Is Morality an Elegant Machine or a Kludge?[1]

Stephen Stich

In a passage in *A Theory of Justice*, which has become increasingly influential in recent years, John Rawls (1971) noted an analogy between moral philosophy and grammar. Moral philosophy, or at least the first stage of moral philosophy, Rawls maintained, can be thought of as the attempt to describe our *moral capacity* – the capacity which underlies "the potentially infinite number and variety of [moral] judgments we are prepared to make." In order to describe that capacity, we must formulate "a set of principles which, when conjoined with our beliefs and knowledge of the circumstances, would lead us to make [the judgments we actually make] were we to apply these principles conscientiously and intelligently." (Rawls 1971, 46) Citing Chomsky's *Aspects of the Theory of Syntax* (1965), Rawls goes on to suggest that describing our moral capacity can be usefully compared to "describing the sense of grammaticalness that we have for the sentences of our native language. In this case the aim is to characterize the ability to recognize well-formed sentences by formulating clearly expressed principles which make the same discriminations as the native speaker." (Ibid. 47)

Though Rawls' analogy between our moral capacity and a Chomskian grammar of our language was new and insightful, the project of describing our moral capacity – of finding a set of principles (or rules or definitions) which entail the potentially infinite number of judgments we actually make – has a history that goes back to Plato. In *The Republic*, Socrates asks Cephalus to provide an account of justice, and then challenges the

[1] I am grateful to Daniel Kelly, Joshua Knobe, Edouard Machery and Ron Mallon for helpful comments on an earlier draft of this paper.

account Cephalus offers by showing that it does not coincide with the judgment we are actually inclined to make about what justice requires in the rather far-fetched case of a man who has stored his weapons with a friend and then asks for them back when he is not in his right mind. (*The Republic*, 1, 131) What unites the Socratic quest for "definitions" and Rawls' project of describing our moral capacity is the assumption that there is *an integrated body of information in the mind* – perhaps something like a generative grammar – which underlies our moral judgments about a potentially infinite number of cases, many of which are unlike anything we have previously encountered.

Though Rawls did not push the analogy between the moral capacity and grammar much further than this, other writers have been bolder. According to Chomsky and his followers, the grammars of all natural languages share important features in common, and since the linguistic information available to children is not rich enough to enable them to learn these features, knowledge of the linguistic universals must be innate. Of course, Chomsky does not deny that there are some very important *differences* among grammars as well. To explain these, Chomsky has suggested that some of the linguistic universals are *disjunctive* – they specify two or more different patterns that some aspect of grammar can exhibit. Which pattern is actually incorporated in the grammar that a child acquires is determined by cues in the child's linguistic environment. But these cues function only to trigger the adoption of one or another innate pattern; they do not contain any information about the pattern. If the moral capacity is like grammar, many writers have noted,[2] we should expect that the moral capacities of all normal humans to have important features in common. And just as knowledge of linguistic universals is innate, so too is knowledge of moral universals. Deep moral differences, if they exist, can be accounted for by the hypothesis that some moral universals, like some linguistic universals, have a disjunctive structure, specifying two or more different patterns that some aspect of the moral rule system can exhibit. Environmental cues encountered during development would determine which of these options is actually incorporated in the "moral grammar" that an adult ultimately acquires.

---

[2] Stich (1993); Harman (1999); Dwyer (1999); Mikhail (forthcoming); Hauser (forthcoming).

Though Chomsky himself, until recently at least, was skeptical of attempts to offer evolutionary explanations of grammar, others have not shared his skepticism. If all humans share an innate, integrated body of linguistic information, they argued, this can hardly be an evolutionary accident. There must be some account of how natural selection, sexual selection and the other processes driving evolution led to this extraordinary innate endowment. And a number of theorists have developed and defended hypotheses about this process. (Dunbar 1988; Pinker & Bloom 1990; Miller 2000, ch. 10) Many have also argued that if the moral capacity is like grammar – or even if it isn't – there must be some account of why this extraordinary psychological capacity evolved. There is a large literature exploring and defending accounts of the evolution of morality. While no brief description of this literature could possibly do it justice, I think it is fair to say that the two most popular ideas are that kin selection and cooperation (or "reciprocal altruism") have played a central role in the evolution of morality. (Trivers 1971; Alexander 1987; Wright 1994; Joyce 2006)

The central ideas of what I propose to call the *elegant machine* view of moral psychology are the three theses I have just sketched.

1. Our moral capacity is subserved by an integrated set of rules or principles which, like the rules of grammar, are designed to work smoothly together.
2. Important features of those rules or principles are innate, and deep differences in moral opinion are to be explained by the fact that some of the innate moral universals are disjunctive.
3. The existence of those innate moral universals has an adaptive explanation.

In depicting the elegant machine view I have, of course, been painting with a very broad brush. Nonetheless, I think (1)-(3) do provide at least a rough sketch of an important position in empirical moral psychology, a position which attracts many people, though some theorists who are broadly sympathetic would emphasize some features of the elegant machine view and downplay others.[3]

[3] The name I've chosen for the view is borrowed from an oft-quoted passage in

While there is much to admire – and to debate – in the papers collected in this issue of *Cognition & Culture*, what I find most striking is that two of them, the papers by Wellman & Miller and by Nichols, suggest a dramatically different view of moral psychology. These papers, and some of Nichols' related work, pose a serious challenge to each component of the elegant machine view. On the alternative view, morality is not subserved by a well integrated collection of rules but by a hodgepodge of psychological mechanisms that are often in competition with one another. While some of these psychological mechanisms may themselves be innate, much of the information that they contain is acquired from the surrounding culture. Culture, on this view, provides much of the *content* of the rules or principles that mechanisms underlying morality store and use.

Thus much moral disagreement is to be explained by appeal to cultural differences rather than by the triggering of one or another branch of an innately specified disjunctive moral universal. Moreover, many of the most important facts about the human moral capacity do not have an adaptive explanation. Rather morality, or at least these central features of morality, are an evolutionary accident. The account of moral psychology that denies all three central elements of the elegant machine view is what I propose to call the thesis that morality is a *kludge*. Though my account of the kludge thesis, like my account of the elegant machine view, is only a rough sketch, I think these two sketches mark out opposite poles in one of the most important theoretical debates in contemporary moral psychology. While the truth may end up somewhere between these two poles, it is my hunch that it will be closer to the kludge account. What I propose to do in the remainder of this commentary is to note some of the ways in which Wellman & Miller's paper and Nichols' encourage this hunch.

---

which Tooby & Cosmides describe the mind as "a confederation of hundreds or thousands of functionally dedicated computers . . . designed to solve adaptive problems endemic to our hunter-gatherer ancestors" and go on to describe these devices as "elegant machines". (Tooby & Cosmides, 1995, pp. xiii-xiv) In fairness to Tooby and Cosmides, I should note that it is far from clear that they are advocates of the elegant machine view of *moral psychology*.

Wellman & Miller report a cluster of findings that "highlight the role of culture in impacting the paths and endpoints of developmental change" (p. 47) as children acquire their moral competence. They emphasize that psychological knowledge, which for them includes knowledge of social roles and moral expectations, results in part from "culturally framed appreciations of social reality" (p. 48). And while they are not advocates of "Pure One-sided Socialization" which rejects nativist accounts entirely, they maintain that "only a very general framework" is innate, and that the framework "does not specify exactly what sorts of intentions (to harm, to help, to satisfy) nor exactly what sorts of responsiveness (duty, self-satisfaction, helping others) are important or how they are seen to interrelate" (p. 45). Moreover, the innate framework itself is not static; it "can and does revise and change." (p. 45) Obviously, this perspective is far removed from the Chomsky-inspired version of nativism found in the second thesis of the elegant machine view. As children develop, Wellman and Miller insist, culture plays a substantive role in shaping and providing content to their moral outlook, rather than simply cueing or triggering one or another innate package of moral rules. Though it is less clear cut, I am inclined to think that Wellman & Miller's paper also poses a challenge to the first component of the elegant machine view – the idea that morality is subserved by an integrated system of rules that work smoothly together. Wellman & Miller emphasize that many aspects of folk psychology are deeply influenced by culture, and they suggest that moral judgments are influenced not just by culturally transmitted moral rules but also by culturally variable inclinations to emphasize or de-emphasize social roles in perceiving and explaining behavior.

Nichols' paper is more emphatic in challenging the idea that our moral competence depends on a single, well integrated system of rules. Rather, he maintains, the data indicate that "the folk view is fractured" (p. 83). "[F]olk intuitions do not present a coherent theory of agency or responsibility. On the contrary, the folk seem to have inconsistent intuitions about agency and responsibility." (p. 81) Moreover, these intuitions are generated by "a strikingly multifarious set of underlying mechanisms" including the mindreading system, a cluster of affective mechanisms, and another cluster of "cold cognitive processes" (78) that may include a commitment to the Kantian maxim that "*ought* implies *can*" and early

emerging beliefs that people have obligations and "*ought* to behave in certain ways" (80). Nichols emphasizes the tentative nature of his hypotheses about the mechanisms underlying the apparently inconsistent intuitions that his research has uncovered. But if anything in the general vicinity of his hypotheses turns out to be correct, it would dramatically undermine the analogy between moral competence and grammatical competence as the latter is understood in the Chomskian tradition. Indeed, if Nichols' account of the mechanisms underlying people's intuitions on moral matters is on the right track, then perhaps the most natural thing to say is that people don't have a moral competence at all.

Though Nichols says little about the role of culture in this paper, in other recent work on moral psychology he has championed an epidemiological approach to morality modeled on the epidemiological account of cultural phenomena pioneered by Dan Sperber (1996). This approach to culture begins with a pair of observations. The first is that in all human cultures a great deal of information is passed from one person to another. The second is that this information transfer is mediated by a variety of psychological mechanisms, some of which may be innate. In order to imitate a dance or a hunting technique, to learn a folk tale or to master a religious ceremony, the learner (or "cultural child" as he or she is sometimes called) must observe more knowledgeable members of the culture ("cultural parents"), infer or reconstruct the intentions and other mental representations that underlie their behavior (including their verbal behavior), and store the reconstructed mental representations in the appropriate place in memory. Neither the mechanisms that underlie the necessary inferences nor those that underlie memory are perfectly accurate, however. They are bound to make mistakes, and those mistakes will often not be random. Rather, because of the way the mechanisms responsible for inference and memory are designed, the mental representations that are reconstructed and stored are more likely to selectively retain some features of the cultural parents' representations, to drop others, and to introduce new features that may not have been present in the cultural parents' representations. The features that are more likely to be retained or added might be thought of as biases or attractors in the transmission process, and over time the transmitted mental representations found in a population will tend to move in the direction of those attractors.

One influential example of research that adopts the epidemiological approach is Pascal Boyer's work on religion (Boyer, 2001). Boyer has shown that people's beliefs about supernatural beings tend to characterize those beings as having just one, or a small number, of bizarre and unfamiliar properties, and otherwise to be pretty much the same as natural beings in that category. Thus a supernatural person may be able to know what is happening in distant places or what will happen in the future, but apart from this his mind will have all the normal characteristics posited by commonsense or folk psychology. The reason for this, Boyer argues, is that the small number of "supernatural" properties make the representations of these beings particularly memorable, while the more mundane features of the supernatural agent's mind are supplied automatically, when people hear accounts of these beings, by the innate mental mechanism that is responsible for attributing mental states to real people.

Nichols' work on etiquette norms provides another illustration of the epidemiological strategy. Nichols has shown that, while a wide variety of behavior has been governed by etiquette norms during the last 500 years, the norms that tend to survive, once they appear, are those that prohibit behavior that evokes innate disgust reactions. The innate disgust mechanism, Nichols argues, biases the transmission process in favor of norms prohibiting disgusting behavior by making those norms more salient and more memorable. (Nichols, 2004) In a related study Nichols has argued that similar processes underlie the transmission and retention of moral norms. In the case of these norms, the affective response to suffering plays the role that disgust plays in the transmission of etiquette norms. Though many different sorts of behavior have been prohibited by moral norms, when rules appear which prescribe transgressions that cause pain and suffering, and thus evoke strong affective reactions, they are more likely to survive than rules whose transgression does not evoke an affective response. Nichols calls these norms "harm norms" and he maintains that harm norms "gained an edge in cultural fitness by prohibiting actions that are likely to elicit negative affect." (2004, p. 154)

The epidemiological approach rejects the contention, central to the elegant machine view, that deep moral differences are restricted to the innate options specified by disjunctive universals. Rather, it insists, a wide variety of behaviors can be, and have been, prescribed by moral rules

at various times in different cultures. Some of those rules are relatively short lived, however. The ones that survive and spread are those that happen to invoke affect or have some other property that biases the transmission process.

Although describing this approach to the explanation of cultural phenomena as "epidemiological" is a metaphor, it is in many ways a very apt metaphor. The norms and other mental representations that are spread by the sorts of processes that are center stage in the epidemiological approach, like the infectious agents tracked by medical epidemiologists, need not contribute anything to their hosts' reproductive success. Those that succeed in spreading through a population do so by exploiting features of their hosts cognitive systems that were designed for very different purposes. The mindreading system that explains why supernatural beings are believed to have familiar psychological properties did not evolve because it enabled people to create religious myths. The core disgust system was presumably in place long before the emergence of rules of etiquette. And the affective response to suffering arguably predates the emergence of morality (DeWaal, 1996, Ch. 2). The epidemiological approach thus gives us insights into some of the quirks of culture, and some of its pathologies. It explains how mental mechanisms which were designed to deal with adaptive problems can, inadvertently as it were, give rise to an efflorescence of cultural phenomena which often contribute nothing to fitness. To the extent that widespread and important aspects of morality can be accounted for by the epidemiological approach, the third component of the elegant machine view of morality will be undermined. Moral universals, if they exist, may not have spread throughout the ancestral population because they contributed to fitness. Rather, like many other widespread memes, they spread and became entrenched because they found ways of exploiting quirks in the process of cultural transmission.

REFERENCES

Alexander R.
1987 *The Biology of Moral Systems.* New York: Aldine De Gruyter.
Boyer, P.
2001 *Religion Explained: The Evolutionary Origins of Religious Thought.* Basic Books.

CHOMSKY, N.
1965 *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
DE WAAL, F.
1996 *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.
DUNBAR, R.
1988 *Grooming, Gossip and the Evolution of Language*. London: Farber & Farber.
DWYER, S.
1999 Moral competence. In *Philosophy and Linguistics*, ed. by K. Murasugi & R. Stainton. Westview Press.
HARMAN, G.
1999 Moral philosophy and linguistics. *Proceedings of the 20th World Congress of Philosophy*, vol. I: *Ethics*, ed. by Klaus Brinkmann. Bowling Green, OH: Philosophy Documentation Center. 107-115.
HAUSER, M.
forthcoming *Moral Minds: The unconscious Voice of Right and Wrong*. New York: Harper Collins.
JOYCE, R.
2006 *The Evolution of Morality*. Cambridge, MA: MIT Press.
MIKHAIL, J.
forthcoming *Rawls Linguistic Analogy*. Cambridge: Cambridge University Press.
MILLER, G.
2000 *The Mating Mind*. New York: Doubleday.
NICHOLS, S.
2004 *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
PINKER, S. & BLOOM, P.
1990 Natural language and natural selection. *Behavioral & Brain Sciences*, 13, 707-784.
RAWLS, J.
1971 *A Theory of Justice*. Cambridge, MA: Harvard University Press.
SPERBER, D.
1996 *Explaining Culture: A Naturalistic Approach*. Oxford: Blackwell.
STICH, S.
1993 Moral philosophy and mental representation. In *The Origin of Values*, ed. by M. Hechter, L. Nadel & R. E. Michod. New York: Aldine de Gruyter. 215-228.
TOOBY, J. & COSMIDES, L.
1995 Foreword. In S. Baron-Cohen, *Mindblindness*. Cambridge, MA: MIT Press.
TRIVERS R.
1971 The evolution of reciprocal altruism. *Quarterly Journal of Biology*, 46, 35-57.
WRIGHT, R.
1994 *The Moral Animal*. New York: Pantheon Books.