Analyse & Kritik 28/2006 (© Lucius & Lucius, Stuttgart) p. 95–98

Fiery Cushman/Liane Young/Marc Hauser

The Psychology of Justice

Abstract: In *Natural Justice* Binmore offers a game-theoretic map to the landscape of human morality. Following a long tradition of such accounts, Binmore's argument concerns the forces of biological and cultural evolution that have shaped our judgments about the appropriate distribution of resources. In this sense, Binmore focuses on the morality of outcomes. This is a valuable perspective to which we add a friendly amendment from our own research: moral judgments appear to depend on process just as much as outcome. What matters is not just that the butler is dead, but who killed him, how, and for what reason. Thus, a complete understanding of natural justice' will entail an account not only of evolutionary pressures, but also of the psychological mechanisms upon which they act.

"What should we be aiming for?" asks Binmore at the outset of *Natural Justice*. At the broadest level, Binmore has set out to recapture morality as a natural phenomenon, divorcing it from metaphysical argumentation "conjured from nowhere". More concretely, Binmore attempts to provide a user-friendly version of his game theoretic account of the biological and cultural evolution of fairness. Frequent detours are permitted for critical treatments of perceived ideological opponents ranging from Kant to Gould to religious leaders and political conservatives; these find their counterweight in praises bestowed upon his ideological heroes: Hume, Nash, Rawls, and Harsanyi. The final chapter contains a prescription for social reform with a healthy dose of game theory. Binmore's ambition is quite plainly to bridge the gap (in his opinion, illusory) between studying moral decision-making and making moral decisions.

In this brief essay we attempt to lay Binmore's game-theoretic analysis of morality side-by-side with the current state of knowledge in moral psychology, including our particular approach to the problem. In doing so we eschew the matter of bridging descriptive claims and normative ones—for the time being, we shall be content merely to learn something about the proper description. We will focus our attention on a class of phenomena of particular concern to moral psychology that is not easily captured in Binmore's framework: concerns with the moral status of actions themselves, as opposed to the consequences of actions.

Binmore's focus on consequences is apparent in the very first sentence mentioned above: "What should we be aiming for?" Throughout his book he is concerned with explaining the fairness norms that emerge to solve questions of the distribution of resources, and considers how biological and cultural evolution act to shape these norms. The range of goods over which such fairness norms can be applied of course extends far beyond money, food, or medicine; they could

apply equally to lives, grooming, or the turns taken with toys. Nevertheless, questions of distributive justice are always questions of what we should aim for, not how we should be aiming.

Binmore's focus on consequences is perfectly in keeping with the long line of research into the evolution of cooperation and fairness in which he follows (Alexander 1987; Axelrod 1984; Fehr/Fischbacher 2003; Trivers 1971), and it is not without good reason that the field has developed in that direction. Firstly, evolution is directly driven by fitness consequences, and only indirectly by the processes through which they emerge. Secondly, the attempt to provide a mathematical foundation for fairness and cooperation is far more likely to succeed when applied to quantifiable outcomes, rather than qualitative descriptions of processes.

So why pay attention to processes? A simple pair of, for many, highly familiar scenarios will illustrate the point. Case 1: Denise is on a trolley when the conductor goes unconscious. The trolley is heading towards five people on the main track where it will hit and kill them. Denise's only course of action is to flip a switch, sending the trolley down a side track where it will hit and kill one individual. Denise flips the switch. Case 2: Frank is standing by the trolley tracks when he witnesses a trolley running out-of-control towards five people. Frank's only course of action to save the five is to push a fat man next to him onto the tracks, killing the man but slowing the trolley sufficiently to save the five. Frank pushes the man.

Philosophers have traditionally used their intuitions about scenarios like these to make arguments about the moral permissibility of actions (Foot 1967; Kamm 1998). Regarding Denise and Frank, many philosophers argue that Denise is justified in saving the five, while Frank is not. We put these intuitions to an empirical test, using the Web to survey thousands of individuals from diverse cultural backgrounds (Hauser et al., in press). The results were unequivocal: nearly 90% of subjects judged Denise's action to be permissible while a mere 10% of subjects judged Frank's action to be permissible. Although subjects were all fluent in English and had access to the Web, they also differed in religious background, exposure to moral philosophy, educational level, and ethnicity. The observed difference between Frank and Denise was consistent among all of these groups.

The important point is that process counts. The outcome of Frank's action and Denise's action is the same, but the means employed by Frank and Denise drive moral intuitions in strikingly different directions. Understanding the psychological mechanisms behind moral decision-making entails a study not only of people's sensitivity to outcomes, but also their sensitivity to the different sorts of actions, intentions, and causes that jointly contribute to the consequence in question.

Following up on our study of Frank and Denise, we developed a set of moral dilemmas arranged into tightly controlled pairs, with each item in a pair differing only by one critical dimension. This allowed us to test individuals' sensitivity to precisely defined moral principles. All three of the principles tested turned up robust effects: subjects judged harmful actions worse than harmful omissions

(*action principle*), harms caused by physical contact worse than harms caused without physical contact (*contact principle*), and harms employed as a means worse than harms produced as a side-effect (*intention principle*) (Cushman et al. in press). Of course, it should be noted that these principles represent just one small province of a potentially vast territory of moral principles (Hauser in press; Mikhail et al. 2002). Important work by others is illuminating other corners of the map, including especially the role of emotions (Greene et al. 2004; Wheatley/Haidt 2005).

Our current work reveals not only that process counts but that different processes may, in fact, count differently. Above, we identified three principles (action, contact, intention), all capturing non-consequential distinctions to which people are sensitive when making moral judgments. These principles are therefore operative in guiding moral judgment, as evidenced by subjects' patterns of judgments. Importantly, these principles were not all equivalently expressed in subjects' justifications of their judgments. The intention principle rarely emerged, while the action and contact principles emerged in the majority of subjects' justifications; further, when subjects noted the contact principle, they also frequently denied that it should carry any moral weight, a comment that rarely emerged for the action principle. We can infer from these results that the intention principle is operative but results in intuitive judgments, whereas the action and contact principles are also operative but appear to be accessible for use in conscious reasoning as evidenced by their emergence in justifications. These results therefore suggest that principles underlying our moral decision-making may operate differently; that is, both conscious reasoning and intuition may contribute to our moral judgments. A more precise characterization of the underlying mechanisms is the subject of ongoing research.

Deriving principles like those we discuss from game-theoretic accounts of distributive justice is not easy. It might be argued that people's preferences for certain processes can be folded into their personal utility function in a quantifiable manner; this is the approach traditionally taken by consequentialist thinkers against the intuitions of their deontologist critics. That argument works fine for philosophers, but it won't do the job for an evolutionary account of fairness norms because it begs the question of where deontological preferences come from in the first place. That is, if we are attempting to explain the source of our moral intuitions, it would be a slight of hand to do so by appealing to a bargaining process that operates over a set moral intuitions!

Of course, there are other solutions to this apparent quandary. Although evolution may be guided by consequences alone, it can select for psychological mechanisms that are sensitive to non-consequentialist features so long as these mechanisms typically result in beneficial consequences. Evolution will always be consequentialist at the ultimate level of selective pressures, but it may give rise to proximate mechanisms that are very non-consequentialist in nature. Alternatively, some mechanisms may be deployed in the context of delivering a moral judgment, but their evolutionary origins may have been selected for non-moral functions, including more general decision-making problems. For example, though our moral judgments depend critically upon our capacity to attribute

intentions and desires to others, mental state attribution, or theory of mind (Baron-Cohen 1995; Frith/Frith 2003; Leslie 1987), is not selectively deployed in the context of moral dilemmas.

Developing a full picture of the evolution of morality, then, requires a careful study not only of ultimate evolutionary forces—the primary focus of Binmore's research program—but of proximate psychological mechanisms, including both their evolutionary and developmental origins. Meanwhile, those of us engaged in the study of psychological mechanisms have much to learn from the many recent models developed of the biological and cultural evolution of fairness and cooperation. *Natural Justice* is a thoughtful and forceful addition to this dialogue, and we look forward to a future in which evolutionary theory and psychological mechanisms contribute jointly to our understanding of the moral mind.

Bibliography

Alexander, R. D. (1987), *The Biology of Moral Systems*, New York

Axelrod, R. (1984), *The Evolution of Cooperation*, New York

Baron-Cohen, S. (1995), *Mindblindness*, Cambridge/MA

Binmore, K. (2005), *Natural Justice*, Oxford-New York

Cushman, F. A./L. Young/M. D. Hauser (in press), The Role of Conscious Reasoning and Intuitions in Moral Judgment: Testing Three Principles of Harm, in: *Psychological Science*

Fehr, E./U. Fischbacher (2003), The Nature of Human Altruism – Proximate and Evolutionary Origins, in: *Nature 425*, 785–791

Foot, P. (1967), The Problem of Abortion and the Doctrine of Double Effect, in: *Oxford Review 5*, 5–15

Frith, U./C. D. Frith (2003), Development and Neurophysiology of Mentalizing, in: *Philosophical Transactions of the Royal Society of London. B 358*, 459–473

Greene, J. D./L. E. Nystrom/A. D. Engell/J. M. Darley/J. D. Cohen (2004), The Neural Bases of Cognitive Conflict and Control in Moral Judgment, in: *Neuron 44*, 389–400

Hauser, M. D. (in press), *Moral Minds: How Nature Designed a Universal Sense Right and Wrong*, New York

— /F. A. Cushman/L. Young (in press), A Dissociation Between Moral Judgment and Justification, in: *Mind and Language*

Kamm, F. M. (1998), *Morality, Mortality: Death and Whom to Save From It*, New York

Leslie, A. M. (1987), Pretense and Representation: The Origins of 'Theory of Mind', in: *Psychological Review 94*, 412–426

Mikhail, J. M./C. Sorrentino/E. Spelke (2002), *Aspects of the Theory of Moral Cognition: Investigating Intuitive Knowledge of the Prohibition of Intentional Battery, the Rescue Principle, the first Principle of Practical Reason, and the Principle of Double Effect*, unpublished manuscript, Stanford

Trivers, R. L. (1971), The Evolution of Reciprocal Altruism, in: *Quarterly Review of Biology 46*, 35–57

Wheatley, T./J. Haidt (2005), Hypnotic Disgust Makes Moral Judgments More Severe, in: *Psychological Science 16(10)*, 780–784