why this book is important. In all chapters one finds an overwhelming number of reasons why economists and social scientists in general should not neglect social personalized interactions in their attempts to explain social reality. In this sense, the book is an instructive introduction to one of the most relevant topics of economic theory in the coming years. On the negative side, the informed reader will find that some of the chapters are hardly up-to-date, especially the references to experimental works. But this is also proof that this is a field continuously developing new ideas and producing new results. In light of all this, it is clear that economics cannot afford to continue neglecting the human side of social relations.

Luis Miguel Miller
Max Planck Institute of Economics & IESA-CSIC

doi:10.1017/S0266267108001880

Natural Justice, Ken Binmore. Oxford University Press, 2005, xiii + 207 pages.

Ken Binmore's *Natural Justice* is a condensed, algebra-free version of his magnum opus *Game Theory and the Social Contract* (Binmore 1994, 1998). Five times shorter than its 1,000 page predecessor, *Natural Justice* preserves both the argument and the style of the main book, except that the author seems now determined to avoid digressions and to develop the main argument as linearly as possible. Brevity comes at a price, and the author admits that in the new book he doesn't "hedge speculations about with reservations and qualifications" and that his "claims aren't proved but illustrated with examples" (p. ix). Some arguments have even been left out entirely, but Binmore usefully provides us with marginal notes referencing the relevant sections of the larger work. Thus, the new book is a compact and efficient presentation of the intricacies of Binmore's previous work. *Natural Justice* is extremely readable, making it ideal for a first-year graduate or upper-level undergraduate course, yet it still provides its readers with a wealth of tools to explore an evolutionary and naturalistic approach to justice, morality and ethics.

"We need to treat morality as a science", says Binmore, recognizing that the moral rules characterizing human societies are "shaped largely by evolutionary forces" (p. 1). To study morality, one must provide scientific explanations for the questions of the origin and evolution of moral rules. *Natural Justice* consists of the application of such a scientific approach to the issues of justice and fairness. The science best suited to the task is the branch of economics that studies social, strategic interactions: game theory. Evolutionary game theory, as well as the theory of repeated games, is a prominent element in Binmore's discourse, since his naturalistic interest lies in the evolution of moral rules. For instance, Binmore suggests that there are analogies between our linguistic and our moral capabilities. As we possess a genetically hardwired ability to acquire language, similarly we possess a genetically hardwired device for moral reasoning.

Philosophically, Binmore's enterprise belongs to three (intersecting) lineages. Firstly, and foremost, being a naturalistic, conventionalist and evolutionary approach to justice and the social contract, Binmore's project is eminently of the Humean ilk, if cast in the modern language of economics and evolutionary game theory. Secondly, the Rawlsian idea of the *original position* lies at the very core of Binmore's theory. It perhaps wouldn't be inaccurate to characterize Binmore's project largely as an attempt to *naturalize* Rawls's notion of the original position. Lastly, because of the important differences between Rawls's and his own conception of the original position, Binmore draws from John Harsanyi's views. In particular, Binmore borrows and adapts Harsanyi's treatment of interpersonal comparisons of utility. The result of these three neatly interwoven threads is Binmore's view of the original position as the genetically evolved mechanism we use to adjudicate issues of fairness in our daily social life (a view first advanced in *Game Theory and the Social Contract* and now restated in *Natural Justice*.) An essential component of Binmore's view on the original position device is Harsanyi's idea of *social indexes* – the "exchange rate" by which players' utilities can be interpersonally compared. However, when it comes to determining the origin of social indexes, Binmore parts from the so-called "Harsanyi doctrine". In his model social indexes are not discovered in the informational vacuum of the original position, but rather are the commonly known (yet ever shifting) product of human cultural evolution. Thus, these are Ken Binmore's coevolutionary views *in nuce*: we are genetically inclined to adjudicate issues of fairness through the device of the original position, and to do so we use as parameters culturally evolved social indexes. To use again the analogy with linguistic theory: as the milieu in which we are reared determines the language we use, likewise cultural evolution shapes the content of our fairness norms.

Binmore's views, however, are not only descriptive. At times, he dons a *whig* prescriptive hat, endorsing "planned decentralization". This prescriptive stance has a pragmatic justification: if Binmore's naturalistic theory of justice and fairness is correct, then one could in principle interpret the reform of the existing social contract as an exercise in mechanism design. Designing a utopian social contract would be pointless, since the contract would be unfeasible. Instead, Binmore's theory would allow the reformer to distinguish between social contracts that are *feasible* (reachable from the *status quo*) and those that are not. Only after feasibility has been addressed can the sensible reformer tackle the question of optimality.

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

Binmore models human social life as a series of strategic interactions. Series of this kind can be modelled as a collection of *repeated games*, which Binmore dubs the “game of life”. If life is a collection of repeated games, then a “social contract” is a collection of equilibria for such games. A social contract, in such an evolutionary brand of contractarianism, cannot be but an equilibrium of the game of life, for otherwise it would not be stable with respect to internal pressures. If stability is essential to the survival of a social contract, so is efficiency, since a society held together by a stable yet inefficient social contract would not be likely to compete successfully with other societies endorsing more efficient social contracts. *Stability* and *efficiency*, however, are not the only relevant dimensions for a game-theoretic study and understanding of social contracts. The folk theorem of game theory shows that there exist a plethora of efficient equilibria in the game of life, and hence presents us with the problem of equilibrium selection. *Fairness* is thus the third substantial element in Binmore’s theory, playing the crucial role of the equilibrium selection device.

The view summarized in the previous paragraph is offered in the first chapter of *Natural Justice*, while the second chapter provides a compact introduction to bargaining theory and to the concepts of Nash, utilitarian and egalitarian bargaining solutions, all of them perspicuously illustrated through geometrical examples.

After a whirlwind review of major philosophical “isms” (chapter 3) conducted in Binmore’s characteristic forceful style, chapter 4 provides an introduction to the game-theoretic notions that are most relevant for the endeavour of *Natural Justice*. The chapter dissects the notion of *Nash equilibrium* – the key notion to understanding the stability of social contracts. Binmore analyses “toy games” of coordination and cooperation, the evolutionary interpretation of mixed strategies, corrects hard-to-erase misunderstandings of the prisoner’s dilemma, and discusses equilibrium selection in games with multiple equilibria. Throughout the chapter, Binmore defends the idea that the deviations from equilibrium behaviour often observed in experimental settings are instances of “downright irrational” (p. 75) behavior, and contends that such irrational behavior can be corrected by providing the subjects with enough time and incentives to learn rational play.

Chapter 5 introduces the theory of repeated games, and in particular the *folk theorem* – the key notion to understanding the efficiency of social contracts. Since the “game of life” is a collection of repeated games, the folk theorem of game theory applies guaranteeing that, for each game, there exist a multiplicity of equilibria. Hence, there exist a vast number of possible social contracts, a large subset of which are efficient. There are objections to the use of the folk theorem in this context. For instance, it would appear that, for the folk theorem to apply, it is necessary that the circumstances allow for perfect monitoring. However, the theorem is

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

invoked to explain the *emergence* of fairness. Since fairness most likely first arose in relatively small groups of our hunter-gatherer ancestors, the assumption of perfect monitoring is not far-fetched after all. (For further criticism on the use of the folk theorem made by Binmore, cf. Gintis (2006) and, for rejoinders, Binmore (2006)). The folk theorem offers the theoretical underpinnings for the idea of *reciprocal altruism*, and through it Binmore (chapters 5 and 6) can recast notions such as right, duty, moral responsibility, etc. in evolutionary terms.

Chapter 7 discusses Hamilton's idea of *kin selection* as a determining factor for the payoffs in the "game of life" when the players are related to each other. Within *Natural Justice*, kin selection plays two roles. On the one hand, it explains (roughly along the lines of Peter Singer's *The Expanding Circle*) that moral behavior is a genetic imperative within the family. This suggests at the same time that it might have evolved outside of the family circle when close, yet unrelated, individuals were treated as if they were relatives. On the other hand, kin selection introduces the intuition behind Harsanyi's idea of empathetic preferences, which is a key component of Harsanyi's account of interpersonal comparisons of utility, in turn a necessary element for the theory of bargaining in the original position. However, as good as reciprocal altruism and Hamilton's rule are for sustaining a cooperative equilibrium and providing the payoffs in the game of life, they cannot by themselves yield a mechanism of equilibrium selection.

To solve the problem of equilibrium selection, Binmore invokes the device of the *original position* – the key notion to understanding fairness in social contracts. The idea, introduced in chapter 9, is that two agents, when confronting each other in some instance of the game of life, can play a (fictitious) "game of morals" if there is no satisfactory conventional solution to a particular interaction. To play the "game of morals", they (fictitiously) repair under the veil of ignorance and start bargaining in the original position. The result of the bargaining process is the equilibrium profile that solves the strategic situation at hand. The solution crystallizes into a convention, and conventions of this kind constitute the notion of fairness held by a society. Relative to this point, two important observations are in order. First, for two agents to be able to bargain in the original position, they must be able to perform interpersonal comparisons of utility. As it is explained in chapter 8, interpersonal comparisons are made possible by Harsanyi's idea of *empathetic preferences*, represented by a "social index" that can be thought of, roughly, as the rate at which player *i*'s utility can be exchanged with player *j*'s. The social indexes of the two agents, according to Binmore, evolve under the pressures of cultural evolution until they are at equilibrium and become common knowledge. Second, Binmore still needs to offer an argument showing that we do in fact recur to such (fictitious) game of morals when we have to bargain a solution to some

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

strategic interaction involved in our social contract. In chapter 9, Binmore argues for the claim that such a use of the original position is genetically evolved from our hunter-gatherer past, and is "written in our genes".

Bargaining in the original position terminates with the solution (equilibrium) to a specific interaction. The use of the veil of ignorance guarantees that the selected solution is considered fair by the parties, and becomes part of the social contract. The question remains: which solution will be chosen? That is, which is the fair solution? Binmore argues that the answer to this question depends on the circumstances in which a society finds itself. In particular he shows that (chapter 10) if there exists an authority that can enforce the outcome of the bargaining process, then the solution to the bargaining problem in the original position is utilitarian, in that the social indexes will be such that they maximize the sum of the weighted payoffs. If (chapter 11) no authority capable of enforcing the bargaining outcome is present, the solution to the bargaining problem must be self-sustaining; in this case, the social indexes in equilibrium will yield the egalitarian solution. In both cases, over the medium run, cultural evolution "leaches out all the moral content of a fairness norm" (p. 158) by shaping social indexes such that they make the solution dictated by the fairness norm coincide with the "brute" Nash solution. In fact, fairness norms are effective only in the short run – for example, when the set of feasible social contracts expands and a new equilibrium is reached by making use of the existing fairness norm. This insight leads naturally to the last chapter (12) of the book – about planned centralization and social reform. Given that a social contract is an equilibrium of a repeated game, Binmore's insight is that social reform should be approached from the vantage point of *mechanism design*. Thus unfeasible social contracts should be ruled out, while possible and desirable ones should be pursued by making sure that agents have the right incentives to move from the current equilibrium to the desired one.

Binmore's recasting of social contract theory into the language of game theory comes at the cost of heroic simplifications and wide shifts of paradigm in the interpretation of terms of art. Both aspects are likely to be contentious, especially to the philosophical readership. His idea of a social contract as the game-theoretic equilibrium of a repeated strategic interaction, as well as his identification of the state of nature and the *status quo*, are unorthodox. Yet, they are important steps in a direction leading towards a naturalistic account of justice and of the social contract.

My comments are articulated in the following four points. First, I consider Binmore's accounts of the original position and the veil of ignorance. Second, I consider an objection to Binmore's notion of justice as *mutual advantage*. My third comment is related to Binmore's views about experiments in economics, while the fourth deals with the role of common knowledge.

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

Which original position? In Binmore's theory, the original position is the equilibrium selection device through which it is possible to select a fair equilibrium among the many that are feasible in the game of life. It is crucial that the original position be *actually* used (albeit fictionally) by the agents when adjudicating issues of fairness. Thus, besides technical differences between Binmore's and Rawls's original position, there is also a substantial philosophical difference. In Binmore's hands the original position ceases to be a sophisticated philosophical argument apt to determine the general principle of justice as fairness; it becomes a naturalized notion, the tool used to settle everyday (mostly picayune) coordination problems. The role of the original position ceases to be normative (the *a priori* selection of criteria of justice) and becomes descriptive (the *empirical* selection of courses of action which are, as such, deemed fair in society). At the same time, the justification of the original position has to shift from Rawlsian reflective equilibrium to a descriptive justification. Binmore proposes, drawing heavily on anthropological research, that the original position device is "written into our genes". The idea is that the original position device started off in our hunter-gatherer past as an insurance device. Uncertain about future hunting outcomes, our ancestors hedged against the possibility of meager future hunts by sharing food from successful hunts. Such an insurance contract evolved then into the original position, where the uncertainty about the outcome of future hunting is replaced with the veil of ignorance hiding one's present identity. While the story is intriguing and plausible, the evidence brought to support it is not plentiful.

A further consideration around Binmore's original position concerns his conception of the *veil of ignorance*. According to Binmore, Rawls operates "an iconoclastic evasion of the logic of the decision problem he creates for [the players] under the veil of ignorance" (p. 151). This is a typical example of a claim defended at length in *Game Theory and the Social Contract* whose supportive argument disappears in *Natural Justice*. But the claim is rather substantial. In fact, the idea that orthodox Bayesian decision theory (rather than the maximin decision rule) must apply in the original position appears to be due to a difference in the conception of the veil of ignorance. Being a *thin* veil of ignorance *à la* Harsanyi (and for the distinction between thin and thick veil of ignorance, cf. in particular pp. 157–159 of Freeman 2007), the players are fully informed about anything but their identity. In *such* a simple situation, as Binmore states (p. 151), "if orthodox decision theory were wrong [. . .] it would always be wrong". But Rawls's veil of ignorance is *not* Harsanyi's, or Binmore's, although Binmore treats it as such in chapter 10. Rawls's decision theory is "iconoclastic" only if the veil of ignorance is thin, but Binmore does not offer any argument at all to defend his thin veil against Rawls's thick version.

The vulnerability objection. One plausible objection to Binmore's conception of justice is that, being based on reciprocal altruism, it

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

presupposes that all parties to the social contract are capable of positively contributing to it. But, as Binmore himself recognizes, "[a] tree or an unborn human is powerless, and so can't be a player in the game of life. Animals, babies, the senile, and the mentally ill are only marginally less helpless, and hence equally unable to take on duties. They are correspondingly unable to exercise any rights under the social contract" (p. 97). The powerless could be looked after by a caring and loving relative, but the issue remains relevant, since a critic could surmise that lacking the good feelings of loving relatives, the powerless in Binmore's theory enjoy no rights whatsoever. To be sure, *Natural Justice* presents an example (cf. pp. 86–7) of a third-party punishment equilibrium model in which an agent cooperates with another even if the former has no warm feelings towards the latter and fears no *direct* punishment from the offended party in case of defection. The model is however an extremely simplified one. The insight on which it is based is generalized (and the vulnerability objection to Binmore's conception of justice as mutual advantage fully answered) in a forthcoming paper by Peter Vanderschraaf (Vanderschraaf 2008).

Behavioral economics and rationality. An important subtext running throughout chapter 4 (and, in a sense, throughout the entire book) is the idea that observed deviations from rational behavior are not satisfactorily explicated by appealing to specifically tailored utility functions – for instance utility functions based on the idea that agents have a "taste for fairness" constructed in their preferences. This solution provides a description, rather than an explication of the phenomenon. Binmore's theory is an attempt to "dig deeper" and explain the evolutionary mechanisms that have resulted in our having a taste for fairness. But Binmore's social contract, as we have seen, relies on properties (stability, efficiency, fairness) that describe the behavior of ideally rational agents. How are we to reconcile the empirically observed discrepancy between human behavior and rational behavior, on the one hand, and a theory that purports to be empirical but relies on an idealized notion of rationality? The answer is that an agent walks into the laboratory bringing along a web of social habits, norms and convention. Such habits of fairness select equilibrium behavior in the real-life equivalents of the laboratory settings, but such behavior need not be an equilibrium of the game played in the laboratory. Binmore claims that, provided that they have enough time and incentives, agents will learn to act rationally, and he suggests that interesting research is to be done relative to such dynamics. Yet, it might be not easy to extricate oneself from such habits (for a thorough analysis of such *habits of the mind* and their importance for an account of social norms, cf. Bicchieri 2006). We know from classic experimental literature that agents adjust very quickly to certain market-like settings, while the adjustment process is "glacially slow" in games like the Ultimatum Game. Binmore acknowledges that "research on this front is progressing steadily,

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

but still has far to go" (p. 75). The question remains, however, whether in actual situations – situations even more complex than the Ultimatum Game – the learning process might be so slow as to invalidate the rationality assumption upon which Binmore's theory of the social contract hinges.

Common knowledge. My main disagreement with Binmore concerns the role that common knowledge plays in conventions and, by extension, in his naturalistic theory of justice. Binmore is not sympathetic to Lewis's account of convention, and in particular to his claim that the clauses of the definition of convention need be common knowledge among the players (for an extended account of the different positions, cf. Binmore 2008, and Sillari 2008). On this very topic, there seems to be an element of discrepancy between *Game theory and the Social Contract* and the recent book. While in *Natural Justice* Binmore dismisses the common knowledge assumption as a mere simplification that could simply be relaxed by adopting a framework of incomplete information, common knowledge seems to be playing a larger role in the older work: "A society's pool of common knowledge – its culture, provides the informational input that individual citizens need to coordinate on *equilibria* in the games that people play" (Binmore 1994: 140). In *Natural Justice*, common knowledge seems to have lost the role of necessary assumption in the equilibrium selection problem, although no argument is provided to explain this apparent change in view. When it comes to studying cultural evolution (*medium run* evolution, in Binmore's account), common knowledge of the current social indexes seems to be a relevant assumption. After all, Binmore is keen to acknowledge that what he says about the dynamics of medium run processes determining the cultural evolution of social indexes "is anything but a crude first stab at a naturalistic theory of interpersonal comparison of utility" (p. 157).

To be sure, many aspects of Binmore's theory of justice need further development, and Binmore is well aware of this. Possibly the most important of such developments is the extension of bargaining in the original position to more than two players. This should be obtained by introducing the possibility of coalition formation (cf. pp. 197–8). But even if *Natural Justice* were to be considered a very speculative "first stab" at a naturalistic theory of justice, it would still be, in the eyes of this reviewer, a welcome one and one of philosophical consequence.

Giacomo Sillari
University of Pennsylvania

REFERENCES

Bicchieri, C. 2006. *The grammar of society*. Cambridge: Cambridge University Press.
Binmore, K. 1994. *Game theory and the social contract. Playing fair*. Boston: MIT Press.
Binmore, K. 1994. *Game theory and the social contract. Just playing*. Boston: MIT Press.

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press

Binmore, K. 2006. Why do people cooperate? *Philosophy, Politics and Economics* 5(1): 81–96.

Binmore, K. 2008. Do conventions need to be common knowledge? Forthcoming in *Topoi*.

Freeman, S. 2007. *Rawls*. London: Routledge.

Gintis, H. 2006. Behavioral ethics meets natural justice. *Philosophy, Politics and Economics* 5(1): 5–32.

Sillari, G. 2008. Knowledge and convention. Forthcoming in *Topoi*.

Vanderschraaf, P. 2008. Justice as mutual advantage and the vulnerable. Forthcoming in *Philosophy, Politics and Economics*.

https://doi.org/10.1017/S0266267108001880 Published online by Cambridge University Press