*Etica & Politica / Ethics & Politics*, XI, 2009, 2, pp. 18–42

# *A Critical Perspective on Moral Neuroscience*

David V. Yokum
Neural Decisions Science Laboratory
Department of Psychology
University of Arizona
dyokum@email.arizona.edu

Filippo Rossi
Neural Decisions Science Laboratory
Department of Psychology
University of Arizona
filippor@email.arizona.edu

**ABSTRACT**

In this paper, we highlight several historical developments in the neuroscience of ethics as well as recent advances that forecast the experimental research to come. We argue, in particular, that our understanding of the moral brain will benefit from the further use of a formal, mathematical approach to the construction and testing of alternative theories, such as that found in the field of neuroeconomics. The use of economic modeling to understand the psychological processes underlying distributional preferences and charitable giving is reviewed to illustrate this potential. We also consider some obstacles to such an approach, notably the challenge of capturing substantive moral values within a mathematical model.

## 0. *Introduction*

The term "neuroethics" references a broad set of issues. It is, as Al Jonsen put it at one of the first conferences to include neuroethics in its title, an "unexplored continent lying between the two populated shores of ethics and of neurosciences."[1]. As we embark into this unchartered territory, it is useful to delineate two general programs of inquiry: (1) the ethics of neuroscience and (2) the neuroscience of ethics (Roskies, 2002). The *ethics of neuroscience* resembles, in part, traditional bioethics. To this extent the issue is how to ensure that neuroscientific research and treatment is carried out in an ethical fashion. Topics such as informed consent, privacy rights,

[1] "Neuroethics: Mapping the Field," hosted by the Dana Foundation on May 13-14, 2002, in San Francisco, CA. Cited in Roskies (2002).

and animal welfare, among others, are addressed.[2] A second aspect, with problems unique to neuroethics, is how to interpret and integrate neuroscientific knowledge within the social and political arena. This includes, on the one hand, the ethical application of neuroscientific technology and, on the other hand, the potential for insights into brain function to alter our conception of how people should act, be treated, or even conceive of their own mental lives. As examples of the former, should neuroimaging evidence be used for forensic purposes (Meegan, 2008; Pettit, 2007) or as part of a job interview (Tovino, 2007)? Should neurosurgery be used to treat antisocial personality disorder or pedophilia (De Ridder, Langguth, Plazier, & Menovsky, 2009)? As for the latter, might the increasingly detailed explanation of the mechanistic underpinnings of moral cognition undermine the general acceptance of free will and moral responsibility (Greene & Cohen, 2004; Morse, 2006)? Or might neuroscientific evidence be used to settle long-standing debates in moral philosophy (Joyce, 2008)?

The *neuroscience of ethics*, in contrast, is the empirical investigation of how the physical brain engenders moral thought and action. What are the cognitive processes and underlying neural processes by which we engage in ethical reasoning or experience moral emotions and intuitions? The moral neuroscientist aims to understand and predict moral behavior as it exists in the real world. There is currently a surging optimism about our ability to make advances in the neuroscience of ethics, and understandably so: technical inventions, such as functional magnetic resonance imaging (fMRI) and neuropharmaceuticals, have substantially increased our ability to examine and manipulate the brain. We will highlight some of the original research exploiting these technologies in the pages to come. Our intention is not, however, to review the literature on moral neuroscience (see Rossi & Yokum, 2009). Rather, we focus on recent empirical advances and make two suggestions for further work. First, an experimental approach that exploits the rigor of mathematical models is needed to orient the empirical investigation of moral cognition. Second, in order for such models to be useful for moral neuroscience, they will have to be enriched to include components other than distributive justice. We begin with a selective historical overview of the neuroscience of ethics. Then, sections two through four discuss the relation between

[2] Indeed, some bioethicists question whether neuroethics is a unique field at all, preferring instead to demote it to a subfield of bioethics. This unduly dismisses the unique aspects of neuroethics to which we are about to turn, in particular the neuroscience of ethics. Nonetheless, the initial skepticism is understandable: there seems to be a recent trend to "neurolize" everything, as in neurolinguistics, neuroeconomics (we admit with a smile, since we are a part of this movement), neuropolitics, neurotheology, neurocriminology, neuromarketing, and even, according to Wikipedia at least, neurocinema.

moral science and economical modeling, at both the behavioral and neuroscientific levels. Our discussion is therefore most relevant for the neuroscience of ethics, and leaves questions within the ethics of neuroscience to be addressed another day.

## 1. *A historical perspective*

Despite the novelty of the technology now available to neuroethicists, the search for the neural substrate of morality is not a new enterprise.[3] Franz Joseph Gall (1758-1828), of phrenology fame, measured the skulls of "eminent benevolent people conspicuous for their very great philanthropy" to reach the conclusion that the organs of benevolence and conscientiousness lay within the middle of the frontal lobe, on either side of the longitudinal fissure. Johann Spurzheim (1776-1832), with a new sample of generous citizens and remorseless criminals, left benevolence in its place but resettled conscientiousness within the parietal lobe.

As phrenology was fading, Viennese neurologist Moritz Benedikt (1835-1920) argued that morality was housed in the occipital lobes. He conceived of the moral sense as, quite literally, a sensory organ, one that allows a person to perceive moral rightness and wrongness as one does visual stimuli. Familiar with the role of the occipital lobes in visual perception, he reasoned that the moral organ would reside there as well. Benedikt complemented this theoretical argument with pieces of pseudoscientific observation. He noted that the brains of several guillotined criminals had relatively retracted occipital lobes that did not cover the cerebellum as usual; this was similar to what is found in gorillas, and fit squarely with the idea that criminals, like lesser species, did not possess a fully developed human brain capable of keeping their animalistic urges in check.

Oskar Vogt (1870-1959), a well-respected scientist and pioneer in neurohistology, examined the cellular architecture of cortical layers and the relationship of such microscopic features with higher-order cognitive traits. His most famous case occurred in 1924, when Russian authorities recruited him to examine the brain of the recently deceased soviet leader in an attempt to explain "the political and moral genius." Vogt noticed that Lenin had an enlarged lamina pyramidalis (a region within the frontal lobe) interposed with abundant association cells. In contrast, this area was narrowed, with sparse association cells, in the brains of several executed criminals that he had examined around the same time. From these observations, Vogt concluded that a properly developed lamina pyramidalis was a necessary underpinning for moral cognition. Notably, Vogt did not exactly conceive of this

[3] The following historical descriptions draw heavily from Verplaetse (2009) and Verplaetse, Braeckman, & De Schrijver (2009).

area as *the* seat of morality, for he believed that morality was not a singular mental phenomenon in the first place. He argued instead that morality resulted from the successful integration of assorted mental representations and their linkage to appropriate emotional responses. Although not fleshed out in the details, Vogt's general idea is remarkably similar to modern theories within the neuroscience of ethics.

These examples are only a small sample from the long-lasting scientific pursuit aimed at understanding the neural substrates of morality. Neuroethics might be a fledgling field of study, but this is more a matter of nomenclature than substance: the neuroscience of ethics has been an interest of scientists for some time. As such, the absence of a compelling theory about how the moral brain works cannot be explained away as a lack of attention or effort. A second lesson is that historical localizations of morality, even when more or less guided by empirical observation, turned out to be simplistic and premature. It is perhaps fitting, then, to approach the neuroscience of ethics with a certain degree of humility and a heightened caution when integrating neuroscientific findings with social and political thought. This is particularly warranted given the public fascination with and quick acceptance of neuroimaging evidence (Racine, Bar-Ilan, & Illes, 2005).

This prudential advice does not mean we should be pessimistic regarding meaningful results. Quite to the contrary, there are empirical tools at our disposal that scientists a century ago could not have even imagined. There are multiple advanced methodological techniques, notably fMRI and transcranial magnetic stimulation (TMS), which allow scientists to examine and even manipulate, in an ethically sound way, the function of living human brains.[4] MRI works, essentially, by using an oscillating electromagnetic field to push hydrogen nuclei, found in abundance throughout the body, out of alignment with a strong magnetic field. When the protons snap back into place, they release a detectable radiofrequency signal. The speed of realignment varies depending upon the surrounding tissue properties, and thus the signal can be used to distinguish different tissue types. Functional MRI exploits the fact that deoxygenated hemoglobin has paramagnetic properties and, as a result, distorts the magnetic field locally. This distortion affects the speed of proton realignment and, in turn, the detected signal (referred to now as the BOLD, or blood-oxygenation-level-dependent, signal). The linkage with cognitive function is that, as neuronal firing increases, the increased metabolic demand is satisfied by a rush of blood into the tissue region, thereby changing the concentration of deoxygenated blood and resulting BOLD signal (see Huettel, Song,

[4] Other important technologies for human neurophysiology include electroencephalography (EEG), positron emission tomography (PET), magnetoencephalography (MEG), transcranial direct current stimulation (tDCS), and diffuse optical imaging.

& McCarthy, 2008).[5] Although important research with TMS is beginning to escalate, results from fMRI dominate the base of empirical evidence currently referenced within the neuroscience of ethics.

An important realization to emerge from neuroimaging work – and admittedly it should also be evident from neurological cases over the last decades – is that there is no unique moral organ or module (Casebeer, 2003; Greene & Haidt, 2002; Lieberman, 2007; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005). There is no known lesion that obliterates only moral cognition, nor is there any known brain region that is metabolically active solely during moral thinking. There are, however, a relatively consistent set of brain areas that become engaged when subjects, for example, make judgments about social dilemmas (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), view morally valenced pictures (Moll, et al., 2002), or make decisions within economic games designed to mimic morally relevant social interactions (Glimcher, Camerer, Fehr, & Poldrack, 2009). Untangling these complex results to generate a coherent theory of when and how various neural regions interact to generate the multitude of behaviors we label as moral is the daunting and exciting task facing scientists interested in the neuroscience of ethics.

To appreciate one trajectory along which much moral research is currently aligned, as well as provide traction for further discussion, let us consider one more example from the history of moral psychology, this time from its recent history. Joshua Greene and colleagues (Greene et al., 2001) published one of the first neuroimaging studies explicitly dedicated to moral cognition. They scanned subjects while they read a series of written moral and non-moral vignettes and, for each one, judged which of two courses of action was preferable. The stimuli included, among others, classic dilemmas within philosophy, such as the so-called trolley problem (Foot, 1978; Thomson, 1976), that present a conflict between maximzing the number of lives saved and not directly harming an otherwise safe person. For example, in the *footbridge* version of the trolley problem, the subject must decide to either let five railway workmen be killed or, alternatively, stop the trolley by pushing a man in front of it, thereby killing him but saving the original five. In the *switch* version, the subject must again decide whether to let five die or kill one to save the five, but in this scenario one pulls a switch to redirect the trolley toward the man, rather than pushing him into it. Behaviorally, subjects overwhelming approve of flipping the switch (typically about 90% approve) but, interestingly, strongly disapprove of pushing the man (only about 10% approve).

[5] Rather counter-intuitively, metabolically active neural tissue actually has *less* deoxygenated hemoglobin. It as if the circulatory system cannot anticipate how much oxygen will be needed, and errs on the side of safety by sending a surplus.

The neural evidence indicated that the minority of subjects who made judgments labeled by the reseachers as characteristically utilitarian (such as pushing the man) had relatively more activation in brain regions correlated, in previous empirical work, with working memory, such as the dorsolateral prefrontal cortex (BA 9/10/46) and parietal cortex (BA 7/40).[6] The other subjects, in contrast, had relatively more activation in brain regions correlated with social and emotional processing, such as the medial frontal gyrus (BA 9/10), posterior cingulate gyrus (BA 31), and bilateral superior temporal sulcus (BA 39).

## 2. *A problem of vagueness: moving beyond simple dichotomies*

Philosophers have devised a veritable universe of moral theories over the centuries, each with unique justifications (see Rawls, 2000) and, often if only implicitly, assumptions about human nature. An interesting empirical question is whether people actually possess beliefs and values that map onto any of these theories, and whether they possess the requisite emotions and rational capacities to satisfy any of their dictates. We will turn to this issue in a moment.

But note first that it is a further question whether people are consciously aware of, or could articulate when asked, the justificatory basis motivating their moral judgments and decisions. There is compelling evidence that in many situations (but not all) people are unable to provide a coherent justification for their behavior (Cushman, Young, & Hauser, 2006; Haidt, 2001). Philosophers, however, do not generally hinge the validity of their theories on whether laypersons consciously accept its tenets, even when making claims about how their theories are instantiated in the real behavior of moral people. People could act *as if* they consciously accepted a theory without actually doing so. From a psychological viewpoint, this could be explained, for example, as evolved innate behaviors (e.g., Sober & Wilson, 1998) or any variety of unconscious, inaccessible processes (e.g., Mikhail, 2007). More to the philosophical point though, laypersons might just be wrong about their initial moral beliefs, values, and justifications. Perhaps they

[6] Korbinian Brodmann, a German neurologist and colleague of Oskar Vogt at the Institute for Brain research in Berlin, proposed a system for classifying cortical areas according to their cytoarchitecture. The resulting numerical nomenclature, which is widely used today, divides the neocortex of each hemisphere into 44 Brodmann areas (BA). Despite the precision of the naming system, however, the areas actually grade into each other by degrees; moreover, the correlation of function with specific anatomical areas is not nearly as precise as originally hoped. One consequence is that authors occasionally have subtle disagreements about which area, exactly, is actually engaged during a task.

need to be taught a moral philosophy. There is a reason, after all, that philosophers endure five to seven years of graduate pay.

Nonetheless, returning to the first empirical question, a first issue is to explain what we will refer to as *moral preferences*. "Preferences" here is used in an economic sense, namely, an ordered relationship between alternatives. A person is said to prefer X over Z if, when presented a choice, he or she behaviorally chooses X rather than Z. A simple descriptive claim about a preference makes no claims about the underlying psychological processes or, for the philosopher, the underlying philosophical principles dictating the preference. It is just that we have observed someone choosing X over Z, and thus make the reasonable assumption that they somehow prefer, for whatever reason, X. Armed with a sufficient variety of observed preferences, theories can be built that predict the future behavior of a person. What is important to realize, however, is that it is possible for a theory of preference, moral or otherwise, to be predictive without accurately explaining the causal mechanism. Economists have long acknowledged this fact (Friedman, 1953), satisfying themselves with the substantial progress that can be made with a focus on predicting behavior irrespective of precise mechanism.

Of course, as scientists, we are also driven to explain the causal mechanism. The work of Greene et al. outlined above (2001; also Greene et al., 2004), as well as other paradigms using moral judgments about written vignettes, fits somewhere between a descriptive account of moral preferences and an attempt at causal explanation. The general strategy was to elicit moral preferences within a variety of dilemmas that were categorized as either personal or impersonal.[7] Many of the dilemmas contained options that fit into one of two philosophical traditions: utilitarianism or deontology. In a notable oversimplification, the utilitarian option was typically operationalized as whichever option results in the most lives saved, while the other option was labeled, by default, as deontological. But not all dilemmas from the standard stimuli set fit cleanly into these categories. For example, options such as cheating on a tax form or hiring a man to rape your wife so she will love you again are clearly not advocated by either utilitarian nor deontological philosophies, and yet such options usually get clumped into the utilitarian category nonetheless, presumably because of a conflation of utilitarianism with selfishness. Regardless, the important results were, on the one hand, the observation of relatively increased activity in brain areas correlated with working memory when subjects reached a utilitarian judgment and, on the other hand, relatively increased activity in brain areas correlated with emotional processing when subjects opted for the labeled deontological option.

[7] There was also an important contrast between moral and non-moral judgments, but this is not relevant for our purposes here.

The interpretive claim that Greene et al. (2001) made next, and which, for better or worse, moves us toward a causal claim, was that these results support a dual-process theory of moral judgment. Within this framework, it is hypothesized that cost-benefit analyses, which underlie characteristically utilitarian judgments, are computed by one cognitive mechanism, whereas prepotent, emotional responses, which underlie characteristically deontological judgments, are computed by a separate cognitive mechanism. The outputs of these two systems, when in conflict, undergo a competition to determine which will drive the overall judgment (Kahneman, 2003).[8] If the prepotent, emotional response is sufficient, it will override the cost-benefit calculation and a deontological judgment will ensue. Assuming that metabolic activity in areas such as the dorsolateral prefrontal cortex and medial frontal gyrus can be taken to indicate, respectively, calculative reasoning and emotion – a not trivial assumption – then the neural evidence is corroborative.

Needless to say, the dual-process interpretation has been controversial, both on empirical and conceptual grounds (Connolly & Hardman, in press; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005; McGuire, Langdon, Coltheart, & Mackenzie, 2009; Mikhail, 2008; see Greene, 2008, for an updated and more elegant version of the theory). This has proven to be a lively and informative debate, one worth examining if unfamiliar. However, one potentially problematic tactic used throughout the literature has been the overreliance, by authors from all sides of the issue, on reverse inference, that is, inference from an observed pattern of brain activity to the conclusion that a specific mental process, such as working memory or emotion, is engaged.

How to interpret neuroimaging evidence in relation to mental function is not, of course, a challenge unique to the cognitive neuroscience of morality; it confronts any scientist examining raw brain activation. The simple fact is that we cannot readily read off cognitive process from brain activity. It is possible, however, to be more or less confident about the reverse inference being made. Poldrack (2006) discusses how our confidence in a given reverse inference can be estimated in a Bayesian framework. In particular, the probability that a cognitive process X is occurring, given activation of brain area(s) Z, is equal to the likelihood of observing

[8] Greenians sometimes discuss this dichotomy between "emotion" and "cognition" in terms of the larger system 1/system 2 literature (Chaiken and Trope, 1999). This terminology is confusing, and the comparision with systems 1 and 2 is misleading. An intuitive system 1 output could be non-emotional, and likewise a deliberative system 2 output could generate an emotion. Nonetheless, once one backs down from the emotion versus cognition language (as has been done; see Cushman et al., in press; Greene, 2008), the notion of multiple systems remains an important empirical idea (see also Sanfey & Chang, 2008).

brain activation in Z given the presence of a cognitive process X, multiplied by the prior probability of observing that cognitive process, and then scaled by the independent probability of finding Z active. One upshot is that the more specific a hypothesized cognitive process, the lower its prior probability will be and, therefore, if the posterior is still high, the more confident we can be in making a reverse inference when activity is observed in the predicted region (such hypotheses are also more risky in that they are more easily falsified in the Popperian sense).

A shortcoming of reverse inferences made to date regarding moral values relates to the vagueness of the hypotheses. The problem, to be more specific, is that precise psychobiological hypotheses simply do not easily fall out of general philosophical theories such as utilitarianism or deontology. Such theories do not make detailed claims about cognitive function and, as such, do not provide enough specificity as to what brain regions are predicted to be active. This is likely to be a problem in general for many philosophical conceptualizations of the right and the wrong, the good and the bad. The consequence is the reverse inferences are systematically plagued with uncertainty, not to mention a degree of vagueness carrying over from the vagueness of the theory itself. We can illustrate this problem by contrasting it with a brief overview of the economic literature on distributional preferences. We review behavioral evidence first, and then turn to neuroscientific evidence; an important issue will be to consider how this work can be integrated with other research programs in moral psychology.

## 3. *The behavioral investigation of moral preferences*

Behavioral economics entails the empirical study of interpersonal, or strategic, behavior using the mathematical tools offered by decision theory and game theory (Osborne & Rubinstein, 1994). Moral behavior relies on basic decision and judgment capacities, not to mention the ability to navigate social interactions, and so not surprisingly many of the topics addressed by behavioral economists are immediately relevant for the neuroscience of ethics as well (Connolly & Hardman, in press). Work on distributional preferences, or the preferences people have about dividing limited resources, clearly overlaps with issues of social justice and, as the name suggests, distributive justice. As will become evident, the formal approach that characterizes economical experimentation distinguishes it from the early work in the neuroscience of ethics.

To begin with, standard game theory starts with the simplifying assumption that people are a species of *Homo economicus*, namely, persons concerned only with maximizing their personal, material payoffs, irrespective of how their behavior

might affect other people. As we have reviewed elsewhere (Yokum & Rossi, 2009), this is an unrealistic assumption. Behavioral evidence clearly indicates that people are sensitive to facts other than their own material payoff. In the ultimatum game, for example, one player, a "proposer," is endowed with a certain amount of money and then asked to share some or none of it with a second player, a "receiver." The receiver, in turn, must decide to either accept the proposed split of money, at which point each player receives the allocated money, or reject it, at which point neither player receives anything. The Nash equilibrium with discrete strategies, or predicted behavior for *Homo economicus*, is that receivers will accept any non-zero offer (since any amount of money is always better than no money) and that proposers, knowing this, will offer the minimum amount possible. Contrary to this prediction, proposers typically offer between 40% and 50% of the endowed amount, while about half of receivers begin to reject offers of less than 20%, with substantially more rejecting as the offer continues to lessen (Camerer, 2003).

What motivations do real people have that *Homo economicus* lacks? A general suggestion is that people perceive unbalanced monetary divisions to be somehow unfair. But what does it mean for something to be unfair? How can we unpack the moral value of fairness? Researchers began to develop models to do precisely this. Candidate models of how distributional preferences might be explained include (see Charness & Rabin, 2001, for a review): inequity aversion, quasi-maximin preferences, competitive preferences, and preferences that are ultimately selfish in one sense or another. The first two are most relevant for our discussion. In both cases, players are interested in their own material payoff, but their utility function is enriched to include a term sensitive to the payoff of other players; they have what are called other-regarding preferences.

If subjects are inequity adverse (Fehr & Schmidt, 1999), their utility will decrease as a function of the distance between the payoffs received by players in the game. For instance, suppose that the proposer has to split \$10 in the Ultimatum Game. His or her utility function would be $U = (10 - X) - \alpha * \max\{((10 - X) - X), 0\}$. What this means, in words, is that provided the proposer offers \$X, he or she will save \$(10 – X) and receive some positive utility from that return; however, that utility will be reduced to some degree by the difference between this personal payoff and the payoff of the other player, that is, (10 – X) – X. The parameter $\alpha$ captures the degree to which the person cares about the inequality. Note that inequity aversion applies regardless of whether the proposer has more and less money than the receiver.[9]

[9] The model proposed by Fehr & Schmidt (1999) is in fact more complex, because it comprises separate terms for whether the subject is above or behind in the game. In particular, if we denote

A quasi-maximin model of preferences (Andreoni & Miller, 2002; Charness & Rabin 2001), in contrast, is based on the theory that players are concerned about their own payoff (as usual), the material payoff to the least well-off in the society, and finally the overall public payoff. In this model inequity *per se* is not necessarily problematic. So long as the payoff to the less well-off increases, or at least does not worsen while the payoffs to others increase, inequities are acceptable.

Inequity aversion and quasi-maximin models have both received empirical support (see Camerer, 2003, for a review), but neither is without problems. Broadly speaking, the models provide impressive predictive power in several specific games, but often perform more poorly when generalized to other situations. For example, Fehr and Schmidt (1999) demonstrated that inequity aversion accurately predicts behavior during an ultimatum game, but Andreoni and Miller (2002) observed a pattern of behavior essentially opposite to that predicted by the model within the closely related dictator game.[10] In the experiment, participants were asked to express eight choices on the allocation of a certain amount of money between themselves and their opponent. The allocations chosen by the subjects were not consistent with the minimization of inequality; for instance, they traded off inequity for efficiency. A second obstacle was pointed out by Charness and Rabin (2001), namely, that certain simple economic games do not provide the means for distinguishing between possible non-selfish motivations. As the authors put it, "the tight fit of these models may merely reflect the fact that in many of the games studied, their prediction happen to be the only way that subjects can depart from self-interest" (p. 818). In the ultimatum game, for example, the receiver might reject a small offer because he or she intrinsically prefers to minimize inequality, but different psychological motivations, such as anger or envy, could produce the same behavioral output. This is particularly important because, in the ultimatum game, the only way in which a receiver can retaliate (and chose a Pareto inefficient strategy) happens to also be consistent with inequity aversion.

A central challenge, for the models above and others not discussed here,[11] is to identify the motivational components underlying choice behavior and then

---

with $\pi_i$ and $\pi_j$ the material payoffs of the two opponents, the utility function of player $i$ is: $U_i(\pi) = \pi_i - \alpha_i \max\{\pi_j - \pi_i, 0\} - \beta_i \max\{\pi_i - \pi_j, 0\}$.

[10] The setup to the dictator game is identical to the ultimatum game, except that the receiver no longer has the option to reject the offer.

[11] We have ignored a different line of modeling, focusing on reciprocity, that is developed within the more general framework of psychological game theory. For two seminal papers on psychological game theory, see Genakopolos et al. (1989) and Battigalli & Dufwenberg (2009). For a development of models of reciprocity within this framework, see Rabin (1993) and Dufwenberg et al. (2004). We discussed some aspects of these models in Yokum & Rossi (2009).

analytically describe the nature and weight of each in the decision-making process. The models to date are, for practical reasons, highly simplified representations of the motivational possibilities, and most of the measures of interest are actually modeled as exogenous parameters. More realistic models will likely require more sophisticated, complex modeling. This is particularly the case for models that aim to capture inter- and intra-individual differences in behavior. It is likely that personal distributional preferences (and any other type of preference, for that matter) shift depending upon the particular decision context. Indeed, the non-existence of a rigid set of ordered preferences might explain some of the difficulty in generalizing any single model of distributional preferences – there might not be such a permanent set, but rather a more flexible array of preferences that rearrange and change across situations. From a modeling perspective, it is tempting at times to introduce more and more parameters in order to mathematically describe the dataset at hand. This might be acceptable if prediction is the sole focus, but such a tactic is clearly problematic for scientists interested in the behavioral mechanisms in operation. The model wanted is one that contains parameters that are theoretically meaningful, for example, by reflecting potentially real cognitive processes. But adding parameters (and hence degrees of freedom) will typically increase the fit of a model regardless of theoretical significance. How to create sufficiently complicated models without succumbing to overparameterization is an important challenge for future work.

Having advocated the benefits of a behavioral economic approach, it is worth noting that economic research thus far has overwhelmingly focused on distributional preferences and, as such, has a rather restricted scope. The moral psychologist is clearly concerned with explaining the existence and function of distributional preferences, but he or she is also concerned with explaining a wide variety of other behaviors that fit into the moral realm. As a general conceptual difference, it can be seen that a theory of distributional preferences addresses what Rawls (1971) would have called a "theory of justice," leaving questions about the "theory of good" largely untouched. Broadly speaking, the former type of normative reasoning would focus on issues pertaining to procedural justice, while the second would appeal to more substantive conceptions of value. As several authors (e.g., Haidt & Joseph, 2008; Shweder, Mahapatra, & Miller, 1987) have argued, most moral psychological research thus far shares this limitation: the scenarios and stimuli used involve, almost exclusively, only justice or harm. This restricted focus ignores a variety of other values that are likely critical players in the moral lives of most people. Haidt and Joseph (2008), for example, argue that the list of moral "foundations," or intuitive moral values, includes not only

fairness/reciprocity and harm/care, but also in-group/loyalty, authority/respect, and purity/sanctity.

An important challenge for future work, therefore, is to expand the scope of moral issues addressed. In particular, how can the mathematical rigor of economic models be used to capture moral values such as loyalty, respect, or sanctity? The goal here would be to develop a utility function somehow enriched so as to be sensitive to a variety of substantive moral values, much as researchers have attempted to enrich utility theories to accommodate distributional preferences. As mentioned above, one challenge in model building is to create a model sophisticated enough to capture real, complex behavior without succumbing to over-parameterization. The more variation in individual differences, the more substantial this problem becomes. Given the variety of candidate moral values, and the potential for individual variation within each, modeling moral values is likely to be a particularly difficult task to accomplish.

## 4. *Moral neuroscience: An economical perspective*

We can investigate, with the aid of technologies such as fMRI, how the assumptions and predictions developed formally in behavioral models might be instantiated in the neurophysiology of the brain. The studies of Greene and colleagues (2001; 2004) were pioneering first steps in moral neuroscience, but they lacked the sort of rigorous mathematical formulation, and hence specificity of hypotheses, which we have in mind here. At the time of those studies, of course, even neuroscientists investigating economic decision-making failed to fully exploit the formal tools available within the closely related field of behavioral economics, such as game theory. This has begun to change, however (see Glimcher et al. 2009, for a review). An important study by Alan Sanfey and colleagues (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003), for example, was one of the first to bring a widely used economic game, the ultimatum game, into the fMRI scanner; they were able to successfully correlate responses to fair and unfair offers with differences in BOLD activity patterns.

Nonetheless, the merger of neuroscience and economic modeling (and game theory, in particular) – what is referred to as neuroeconomics – remains a nascent approach in the overall moral neuroscience research program. Despite the excitement surrounding neuroeconomics, there are actually only a handful of studies to date which use economical models to analyze brain data (Glimcher et al., 2009). Moreover, there are even less studies which use brain data to compare the performances of competing models (e.g., Hampton et al., 2008)

It is worth considering why we need formal modeling rather than, for example, a more refined conceptual analysis of the philosophical positions being determined. To begin with, social judgment and decision-making is, obviously, a complex psychological process (or processes) entailing a wide variety of neural computations spanning a broad set of brain regions. This is not surprising, given the diversity of components that might factor into even the simplest of behavioral interactions with another person, such as working and episodic memory, value representation, emotion, motivation, theory of mind, and reward and punishment learning. Several studies are beginning to provide converging evidence of a set of brain regions correlated with each of these various components (Frank, Cohen, & Sanfey, 2009; Lieberman, 2007; Moll, et al., 2005), although much work remains to be done. Success in this brain mapping endeavor has relied on, and will continue to rely on, the ability to describe the targeted psychological phenomenon in a manner that is sufficiently precise and plausible at the biological level.

We are interested, after all, in not only finding a pattern of BOLD activity that differs across experimental conditions, but also in explaining, in detail, the *process* by which this activity instantiates the hypothesized psychological processes underlying different behavioral responses (see Friston, 2002). A general contrast between utilitarian and deontological dilemmas, for example, is not very helpful in teasing apart this more fine grained question. This is because there are no "utilitarian" or "deontological" reasoning processes in the brain; there are only interactions between various neural components that, under the right conditions, give rise to behavior that can be described, with varying degrees of success, as characteristically utilitarian or deontological in one sense or another. Researchers have not, of course (or at least presumably), actually believed in such a thing as a straightforward utilitarian or deontological brain mechanism, but nonetheless experimental designs often seem to assume that something like utilitarianism or deontology could be directly and cleanly mapped onto disparate brain mechanisms. One problem is that overlapping brain regions might participate in the generation of both types of abstractly described behavior. Emotion might lead to a characteristically deontological judgment in one circumstance, but in a different circumstance draw attention to the merits of a characteristically utilitarian judgment. Moreover, even if a single, unique region were found, a high-level concept such as utilitarianism is not a very helpful primitive for use in explaining process. It seems we would still want to explain the underlying computation that renders the mental phenomenon "utilitarian" in the first place.

Now the researcher could respond to this problem with a more refined conceptual analysis, that is, he or she could decompose the philosophical term into subcomponents that might more realistically map onto brain function. What is

difficult to imagine, however, is that such an enterprise could ever operationalize the candidate subcomponents with the rigor and precision offered by mathematical models or, by extension, empirically test theoretical predictions with the same vigorousness. The problem relates back to the complexity of the task at hand. There is every reason to expect that the neurophysiology of moral cognition is extremely complex, consisting of numerous, massively dynamic, neural networks. Describing such mechanisms at a highly detailed level with only conceptual language is likely impossible. It would be akin to an engineer trying to successfully explain how to build a space shuttle without the use of physics – only probably much worse.

Fortunately, most research programs seem to be abandoning an exclusively conceptual analysis, opting instead for the sort of modeling approach we advocate. In particular, the basic psychological components hypothesized within moral psychology are beginning to mimic those found in behavioral economics and adopted in neuroeconomics, for example, the use of a utility function that people seek to maximize, as well as a learning algorithm for updating beliefs about how others will act. This theoretical framework, importantly, allows for precise mathematical description (Sanfey, Loewenstein, McClure, & Cohen, 2006; Glimcher et al., 2009).

Note that we are still, in an important sense, targeting the broader theoretical issues assumed in, for example, Greene et al.'s (2001, 2004) work on utilitarianism and deontology. The primitive categories under investigation (i.e., those captured by the parameters of the utility function), can ultimately be used to inform the larger theoretical debate about abstract concepts. Nonetheless, when the focus is on detailed mechanistic explanation, remaining at the modeling level will likely be necessary, since computational models will be necessary to keep tract of the neurophysiological complexity. To give a sense of how researchers have begun to shift from philosophical categories toward more precise modeling, as well as illustrate the relative benefits of the latter, two lines of study are worth reviewing: the work of Hsu and colleagues (2008) on distributional justice and the research of Harbaugh et al. (2007) on charitable giving.

One of the challenging questions in moral philosophy and social economics concerns the conflict that may emerge between efficiency and equity. Suppose, for example, that we must allocate a limited resource such as an expensive medicinal pharmaceutical. In particular, assume there are two patients who will die without the medication, but there are only 30 dosages remaining. The first patient, Sandra, needs only one dose per day in order to survive; Susan, on the other hand, requires three doses per day. How should we allocate the medicine? If we split the doses evenly, then Sandra will survive 15 days, but Susan will only survive five days

(since she requires more dosage per day). This amounts to 20 extra days of collective life between the two patients. If, on the other hand, we give all of the medicine to the Susan, she will survive 30 days while Susan will die right away. This is a larger gain, collectively, in extra days of life (viz., 30 rather than 20). Does this added efficiency conflict with our sense of equity (Segal, 2006)?

Hsu and co-workers (2008) presented allocation problems of this sort to subjects in an fMRI experiment. In particular, subjects had to decide between possible allocations of food between children in an orphanage in Uganda. For each trial of the task, the participants saw a picture of three children and told that the monetary equivalent of up to 24 meals could be given to each. However, depending on the type of allocation, there was a cost – a certain number of meals would be subtracted from certain children. For example, one allocation option would subtract 15 meals from one child but none from the other two, while the alternative allocation would subtract 13 meals from one child, five meals from the second child, and no meals from the third. Similar to the case of Sandra and Susan, the former case is more efficient (less meals are lost overall), but the latter case perhaps seems more equitable (one child does not bear all the cost alone).

The choice was modeled as follows. Define the marginal efficiency of an allocation as the difference between the summation of meals (M) in the chosen (c) and unchosen (u) distribution ($\Delta M = M_c - M_u$). A Gini coefficient was used to measure inequity (Gini, 1921); that is, a coefficient [0, 1] was computed which measures the distance, under the chosen allocation, between a uniform distribution and the Lorenz curve (i.e., the empirical, cumulative distribution of wealth.) The Gini coefficient is zero if the payoffs are equal amongst all persons, but increases toward one otherwise; it is equal to one if a single individual were to have all of the income. The marginal utility between the chosen and unchosen allocations can then be modeled as: $\Delta U(x) = (M_c - \alpha * G_c) - (M_u - \alpha * G_u)$, where $\alpha$ is a free parameter and G is the Gini coefficient for that choice. Note that as $\alpha$ increases the individual is more inequity averse.

Hsu et al. (2008) estimated the parameters for the model from the behavioral data, and then used these results to analyze the brain data. The caudate nucleus (head region) and septal-subgenual area were found to correlate with the marginal utility $\Delta U$. Moreover, the participant-wise inequity aversion parameter ($\alpha$) correlated with the magnitude of the activation in these regions. It was also possible to dissociate brain signals correlated with efficiency from those correlated with equity. Bilateral putamen activity was positively correlated with only the $M_c$ term, while the insular cortex was negatively correlated with $\Delta G$ (i.e., the marginal equity, $G_c - G_u$). This latter correlation means there is more insular activity when the inequitable option is chosen. One could make the reverence inference that such

activity indicates an emotional aversive response, much as how insular activity has been interpreted in previous studies (such as the ones by Greene et al., 2001, 2004), but to inequity specifically rather than some abstract notion such as deontology. The upshot is that the cognitive processes being implicated in this reverse inference are far more specific: rather than making generic claims about deontological reasoning, it is specified that the marginal equity of the available allocation options is what is being tracked by the insula. It is difficult to imagine how this theoretical advance could be made without the benefit of a modeling toolbox.

A second line of investigation attempts to understand the psychological motivations underlying charitable giving and, in particular, the distinction between pure altruism and impure altruism. (Andreoni, 1989, 1990). This relates back to the question broached earlier about why real humans do not act like *Homo economicus*. Why would a person voluntarily donate rather than selfishly retain his or her own payoffs? We have discussed candidate behavioral models at length elsewhere (Yokum & Rossi, 2009), and also mentioned a few possible explanations in the previous section. Here, however, we turn to consider some of the recent neuroscientific work on the issue.

James Andreoni (1989, 1990) pioneered much of the behavioral, game theoretic work on voluntary donation, typically using an experimental paradigm known as the public goods game (Leydard, 1995). The pure/impure altruism dichotomy is meant to capture the possibility that a person might donate either because an increase in a public good is desirable *per se* or, alternatively, because he or she experiences a sort of selfish, personal satisfaction from the very act of giving. To see how this can be formally modeled, consider a citizen $i$ who is endowed with wealth $w_i$, and the simplifying assumption of only one private good and one public good. The citizen can spend his or her wealth on private consumption or charitably donate it to the public good. Let $G$ represent the total contribution to the public good from all persons, including the citizen being considered. We can then define a utility function on the above terms: $U_i = U(x_i, g_i, G)$. What this means is that people derive satisfaction from their private consumption ($x_i$), their personal contribution to the public good ($g_i$), and the overall level of the public good ($G$). A person is purely altruistic if only $x_i$ and $G$ enter as arguments, while he or she is impurely altruistic if $x_i$, $G$, and $g_i$ all enter as arguments. The term $g_i$ captures what is referred to as "warm-glow." It is a feeling of personal satisfaction about one's contribution *per se*, a feeling that is independent of the actual benefit to the public good. As such, even though the behavioral output is altruistic, it is actually selfish at bottom – it is a sort of egoistic altruism.

The pure/impure dichotomy may seem marginal, but the consequences on behavior between a subject endowed with a utility function of the form $U_i = (x_i, g_i)$

and one endowed with $U_j = (x_j, G)$ can be quite dramatic. Suppose, for example, that subjects are allowed to choose between the following: (a) a third party donates $100,000 to support some charitable organization or (b) the subject personally donates $1 to support some charitable organization. The pure altruist will strictly prefer (a) to (b), wanting to ensure the charitable organization receives the most amount of money, but the subject motivated by warm glow may prefer (b) to (a)!

Jorge Moll and collaborators (2006) ran the first fMRI study that tried to identify the motivational components of voluntary donation. Subjects were endowed with $128, and then presented with several charitable organizations. For each, they were given a binary choice: donate or oppose donation. The experimenters manipulated the specific details of the choices in order to create three different conditions, namely, ones entailing (1) pure monetary reward, (2) non-costly donation, and (3) costly donation. In the first condition, subjects were asked to accept or reject a monetary reward for themselves, without any consequence for the charitable organization; in this case, therefore, only egoistic considerations matter. In the non-costly donation condition, subjects were asked to accept or reject a monetary reward on behalf of the organization, without any consequence for themselves; in this case, therefore, only pure altruistic considerations matter. Finally, in the costly condition, the experimental subjects had to sustain a certain cost in order to benefit the charitable organization. For example, they might lose $2 while the organization gains $5. In this case, there is a tradeoff between egoistic and altruistic considerations.

Analysis of the fMRI data revealed that activity in components of the brain reward system, notably the ventral tegmental area and striatum, were correlated with both the pure monetary donation condition and cases when the subjects decided to donate money to the organization. In a contrast between donation and pure reward conditions, the researchers observed enhanced brain activity in the ventral striatum and the subgenual area for costly donations; activity in the subgenual area was actually unique for costly donations. These results are consistent with the idea that charitable giving can be a personally rewarding experience. The fact that subgenual activity was specific to decisions to donation also presents the tantalizing possibility of a structure that is relatively unique to altruistic impulses.

Harbaugh and colleagues (2007) recently investigated the distinction between pure and impure altruism in more detail. Participants were again asked to accept or reject a series of monetary allocations between themselves and a charitable organization. They were also, however, presented with three conditions that did not require a decision. The first, meant to measure pure personal gain, entailed a payment to the subject at no cost to the organization. The second measured pure

altruism, and entailed a payment to the organization at no cost to the subject. The final "mandatory tax" condition entailed a compulsory transfer of money from the subject's account to the charity. Together with the voluntary decisions, these conditions allowed the researchers to more precisely examine the various components of the Andreoni model we discussed above, namely, $U_i = U(x_i, g_i, G)$. For example, by contrasting the mandatory tax and voluntary donation conditions, the researchers could observe whether there was unique brain activity correlated with the act of giving *per se* – whether there might be a brain correlate to warm-glow altruism. Note that in both cases the organization gains money from the subject (and thus overall benefit to the public good, $G$, and cost to the subject, $x_i$, is the same), but only in the voluntary condition is the subject responsible for the decision (and hence $g_i$ relevant).

Harbaugh et al. (2007) went further – and this next step is exemplary of the sort of power economic modeling can bring to the table. It should, in principle, be possible to predict subjects behavior if we had some way to measure of the various components of the $U_i = U(x_i, g_i, G)$ model. The contrast conditions above were designed to provide precisely this data. Harbaugh et al. used the magnitude of reward-related BOLD activity during the non-costly payments to the subject as an indicator of his or her marginal utility from personal payment ($x_i$), and activity during the mandatory tax condition to measure the marginal utility from increases to the public good ($G$). Using a different set of voluntary donation decisions, it was possible to successfully predict real giving behavior. In particular, participants who had relatively higher reward-related activation during non-costly personal payments (indicative of a relatively larger $x_i$) were less likely to donate. Likewise, subjects with relatively higher ventral striatal and insular activity in the mandatory tax condition were more likely to donate.

The predictive power of the model that was observed in the out-of-sample test greatly increases our confidence in the validity of theories of pure altruism and warm-glow. The results are still correlational, but the ability to use independent brain measures – measures across different trials and conditions – in order to enhance predictive power suggests that we are actually grasping something about the underlying causal mechanism. This is further corroborated by the consistency between the Harbaugh et al. (2007) and Moll et al. (2006) experiments and our broader understanding of the brain reward system.

The experiments outlined above illustrate the benefits of complementing brain imaging with formal modeling. This obviously does not minimize the importance of previous studies in moral psychology that lacked such modeling, but it does suggest a methodological direction for future work. There is an evolution in moral neuroscience that is occurring and should be encouraged, namely, the adoption of

neuroeconomic models to help interpret and understand the neurophysiological bases of moral cognition. This will not, however, be an easy task. The formal toolbox of economic modeling is not, at present, developed enough to capture the diversity of substantive moral values that likely exist. As mentioned previously, models to date focus on theories of justice rather than theories of the good. This was true, for example, of the Hsu et al. (2008), Moll et al. (2006), and Harbaugh et al. (2007) models just discussed. We therefore confront, once again, the challenge of whether mathematical modeling can be expanded enough to cover the broad spectrum of moral problems of interest to the neuroscientist of ethics. This is, we believe, ultimately an empirical question. And it is an empirical question well worth exploring, given the benefits that modeling, if successful, has to offer.

## 5. *Conclusion*

Important advances have been made by investigating whether high-level philosophical theories, such as utilitarianism and deontology, might *directly* correspond with individual moral preferences and the neurophysiological bases of those preferences. Unfortunately, we argued, this framing of the empirical strategy encounters a critical barrier when confronted with the complexity of interpreting both behavioral and brain data. Abstract conceptual language is not precise enough to capture the casual mechanisms at work in sufficient detail; such an endeavor is similar, we made the analogy, to an engineer trying to build a space shuttle without relying on physics.

We should not, however, be discouraged by the shortcomings to date. The neuroscience of ethics is a growing discipline with promising avenues for future work. We argued that an *indirect* investigation of moral psychology, one that exploits the formalism of mathematical models, is especially important and likely to be fruitful. This would entail, in particular, that the original conceptual categories be decomposed and operationalized in mathematical terms that have a plausible psychobiological counterpart. These terms can then be used to build formal models which allow for the interpretation of complex data and the generation of testable predictions about real behavior. We illustrated this potential by reviewing some of the early neuroeconomic research aimed at understanding the neural bases of distributional preferences in the contexts of allocation decisions and charitable giving. It was also acknowledged, however, that building models which are sophisticated enough to capture substantive moral values will not be a straightforward task.

At the end of the day, the usefulness of the methodological toolbox advocated in this paper must be assessed empirically. Nonetheless, the results obtained thus far justify a sense of optimism regarding the work to come. And although the foreseeable obstacles are significant, notably the challenge of how to model substantive moral values, they seem to be practical issues to be overcome, rather than insurmountable theoretical flaws in the method. The moral neuroscience research program, therefore, is in a position to expand: a new methodology is on the block, and there are a multitude of possibilities for putting it to use. Particularly exciting is the possibility of expanding our focus beyond theories of justice into theories of the good.

*Bibliography*

Andreoni, J. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *Journal of Political Economy, 97*, 1449-1458.
Andreoni, J., & Miller, J. (2002). Giving according to GARP: an experimental test of consistency of preferences for altruism. *Econometrica, LXX*, 737-753.
Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory, 144*, 1-35.
Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.
Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience, 4*, 841-846.
Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.
Charness, G., & Rabin, M. (2001). Understanding social preferences with simple tests. *The Quarterly Journal of Economics, 117* (3), 817-869.
Connolly, T., & Hardman, D. (in press). "Fools rush in": A JDM perspective on the role of emotions in decisions, moral and otherwise.
Cushman, F., Greene, J., & Young, L. (in press). Our multi-system moral psychology: towards a consensus view. In J. Doris, G. Harman, S. Nichols, J. Prinz, W. Sinnott-Armstrong, & S. Stich (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford: Oxford University Press.
Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17* (12), 1082-1089.

De Ridder, D., Langguth, B., Plazier, M., & Menovsky, T. (2009). Moral dysfunction: Theoretical model and potential neurosurgical treatments. In J. Verplaetse, J. De Schrijver, S. Vanneste, & J. Braeckman (Eds.), *The Moral Brain: Essays on the Evolutionary and Neuroscientific Aspects of Morality* (pp. 155-183). New York, NY: Springer.
Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*, 268-298.
Fehr, E., & Schmidt, K. (1999). Theories of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817-868.
Foot, P. (1978). The problem of abortion and the doctrine of the double effect. In P. Foot, *Virtues and Vices: And Other Essays in Moral Philosophy*. Oxford: Blackwell Publishers.
Frank, M. J., Cohen, M. X., & Sanfey, A. G. (2009). Multiple systems in decision making: A neurocomputational perspective. *Current Directions in Psychological Science, 18*, 73-77.
Friedman, M. (1953). Methodology of positive economics. In D. Hausman, *The Philosophy of Economics* (pp. 145-173). Cambridge, MA: Cambridge University Press.
Friston, K. (2002). Beyond phrenology: What can neuroimaging tell us about distributed circuitry. *Annual Review of Neuroscience, 25*, 221-250.
Genakopolos, J., & Pearce, D. S. (1989). Psychological games and sequential rationality. *Games and Economic Behavior, 1*, 60-79.
Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal, 31*, 124-126.
Glimcher, P. W., Camerer, C. F., Fehr, E., & Poldrack, R. A. (Eds.). (2009). *Neuroeconomics: Decision making and the brain*. San Diego, CA: Elsevier.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389-400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.
Greene, J. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vol. 3: The Neuroscience of morality: Emotion, brain disorders, and development, pp. 35-79). Cambridge, MA: MIT Press.
Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London, 359*, 1775-1785.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6* (12), 517-523.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 316*, 814-834.
Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The Innate Mind* (Vol. 3: Foundations and the Future, pp. 367-391). New York, NY: Oxford University Press.
Hampton, A., Bossaert, P., & O'Doherty, J. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *PNAS, 105* (18), 6741-6746.
Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316*, 1622-1625.
Hsu, M., Anen, C., & Quartz, S. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science, 320*, 1092-1095.
Huettel, S. A., Song, A. W., & McCarthy, G. (2008). *Functional Magnetic Resonance Imaging* (2nd ed.). Sunderland, MA: Sinauer Associates, Inc.
Joyce, R. (2008). What neuroscience can (and cannot) contribute to metaethics. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vols. 3: The neuroscience of morality: Emotion, brain disorders, and development, pp. 371-394). Cambridge, MA: MIT Press.
Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review, 93* (5), 1449-1475.
Leydard, J. (1995). Public goods: A survey of experimental evidence. In J. H. Hagel, & A. E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 111-194). Princeton, NJ: Princeton University Press.
Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annural Review of Psychology, 58*, 259-289.
McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45* (3), 577-580.
Meegan, D. V. (2008). Neuroimaging techniques for memory detection: Scientific, ethical, and legal issues. *American Journal of Bioethics, 8* (1), 9-20.
Mikhail, J. (2008). Moral cognition and computational theory. In W. Sinnott-Armstrong (Ed.), *Moral Psychology* (Vols. 3: The Neuroscience of morality: Emotion, Brain Disorders, and Development, pp. 91-91). Cambridge, MA: MIT Press.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences, 11*, 143-152.
Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourao-Miranda, J., Andreiuolo, P. A., et al. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience, 22* (7), 2730-2736.
Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences, 103* (42), 15623-15628.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, 799-809.
Morse, S. (2006). Brain overclaim syndrome and criminal responsibility: A diagnostic note. *Ohio State Journal of Criminal Law, 3* (397), 397-412.
Osborne, M., & Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA: MIT Press.
Parens, E. (1998). Is better always good? The enhancement project. *The Hasting Center Report, 28* (1), S7.
Pettit, M. (2007). FMRI and BF meet FRE: Brain imaging and the federal rules of evidence. *American Journal of Law and Medicine, 33*, 319-340.
Poldrack, R. A. (2006). Can cognitive process be inferred from neuroimaging data. *Trends in Cognitive Science, 10*, 59-63.
Rabin, M. (1993). Importing fairness into game theory and economics. *American Economic Review, 83*, 1281-1302.
Racine, E., Bar-Ilan, O., & Illes, J. (2005). fMRI in the public eye. *Nature Reviews Neuroscience, 6*, 159-164.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, J. (2000). *Lectures on the History of Moral Philosophy*. (B. Herman, Ed.) Cambridge, MA: Harvard University Press.
Roskies, A. (2002). Neuroethics for the new millennium. *Neuron, 35* (1), 21-23.
Rossi, F., & Yokum, D. (2009). Sulla natura della moralita': una brevissima rassegna. *Annuario di Etica, 6*.
Sanfey, A. G., & Chang, L. J. (2008). Multiple systems in decision making. *Annals of the New York Academy of Sciences, 1128*, 53-62.
Sanfey, A. G., Loewenstein, G., McGlure, S. M., & Cohen, J. D. (2006). Neuroeconomics: Cross-currents in research on decision-making. *Trends in Cognitive Sciences, 10* (3), 108-116.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300* (5626), 1755-1758.

Segal, U. (2006). Fair Bias. *Economics and Philosophy, 22*, 213-229.
Shweder, R. A., Mahapatra, M., & Miller, J. G. (1987). Culture and moral development. In J. Kagan, & S. Lamb (Eds.), *The Emergence of Morality in Young Children* (pp. 1-83). Chicago, IL: University of Chicago Press.
Sober, E., & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 204-217.
Tovino, S. (2007). Functional neuroimaging and the law: Trends and directions for future scholarship. *American Journal of Bioethics, 7* (9), 44-56.
Verplaetse, J. (2009). *Localizing the moral sense: Neuroscience and the search for the cerebral seat of morality, 1800-1930*. New York, NY: Springer.
Verplaetse, J., Braeckman, J., & De Schrijver, J. (2009). Introduction. In J. Verplaetse, J. De Schrijver, S. Vanneste, & J. Braeckman (Eds.), *The Moral Brain: Essays on the Evolutionary and Neuroscientific Aspects of Morality* (pp. 1-43). New York, NY: Springer.
Yokum, D., & Rossi, F. (2009). A neuroeconomic perspective on charitable giving. *Humana Mente* (10), 59-72.