Ken Binmore
Natural Justice
Oxford and New York: Oxford University Press 2011
224 pages
US$24.95 (paper ISBN 978-0-19-979148-4)

Ken Binmore's *Natural Justice* is a worthwhile addition to the growing literature on the evolution and nature of fairness norms. Though the presentation is a bit uneven, and even downright grouchy at times, the material presented is worth careful study, and is considerably more accessible than his daunting two-volume *Game Theory and the Social Contract* (MIT Press, 1992/1994), of which this book serves as a sort of *précis*. The book will also be of interest to students of John Rawl's 'Justice as Fairness'.

Though this book appears to be a rather theoretical work, Binmore considers himself a social reformer in the tradition of Jeremy Bentham, or better yet, the 'Whigs' of seventeenth century Britain. Binmore presents a sustained and systematic presentation of his position on the large questions of social justice. He provides us with focused arguments toward his general conclusions, and the tools needed to present those arguments in a precise manner.

Bracketing Binmore's narrow, technical project are his broader motivation and his general message. The motivation will be familiar to fans of economics and decision and game theory. In brief, he insists that it makes no sense to plan or institute social reforms that are fundamentally unstable, that are 'inconsistent with human nature', which are not 'equilibria in the game of life'. This means that if you are serious about bringing about real social reform—rather than just grouching and marching—you need the tools of game theory and economics, in order a) to understand why things are the way they are (the structure of the current 'status quo' equilibrium), and b) to have any realistic expectation that there is an alternative equilibrium attainable by the proposed reform.

The general message, which comes at the end (and with relatively little warning), is that egalitarian reforms that are economically feasible and productive are possible only through what he calls 'planned decentralization'. The idea behind this is that one doesn't leave everything to market forces, simply hoping for the best based on the convictions of libertarian economists; nor does one plan and control reforms from a central government in the socialist manner. Rather, in the spirit of the 'mechanism design' field of economics, one plans a synergistically stable set of motivations and institutions that are self-regulating in the manner of classical markets, yet that are also somehow immune to their pathologies.

In between, Binmore presents his tools and makes his arguments. The formal tools are game theory (where one finds the Nash equilibria), bargaining theory (where agents choose among the set of feasible and possibly efficient equilibria), and the standard theory of rational choice underlying both. Less formal though no less central to his argument are evolutionary theory—in particular Hamilton's theory of kin selection

(accounting for our fairness instincts)—and the Rawls/Harsanyi concept of the Original Position (OP), from which a rational agent is imagined to choose what kind of society to live in, given that she doesn't know what station in life she will end up occupying.

Students of Rawls may find Binmore's use of the OP puzzling at first. Whereas Rawls used this theoretical device in a prescriptive way to appeal to *our* sense of fairness, Binmore approaches it as an economic tool—a scenario that just happens to appeal to everyone's sense of fairness. In other words, in Binmore's hands the OP is a predictive tool that helps us in the task of understanding why people accept as fair the things they do, and it helps us predict whether people will accept a new reform as fair, rather than being a tool to convince us of what is *really* fair.

The core of Bimore's argument is that Rawls very nearly gets the right 'egalitarian' conclusions about the OP, even though Rawls' analysis is faulty, since he 'denies orthodox decision theory' in assuming that choosers are extremely risk averse. He contrasts Rawls' reasoning to that of John Harsanyi, the other main proponent of the OP. Harsanyi concluded that rational choosers in the OP would calculate their expected utility in each society, and then choose the society with the highest average utility—the utilitarian solution. Binmore agrees that this is how a rational agent would calculate, but argues that the utilitarian solution is stable only in the presence of an absolute and omniscient enforcer. That is, if an agent chooses a society but is then, after the flip of the 'phantom coin', motivated to cheat on the outcome or to ask for a re-flip, the arrangement is not stable. Only in the presence of a God-like enforcer (or an absolutely binding sense of duty) would one abide by the rules of a utilitarian utopia in which one got the short end of the stick. But rational agents would not choose an unstable society. Rather, they would choose the best stable society, which turns out to be (roughly) one where no one would rather be someone else, all things considered—the egalitarian solution.

The 'all things considered' turns out to be a source of some difficulty, since (if you are an economist) it highlights the issue of interpersonal comparisons, e.g., how to measure your degrees of happiness given your allotment of resources against mine and everyone else's. But Binmore thinks this dragon is slayable, since people *do* guess at the happiness of others, even given that those others may have very different preferences. (These are called empathetic preferences.) Moreover, communities come to consensuses about who is happier than whom, given the preferences of each (empathy equilibria). These are separate from 'social indices', which let us accommodate the variety of ways in which real communities arrive at agreements that some people are more deserving than others, which gets calculated into various bargaining solutions.

Along the way, Binmore reiterates his conviction that our sense of fairness and the fairness norms that they express in different cultural contexts evolved as 'equilibrium selection devices'. This particular phrasing captures two characteristic concerns. The first was mentioned above, that one cannot actually select non-equilibrium outcomes, or in the evolutionary context, it cannot be the function of a social instinct to select unstable solutions to problems. The second is that, in game theory, many games have multiple equilibria, and one of the standard worries for rational choice theorists is finding some

non-arbitrary (rational) way of picking one of several equilibria as optimal or preferred, or anything other than arbitrary. Some theorists such as Brian Skyrms, have simply switched over to evolutionary game theory and let the messiness of the physical world break the symmetry. Binmore also helps himself to evolutionary history, but stays within the rational choice framework, though emphasizing that people are equipped with non-rational social instincts that help to solve the equilibrium selection problem. In particular, he postulates that the biological function of our sense of fairness is to allow us quickly to adapt to surpluses generated by technological innovation or environmental chance by choosing one of the many arrangements that *would* be motivationally stable. Something like the OP kicks in, allowing us to quickly agree on a particular distribution. We call it 'fair', all things considered, not because it flows from some matter of principle, but because that is just what we call the strategy that evolution created for this kind of equilibrium selection problem.

Whether or not Binmore's argument ultimately works is not something I am prepared to pronounce on at this point. Do the bargaining outcomes actually work the way he claims? Do his simplifying assumptions invalidate his conclusions? Can one really analyze games and bargaining and social and biological evolution all interacting in the way he does? One probably needs to read this book a couple of times in order to be fully satisfied. Or better yet, a graduate seminar, one chapter a week, with experts on the white board leading us through the nitty gritty, might be just the thing. But I do think that Binmore is right about a number of things: that social reforms should be based on an understanding of the status quo and on a realistic expectation that things won't just roll back to where they are, or end up somewhere worse. And I do think that mathematical tools of economics and decision and game theory are really the only way to address these concerns with any precision. They may be oversimplified, but not as much as non-mathematical models. Moreover, I think that at some point economists *et al.* need to start incorporating our social instincts into our idealization of rational agency. If Bimore hasn't got it exactly right, he may at least have done a good bit of the work needed to get there.

William F. Harms
Seattle Central Community College