rooted in reason, construed narrowly, but as long as such judgments, insofar as they are correct, follow from such considerations, that is enough to vindicate moral rationalism.

An obvious problem with this response is that the different theorists who claim to be able to ground moral requirements in derivations from principles supported by reason alone reach very different, indeed inconsistent conclusions about those requirements. Some hold that some form of utilitarianism is uniquely favored by reason, others that one or another form of deontology is. Each purported derivation convinces, at best, a minority of discussants. Everyone else regards the derivation as unsound. This is very unlike the situation that holds for valid arguments in logic and mathematics.

The ideas advanced in this paper provide a simple and plausible explanation for this state of affairs. Reason, narrowly conceived, is too weak—too lacking in substantive content—to supply sound derivations of substantive moral requirements of the sort required. This is why no one has been able to produce them. The additional processing that is at work in generating the moral judgments we make is supplied by the "emotional" structures discussed in this essay. When we leave these out of the picture, not enough remains (just in "reason" itself) to generate, either causally or as a matter of logic, the content of our moral judgments.

For some this will be bleak conclusion. That we possess the emotional/evaluative processing structures we do is obviously a contingent matter—these structures are shaped by natural selection and, as I have emphasized, are apparently in some respects specific to human beings. So what seems to follow is something like this: moral and other values are values *for us* or values in relation to the sort of creatures we are, rather than values that are "objective" in the sense of being values for all rational creatures or mattering independently of the emotional processing to which human beings as a species are susceptible.

For both reasons of space and competence, I will not try to discuss this conclusion in any detail, confining myself to just a few remarks. First, and most fundamentally, one might wonder why "ethics for us" is not ethics enough (for us, that is). Our emotional/evaluative capacities are, as I have emphasized, plastic and influenced by learning, but it is not as though we can collectively simply excise them or replace them with something totally different. Whatever contingency in moral evaluation is present simply as a result of our possession of these structures may not reflect alternative possibilities actually available to neurologically normal subjects. So why should we be concerned about them? Even if it is true that highly intelligent social insects or chimpanzees would find very different moral requirements appealing (see below for more

on this possibility), why should this matter to human beings, who cannot simply assume the affective and motivational capacities of these creatures?

Second, a number of influential moral and political philosophers are sympathetic to the idea that the derivation of substantive moral conclusions requires a notion of reason or reasoning that is thicker or more contentful than the thin notion described above. For example, one might think of the notion of what is "reasonable" as this occurs in Rawls and Scanlon as embodying such a thicker notion. From the perspective adopted in this essay, one might attempt to flesh out the additional content possessed by this thicker notion of the "reasonable" in terms of the role played by emotional and affective structures and their interaction with other more cognitive structures. It is this additional content that would allow us to say such things as: "although it may not be contrary to reason, narrowly conceived, to prefer the destruction of the rest of the world to the pricking of my finger, it is certainly unreasonable to have such a preference, and others could reasonably reject my acting on such a preference". Here part of the content of "reasonable" would be supplied by the pro-social emotions and affect that are the output of structures like VMPFC/OFC etc. One consequence of making use of this thicker, more affect-laden notion of reason in moral argument may be that one is also forced to recognize that our possession of it is the result of various biological contingencies. However, wishing that matters were otherwise, is not an argument that they *are* otherwise.

As I noted above, the suggestion that our moral commitments reflect features of our emotional processing is presumably troubling to some in part because it seems to raise the possibility of a group of creatures who are as rational and intelligent as human beings with respect to activities like logic and mathematics but very different emotionally. For example, we may be tempted to imagine a population of primates with the emotional/affective/motivational capacities of chimpanzees who are nonetheless as intelligent as humans. Although "rational", they may be (because of their emotional capacities and the valuations to which this leads), far less cooperative than we are, much more hierarchical, disinclined to treat one another as equals except when it is in their self-interest to do so and so on. This seems to raise the troubling question of whether, in view of these differences, we should think of this population as subject to very different moral requirements from those which we think govern human populations. We might also find it tempting to ask whether, in view of their shared rationality, this population would discover and come to regard as binding a set of moral requirements similar to those we endorse, despite their affective differences, as some optimistic moral rationalists might suppose. (And we might also ask whether they would be subject to these requirements even if, because of their affective differences, they were rarely motivated to conform to them.)

I think this thought experiment rests on a very problematic assumption-- that cognitive/rational abilities can evolve independently of emotional/affective/motivational capacities. In my view, we are as smart as we are at logic and related activities in part *because* of the emotional/affective differences that separate us from chimps and that make possible for us forms of social life and interaction that are unavailable to chimps. And those different emotional capacities would not have been useful to us or favored by natural selection if we were not also able to use them in conjunction with cognitive capacities that differentiate us from chimps. In short, there is every reason to suppose that

the emotional processing subserved by structures like the VMPFC/OFC co-evolve with capacities associated with structures like DLPFC. I certainly don't claim that it follows that creatures with cognitive capacities like ours would automatically also possess emotional/affective processing just like ours, but I do claim that such creatures are unlikely to have the emotional lives of chimps, honeybees or, for that matter, human psychopaths—such creatures will instead possess forms or emotional processing or valuation that lead to some degree of cooperation and concern for others[12].

**6. Internalism versus Externalism**

Let me now turn to some additional issues having to do with "internalism". Philosophers like to debate whether moral judgment (or perhaps "sincere" moral judgment) is "intrinsically motivating". If the account I have been defending is on the right track, the following picture seems correct. If we confine attention to neurotypical subjects, then, as we have seen, moral judgment is (at least often or usually) causally influenced by a value signal in OFC /VMPFC—this is true both for subjects who make deontological and subjects who make utilitarian choices. People can of course make moral judgments in circumstances in which there is no opportunity to undertake actions associated with those judgments, but when action or behavior is possible, these value signals also typically causally influence choice of action. Thus, as a matter of empirical fact, in neurotyicals the evaluations that causally influence moral judgment often also causally influence choice of action, when there is opportunity for the latter. This empirical fact apparently holds whether or not it is also true that it is a conceptual or semantic truth that sincere moral judgment *always* motivates or influences action.

Some recent work on internalism and related issues has focused on the implications of various special or atypical populations (psychopaths, patients with VMPFC damage, autistics etc) with compromised processing in emotional areas for these issues. (See, e.g., Roskies, 2003). Suppose, for example, we come to believe that psychopaths can make genuine moral judgments or recognize that some actions are morally wrong without being motivated at all to act in accord with these judgments. Would this show that internalism is mistaken? If, on the other hand, we decide that psychopaths cannot make genuine moral judgments because they do not engage in the emotional processing that typically accompanies this in neurotypicals, does this provide support for internalism or undercut moral rationalism?

I think that my remarks above can help cast some light on these questions. At the risk of oversimplifying, I take the general picture of these special populations that is emerging to be the following. (Cf. Glenn et al., 2009, Young et al. 2010, 2012) Psychopaths, VMPFC patients, and autistics can use sentences expressing moral

[12] A referee asks whether I am claiming that it somehow conveniently works out that our emotional apparatus has been moulded by natural selection in such a way that we make moral judgments that are "correct" by some standard entirely independent of that apparatus. This is not what I am claiming; my talk of "ethics for us" is meant to suggest there is no such independent standard.

judgments like "Stealing is wrong" in ways that show normal linguistic competence. For some range of cases, such subjects will also make judgments that are not very different from the judgments made by neurotyicals—they will agree or at least say they agree that stealing is wrong. Nonetheless, subjects in these special populations tend to differ from judgments made by neurotypicals in more subtle ways, with the nature of the difference depending on the character of the impairments in the different special populations. For example, VMPFC patients have been found to differ from neurotyicals in their judgments of the wrongfulness of actions that involve *attempts* to produce harm and, as noted above, they also tend to make different (more "utilitarian") judgments with respect to certain moral dilemmas such as the trolley problem. Psychopaths judge non-intentional harms more leniently than normals. In contrast to neurotypicals, high functioning autistics are reported to not reliably judge accidental and attempted harms as morally different. To the extent that such subjects have impaired functioning in OFC/VMPFC or in structures like the amygdala that provide input to these, we can conclude (quite independently of any differences in the content of their judgments) that it must be the case that the causal pathways leading to their moral judgments are different from or are functioning differently from those employed by neurotypical subjects. It is presumably these differences that lead to the subtle differences in judgment described above. Thus one possibility – arguably a natural conjecture—is that although subjects in special populations are sometimes able to produce moral judgments that are similar to those produced by neurotypicals, they do this by making use of processing that is quite different from that employed by neurotypicals. (See Glenn et al. 2009 for a similar suggestion). For example, someone with impaired or abnormal processing in structures like OFC/VMPFC might nonetheless have been able to learn a set of moral rules or maxims that are generally accepted in her society—that it is wrong to steal, kill and so on. These might be committed to memory and then be deployed to answer moral judgment questions through the activity of structures like DLPFC without any involvement of emotional structures. This may result in judgments that exhibit considerable overlap with those made in more normal ways by neurotypicals for some range of cases, while differing in more subtle ways in tasks that require deployment of structures like OFC/VMPFC (such as trolley problems). To the extent that such subjects make moral judgments without making use of structures like OFC/VMPFC, it seems entirely possible that there may be no associated motivation to act or at least a weaker or different motivation than in neurotypicals.

Philosophers who are sympathetic to the idea that moral judgments must, for conceptual or semantic reasons, be motivating or involve emotional processing are often tempted to suppose that moral judgments of psychopaths and other special population subjects are not "genuine" or "sincere" moral judgments because they lack the elements of normal processing just described. This move strikes me as question-begging (or at least unfruitful) in the absence of some independent characterization of what makes a moral judgment real or sincere. On the other side, philosophers sympathetic to rationalism and externalism sometimes take the apparent possibility of moral judgment in special populations to show that there is no intrinsic or essential connection between moral judgment and motivation. Whatever one thinks of this last claim, it is important that it not lead us to overlook the causal connection between emotional processing, valuation, judgment, and motivation that *is* usually present in neurotypicals. In other

words, when construed not as a universally true conceptual/semantic claim, but rather as an empirical claim about what is usual in neurotypical populations, internalism has much to recommend it. Moreover, if the conjectures advanced above are correct, it also seems to follow that the moral judgments endorsed by special populations are to a substantial extent parasitic on the judgments of neurotypicals in the sense that those in the special populations learn to make these judgments by mimicking those of surrounding neurotypicals. If this is correct, there is no reason to suppose a population consisting entirely of psychopaths without any contact with neurotyicals would acquire on its own tendencies to moral judgment that even loosely resemble those of neurotypicals.

So far I have neglected an issue that may seem central to debates about internalism and moral rationalism. This has to with the significance of differences, not between human beings and other species, or between neurotypicals and special populations, but *among* neurotypicals in emotional processing and evaluative assessment. Both casual observation and more careful empirical observation support the claim that neurotypical humans vary considerably in emotional/evaluative response, particularly in the moral sphere. Some people are very empathetic and find helping others rewarding (at least for some others), other people much less so. Does it follow that, e.g., the extent to which moral requirements are binding on people (or the extent to which they have "reasons" to conform to such requirements) depends on the extent to which their affective processing is such that they value acting in conformity to these requirements?

In my view, nothing I have said above requires a positive answer to this question. Although the matter deserves more discussion than I can give it here, my view, very roughly, is that it is generic facts about human emotional processing and valuation, characteristic of us as a species, rather than individual differences, that are regulative when it comes to moral requirements and reasons. For example, in many situations in which outcomes would be improved by cooperation, very substantial numbers of people (although far from all) find cooperation and reciprocation rewarding and non-reciprocation sufficiently adverse that they are willing to punish free-riders. When cooperation would provide goods (evaluated as such by nearly all those affected) that would not otherwise be provided, these facts about the emotional response and evaluation underlie our willingness to require cooperation from those who do not find cooperation rewarding in itself – a requirement that we may be willing to enforce with sanctions. Moral requirements have to do, after all, with regulating our common lives together and there are very obvious considerations (having to do, e.g. with incentives) for not allowing people to evade those requirements simply because they say, even truthfully, that their preferences and emotions are such that they do not value conforming to them. So while some particular person, Jones, can be subject to a requirement (and have a reason) to help others even if he does not feel like doing so, in my view it would be a mistake to conclude from this observation that recognition of this moral requirement is completely independent of *any* emotionally mediated evaluation that is widely shared by human beings.

## 7. Conclusion

I have argued that areas commonly identified as involved in emotional processing play a central role in moral judgment in neurotypical human beings because such areas

contribute causally to the construction of moral evaluations. The moral judgments that humans find appealing or compelling or intuitive reflect features of the operation of such areas, and this is so whether the judgments in question are "utilitarian" or "deontological". It is because of the involvement of these areas in moral judgment that attempts to derive moral requirements from "reason", conceived as something totally distinct from anything affective and just involving the kind of processing that is operative when humans work on logic and mathematics, are likely to be unsuccessful. The way forward is not to look for moral requirements that are binding on all rational creatures just in virtue of their rationality but rather to recognize that moral requirements that we find appealing and that are suitable for regulating our lives will reflect facts about our affective and motivational commitments, as well as our capacities for reasoning.

References


Allman, J., Watson, K. , Tetreault, N., Hakeem, A. 2005. "Intuition and autism: a possible role for Von Economo neurons". *Trends in Cognitive Science.* 9, 367-373.

Allman, J., Tetreault, N. , Hakeem, A. , Manaye, K. , Semendeferi, K. , Erwin, J. Park, S. Goubert, V. Hof, P. (2010) " The von Economo neurons in frontoinsular and anterior cingulate cortex in great apes and humans" *Brain Structure and Function* 214:495–517.

Anderson, S., 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.

Blair, J. Mitchell, D. , Blair, K. (2005) *The Psychopath: Emotion and the Brain*. Oxford: Blackwell.

Decety, J., Michalska, K. and Kinzler, K. (2012) "The Contribution of Emotion and Cognition to Moral Sensitivity: A Neurodevelopmental Study" *Cerebral Cortex* 22: 209-20.

Escobedo, J. and Adophs, R. (2010)"Becoming a better person: temporal remoteness biases autobiographical memories for moral events" *Emotion* 10: 511-18.

Glenn, A, Raine, A., Schug, R., Young, L. and Hauser, M. (2009) "Increased DLPFC activity during moral decision- making in psychopathy" *Molecular Psychiatry* 14: 909–911.

Greene, J. Nystrom, L., Engell, A., Darley, J. , Cohen, J. (2004) "The neural bases of cognitive conflict and control in moral judgment". *Neuron*, Vol. 44, 389-400.

Hamann, K., Warneken, F., Greenberg, J., & Tomasello, M. (2011). "Collaboration encourages equal sharing in children but not chimpanzees". *Nature, 476,* 328-331.

Hare, T., Camerer, C. and Rangel, A. (2009) "Self-control in decision-making involves modulation of the vmPFC valuation system" *Science* 324:646-648.

Hare, T., Malmaud, J. Rangel, A. (2011) "Focusing attention on the health aspects of food changes value signals in the vmPFC and improves dietary choice" *Journal of Neuroscience* 31:11077-11087.

Hare, T., Camerer, C. , Knoepfle, D. , O'Doherty, J. and Rangel, A. (2010) "Value Computations in Ventral Medial Prefrontal Cortex during Charitable Decision Making Incorporate Input from Regions Involved in Social Cognition" *The Journal of Neuroscience* 30 583–590

Hoffman, M. (2001) *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge: Cambridge University Press.

Kahane, G. (2012) "On the Wrong Track: Process and Content in Moral Psychology" *Mind and Language* 27: 519-45.

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

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). "Damage to the prefrontal cortex increases utilitarian moral judgements". *Nature, 446* (7138), 908-911.

Maclean, P (1990) *The triune brain in evolution: role in paleocerebral functions*. New York: Plenum Press.

Parfit. D. (2011): *On What Matters* Oxford: Oxford University Press.

Quartz, S. (2009). "Reason, Emotion, and Decision-Making: Risk and Reward Computation with Feeling". *Trends in Cognitive Sciences*. 13(5). 209-215.

Rolls, E. (2005) *Emotion Explained*. Oxford: Oxford University Press.

Roskies, A. (2003) "Are ethical judgments intrinsically motivational? Lessons from acquired sociopathy" *Philosophical Psychology,* 16: 51-66.

Scanlon, T. (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press.

Schoenbaum, G., Roesch, M., Stalnaker, T., and Takahashi, Y. (2011) " Orbital Frontal Cortex and Outcome Expectancies: Optimizing Behavior and Sensory Perception" in J. Gottfried (ed.) *Neurobiology of Sensation and Reward*. Boca Raton, FL: Taylor and Francis, 329-350.

Shenhav, A., Greene, J. (2010) “Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude”. *Neuron*, 67. 667-677.

Singer, P. (2005) “Ethics and Intuitions” *The Journal of Ethics*, 9: 331-352.

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., Damasio, A. (2010). “Damage to ventromedial prefrontal cortex impairs judgment of harmful intent”. *Neuron*, *65*, 845-851.

Young, L. , Koenigs, M. , Kruepke, M., Newman, J. (2012). “Psychopathy increases perceived moral permissibility of accidents”. *Journal of Abnormal Psychology*

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