2007

Moral Cognition and Computational Theory

John Mikhail
Georgetown University Law Center, jm455@law.georgetown.edu

November 2007

Moral Cognition and Computational Theory

John Mikhail
Associate Professor of Law
Georgetown University Law Center
jm455@law.georgetown.edu

Walter Sinnott-Armstrong, ed., *Moral Psychology, Vol. 3: The Neuroscience of Morality* (Cambridge: MIT Press forthcoming)

To appear in Walter Sinnott-Armstrong, ed., *Moral Psychology, Vol. 3: The Neuroscience of Morality* (Cambridge: MIT Press)

Moral Cognition and Computational Theory

John Mikhail

1. In his path-breaking work on the foundations of visual perception, David Marr distinguished three levels at which any information-processing task can be understood and emphasized the first of these:

> Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the nature of the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented (Marr, 1982, p. 27).

I begin with Marr to call attention to a notable weakness of Joshua Greene's ambitious and provocative essay: its neglect of computational theory. A central problem moral cognition must solve is to recognize (i.e., compute representations of) the deontic status of human acts and omissions. How do people actually do this? What is the theory which explains their practice?

Greene claims that "emotional response . . . predicts deontological judgment" (Greene p. 42), but his own explanation of a subset of the simplest and most extensively studied of these judgments—trolley problem intuitions—in terms of a personal/impersonal distinction is neither complete nor descriptively adequate (Mikhail, 2002), as Greene now acknowledges in a revealing footnote. As I suggest below, a more plausible explanation of these intuitions implies that the human brain contains a computationally complex "moral grammar" (e.g., Dwyer, 1999; Harman, 2000; Mikhail, 2000; Mikhail et al., 1998), analogous in certain respects to the mental grammars operative in other domains, such as language, vision, music, and face recognition

(Jackendoff, 1994). If this is correct, then Greene's emphasis on emotion may be misplaced, and at least some of his arguments may need to be reformulated.

2. Consider the following trolley problem variations, which I designed to study the computations underlying moral judgment (Mikhail, 2000).

Bystander
Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. Unfortunately, there is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Hank to throw the switch?

Footbridge
Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. Unfortunately, the heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to throw the man?

Consensual Contact
Luke is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Luke sees what has happened: the driver of the train saw a man walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man. It is moving so fast that he will not be able to get off the track in time. Fortunately, Luke is standing next to the man, whom he can throw off the track out of the path of the train, thereby preventing it from killing the man. Unfortunately, the man is frail and standing with his back turned. Luke can throw the man, injuring him; or he can refrain from doing this, letting the man die. Is it morally permissible for Luke to throw the man?

Disproportional Death

Steve is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Steve sees what has happened: the driver of the train saw a man walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man. It is moving so fast that he will not be able to get off the track in time. Fortunately, Steve is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the man. Unfortunately, there are five men standing on the side track with their backs turned. Steve can throw the switch, killing the five men; or he can refrain from doing this, letting the one man die. Is it morally permissible for Steve to throw the switch?

As is well known, problems like these can be shown to trigger widely shared deontic intuitions among demographically diverse populations, including young children (Gazzaniga, 2005; Greene et al., 2001; Hauser et al., in press; Mikhail, 2002; Mikhail et al., 1998; Petrinovich & O'Neill, 1996; Petrinovich et al., 1993; Waldmann, under review). Here I wish to draw attention to some of their theoretical implications.

3. It is clear that it is difficult if not impossible to construct a descriptively adequate theory of these intuitions—and others like them in a potentially infinite series—based exclusively on the information given (Mikhail, 2000). Although each of these intuitions is triggered by an identifiable stimulus, how the mind goes about interpreting these hypothetical fact patterns, and assigning a deontic status to the acts they depict, is not something revealed in any obvious way by the scenarios themselves. Instead, an intervening step must be postulated: a pattern of organization of some sort that is imposed on the stimulus by the mind itself. Hence a simple perceptual model, such as the one implicit in Haidt's (2001) influential account of moral judgment, is inadequate for explaining these intuitions.[1] Instead, as is the case with language perception (Chomsky, 1964), an adequate perceptual model must be more complex (Figure 1).

The expanded perceptual model in Figure 1 implies that, like grammaticality judgments, permissibility judgments do not necessarily depend only on the superficial properties of an action-description, but also on how that action is mentally represented. Additionally, it suggests that the problem of descriptive adequacy in the theory of moral cognition may be divided into at least three parts: (1) the problem of describing the computational principles ("deontic rules") operative in the exercise of moral judgment, (2) the problem of describing the unconscious mental representations ("structural descriptions") over which those computational operations are defined, and (3) the problem of describing the chain of inferences ("conversion rules") by which the stimulus is converted into an appropriate structural description.

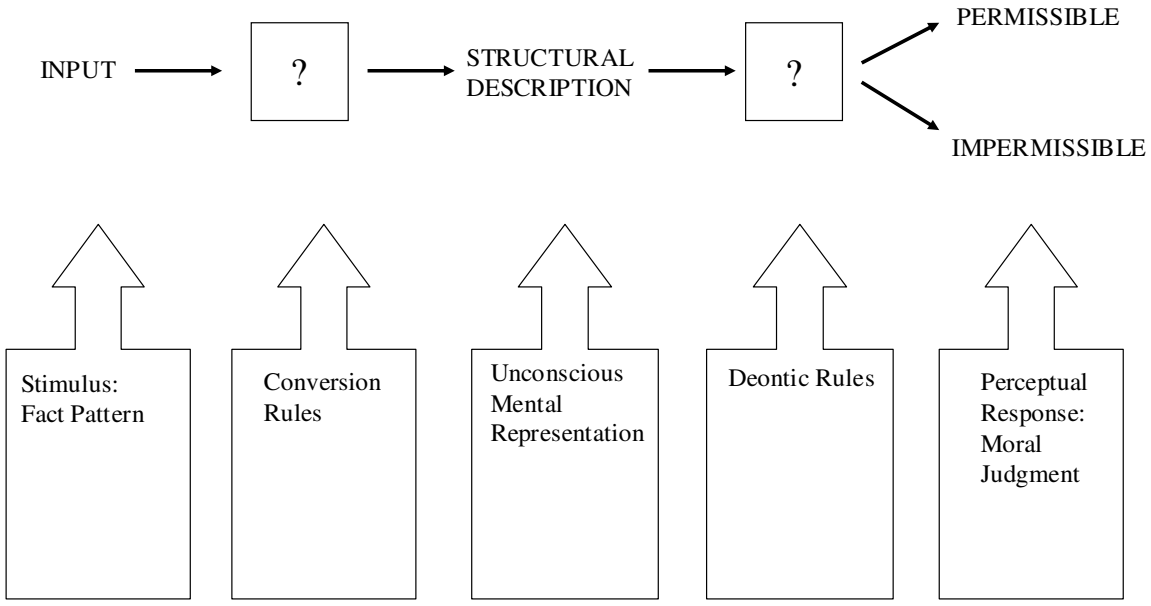

Fig. 1: Expanded Perceptual Model for Moral Judgment (Mikhail, 2000)

4. It is equally clear that Greene's own explanation of these intuitions is neither complete nor descriptively adequate. In a series of papers, Greene argues that people rely on three features to distinguish the Bystander and Footbridge problems: "whether the action in question (a) could reasonably be expected to lead to serious bodily harm, (b) to a particular person or a member or members of a particular group of people (c) where this harm is not the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107; see also Greene, 2005; Greene et al., 2004; Greene & Haidt, 2002). Greene claims to predict trolley intuitions and patterns of brain activity on this basis. However, this explanation is incomplete, because we are not told how people manage to interpret the stimulus in terms of these features; surprisingly, Greene leaves this crucial first step in the perceptual process (the step involving conversion rules) unanalyzed. Additionally, Greene's account is descriptively inadequate, because it cannot explain even simple counterexamples like the Consensual Contact and Disproportional Death problems[2]—let alone countless real-life examples which can be found in any casebook of torts or criminal law (Mikhail, 2002; Nichols & Mallon, 2006). Hence Greene has not shown that emotional response predicts these moral intuitions in any significant sense. Rather, his studies suggest that some perceived deontological violations are associated with strong emotional responses, something few would doubt or deny.

5. A better explanation of these intuitions is ready to hand, one that grows out of the computational approach Greene implicitly rejects. We need only assume people are "intuitive lawyers" (Haidt, 2001) and have a "natural readiness" (Rawls, 1971) to compute mental representations of human acts in legally cognizable terms. The Footbridge and Bystander problems, for example, can be explained by assuming that these problems trigger distinct mental

representations whose relevant temporal, causal, moral, and intentional properties can be described in the form of a two-dimensional tree diagram, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive, and transitive (Goldman, 1970; Mikhail, 2000). As these diagrams reveal, the key structural difference between these problems is that the agent commits multiple counts of battery prior to and as a means of achieving his good end in the Footbridge condition (Figure 2), whereas in the Bystander condition, these violations are subsequent and foreseen side effects (Figure 3).

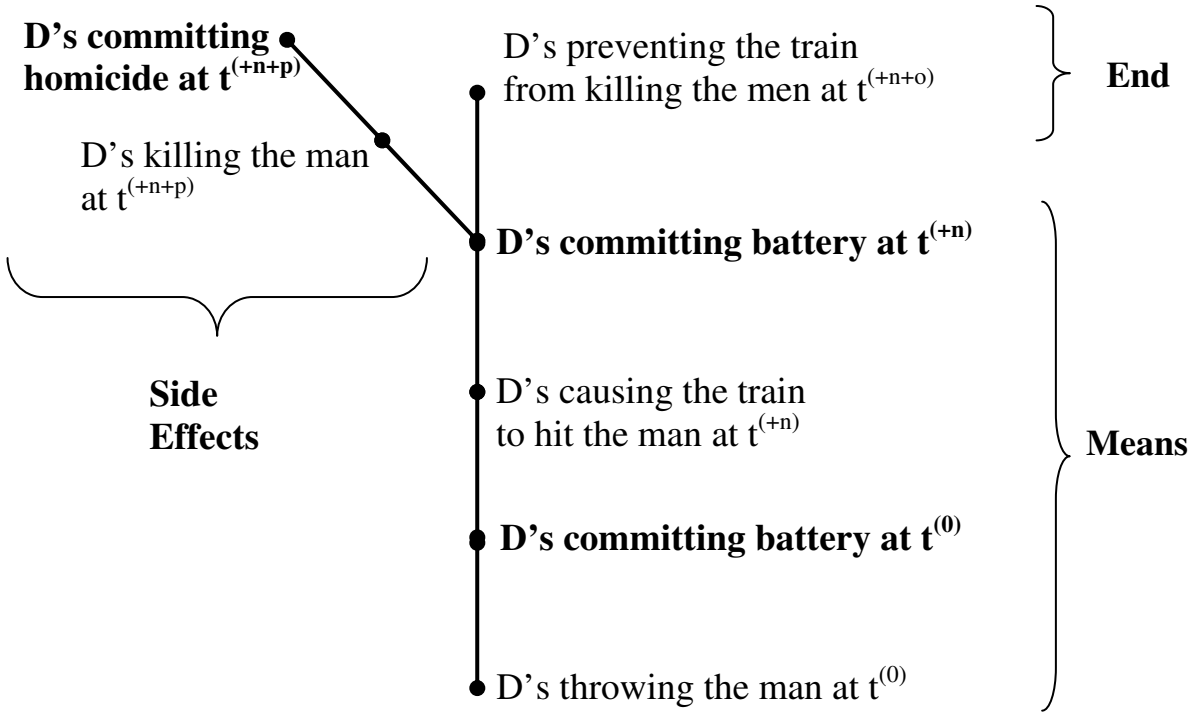

Fig. 2: Mental Representation of Footbridge Problem (Mikhail, in press)

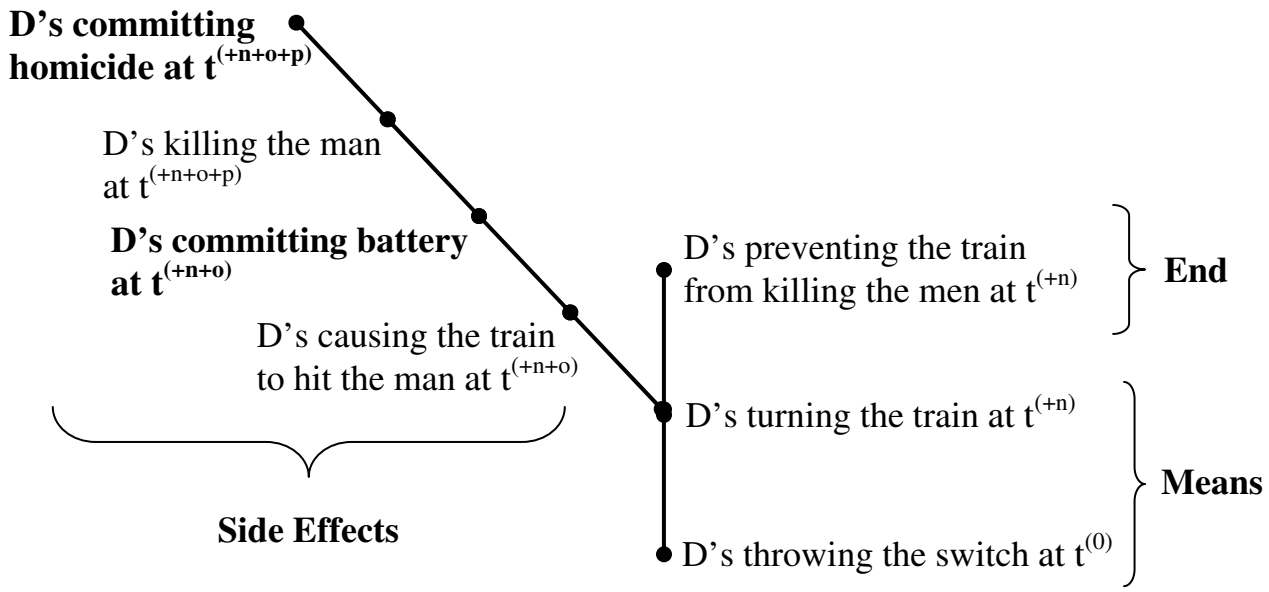

Fig. 3: Mental Representation of Bystander Problem (Mikhail, in press)

The computational or "moral grammar" hypothesis holds that when people encounter the Footbridge and Bystander problems, they spontaneously generate unconscious representations like those in Figures 2-3. Note that in addition to explaining the relevant intuitions, this hypothesis has further testable implications. For example, we can investigate the structural properties of these representations by asking subjects to evaluate probative descriptions of these actions. Descriptions using the word "by" to connect individual nodes of the tree in the downward direction (e.g., "D turned the train by throwing the switch," "D killed the man by turning the train") will be deemed acceptable; by contrast, causal reversals using "by" to connect nodes in the upward direction ("D threw the switch by turning the train," "D turned the train by

killing the man") will be deemed unacceptable. Likewise, descriptions using the phrase "in order to" to connect nodes in the upward direction along the vertical chain of means and ends ("D threw the switch in order to turn the train") will be deemed acceptable. By contrast, descriptions linking means with side effects ("D threw the switch in order to kill the man") will be deemed unacceptable. In short, there is an implicit geometry to these representations, which Greene and others (e.g., Sunstein, 2005) neglect but an adequate theory must account for (Mikhail, 2005).[3]

6. The main theoretical problem raised by the computational hypothesis is how people manage to compute a full structural description of the relevant action that incorporates certain properties, such as ends, means, side effects, and *prima facie* wrongs like battery, when the stimulus contains no direct evidence for these properties. This is a poverty of the stimulus problem (Mikhail, 2006), similar in principle to determining how people manage to extract a three-dimensional representation from a two-dimensional stimulus in the theory of vision (e.g., Marr, 1982), or to determining how people recognize the word boundaries in an undifferentiated auditory stimulus in the theory of language (e.g., Chomsky & Halle, 1968). Elsewhere, I describe how these properties can be recovered from the stimulus by a sequence of operations which are largely mechanical (Mikhail, in press). These operations include (1) identifying the various action descriptions in the stimulus and placing them in an appropriate temporal and causal order, (2) applying certain moral and logical principles to their underlying semantic structures to generate representations of good and bad effects, (3) computing the intentional structure of the relevant acts and omissions by inferring (in the absence of conflicting evidence) that agents intend good effects and avoid bad ones, and (4) deriving representations of morally salient acts like battery and situating them in the correct location of one's act tree (Mikhail, 2000,

2002).[4] While each of these operations is relatively simple, the length, complexity, and abstract character of the process as a whole belies Greene's claim that deontological intuitions do not depend on "genuine" (p. 3), "complex" (p. 10), or "sophisticated abstract" (Greene & Haidt, 2002, p. 519) moral reasoning. In light of this and of Greene's failure to provide an adequate description of the computations which must be attributed to individuals to explain their moral intuitions, his reliance on characterizations like these seems unwarranted.

7. Greene rejects the computational hypothesis largely on the strength of a single counterexample, namely, Thomson's (1986) ingenious loop case. "The consensus here," he says, "is that it is morally acceptable to turn the trolley…despite the fact that here, as in the footbridge case, a person will be used as a means" (Greene, p. 10; see also Greene et al., 2001, p. 2106). To test this assumption, I devised the following two scenarios (Mikhail, 2000) and discovered that no such consensus exists.

> **Loop #1 (Ned)**
> Ned is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ned sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. Unfortunately, the heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Ned to throw the switch?

> **Loop #2: (Oscar)**
> Oscar is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Oscar sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that

> they will not be able to get off the track in time. Fortunately, Oscar is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. Unfortunately, there is a man standing on the side track in front of the heavy object with his back turned. Oscar can throw the switch, preventing the train from killing the men, but killing the man; or he can refrain from doing this, letting the five die. Is it morally permissible for Oscar to throw the switch?

Unlike other trolley problems, on which roughly 85%-95% of individuals agree, there is substantial disagreement over the permissibility of intervening in the two loop cases. For example, in the initial study utilizing these problems, only 48% of individuals judged Ned's throwing the switch to be permissible, whereas 62% judged Oscar's throwing the switch to be permissible (Mikhail, 2002; see also Mikhail, 2000; Mikhail et al., 1998). However, as these figures suggests, individuals did distinguish "Ned" and "Oscar" at statistically significant levels. These findings have since been replicated in a web-based experiment with several thousand subjects drawn from over 120 countries (Hauser, Cushman, Young, Jin & Mikhail, in press; see also Gazzaniga, 2005). Greene's account has difficulty explaining these findings, just as it has difficulty explaining the Consensual Contact and Disproportionate Death problems. All of these results, however, can be readily explained within a moral grammar framework (Mikhail, 2002).

8. In many respects, Greene's positive argument for an emotion-based approach to moral cognition has considerable plausibility. Nevertheless, some of the evidence he adduces in its favor appears to be weaker than he assumes. His reaction-time data, for instance, are inconclusive, because the moral grammar framework makes the same predictions regarding people's reaction times and arguably provides a better explanation of them. One who permits throwing the man in Footbridge must in effect overcome the prior recognition that this action constitutes an immediate and purposeful battery, and this process takes time; but one who

prohibits throwing the switch in Bystander need not override any such representation. Furthermore, while both the doing and forbearing of an action can be permissible without contradiction, the same is not true of the other two primary deontic operators (Mikhail, 2004). Hence Greene's reaction-time data can be explained by appealing to the cognitive dissonance resulting from the presence of a genuinely contradictory intuition in Footbridge which is not present in Bystander. By contrast, labeling the conflicting intuition a "prepotent negative emotional response" (Greene, p. 14) does not seem explanatory, for reasons already discussed.

Some features of Greene's experimental design also may be questioned. For example, the fact that it takes longer to approve killing one's *own* child (Crying Baby) than it does to condemn "a teenage girl" for killing *hers* (Infanticide) may not be entirely probative; Greene (2001) appears to co-vary multiple parameters here (cost/benefit and first-person/third person), undermining confidence in his results. More significantly, Greene does not appear to investigate considered judgments in Rawls' sense, that is, judgments "in which our moral capacities are most likely to be displayed without distortion" (Rawls, 1971, p. 47), in part because most of his dilemmas are presented in the second-person (e.g., "*You* are standing on a footbridge … Is it appropriate for *you* to push the man?"). This presumably raises the emotional index of his scenarios and risks magnifying the role of exogenous factors.[5]

Additionally, Greene does not appear to investigate deontic knowledge as such, because he asks whether actions are "appropriate" instead of whether they are morally permissible.[6] That this question appears inapposite can be seen by considering the analogous inquiry in linguistics: asking whether an expression is "appropriate" rather than "grammatical." Chomsky (1957, p.15) emphasized the importance of distinguishing *grammatical* from closely related but distinct notions like *significant* or *meaningful*, and the same logic applies here. Finally, whether one

"ought" to perform a given action is distinct from whether the action is morally permissible, and Greene occasionally conflates this crucial distinction (see, e.g., Greene et al., 2001, p. 2105).

9. These brief remarks are not meant to imply that Greene's project is without merit. On the contrary, I think his ideas are interesting, powerful, and at times even brilliant. His insight and creativity, clearly on display here, have helped give the field of moral psychology a much-needed boost. I would encourage him, however, to devote more effort to understanding the computational properties of moral cognition, in addition to its underlying mechanisms. Marr warned that "one has to exercise extreme caution in making inferences from neurophysiological findings about the algorithms and representations being used, particularly until one has a clear idea about what information needs to be represented and what processes need to be implemented" (Marr, 1982, p. 26). Without a better understanding of the rules and representations needed to explain widely shared moral intuitions, more caution would seem to be in order.

References

Chomsky, N. (1957). *Syntactic Structures*. Mouton: The Hague.

Chomsky, N. (1964). *Current Issues in Linguistic Theory*. New York: Pantheon.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Cambridge: MIT Press.

Dwyer, S. (1999). Moral competence. In R. Stainton, ed. *Philosophy and Linguistics*. Boulder, CO: Westview Press.

Jackendoff, R. (1994). *Patterns in the Mind: Language and Human Nature*. New York: Basic Books.

Gazzaniga, M. (2005). *The Ethical Brain*. New York: Dana Press.

Goldman, A. (1970). *A Theory of Human Action*. Princeton: Princeton University Press.

Greene, J. (2004). Cognitive neuroscience and the structure of the moral mind. In P. Carruthers, S. Laurence, and S. Stich, (eds.), *The Innate Mind: Structure and Contents*. Oxford: Oxford University Press.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6 (12), 517-523.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293.

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

Harman, G. (2000). *Explaining Value and Other Essays in Moral Philosophy*. New York: Oxford University Press.

Hauser, M., Cushman, F., Young, L., & Mikhail, J. (in press). A dissociation between moral

judgments and justifications. *Mind & Language*.

Marr, D. (1982). *Vison: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman and Co.

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A Theory of Justice'. PhD Dissertation, Cornell University.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect. Georgetown University Law Center Public Law & Legal Theory Working Paper No. 762385. Available at http://ssrn.com/abstract=762385

Mikhail, J. (2004). Islamic rationalism and the foundation of human rights. In A. Soeteman, ed *Pluralism and the Law*. Franz Steiner Verlag.

Mikhail, J. (2005). Moral heuristics or moral competence? Reflections on Sunstein. *Behavioral and Brain Sciences*, 28, 557-558.

Mikhail, J. (2006). The poverty of the moral stimulus. In W. Sinnott-Armstrong, ed., *The Psychology and Biology of Morality*. Oxford University Press.

Mikhail, J. (in press). *Rawls' Linguistic Analogy*. Cambridge University Press.

Mikhail, J., Sorrentino, C., and Spelke, E. (1998). Toward a universal moral grammar. In M.A. Gernsbacher and S.J. Derry (Eds.), *Proceedings, Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1250.

Nichols, S. and Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition* __.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions.

Journal of Personality and Social Psychology. Vol. 64. No. 3, 467-478.

Rawls, J. (1971). *A Theory of Justice.* Cambridge: Harvard University Press.

Schnall, S., Haidt, J., & Clore, G. (2004). Irrelevant disgust makes moral judgment more severe, for those who listen to their bodies [cite].

Sunstein, C. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531-573.

Tuck, R. (1979). *Natural Rights Theories: Their Origin and Development*. Cambridge: Cambridge University Press.

Waldmann, M., & Dieterich, J. (unpublished). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in bioethical intuitions.

Wheatley, T., & Haidt, J. (2004). The wisdom of repugnance: Hypnotically induced disgust makes moral judgment more severe. Psychological Science [cite].

Notes

[1] A notable feature of Haidt's "social intuitionist" model is that it provides no sustained analysis of the link between eliciting situation and intuitive response (see Haidt, 2001, p. 814, Figure 2).

[2] Throwing the man in Consensual Contact is an action which "could reasonably be expected to lead to serious bodily harm to a particular person . . . where this harm is not the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107). On Greene's account, therefore, if I understand it correctly, this case should be assigned to his "moral-personal" category and judged impermissible. Yet, in one experimental study, 93% of participants found this action to be permissible (Mikhail, 2002). Conversely, while throwing the switch in Disproportional Death is an action which "could reasonably be expected to lead to serious bodily harm to . . . a particular group of people," it is also "the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107). On Greene's account, therefore, it should be assigned to his "moral-impersonal" category and judged permissible. Yet, in the same study, 85% of participants found this action to be impermissible. How do individuals manage to come to these conclusions? The answer cannot be the one proposed by Greene et al. (2001). However, it may be that I am misinterpreting the intended scope of Greene's personal-impersonal distinction, in which case clarification would be welcome.

[3] Figures 2-3 also raise the possibility, which Greene does not consider, that deontic intuitions can be explained on broadly deontological (i.e. rule-based) grounds without reference to rights or duties. Put differently, they suggest that these concepts (and statements incorporating them, e.g., "Hank has a right to throw the switch," "Ian has a duty not to throw the man," "The man has a right not to be thrown by Ian," etc.), while playing an important perspectival role in deontological systems, are conceptually derivative, in a manner similar to that maintained by Bentham and other utilitarian theorists (Mikhail, 2000, 2004; Tuck, 1979).

[4] In the Footbridge Problem, for example, one must infer that the agent must *touch* and *move* the man in order to throw him onto the track in the path of the train, and the man would not *consent* to being touched and moved in this manner, because of his interest in self-preservation (and because no contrary evidence is given). By contrast, in the Consensual Contact problem one naturally assumes that the man *would* consent to being thrown out of the way of the train, even though doing so will injure him. The computational hypothesis holds that when people respond intuitively to these problems, they do in effect make these inferences, albeit unconsciously.

[5] Of course, if one wishes to study performance errors as such, then it may make sense to manipulate and enhance the influence of exogenous factors. This seems to be the approach adopted by Haidt and his colleagues (e.g., Wheatley & Haidt, 2004; Schnall et al., 2004) in the studies of theirs Greene relies upon.

[6] See Greene et al. 293 (5537): 2105 Data Supplement—Supplemental Data at http://www.sciencemag.org/cgi/content/full/293/5537/2105/DC1 (last visited 9/25/2001).