Analyse & Kritik 28/2006 (© Lucius & Lucius, Stuttgart) p. 83–88

Fabienne Peter

Justice: Political Not Natural

Abstract: Ken Binmore casts his naturalist theory of justice in opposition to theories of justice that claim authority on the grounds of some religious or moral doctrine. He thereby overlooks the possibility of a political conception of justice—a theory of justice based on the premise that there is an irreducible pluralism of metaphysical, epistemological, and moral doctrines. In my brief comment I shall argue that the naturalist theory of justice advocated by Binmore should be conceived of as belonging to one family of such doctrines, but not as overriding a political conception of justice.

A political conception of justice, as famously put forward by John Rawls, rests on fundamental democratic values. The premise is that an irreducible pluralism of views about what justice requires and about what constitutes the relationship between individuals and the society they live in renders it impossible to base justice on any single comprehensive philosophical doctrine. In my brief comment I shall argue that the naturalist theory of justice advocated by Ken Binmore should be seen as belonging to one family of such doctrines. Naturalist theories are not written by nature, but are scholarly attempts to reflect on a select set of data about social life. They are part of a particular (and venerable) philosophical tradition of thinking about justice. The theories put forward are contested by fellow naturalists as well as by adherents of other philosophical traditions. I agree with Binmore that we should theorize about how the social world is structured and, based on this, about what constitutes justice. But he interprets this endeavor too narrowly. I shall argue naturalist theories go wrong when they are conceived of as overriding a political conception of justice. I find Binmore's book very intelligent and I would recommend it to everyone as an extremely stimulating and enjoyable read. But contrary to what he suggests, it is best interpreted as only one among many contestable philosophical doctrines about justice.

Binmore is interested in the distributive aspects of social contracts. He takes the traditional approach in moral philosophy to be to advocate a particular sharing rule—a theory of justice—without asking how it could have come about and be sustained in actual societies. This, he contends, is a hopeless and futile endeavor. According to him (and many others), justice happens behind our backs. Moral rules, so the claim goes, are shaped by evolutionary forces and they should be studied accordingly:

> "If one wishes to study such rules, it doesn't help to ask how they advance the Good or preserve the Right. One must ask instead how they evolved and why they survive. That is to say, we need to treat morality as a science." (1)

Naturalist theories of justice seek to explain why particular sharing rules have evolved. Binmore's main thesis is that John Rawls' justice as fairness is right in some of its conclusions, but for the wrong reasons. According to Binmore, Rawls mistakenly adheres to the traditional approach in moral theory, but his device of the original position receives support from an evolutionary perspective. In the book he argues that the original position reveals the "common deep structure of human fairness norms" (15).

For my purposes here it is not important to discuss the details of the convergence that Binmore makes out between his naturalist approach and Rawls' theory of justice as fairness. It suffices to say that he associates Rawls' idea of the original position with an equilibrium selection device in rational bargaining. By combining the folk theorem of game theory with an evolutionary account of fairness norms, Binmore argues that a fair social contract can emerge as an equilibrium in repeated games (chapter 11). Using a Humean metaphor, he argues that a fair social contract is like a masonry arch: it holds together without the glue of moral authority since it succeeds in coordinating individual behavior on a stable equilibrium.

Binmore criticizes social contract theories that focus on culture only and pleads for the need to study them within the constraints of human biology. But, as Binmore is well aware, accepting the constraints of biology does not imply that what matters with regard to justice is determined by biology. The concern with justice appears as a distinctively social phenomenon and thus needs to be studied as such. The question is how.

Every naturalistic approach rests on some model of human behavior that attempts to describe observed behavior and there is an inevitable gap between behavior itself and the model of behavior applied. Of course, there are better and worse models of behavior. Most people would agree, for example, that a model of behavior that postulates that behavior is the result of an explicit calculation of the benefits and costs of alternative actions is not a very plausible one. But the point is that there is scope for argument about which model of behavior is best. Different models will impose different frameworks of analysis. The selection of a particular model of behavior is thus not neutral and needs justification.

Binmore uses the standard economic model of behavior and defends its assumptions about rationality. According to this model, someone acts rationally if the behavior corresponds to the maximization of an expected utility function. Binmore rightly insists that the standard model leaves the content of the utility function unspecified (65). But this is not to say that the behavioral model is normatively neutral. As Binmore is aware, experiments in economics show that the consistency premise of expected utility theory is regularly violated, and one might argue that this evidence puts pressure on the premise. In addition, Binmore favors explanations that start from utility functions expressing nar-

row self-interest. He uses the folk theorem to show how reciprocal altruism can emerge from selfish behavior (79 ff.).

One argument Binmore gives in support of his approach is that motivations other than self-interest might be there but are scarce, especially when there is material deprivation or other kinds of discouraging influences. But the fact that certain circumstances favor certain behaviors in many of us does not show that the behavior that tends to be observed most frequently in the most adverse circumstances is the adequate benchmark for all behavior in all circumstances.[1]

And indeed, some behavioral economists criticize Binmore for using the standard model, and specifically for trying to do too much work with the self-interest assumption.[2] They point to findings which suggest not only that, in a wide range of circumstances, behavior tends to deviate from the maximization of self-interest, but also that there is a systematic structure to this observed behavior. Perhaps most importantly, they report widespread behavior based on other-regarding preferences. They conclude that a different explanatory model is required to account for such behavior. The model that Herbert Gintis, for example, favors is one that upholds the maximizing assumption but combines it with an empirically corroborated theory about the content of the utility functions.

Binmore seems unconvinced. According to him, "trying to explain human behavior in this way is rather like trying to explain the orbits of planets with Ptolemaic epicycles. You can get a really good fit if you juggle with enough epicycles, but what would be the point?" (75) He defends his approach with an argument from irrationality: a model of rational action must allow for the possibility of irrational action. He accuses those behavioral economists who seek to expand the standard model for neglecting this distinction and for failing to acknowledge that, often, observed behavior is simply irrational.

Binmore is right on this issue: there is no point in having a theory of rational behavior that does not allow for irrationality. But this move only reinforces my earlier point about the normative content of behavioral models. Binmore elevates the interpretation of rationality espoused by the standard economic model to an absolute benchmark and defends it stubbornly—notwithstanding the vast amount of evidence that suggests a need to rethink rationality. In comparison with the behavioral economists he criticizes, his approach can be accused of making the opposite mistake: it sticks to an overly narrow definition of rationality and insists on the irrationality of behavior that is at odds with his model. Let me call his approach dogmatic.

The approach favored by many experimental economists I would call 'nostalgic'. It is nostalgic in that it responds to objections to the economic model of rational behavior by tinkering with the substantive content of utility functions, while maintaining the formal structure of the standard model. These two approaches do not exhaust the possibilities, however. A progressive alternative takes account of Amartya Sen's 'Rational Fools' critique of the standard

[1] Binmore recognizes the scope for non-selfish behavior in family contexts or contexts that are perceived as such. Cf. his chapter 7.

[2] See, for example, the discussion of *Natural Justice* in Gintis 2006. In his review essay, Gintis argues that a naturalistic theory of justice should use a richer behavioral model.

model (Sen 1977). The critique distinguishes between three types of motivation: egoism, sympathy, and commitment. Binmore puts particular emphasis on the motivational category of egoism. Sympathy is the category that has experimental economists most interested. It captures behavior that is influenced by the effects other people's welfare has on the person's own. The difference between the two is that explanations based on sympathy violate what Sen (1985) calls the "self-centered welfare" premise of the standard economic model because they refer to other-regarding preferences. Explanations based on egoism respect this premise. And both types of explanations respect what Sen (1985) calls the "self-welfare goal" premise: behavior is linked to increases in one's welfare. The third category of possible motivations that Sen distinguishes, however, violates this premise as well. Committed action refers to a kind of behavior that is motivationally unrelated to the agent's welfare, however broadly conceived. The clearest case of action from commitment is when one feels compelled to intervene in a certain matter, even if doing so leaves one worse off. In other cases of committed action, there might not be a negative impact on one's welfare. What matters, however, is that increasing one's welfare is not the central motive. When one acts from commitment, one acts out of a sense that something is right or wrong; one has a reason for action that is not tied to personal gains and losses. Such reasons may stem from a sense of social identity, from social rules or from general principles one endorses.

These categories are first of all descriptive. Sen (1977) argues that all three types of motivations are regularly observed. But taking them seriously also points to an alternative behavioral model and thus has a normative component. I cannot go into the details of how a progressive approach, one that includes committed action, would interpret rationality—not last because this is still an open question (cf. Anderson 2001; Peter/Schmid (eds.) forthcoming). Let me merely point out what the main difference is between the dogmatic and the nostalgic approaches to rationality, on the one hand, and the progressive approach on the other. The former endorse while the latter rejects Humean skepticism about practical reason. Whereas the former treat ends—be they self- or other-regarding—as given to the individuals, the latter treats them as accessible to deliberation—both individual and collective. In Sen's words, "[r]ationality cannot be just an instrumental requirement for the pursuit of some given—and unscrutinized—set of objectives and values" Sen (2002, 39). According to the progressive alternative to the conservatism of the dogmatic and nostalgic approaches, rationality is not just about how to pursue given ends, but also about what to pursue.[3]

What I call a progressive approach suggests a political justification of the social contract. By political justification I mean a process of collective deliberation about the principles that shape the social contract. Such a political justification stands out against the evolutionary one that Binmore seeks to give using the standard economic model of behavior. Following Hume, Binmore argues that "a social contract can be seen as a largely unrecognized consensus to coordinate

[3] For an excellent discussion of the link between Sen's critique of rational choice theory and a moderate Kantian conception of practical reason, see Pauer-Studer 2006.

on a particular equilibrium of the game of life that we play together" (4). From the perspective of a progressive approach to behavior, one that covers individual and social deliberation about ends, this implicitness is suspect and unwarranted. People demand justification for the sharing rules adopted and are prepared to explain why they favor particular rules. In Rawls' words, a just society is not simply a stable society, but a society that is "stable for the right reasons" (Rawls 1999, 29).

Binmore casts his naturalist theory of justice in opposition to theories of justice that claim authority on the grounds of some religious or moral doctrine. He argues that unlike the "traditionalists", the naturalists "don't try to force their aspirations on others by appealing to some invented source of absolute morality" (19). I agree with Binmore that in thinking about justice, this should be avoided. But there is quite a gap in Binmore's argument between disputing the decisiveness of any moral authority and advocating social coordination via an unrecognized consensus. The behavioral approach he selects leads him to neglect the potential contestedness of any comprehensive philosophical doctrine about justice and the need to include such contestation at the heart of theorizing about justice. His approach allows him to analyze coordination as an unintended consequence of individual behavior. This is valuable as far as it goes. But it is inadequate to analyze coordination that results from deliberation about ends. Individuals are not just blind rule-followers. They can think about the principles that set the constraints to their actions, and they have views about what are good and what are bad principles. In short, Binmore's approach misses out on the deliberative aspect of the 'evolution' of social contracts.

To put the point differently: by focusing on the divide between naturalists and traditionalists, Binmore overlooks the possibility of a political conception of justice—a theory of justice based on the premise that there is an irreducible pluralism of metaphysical, epistemological, and moral doctrines. Taking this pluralism as his starting-point, Rawls argues for a conception of justice that is "political not metaphysical".[4] In making his case, Rawls' language sometimes echoes Binmore's. In *Political Liberalism*, for example, he writes: "given the fact of reasonable pluralism, citizens cannot agree on any moral authority, whether a sacred text, or institution. Nor do they agree about the order of moral values, or the dictates of what some regard as natural law." (Rawls 1993, 97) One might add, that neither them as a collective nor those among them who are scientists agree on an evolutionary account of the emergence of social and moral norms which could take the place of a moral authority. Binmore's belief in the deciding power of a science of morality thus represents one such comprehensive philosophical doctrine, but cannot provide the foundation for thinking about justice. To borrow Binmore's Humean metaphor: we should try to understand what holds an arch together and also, given this understanding, think about which arch we might want to build. No authority—scientific or moral—can relieve us of this task.

[4] Compare with the title of his 1985 essay "Justice as Fairness: Political not Metaphysical".

Bibliography

Anderson, E. (2001), Unstrapping the Straitjacket of 'Preference', in: *Economics and Philosophy 17*, 21–38

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

Gintis, H. (2006), Behavioral Ethics Meets Natural Justice, in: *Politics, Philosophy, and Economics 5(1)*, 5–32

Rawls, J. (1985), Justice as Fairness: Political not Metaphysical, in: *Philosophy and Public Affairs 14(3)*, 223–51

— (1993), *Political Liberalism*, New York

— (1999), *Law of Peoples*, Cambridge/MA

Peter, F./H. B. Schmid (eds.) (forthcoming), *Rationality and Commitment*, Oxford

Pauer-Studer, H. (2006), Instrumental Rationality versus Practical Reason: Desires, Ends, and Commitment, in: Peter/Schmid (eds.) forthcoming

Sen, A. (1977), Rational Fools: A Critique of the Behavioral Foundations of Economic Theory, in: *Philosophy and Public Affairs 6*, 317–344

— (1985), Goals, Commitment, and Identity, in: *Journal of Law, Economics, and Organization 1(2)*, 341–55

— (2002), *Rationality and Freedom*, Cambridge/MA