Cognitive Models of Moral Decision Making

Wendell Wallach

Interdisciplinary Center for Bioethics, Yale University

1. Introduction

The study of moral decision making by cognitive scientists is coming of age. The earlier years of research were characterized by bold claims regarding the evolutionary origins of morality, the primacy of unconscious mechanisms over reasoning in determining moral behavior, and the relationship between various features of neurophysiology and moral agency. In turn, these claims generated criticism from other scholars as being premature, based on inherently confusing methodologies, or founded in misunderstanding of moral philosophy. In any young science, collegial criticism can be expected to improve the quality of empirical investigations. That has certainly been the case for cognitive scientists interested in deepening our understanding of the psychology underlying moral decisions. Richer and more focused experiments are being designed. Carefully constructed philosophical arguments accompany the claims made for empirical findings.

In recent years, there has been a steady stream of research on the ways in which emotions and moral sentiments, intuitions, framing effects and implicit biases, heuristics, or an innate moral grammar influence behavior in morally significant situations. There remains, however, some confusion as to when these categories should be seen as overlapping or as distinct collections of cognitive processes. Furthermore, this early research is by necessity piecemeal. Well-designed experiments that focus on the role, structure, or function of a particular aspect of cognition can overemphasize the importance of those features.

Is there a specific set of cognitive processes responsible for the explicitly ethical dimension of decisions? Will, for example, the understanding of how judgments link
directly to moral intuitions account for much of moral behavior? Or does the choice of morally responsible actions require input from a broad array of cognitive competencies? Moral philosophers generally emphasize the importance of reason in the determination of what is right, good, and just. But virtue theorists perceive good actions as flowing from a good character, which implies the acquisition of many attributes through experience and habit. Most of the contributors to this issue lean toward the contention that a specific set of cognitive processes play a central role in determining moral behavior; however, they differ on what those processes might be.

Research on the prospects for creating artificial agents capable of making moral judgments by the editors of this issue suggests the importance of developing a comprehensive model of the manner in which humans make moral decisions (Wallach & Allen, 2009; Wallach, Franklin, & Allen, this issue). A comprehensive model might encompass all the differing collections of cognitive mechanisms that influence behavior.

In producing a special issue for topics on Cognitive Models of Moral Decision Making, we were interested in how a few leading researchers construe their work in light of the rigorous collegial debate that has characterized the empirical study of moral psychology. The articles in this volume illustrate that our contributors continue to perceive the development of cognitive models of moral decision making in very different ways. A truly comprehensive model may still lie somewhere in the future. Nevertheless, these articles represent key elements that would make up such a model, as well as excellent examples of the state of ongoing research.

2. Moral psychology and moral philosophy

The study of moral decision making underscores a long-standing tension in the study of ethics, the tension between is and ought, concerning the relevance of constraints from human nature and human psychology upon the role that normative ideals should play in the choice of actions. Moral philosophers place particular emphasis on illuminating ideals and in the determination of shoulds in the choice of actions. There has been a presumed gap between all normative and all descriptive approaches to ethics. While the influence of psychological mechanisms on moral judgments is significant, moral philosophers suspect that many of these mechanisms are self-serving.

The suspicion with which ethicists view the relevance of psychological mechanisms particularly emotions, to morality can be traced back to the Greek and Roman Stoic philosophers. The Stoics emphasized the importance of dispassionate reasoning in determining the right course of action. Plato, Descartes, Spinoza, Kant, and Nietzsche also shared in the attitude that emotions hinder moral judgment.

Not all moral philosophers have agreed with the Stoics regarding the destructive role of emotions. Aristotle appreciated how emotions could both reinforce and interfere with the honing of a virtuous character. For Aristotle, emotions were a component of the good life and intrinsically relevant to ethics. Blaise Pascal (2004, [1670]) was to propose in the Penseées that emotions could serve moral purposes, but this viewpoint did not find many
adherents until it was more fully developed as a theory of moral sentiments by David Hume (2000, [1739–40]) and Adam Smith (2002, [1759]). Rousseau, Hutcheson, and Schopenhauer were to build upon the contention that some emotions (compassion, pity, care, and love) are conducive to moral behavior.

For more than 2,000 years moral philosophers had demonstrated interest in what we now call moral psychology, and the two fields of study developed hand in hand. But that link was to be severed, initially by David Hume, and much later by G. E. Moore. Hume (2000, [1739–40]) provided the first formulation of the *is-ought* problem, when he famously claimed that one could not derive an *ought* from an *is*. Moore (2004, [1903]) condemned what he called the *naturalistic fallacy*. He argued that ethical concepts such as ‘‘good’’ are indefinable and irreducible to natural properties.

Throughout much of the 20th century the *is-ought* problem, the influence of the *naturalistic fallacy* on the thinking of moral philosophers, and the domination of psychology by behaviorism all placed a damper on the empirical study of moral psychology. Challenges to behaviorism were of course important to the emergence of cognitive science during the later half of the century. Challenges to the *is-ought* problem and the *naturalistic fallacy* (Kohlberg, 1971; Searle, 1964, 1969) encouraged the emergence of empirical research on moral psychology.

Today, a growing community of naturalistic philosophers, strong in their appreciation of scientific methodologies, is actively engaged with scientists researching moral psychology (Sinnott-Armstrong, 2007, 2008a,b). Some of these philosophers are friendly to the naturalizing of ethics (Casebeer, 2003; Flanagan, Sarkissian, & Wong, 2008). Others nurture the interdisciplinary study of moral psychology by critically appraising what the empirical research does and does not demonstrate. A third group is actively developing the discipline of experimental philosophy (Knobe, 2004; Knobe & Nichols, 2008) as a new tool for testing the validity of a variety of presumptions about human behavior. In different ways, these philosophers see themselves as returning to the historical roots of ethics and merging it again with the study of psychology.

Nevertheless, the bridge between moral philosophy and the scientific study of moral psychology remains rather tenuous. For many moral philosophers the division between *is* and *ought* is impenetrable. Cognitive scientists who wish to make normative claims for their descriptions of influences on moral judgment have found that they have a rather high mountain to climb.

Less controversial are contentions that the study of moral psychology will help in evaluating which of the competing ethical theories (consequentialist, deontological, or virtue-based) is best accommodated by the existing empirical evidence. ‘‘[T]here has been in recent years a growing consensus to the effect that an ethical theory committed to an impoverished or inaccurate conception of moral psychology is at a serious competitive disadvantage’’ (Doris & Stich, 2008).

The entanglement of the empirical study of moral psychology with fascinating philosophic issues contributes to much of the excitement generated by the research. But it can also place a special burden upon researchers to demonstrate that their work goes beyond philosophy and is grounded in good science.
3. Morality as a subject for empirical research

Well before the consolidation of the cognitive sciences, anthropologists and a few sociologists and psychologists initiated empirical research revealing of ethical dimensions of human psychology (Piaget, 1932 [1999]; Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950; Brandt, 1954; Kohlberg, 1969; Milgram, 1963, 1974). But research on moral behavior has picked up dramatically during the past decades with the advent of new tools, new methodologies, and new approaches.

The development of neuroscience has been particularly important for grounding the study of moral psychology. Benjamin Libet (Libet, Gleason, Wright, & Pearl, 1983) placed neuroscience at the center of any future discussion of free will when he demonstrated that the readiness potential, a neocortical indication of an impulse to act, appears roughly 800 ms before the action, and approximately 350 ms before subjects reported that they had decided to act. Antonio Damasio’s (1994) work with brain-damaged patients laid foundations for a new understanding of the contribution of emotions to reasoning, and helped give birth to a neuroscience of moral emotions (Moll, de Oliveira-Souza, Zahn, & Grafman, 2008). Neuroscientists are also contributing to knowledge about the absence of morality in psychopaths (Glenn, Iyer, Graham, Koleva, & Haidt, 2009a; Glenn, Raine, & Schug, 2009b; Kiehl, 2008; Kiehl et al., 2001), to a deeper understanding of moral development during adolescence (Baird, 2008; Giedd et al., 1999), and to disturbances in moral development among young children who have suffered neural damage (Anderson, Antoine, Damasio, Tranel, & Damasio, 1999; Eslinger, Flaherty-Craig, & Benton, 2004).

CAT, SPECT, PET, and MRI opened windows into the neural correlates of mental states. Neuroethicists are recording brain activity while subjects ponder ethical dilemmas (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003), hear morally charged sentences (Moll, Eslinger, & de Oliveira-Souza, 2001), or view morally significant images (Moll et al., 2002). Evidence gathered from neuroimaging studies figure prominently in two articles in this volume (Paxton & Greene; Sanfey & Kvaran).

Darwin (2010, [1871]) noted in The Descent of Man that the ‘‘origin of the moral sense, lies in the social instincts, including sympathy. And these instincts no doubt, were primarily gained as in the case of the lower animals, through natural selection.’’ But it would take relatively recent scientific approaches developed by game theorists, sociobiologists, and evolutionary psychologists to begin grounding Darwin’s insight in evidence. For example, the application of game theory to animal behavior led to a theoretical framework regarding the evolution of cooperative behavior in the form of reciprocal altruism (Axelrod & Hamilton, 1981; Hamilton, 1964a,b; Trivers, 1971) and the evolution of fairness and a social contract (Skyrms, 1996, 2000). Claims for the evolutionary origins of morality have also been supported by similarities between the social mechanisms that reinforce communal behavior in nonhuman animals, and the patterns of rules and behaviors evident in human cultures (Flack & de Waal, 2000; de Waal, 1996). However, more direct support for evolutionary influences on human morality will require establishing links between innate dispositions, or innate computational mechanisms, and moral judgment. Two of our articles
(Flanagan & Williams; Paxton & Greene) discuss the Social Intuitionist Model (SIM) formulated by Jonathan Haidt and colleagues (Haidt, 2001; Haidt & Bjorklund, 2008; Haidt & Joseph, 2007), which posits the existence of innate intuitions. A third article (Dwyer et al.) provides a summary of current research directed at elucidating the rules and structures that would make up an innate moral grammar (Rawls, 1999, [1971]).

One other theme that has emerged from the research, and which is central to an appreciation of articles in this issue, is the diminishment of the role reason plays in determining moral behavior. The SIM model of moral cognition, for example, allows a role for reason, but SIM is generally understood as arguing for the primacy of emotionally activated intuitions over reasoned reflection in the making of most moral judgments. This is but one example of how cognitive scientists perceive a diminished role for reason. The idealized notion of a rational agent, and thus a rational moral actor, has been undermined by evidence that the reasoning capacity of the individual is bounded (Simon, 1955), limited (Miller, 1956), subject to cognitive biases (Tversky & Kahneman, 1974; Kahneman, Slovic, & Tversky, 1982; Gilovich, Griffin, & Kahneman, 2002), and fraught with implicit biases (Greenwald & Banaji, 1995). Experiments in psychology and social psychology established a view that moral behavior could be altered by relatively minor changes to the situation, by the effects of priming, and by other unconscious or nonconscious influences (Isen & Levin, 1972; Darley & Batson, 1973; Haney, Banks, & Zimbardo, 1973; Uleman & Bargh, 1989; Hassin, Uleman, & Bargh, 2006). But the pendulum may be swinging back. Models that emphasize important aspects of moral reasoning are central to three of the articles summarized below (Gigerenzer; Knobe; Paxton & Greene).

Cataloging the explosion of directions research on moral psychology has taken during the past 20 years is beyond the scope of this introduction. But the major research trajectories are well represented by contributions to this volume, as is some of the excitement and philosophical controversy.

4. Mechanisms, modularity, and morality

Most of the essays in this issue of *topiCS* offer models of moral decision making that are modular in the sense that the moral faculty is composed of a collection of individual competencies keyed as responses to specific social contexts. Owen Flanagan and Robert Williams compare two modular theories of morality, one proposed by the ancient Chinese philosopher Mencius and the more recent Social Intuitionist Model (SIM). Both theories are based upon the existence of innate modules, four for Mencius and five for SIM. With an emphasis on emotionally activated intuitions as the foundation for moral judgments, SIM has developed into one of the more empirically supported models of moral cognition. In addition to comparing these modular theories, Flanagan and Williams evaluate whether the normative claims for Darwinian-based modular theories alter our understanding of the *ought* of ethics. In “What Does the Modularity of Morals Have to Do With Ethics? Four Moral Sprouts Plus or Minus a Few,” they explain the *is-ought* problem from a new perspective and conclude that the *ought* of ethics has not yielded to the *is* of moral psychology.
Assembling a system from the bottom-up that is capable of accommodating moral considerations draws attention to the importance of a wide array of mechanisms in honing moral intelligence. “A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents” outlines a comprehensive, apparently modular, but actually quite interactive, approach to moral decision making. The cognitive model Wendell Wallach, Stan Franklin, and Colin Allen propose places moral decision making within the broader context of an agent’s general capacity to make decisions. Stan Franklin and colleagues at the University of Memphis are developing LIDA, a computer-based model of human-like cognition. They are particularly concerned with insuring that the model is compatible with recent research findings by neuroscientists. In this essay, Franklin collaborates with Wallach and Allen to explore how a LIDA-based agent might be capable of accommodating moral considerations in its choices and actions.

“The Linguistic Analogy: Motivations, Results, and Speculations” returns to the theme of teasing out an innate moral faculty. The contention that there is an innate moral grammar (Hauser, 2006; Mikhail, 2000; Rawls, 1999, [1971]) is inspired by Noam Chomsky’s (1985 [1957]) generative linguistic grammar. Susan Dwyer, Bryce Huebner, and Marc Hauser provide an overview of research to date directed at revealing the small set of implicit rules and structures that would make up a moral equivalent of a linguistic grammar. These rules and structures are viewed as being the computational and representational capacity that structure the making of moral judgments. They conclude by arguing that adoption of the linguistic analogy as a working hypothesis provides a particularly plausible strategy for making a moral organ the object of rigorous empirical study.

In “Moral Reasoning: Hints and Allegations,” Joseph Paxton and Joshua Greene review some recent research that indicates certain forms of moral reasoning, including the application of deontological and consequentialist moral principles, can play a significant role in overriding or defusing questionable moral intuitions and biases. They contend that the “dual-process” theory of moral judgments, which Greene has formulated with colleagues (Greene, 2007; Greene et al., 2001, 2004), differs in two respects from SIM, and is more in line than SIM with this recent evidence regarding the role of moral reasoning. Greene’s theory posits two different modes of moral thinking. One mode driven by emotional responses naturally supports judgments that are usually considered to be about rights and duties, while the other entails controlled cognitive processes that determine what course of action promotes the “greater good.”

Rationality is often supported by simple heuristics that are satisficing rather than optimizing (Gigerenzer & Selten, 2002; Gigerenzer et al., 1999; Muramatsu & Hanoch, 2005; Simon, 1955). Gerd Gigerenzer’s essay in this issue extends his model of heuristics as a fast and frugal toolbox of satisficing solutions to decisions that are explicitly about moral considerations. He asks, “What vision of moral behavior emerges from the perspective of bounded rationality?” In “Moral Satisfying: Rethinking Behavior as Bounded Rationality,” Gigerenzer emphasizes that “behavior is a function of both mind and environment” while challenging the contention that humans have a special moral grammar. His ecological approach points toward improving behavior by changing environments rather than through policies that attempt to alter beliefs or inner virtues. In concluding comments, Gigerenzer
discusses how his model of morally satisficing heuristics differs from that of contributors to a recent volume on the subject (Byron, 2004).

The idea that people represent the relationships between actions in terms of action trees has become a helpful device for unpacking the way in which moral decisions are made. Joshua Knobe probes the presumption that preexisting action trees are used to make moral judgments. In "Action Trees and Moral Judgment," Knobe, whose work has been closely identified with the development of experimental philosophy, discusses research which indicates that people’s moral judgments have an impact on the very structure of the action tree itself. But Knobe notes that researchers have also demonstrated that the action trees people construct can influence their moral judgments. While there may be a number of plausible theoretical models as to how such different findings can be reconciled, determining which of these possible solutions is correct will be a topic for future research.

Alan Sanfey and Trevor Kvaran are also interested in proposing a fruitful trajectory for future research. Their article, "Towards an Integrated Neuroscience of Morality: The Contribution of Neuroeconomics to Moral Cognition," outlines research to date in both neuroeconomics and in the neuroscience of morality. Both fields deal with the neural processes that underlie decision making, while differing largely on the values that inform the decisions being made. The authors propose that more explicit attention to the role of values and norms in economic decision making will facilitate the building of more accurate models in neuroeconomics. Furthermore, adopting the mathematical models favored by neuroeconomists might help moral psychologists overcome methodological limitations that have hindered that field’s progress. Sanfey and Kvaran conclude by describing a few recent studies that integrate approaches from both fields of research.

It goes without saying that summaries always fail to captures the richness of articles and the subtleties of arguments, but perhaps they will serve as enticements to read the full text.

5. Conclusions

The jury is out on whether ethics can or should be naturalized, on what descriptions of neural correlates actually explain, or on the determinative character of unconscious processes and the kind of conscious control people have over their choices and actions. Debate arising from such issues energizes the study of moral psychology while demanding increasing sophistication in the design of experiments and in the application of research findings.

While a rich field of interdisciplinary inquiry has emerged, the cognitive models of moral decision making that have been developed are quite sketchy. Readers of the articles in this volume may at times be under the impression that the scientists studying moral psychology are similar to the blind men studying different features of an elephant. Even the comprehensive framework outlined by Wallach, Franklin, and Allen is far from being instantiated. Furthermore, any comprehensive framework would require considerable work to accommodate the details of a moral grammar, the mechanisms through which emotions give form to intuitions, and the role of reason, rules, and principles in the making of moral decisions.
presuming that these have been substantially worked out. In other words, the individual
cognitive models discussed in this volume each require considerable enrichment, and then
need to be integrated into one cognitive model. There are exciting years of research ahead.
The development of a comprehensive model of moral decision making lies in the hopefully
not-too-distant future.

References

Anderson, S. W., Antoine, B., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and
Armstrong (Ed.), *Moral psychology volume 3: The neuroscience of morality: Emotion, brain disorders, and
Cambridge University Press.
MIT Press.
(Fall 2008 edition), Edward N. Zalta (Ed.). Available at: http://plato.stanford.edu/entries/moral-psych-emp/
Flack, J., & de Waal, F. (2000). ‘Any animal whatever’: Darwinian building blocks of morality in monkeys and
psychology, volume 1: The evolution of morality: Adaptations and innateness* (pp. 1–26). Cambridge, MA:  
MIT Press.
Neuroscience, 2*(10), 861–863.
dition.
Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* New
York: Oxford University Press.
judgment.* Cambridge, UK: Cambridge University Press.


