Erasmus Journal for Philosophy and Economics,
Volume 11, Issue 1,
Spring 2018, pp. 50-79.
https://doi.org/10.23941/ejpe.v11i1.246

Naturalism and Moral Conventionalism: A Critical Appraisal of Binmore's Account of Fairness

CYRIL HÉDOIN
REGARDS Research Center, University of Reims Champagne-Ardenne

Abstract: This article provides a critical examination of Ken Binmore's theory of the social contract in light of philosophical discussions about moral naturalism and moral conventionalism. Binmore's account builds on the popular philosophical device of the *original position* but gives it a naturalistic twist. I argue that this makes it vulnerable to moral skepticism. I explore a possible answer to the moral skeptic's challenge, building on the fact that Binmore's account displays a variant of moral conventionalism. I ultimately conclude however that the conventionalist answer leads to a purely behaviorist view of morality, which implies that there is nothing special about morality and fairness norms. I propose alternative interpretations of conventionalism. These accounts escape most of the difficulties because they place emphasis on the reasons that establish a moral convention.

Keywords: Binmore, moral naturalism, moral conventionalism, original position, fairness

JEL Classification: B00, B31, C73, D02

1. INTRODUCTION

Economists have demonstrated during the last three decades a growing interest in issues related to fairness and morality. Indeed, the rise of game theory has considerably changed the disciplinary landscape between economics and moral philosophy: economists now have a tool at their disposal directly relevant to making significant contributions to moral philosophy. This article provides a critical examination of a specific attempt to produce a theory of fairness through the game-theoretic lens, namely Ken Binmore's theory of the social contract

AUTHOR'S NOTE: I would like to thank the two anonymous referees for their challenging and very useful comments on the previous version of this article. I also thank O. Çağlar Dede for his editorial work. All errors and omissions are of course mine.

(Binmore 1994, 1998, 2005). Binmore presents his account as an attempt to "treat morality as a science" (Binmore 2005, 1). It pursues two goals: first, to account for the origins and the content of our fairness judgments; second, to argue for an egalitarian view of fairness. Clearly, the justifiability of the second prescriptive goal depends on the success of the first descriptive goal. However, several philosophers have argued that pursuing the first goal might undermine the justifiability of the second (see, for instance, Joyce 2006). My examination of Binmore's account responds to this general philosophical worry.

Binmore's theory of fairness builds on the popular philosophical device of the *original position*, independently developed by John Rawls (1971) and John Harsanyi (1953). However, Binmore gives a naturalistic twist to this device. He naturalizes it through two related claims: first, he argues that genetic and biological evolution has encoded the original position in our genes. In particular, he claims, biological evolution has endowed us with the ability to sympathize and empathize with others, regardless of genetic relatedness. Second, Binmore argues that cultural evolution has led to the emergence of standards of fairness under the forms of empathetic preferences that make interpersonal comparisons of utility possible. The original position is then conceived by Binmore as a genetically encoded but culturally loaded algorithm, which humans use to coordinate in the "game of life", i.e. the game whose "rules are determined by the laws of physics and biology; by geographical and demographic facts; by technological and physiological constraints" (Binmore 1998, 4). The game of life has a multiplicity of Pareto-efficient equilibria. The original position device is instantiated in what Binmore calls the 'game of morals' and selects one equilibrium on the basis of an egalitarian standard of fairness.

My goal in this paper is to clarify the implications of the naturalization of the original position for the status and the significance of fairness claims and judgments. I shall argue that the means by which the original position is naturalized makes it vulnerable to moral skepticism. Specifically, I argue that Binmore's naturalization of the original position implies that fairness judgments are grounded on the power structure of the society. A moral skeptic can then argue that these judgments do not have any moral content and authority, and thus, cannot be objectively true. I explore a possible answer to the moral skeptic's challenge by arguing that Binmore's account displays a variant of moral conventionalism. However, I conclude that Binmore's

conventionalist answer leads to a purely behaviorist view of morality, which implies that there is nothing special about morality and fairness norms. In response, I consider alternative accounts of moral conventionalism which emphasize the importance of the reasons that establish moral conventions. These alternatives escape most of the difficulties which are associated with Binmore's account.

The article is organized as follows. Section 2 presents Binmore's account by explaining the naturalization of the original position as a device to coordinate in the game of life. Section 3 raises the critique from moral skepticism against Binmore's account, as the latter is understood as an instance of moral naturalism. Section 4 examines a possible answer to this critique by characterizing Binmore's account as an instance of moral conventionalism. Section 5 argues that Binmore's moral conventionalism nevertheless fails to answer the skeptic's critique, while also demonstrating that other forms of moral conventionalism are immune to it.

2. BINMORE'S NATURALISTIC ACCOUNT OF FAIRNESS

Ken Binmore's naturalistic account of the social contract and fairness norms is developed in a two-volume book *Game Theory and the Social Contract* (Binmore 1994, 1998).[1] It builds on the ideas of three influential authors in moral and political philosophy: David Hume, John Harsanyi and John Rawls. It ultimately leads to a vindication of Rawls's egalitarianism against Harsanyi's utilitarianism. Binmore sees in Hume the foundations of a 'conventionalist' view of justice in which fairness norms are taken to be the product of an evolutionary process. While they are initially conceived as conventional devices to solve coordination problems, fairness norms progressively acquire normative power as they become the commonly understood standard for determining whether a situation is just or unjust. From Rawls and Harsanyi, Binmore's account retains the philosophical concept of the *original position* that both authors simultaneously developed in the 1950s.

Binmore's naturalistic account of the social contract is in line with the growing body of scholarship that applies tools and models from

[1] Binmore exposes his account in a less mathematical and less detailed fashion in a later book, *Natural Justice* (2005). This book does not add anything substantive to the preceding two volumes except for its more straightforward presentation of the main ideas. Therefore, I will not refer to it except in the few instances where it indicates that Binmore has changed his mind with regard to what is written in *Game Theory and the Social Contract* (1998).

natural and social sciences to issues of moral and political philosophy. More specifically, it is a representative contribution of a set of approaches combining the mathematical principles and concepts of game theory with theories of natural and cultural evolution to study the origins of morality.[2] In the present case, Binmore's naturalism develops as an attempt to naturalize Rawls's and Harsanyi's original position (henceforth, OP). The term 'naturalize' and the derivative 'naturalization' here refer to the fact that Binmore attempts to show that the OP is not merely a philosophical thought experiment. It is actually part of the natural world in the sense that it corresponds to a device—or an algorithm—that humans are using to solve coordination problems. Indeed, Binmore argues that the device of the OP is actually a genetically-encoded algorithm used by people to make fairness judgments because of our natural history. Moreover, the use of the OP algorithm depends on standards for making interpersonal comparisons of utility that evolve from *cultural* evolutionary processes. Binmore substantiates these claims through a game-theoretic formalization of the bargaining that takes place behind the veil of ignorance constitutive of the OP.

Binmore defines a social contract as "the set of common understandings that allow the citizens of a society to coordinate their efforts" (2005, 3). He claims that any social contract must satisfy three requirements: stability, efficiency and fairness. The first is the most important. Binmore rejects ad hoc assumptions that moral philosophers have sometimes invoked to make the agreement concluded under the veil of ignorance binding (see, for example, Gauthier 1986). Since every member of a society is part of the social contract (including government members and law enforcers), a stable agreement must be self-enforcing. Arguing that any social contract relies on a *repeated* game, Binmore makes use of the folk theorem of repeated game theory according to which multiple equilibria exist across all different sorts of games as soon as a given game is indefinitely repeated between the same players. The folk theorem shows that the evolution of cooperation is not dependent on the existence of prosocial preferences. Several stable social contracts are then possible, without necessarily depending on prosocial preferences.

[2] Other representative works include Alexander (2010), Skyrms (1996, 2004), Sugden (2005), and Young (2001).

Next to the stability requirement, the two other conditions for the viability of a social contract are efficiency and fairness. Efficiency is defined by the Pareto criterion and is based on a simple argument about group selection. Different communities may agree on different social contracts. If we assume that communities' expansion is a function of the efficiency of their social contracts, then communities operating under a sub-optimal one will progressively be trumped over by those operating with efficient ones.[3] If we assume, like Binmore, that the negotiation of a social contract takes the form of a bargaining game between two players, then any viable social contract is contained in the area between the disagreement point (which corresponds to the minimax gain for each player) and the maximum that each player can gain. An efficient agreement is by definition placed on the Pareto frontier, delimiting the set of feasible social contracts. However, numerous agreements are still possible. According to Binmore, the fairness criterion serves as a device to select one of the efficient equilibria. Fairness norms, therefore, help individuals to coordinate on a particular outcome through the expectation that everyone will choose it. Since individuals agree to coordinate on a particular equilibrium, it is deemed to be fair in a sense that I will now explain.

How the fair equilibrium is determined, and thus which equilibrium will be chosen, are the central questions answered by Binmore's naturalistic account of justice. This lays the groundwork for the naturalization of the OP. According to Binmore, humans are engaged in an ancestral 'game of life' whose rules are defined by biological constraints. Binmore describes it as a (repeated) bargaining game played by two players, Adam and Eve. He takes the problem of food-sharing in human foraging communities as the hallmark problem encoded in the game of life. From the very start of human history, sharing food has been an allocation problem that any community had to solve. Since the same problem is also faced by animals such as chimpanzees, Binmore makes the conjecture that humans have been genetically programmed to

[3] As Binmore notes, this argument is immunized against the traditional critiques of group selection explanations (1998, 185). In fact, the stability requirement assures that any existing social contract is an equilibrium and therefore that individuals have an interest to enforce it. Sugden (2001b) notes however that Binmore does not link reproduction with utility. As in evolutionary game-theoretic models in economics, utility describes only the propensity for a strategy to be replicated in the society. If Pareto optimality is defined in terms of utility, then assuming that Pareto improvements promote survival or competitiveness is a *non sequitur*. Some behavioral patterns with strong replication propensity, such as addictive behaviors, can be destructive in the long run.

play the game of life. In the context of genetically related individuals, it is easy to show that sharing food with one's relatives is an equilibrium (not the only one, however), which might be selected and implemented through a food-sharing insurance contract. Under such a contract, unlucky relatives who have failed to get any food receive some food from more lucky relatives. Indeed, this kind of mechanism makes it more likely that genes shared by relatives will spread. However, in human societies, cooperation expands well beyond the circle of genetic relatives: food-sharing insurance contracts also take place in the context of genetically unrelated individuals facing uncertainty about the results of their hunt. Under this 'veil of uncertainty' where one does not know whether he will be able to catch any food in the future, each person must sympathize with her possible future selves ('Mr. Lucky' or 'Ms. Unlucky') by anticipating how much their future preferences would be satisfied in the different possible scenarios. A food-sharing insurance contract represents a Pareto-improvement for agents facing such kind of uncertainty. Moreover, the folk theorem of repeated game theory indicates that such contracts are sustainable as equilibria. Binmore contends that the device of the OP first evolved as a way to negotiate such contracts (1998, 219). On this basis, it progressively became a genetically-encoded algorithm used to solve more general and larger fairness issues:

> people take a technique used within one circle of social problems and unthinkingly apply it to a wider domain of problems. In so doing, they continue to play by the rules of the game for which the technique originally evolved, not noticing—or pretending not to notice—that the rules of the game played in the wider circle may be quite different (Binmore 1998, 219).

There is, however, a clear difference between the veil of uncertainty that members of hunter-gatherer groups had to deal with and the veil of ignorance of the OP as it was initially conceived by Harsanyi (1953): in the latter, persons under the veil of ignorance have to put themselves in *others*' position to determine what they have to do. They must *empathize* with other persons by pretending that they have *the other's* preferences; they must assume that they literally *are* these other persons. This implies that each person has the ability to make comparisons (either at the level of utility or the level of preference satisfaction) between each other member of the population. According

to Binmore, the use of the OP as a device to make fairness judgments thus evolved from the combination of food-sharing negotiations between genetically unrelated individuals and interactions between genetically related individuals:

> all that is then needed is for us to hybridize these two processes by allowing a player to replace one of the future persons that a roll of dice might reveal him to be, by a person in another body who is to be treated in much the same way that he treats his sisters, his cousins or his aunts (Binmore 1998, 220).

Cooperation with non-relatives through the OP is thus partially the product of natural evolution: first, kin selection has 'programmed' organisms to cooperate with genetic relatives; second, natural pressures due to the uncertainty regarding feeding in hunter-gatherer societies have favored the selection of an innate ability to empathize with others. However, while natural selection has endowed us with an innate ability to make fairness judgments through the OP device, it has not determined the *content* of these judgments. Following Harsanyi, Binmore considers that individuals possess *empathetic preferences* allowing them to determine if they prefer to be 'person *i* in situation *x*' or 'person *j* in situation *y*'. These empathetic preferences make the agents' utility functions commensurable and determine the rate at which the utilities of each individual will be traded with those of others (Binmore speaks of 'social indices'). The content of empathetic preferences and therefore the value of the social indices are determined by cultural evolution.[4] On this basis, the fairness of the social contract is established by adding a device to the game of life through which individuals will be able to coordinate their action: the 'game of morals'.

The game of morals is purely conventional. Binmore interprets it as a heuristic through which individuals reflect on and anticipate the reaction of every member of a society when a new social contract is established. Like the game of life, the game of morals leads individuals to make use of empathy. Binmore contends that any individual can at every moment appeal to the game of morals if he is not satisfied with his situation. Appealing to the game of morals is like rolling the dice again and negotiating a new social contract under a veil of ignorance.

[4] Binmore makes an analogy with the evolution of language. The capacity to develop and to learn a language has a biological origin. However, the content of any language comes from the cultural history of each society and is independent of any biological factor.

The clause of unlimited appeal has radical implications: first, it means that individuals must have the same empathetic preferences—Binmore calls it *symmetric empathetic equilibrium*. On the contrary case, an agreement would not be reached, leading agents to play the game of morals again and again. Second, the agreement must be considered fair by each individual according to the existing social indices. As Binmore puts it:

> A *fair social contract* is then taken to be an equilibrium in the game of life that calls for a use of strategies which, if used in the game of morals, would never leave a player with an incentive to exercise his right of appeal to the device of the original position [...] the game of morals is nothing more than a coordination device for selecting one of the equilibria in the game of life (2005, 172, his emphasis)

Without the existence of any enforcing authority, Binmore shows that the agreed social contract will correspond to the proportional (or egalitarian) bargaining solution. The solution ensures that the players' utility functions are suitably rescaled according to the social indices that correspond to the prevailing empathetic equilibrium. Indeed, it is clear that any social contract which is not egalitarian will lead the worst-off individuals to appeal for a new game of morals. If everyone uses the game of morals to choose a fair social contract, and if this becomes common knowledge, then this necessitates an egalitarian social contract.

While Binmore is rather vague about the precise mechanisms responsible for the evolution of empathetic preferences,[5] he provides an interesting argument for their role in determining the equilibrium reached in the game of morals. The argument is somewhat complex but relies on two key assumptions: the first assumption is that *real* bargains always converge toward the Nash solution; the second assumption is that the enforcement of the agreement reached behind the veil of ignorance is ultimately always self-enforcing (for example, there is no external enforcer). The former is indeed essential in the algorithm Binmore proposes to compute the value of the "social indices" (Binmore 1998, 441). He considers three temporal scales in the evolution of fairness norms (1998, 226-227). In the *short run*, both individuals' personal and empathetic preferences are fixed and only their choices made through the device of the OP can change. In the *medium run*, the

[5] Binmore had initially made use of Richard Dawkins's (1975) concept of 'meme' to account for cultural revolution. He has retreated, however, in *Natural Justice*, acknowledging the huge difficulties related to this concept.

individuals' personal preferences are fixed, but their empathetic preferences are susceptible to change, as the Pareto-frontier of the bargaining game they are playing moves. In the *long run*, both empathetic and personal preferences can change through the forces of cultural evolution and biological evolution respectively. The temporal scale of cultural evolution is thus the medium run, while biological evolution operates in the long run. In the short run, the agents play the game of morals—using the OP device to select one of the Pareto-efficient equilibria. The result of the bargain is perceived as 'fair' by the participants because it selects the egalitarian equilibrium *given the (symmetric) empathetic preferences* of the players. However, these empathetic preferences find their origins in the bargaining relationships that take place in the game of life. In the latter, the players only rely on their bargaining skills and it is assumed that the resulting outcome corresponds to the Nash bargaining solution. As is well-known, to which point on the contract curve this solution corresponds to depends on the players' risk and/or time preferences. Relatively risk-tolerant and patient players will have an advantage in the bargaining process and will obtain the lion's share of the available resources. What happens then is that empathetic preferences are set such that the outcome corresponding to the Nash solution is selected as the egalitarian solution in the game of morals. In other words, the equilibrium in the game of morals corresponds to the egalitarian solution, taking the players' empathetic preferences to measure utility. However, at the equilibrium in the game of life, the players' utilities determined on the basis of their personal preferences are such that the Nash solution is satisfied. Obviously, the latter need not be egalitarian.[6]

Actually, Binmore's claim seems to be that empathetic preferences serve as *a posteriori* egalitarian rationalization of previous bargaining arrangements reached in the game of life. These arrangements need not be egalitarian (this depends on the players' time preferences or risk preferences, as well as their bargaining abilities), but in the context of the game of morals they must be seen as *fair* by the players; otherwise, one of them would want to 'roll the dice' again. This explains Binmore's conclusion that existing social contracts must be egalitarian when evaluated according to the players' empathetic preferences.

[6] Moreover, the very notion of 'equality' in the context of the Nash solution is meaningless since the latter does not assume that interpersonal comparisons of utility are possible.

3. FROM DESCRIPTIVE TO PRESCRIPTIVE ETHICS: NATURALISM AND MORAL SKEPTICISM

The preceding section has shown that Binmore's naturalistic account of fairness leads to a substantive moral conclusion. Assuming that the OP is a device that has been historically used to solve coordination problems in the game of life, social contracts must actually be egalitarian, at least when judged according to the prevailing empathetic preferences in the population. In this section, however, I argue that independently of what one may think of its substantive conclusion, this account has to face the same meta-ethical challenge that confronts all forms of moral naturalism. This challenge corresponds to what is generally labeled 'moral skepticism': the view according to which the naturalistic foundations of morality raise doubts about the justification of moral judgments.

The naturalization of the OP is used by Binmore as part of the larger project of treating "morality as a science" (2005, 1). However, the naturalistic project of treating morality as a science may have several meanings. The most modest interpretation is restricted to the domain of what is sometimes called 'descriptive ethics'. Naturalism then corresponds to the general endeavor of providing a scientific account of moral practices and institutions and the genealogy of moral judgments. As it will be clearer below, such a form of naturalism does not imply any commitment regarding either the existence of moral facts or properties, or the truth-value of moral beliefs. Two stronger forms of naturalism intend to provide some articulation between descriptive ethics and the domain of 'prescriptive' ethics. Indeed, they suggest that knowledge about the way moral judgments and practices evolved vindicate these judgments and practices, and more generally, moral theories. In other words, they are committed to the claim that the naturalistic origins of moral judgments and practices *justify* these judgments and practices in some well-defined sense. The first of these two forms of naturalism, *moral naturalism*, builds upon the postulate that moral properties and facts can ultimately be reduced to naturalistic properties and facts. The second form, *moral conventionalism*, takes morality to consist of nothing but conventional practices. As Binmore uses his naturalistic account as a defense of some version of Rawlsian egalitarianism, it is clear that it should be regarded as an exercise in both descriptive and prescriptive ethics, either as an instance of moral naturalism or of moral

conventionalism. I focus in this section on the objections made to moral naturalism as they will prove useful to discuss moral conventionalism in the next two sections.

The main objections made against moral naturalism can be summarized in the following way: by showing that fairness and, more generally, morality have naturalistic foundations, naturalistic approaches undermine the very ground on which the normative force of morality and fairness are built upon. This 'very ground' is constituted by the naturalistic origins of morality. Far from vindicating morality, these origins make it illusory or even non-existent. This objection in particular has been made against *evolutionary* moral naturalism (that is, the set of views according to which moral values and obligations are grounded by facts about biological evolution); but it is also relevant to other forms of moral naturalism (Joyce 2006). I shall argue in this section that the objection is even more compelling with respect to Binmore's naturalization of the OP. This leads to the following problem: if Binmore's account is empirically relevant, then this leads to doubt about the moral force of fairness norms. More precisely, once one knows and accepts Binmore's account of fairness norms, then it is not clear why one should maintain that his beliefs about what is fair are justified.

Joyce (2006) develops a strong argument that moral naturalism leads to moral skepticism, the meta-ethical view according to which it is doubtful that our moral judgments and beliefs can ever be justified (Sinnot-Armstrong 2015). In the specific case of evolutionary moral naturalism, Joyce's main point is that the empirical knowledge of the genealogy of our moral judgments and beliefs (the fact that these judgments and beliefs emanate from dispositions that have evolved through natural selection) fails to justify them. The reason is that this knowledge does not entail any confidence in the idea that natural selection is likely to have produced true beliefs. As a consequence, this knowledge should lead to moral skepticism, or even to moral nihilism.[7] Consider the following analogy:

> Suppose that there were a pill that makes you believe that Napoleon won Waterloo, and another that makes you believe that he lost. Suppose also that there were an antidote that can be taken for either

[7] A moral nihilist argues that the empirical knowledge of the genealogy of moral beliefs render them *unjustifiable*—rather than merely failing to provide a justification. In this case, it is contended that we cannot provide any justification for our moral beliefs ever and therefore that we should not accept any moral claim. An obvious implication is that nothing can be morally wrong according to a moral nihilist.

> pill. Now imagine that you are proceeding through life happily believing that Napoleon lost Waterloo (as, indeed, you are), and then you discover that at some point in your past someone slipped you a 'Napoleon lost Waterloo' pill… Should this undermine your faith in your belief that Napoleon lost Waterloo? (Joyce 2006, 179).

Joyce argues—quite reasonably—that the answer to this last question should be 'yes'. Correspondingly, your knowledge of the genealogy of your belief 'Napoleon lost Waterloo' should encourage you to take the antidote. Now, if we substitute the belief 'Napoleon lost Waterloo' for any moral belief or judgment and the belief pills for natural selection, then the moral skeptic's argument is easy to understand:

> Were it not for a certain social ancestry affecting our biology, the argument goes, we wouldn't have concepts like *obligation*, *virtue*, *property*, *desert*, and *fairness* at all. If the analogy is reasonable, therefore, it would appear that once we become aware of this genealogy of morals we should (epistemically) do something analogous to taking the antidote pill: cultivate agnosticism regarding all positive beliefs involving these concepts until we find some solid evidence either for or against them (Joyce 2006, 181, emphasis in original).

The analogy works on the basis of the postulate that there is absolutely no reason to think that natural selection is likely to have produced *true* beliefs. Assuming that there are independent moral facts or that moral facts can be reduced to non-moral facts, descriptive evolutionary ethics (the scientific works examining to what extent human morality is the product of natural selection) does not provide a basis for believing that our beliefs about these facts are true. Quite the contrary, moral skepticism argues that descriptive evolutionary ethics undermines morality. Knowing the non-moral genealogy of our moral beliefs can only foster doubt about their possible truth. Moreover, moral skeptics argue that it is implausible to find in the non-moral genealogy of moral beliefs any source for the *necessary practical authority* of moral prescriptions (Joyce 2006, 190-9). In a nutshell, even though natural selection may have led to the existence of moralized social and psychological pressures R for person A to do φ, this does not imply that everything else, being equal, he *ought* to do φ. Actually, what A ought to do also depends on his desires and non-moral beliefs. In other words, it seems that there is no desire-independent practical reasoning that can

endow moral beliefs with the required practical authority. Moral prescriptions would then be followed not because they are 'moral', but only because individuals have the psychological urge to conform to them due to unrelated (amoral) factors. For moral skeptics, this is a significant reason to doubt that our moral beliefs are justified.

This is not the place to pursue the issue of the plausibility of moral naturalism or skeptical critiques further. However, by introducing the argument from moral skepticism against moral naturalism, I want to show that it is directly relevant for Binmore's naturalistic account of fairness. Indeed, this argument can be reconstructed as follows: if Binmore's account is empirically relevant, then we should have doubt about the moral force of fairness norms. More precisely, once one knows and accepts Binmore's account of fairness norms, then it is not clear why one should consider that his beliefs about what is fair are justified. The moral skeptic's critique seems to be even stronger in Binmore's case, because Binmore's naturalism emphasizes the essential role played by *bargaining power* in the evolution of fairness norms. To fully establish this claim, it is first required to show that Binmore seeks to provide a non-moral genealogy to our fairness claims.[8] This is not difficult since he is quite explicit about this point. The OP is a device for making fairness judgments. It has two distinct naturalistic origins: first, it evolves from a biological and genetic genealogy that starts with the family games played by our ancestors, which continues with food-sharing insurance contracts that had to be negotiated in hunter-gatherer societies. This is what Binmore calls the 'game of life'. Second, the use of the OP in concrete cases necessitates making interpersonal comparisons of utility in what Binmore calls the 'game of morals'. This depends on the existence of empathetic preferences whose content (which materializes through 'social indices') evolves through a potential cultural genealogy. The former kind of naturalistic origins obviously makes Binmore's account a target for moral skepticism. But I would argue that moral skepticism has more bite on the latter.

Indeed, as I explain in the preceding section, empathetic preferences operate as *a posteriori* rationalizations of previous bargaining results. One may then wonder whether there is anything fair in the resulting fairness norms that select among the multiple efficient equilibria in our

[8] Since Binmore is not concerned with the naturalistic foundations of morality as a whole, but only with those of our conception of fairness, the discussion will now be restricted to the latter.

daily interactions. We thus recover the skeptic's point: once you realize that moral or fairness judgments are grounded on norms that have naturalistic origins (biological and/or social), this should raise doubts about their justification. To Binmore's credit, he is not shy about this, since he explicitly acknowledges the importance his account gives to bargaining power in the evolution of fairness.[9] Bargaining power may have several origins. As indicated above, it may result from the shape of the individual's personal preferences, the latter being a function of the individual's social position (or genealogy, considering that preferences are partially genetically transmitted). It may also result from the position of the disagreement point in the bargaining game since, by construction, the Nash solution will then favor the agent with the larger initial endowment.[10]

It seems that the moral skeptic is entitled to ask whether fairness judgments resulting from such asymmetries in bargaining power should count at all as authentic justified moral beliefs.[11] Indeed, if I know that my judgment for evaluating the fairness of a situation depends on preferences that have been shaped by the power structure of society, should I give it any more credence than my belief that Napoleon lost Waterloo when I know that it results from the fact that I have taken the appropriate pill? Moreover, since fairness judgments are a kind of moral judgment, they should have the same normative and practical force as any other moral judgments. However, unless one recognizes that the power structure of the society is itself 'fair' (whatever that may mean), it is not clear why one should grant any normative significance to his fairness judgments. I may indeed honestly judge that the current situation is fair in spite of the fact that I am disadvantaged relative to others in the population (and possibly advantaged relatively to some other persons). But why should I trust my judgment and have *any*

[9] "We have to live with the unwelcome truth that the interpersonal comparisons of utility necessary to make fairness judgments meaningful are ultimately determined by the underlying balance of power" (Binmore 1998, 425).

[10] Recall that the Nash solution corresponds to the point that maximizes the product of each player's utility at the bargaining outcome minus their utility at the disagreement point.

[11] Note that the skeptic's query is left unanswered even if one assumes some external authority and thus endorses the utilitarian solution rather than the egalitarian one in Binmore's account. The point is that fairness judgments made on the basis of empathetic preferences do not express justified moral beliefs from the skeptic's point of view.

normative reason to adhere to it knowing that it results from the fact that others were indeed advantaged in the past?[12]

It appears, then, that Binmore's account of fairness is vulnerable to the moral skeptic's rebuttal of moral naturalism. The descriptive claim that fairness judgments have naturalistic origins provides a strong reason to doubt their prescriptive validity or force. At this point, a possible response is to concede and to accept the skeptic's conclusion that fairness judgments cannot be justified. However, an alternative path is available by endorsing moral conventionalism. The next two sections evaluate whether interpreting Binmore's account as an instance of moral conventionalism can escape the skeptic's conclusion.

4. A THIRD PATH: BINMORE'S ACCOUNT AS AN INSTANCE OF MORAL CONVENTIONALISM

Though the skeptic's critique of moral naturalism is powerful, it is not plausible to assume that Binmore agrees with the skeptic's conclusions, as that would make his defense of Rawlsian egalitarianism meaningless. This section, as well as the next, investigates the third path between moral naturalism and moral skepticism, which I refer to as moral conventionalism. While I find moral conventionalism a plausible and highly attractive meta-ethical stance that potentially avoids the skeptic's conclusion, I shall argue that Binmore's naturalistic account offers a variant of moral conventionalism that falls short of vindicating fairness and morality more generally. The main reason for this is the lack of reflexivity that individuals have over their empathetic preferences in Binmore's account. In this section, I provide a characterization of moral conventionalism and explain why it may answer the skeptic's critique. The next section explains why the kind of moral conventionalism endorsed by Binmore is nonetheless unsatisfactory in this regard.

Broadly speaking, moral conventionalism can be characterized as *the meta-ethical view according to which morality is conventional*. On this view, morality is constituted by conventional rules which (by definition) (i) depend on social practices, (ii) are historically contingent and (iii) are arbitrary in some sense (see, for instance, Marmor 2009). There are several variants of moral conventionalism, all of them combining in one way or another Hume's account of justice as an artificial virtue (a virtue

[12] Note that this is a pretty weak normative requirement for a moral judgment. Most moral thinkers would require for a moral judgment to give one a *decisive* reason to abide by it.

that depends on conventional rules) with David Lewis's (2002) theory of conventions (Verbeek 2008). Let me first explain the concept of conventional rules. Clarifying this concept is indeed essential to understand why moral conventionalism is unable to avoid moral skepticism. Feature (i) is intended to capture that a convention exists in some community if and only if it is actually followed. By 'being actually followed', I mean that a convention *C* necessarily corresponds to the regularity of behavior *R* that occurs in a given community *G* under a given set of circumstances *S*. Another way to characterize this property is to say that a convention is *practice-dependent*. Feature (ii) indicates that a convention *C* has emerged and evolved through a process taking place in historical time, but that an alternative convention *C'* would have emerged and evolved had historical circumstances been different at some moment. This is the same as saying that a different convention *C'* (and thus a different regularity of behavior *R'*) could have existed in the very same community *G* under the very same set of circumstances *S*. Finally, feature (iii) is that there is no categorically imperative reason for following a convention *C*. By this, I mean that there are in principle reasons for following an alternative convention *C'* in the very same set of circumstances *S*.[13] In principle, a minimal reason for following a convention *C* is that each individual expects others to follow *C*. On this basis, I propose to characterize a convention in the following way:

> A rule *C* is conventional if and only if, for a community *G* and a set of circumstances S:
> 1) *C* is practice-dependent, historically contingent and arbitrary.
> 2) There is some *k*-order of mutual belief in *G* that *C* is followed in *S*.

The first condition follows from the three features stated above. The second is needed to ensure that the behavioral regularity *R* is not the result of pure randomness but rather of *intentional* behavior. Depending on one's preferred account of conventions, the *k*-order of mutual belief may vary between first-order mutual belief and common belief (as is the

[13] Available definitions of conventions in the literature (e.g., Marmor 2009) generally do not distinguish between historical contingency and arbitrariness. I think however that it is important not to conflate these two features. Indeed, the former feature refers to the *causal origins* of conventions while the latter rather refers to the *nature* of conventions. I return to the distinction between causal and constitutive dependency of morality upon conventions below.

case in Lewis's account). It is not needed to take a position on this last issue here.

On this basis, moral conventionalism can here be precisely defined as the view according to which the conventional nature/origin of morality concerns those rules that allow persons to coordinate and to cooperate.

Moral conventionalism has been endorsed by several economists and philosophers finding their inspiration in Hume's scholarship. In addition to Binmore (1998), Skyrms (1996) and Sugden (2005) have made significant contributions by attempting to show that fairness and morality are ultimately conventional—though they largely differ in their details. Asked to answer the skeptic's critique of moral naturalism, the moral conventionalist is most likely to simply reject the skeptic's two fundamental premises that 1) moral claims must depend on justifiably true beliefs and 2) moral claims have an unconditional normative force. The conventionalist's best defense consists in denying that there are the kinds of moral facts and moral claims of the sort that the skeptic argues for: facts and claims that depend on justifiably true beliefs and have unconditional normative forces. This is not a problem for the conventionalist though: there are other kinds of facts (let's call them 'conventional facts'), and according to the conventionalist these are the sole ones that constitute morality. These facts refer to tacit and arbitrary agreements between persons that solve coordination problems. This is clearly a view espoused by Binmore who emphasizes that fairness norms have been primarily designed to solve small-scale coordination problems. Though arbitrary, on some moral conventionalist accounts, conventions would progressively acquire a normative force in the population through a psychological process of habituation. Morality would then be nothing more than a set of conventions combined with some specific subjective feelings that people have toward them (Sugden 2005). According to Binmore, fairness norms are actually 'mere' conventions solving coordination problems. What makes these conventions moral is the nature of the coordination problems they are designed to solve. The choice of resource allocation in the game of life is the kind of coordination problems that falls in the realm of morality. Thus, the conventions established in the game of morals on the basis of the OP algorithm are moral in this sense.

Two general kinds of objections can be made against moral conventionalism, one empirical, the other philosophical. I do not regard

any of them sufficiently conclusive, which makes moral conventionalism an attractive meta-ethical stance against moral skepticism. To start with the empirical objection: moral conventionalism goes against a significant body of literature in empirical moral psychology which demonstrates (via experiments) that moral and conventional judgments differ in nature. Conventional judgments depend on conventions and thus respond to social practices, are arbitrary, and historically contingent. By contrast, moral judgments are generally regarded as lacking these three properties. Relatedly, moral and conventional rules are grounded on different kinds of judgments, and the conventionalist claims that morality is conventional is thus empirically false.[14] In particular, the empirical literature seems to establish that children of three years old, faced with some specifically designed tasks, exhibit an ability to distinguish between moral and conventional rules. Philosophical accounts interpreting these experimental results locate the distinction between morality and conventions both in *form* and in *content* (Southwood 2011). Regarding their form, moral rules tend to be characterized as non-contingent and global in scope, while conventional rules are characterized as contingent and more local. Regarding their content, it is suggested that moral rules deal with essentially other-regarding as well as impartial behavior and issues related to welfare, harm, fairness and trust. Conventional rules instead regulate self-interested behavior in the context of agreed-upon social practices.[15] Ultimately, it has been argued that the empirical evidence supports a conception of morality which has four constitutive properties—these are: seriousness, generality, authority-independence and objectivity. Conventional judgments and rules are believed not to have these properties (Kumar 2015).

It could be argued that the empirical evidence and its philosophical interpretations cast doubt over the relevance of moral conventionalism. It might be argued that the salience of the moral/conventional distinction for children or even adults is due to moral naiveté or

[14] The literature in development and moral psychology on the moral/conventional distinction is relatively abundant. The work of Elliot Turiel (1983) is generally regarded as seminal. Other important references are Smetana (1993) and Nucci (2001).

[15] Southwood (2011) argues that a philosophically more convincing way to ground the morality/convention distinction, still compatible with the empirical evidence, is by reference to whether or not a rule is practice-dependent. Specifically, contrary to conventional judgments, moral judgments are claimed to be practice-independent, i.e. they do not depend on the existence of a socially-agreed upon practice in the relevant community.

cognitive error. But, as a proponent of a Humean account of justice and morality would recognize, the evidence in support of that the distinction appears to be ignored *only* by persons with psychopathological tendencies "gives pause for thought" (Sugden 2008, 3). Which is to say, "It would be disturbing to have to conclude that psychopaths have a better understanding of the nature of morality than psychologically normal people do." (Sugden 2008, 3). Still, I think that the moral conventionalists can answer the empirical challenge of moral psychology in a way similar to Sugden (2008), who suggests that there are reasons to think that the very moral/conventional distinction is itself conventional. First, it should be noted that the empirical evidence is not as straightforward as some moral and development psychologists suggest. For instance, while gratuitous physical aggressions are virtually universally perceived as morally wrong, which actions belong to the category 'gratuitous physical aggressions' seem to vary across contexts and cultures (Haidt, Koller, and Dias 1993). In other words, moral judgments seem to be conventional after all. Other empirical studies establish that the transgression of some rules that are regarded as conventional (for example, rules of politeness, etiquette, and respect) in the western world, is considered to be harmful and serious in other cultural contexts (Sugden 2008). A second and related reason to doubt the empirical relevance of the moral/convention distinction is that the importance given to concepts of welfare, fairness, and trust, and which are supposed to be the objects of moral judgments is itself constitutive of western philosophy and liberal societies. As Sugden (2008: 20) notes, even proponents of the distinction tend to recognize that the concepts of welfare, fairness and trust should be understood subjectively. Of course, taken seriously, such a claim would entail that the very definitions of welfare and fairness can be the subject of conventional judgments, thus ultimately undermining the moral/conventional distinction.

The second objection against moral conventionalism is more philosophical and targets another distinction, that is: How can we distinguish moral from non-moral conventions? We should be able to discriminate between conventions that 'merely' solve coordination problems without any moral significance (e.g., on which sides of the road should we drive?) and morally loaded conventions (e.g., how should we punish murder? How should wealth be distributed in a population?). Binmore's account does not offer such a distinction but we can develop

some considerations on this issue. In particular, it may be argued that there are indeed authentic *moral* conventions. This is plausible even though there is a trap here: the fact that there are *moral reasons* to follow a convention does not make the convention a *moral* convention (Marmor 2009). On some accounts, I clearly have a moral reason to follow the convention about which side of the road one has to drive on, since not following it could lead to injuries or even deaths. What would be moral conventions then? Marmor suggests that the role of such conventions "is to mediate between abstract moral ideals and their concrete realization in our social interactions" (2009, 149). Consider the fact of giving to a charity. The latter is a moral ideal that gives indications and reasons for action. However, this is a very abstract ideal which leaves many issues unanswered: How much to give? To whom? How often? Marmor suggests that "[i]n such cases, conventions may evolve that specify norms of behavior that instantiate the moral principle of charity" (2009, 150). This definition is somewhat in accordance with our discussion of the moral/conventional distinction above. I have noted that while some actions, such as arbitrary physical aggressions, are universally condemned as morally wrong, the very characterization of an arbitrary physical aggression is itself conventional. It might be objected that on this account of moral conventions, conventions do not *create* morality but rather instantiate it. This seems to be quite different from the strongest forms of moral conventionalism (Binmore's included), which claim that morality is constituted by conventions. This is not really convincing, however, because the creation/instantiation distinction is actually illusory. Either we can maintain that moral properties and facts exist independently of social practices and are not created by them; in which case the moral skeptic's critique applies. Alternatively, we can maintain that moral properties and facts are practice-dependent; in which case morality is created at the same time that it is instantiated through social practices.

A stronger objection can be made, however. Indeed, the skeptic's critique can be reformulated along the following lines: why should we give any moral significance to conventions that (by definition) are ultimately arbitrary and contingent? The next section deals with this objection and argues that while moral conventionalism can eventually answer it, Binmore's specific account does not.

5. WHY BINMORE'S MORAL CONVENTIONALISM DOES NOT ANSWER THE SKEPTIC'S CRITIQUE

The skeptic's strongest objection to moral conventionalism relies on the claim that because moral conventions have amoral origins, they cannot have the kind of normative authority that any moral prescription is thought to have. Recall that one of the constitutive features of conventions is their arbitrariness. A minimal reason to follow a (moral) convention *C* is the expectation that others will also follow *C*, while there would be a reason to follow an alternative (moral) convention *C'* where one would expect others to follow *C'*. This has at least two implications. First, the reasons to follow a convention are *never* fully desire-*independent*. It depends on having appropriate preferences such that conforming to the social practice is best for the individual. Second, while I may have a desire-independent reason to follow a moral convention *C*, this reason can in principle be dominated by other desire-dependent reasons. The prisoner's dilemma is, of course, the prototypical case of such a situation. But it may also occur in pure coordination games, where while one may have a desire-independent reason to follow a moral convention *C*, the sole fact of expecting others to follow *C'* is sufficient to lead one also to follow *C'*. It follows that moral conventions do not have any necessary practical authority. As I noted in section 3, the lack of necessary practical authority is one of the skeptic's arguments against moral naturalism, and the very same argument could be used against moral conventionalism.

Of course, the moral conventionalist can respond in a way that is not available to the moral naturalist. The conventionalist may answer by claiming that moral conventions need not be endowed with any particular force and need not generate unconditionally dominant desire-independent reasons for action. The fact that people follow conventions to resolve issues related to morality or fairness should be taken as such, and there is nothing special about it. It might be argued that *why* people follow such conventions is irrelevant to our understanding of morality. I think Binmore, as well as other moral conventionalists, would be perfectly satisfied with this answer. Fairness norms have evolved as a coordination device in morally loaded coordination problems, and it is a fact that people follow them, which is, in itself, the evidence for the belief that they accept them. It is possible that there is nothing more to say about morality or fairness. In essence, this is very similar to Daniel Dennett's claim that moral norms function as "conversation-stoppers"

(1996, 506): they put an end to debates that otherwise cannot possibly be solved by finite computing machines.

This, however, leads to a further and ultimate difficulty. Suppose we accept all the conventionalist's claims and arguments. Together, they form a set of propositions about morality and fairness that we can denote as theory *T*. Binmore's account is a specific variant of *T*, but other similar conventionalist accounts are also instances of *T*. Suppose now, analogically to what macroeconomists are routinely assuming, that people form 'rational expectations'. By this, I mean that their beliefs and preferences about moral issues and matters of fairness are well-informed, *i.e.* they are generated on the basis of all the available and relevant information *I*. Suppose that people follow a set of moral norms and conventions *N* without necessarily ascribing to them a particular moral or normative value. Now, a critical test for conventionalism depends on the plausible answers we could give to the following question: *should people following N without knowing T continue to follow N once T is included in I?* For instance, learning that my belief that the current wealth distribution is fair is grounded on a norm that results from past bargains, where some agents had bargaining advantages (say, they were more skilled), should I continue to use this norm to form my beliefs about the fairness of the wealth distribution?

This question builds on the same intuition as Joyce's belief pills analogy, but is more about practical than theoretical reason. In essence, why should I continue to accept and act upon a particular claim or judgment about an issue once I realize that it originates from circumstances that have nothing to do with the issue at stake? It is plausible that a person introduced to Binmore's account, realizing that the fairness norms she is following result from power relations, should at least start to reflect on whether there are relevant reasons to continue to abide by the norms. Of course, since fairness norms are equilibria in the game of life and in the game of morals, the unilateral deviation is impossible (or at least irrational). By the very definition of the equilibrium concept, a player cannot increase his utility (measured according either to his personal or empathetic preferences) by using a different strategy. But, a coalition of disadvantaged individuals could in principle rationally deviate from the current equilibrium if they succeed

in coordinating to change their behavior simultaneously. This would lead, in turn, to a shift in the corresponding empathetic equilibrium.[16]

I do not think this problem necessarily undermines moral conventionalism, though. However, at this point, I would like to distinguish between Binmore's naturalistic and conventionalist account of fairness and another form of moral conventionalism that, taking inspiration from Gauss (2013), I will call 'Moral Conventionalism with Public Justification'. To understand the point of the distinction, it is useful to give a numerical example to illustrate how fairness norms solve coordination problems in Binmore's account. The example will make it clear why Binmore's account is vulnerable to the above critical test. Consider two individuals bargaining in the context of the game of life over the allocation of some divisible asset. Figure 1 below gives the payoffs (expressed in terms of von Neumann-Morgenstern utilities) of the two players (that, following Binmore, I name Adam and Eve) as a function of the asset distribution. The players' utilities are arbitrarily set on a 0-100 scale, and I assume that in cases where players fail to agree over an allocation, the asset is lost and both obtain a payoff of 0:

Asset distribution (Adam/Eve)	**90/10**	**80/20**	**70/30**	**60/40**	**50/50**	**40/60**	**30/70**	**20/80**	**10/90**
u_{Adam}	85	75	70	63	51	44	31	22	8
u_{Eve}	15	33	41	52	57	64	71	73	77

Figure 1

As indicated in figure 2 below, the Nash bargaining solution **N** corresponds to the allocation where Adam obtains 60 percent of the asset and Eve 40 percent (**D** is the disagreement point). Now, suppose that two individuals, John and Oskar, have to bargain over the asset and use the prevailing fairness norms to coordinate. In Binmore's framework, that means that John and Oskar are playing the game of morals and are using the OP device to solve their coordination problem. Following Binmore, we assume that no external authority can enforce the agreement obtained behind the veil of ignorance. Both players have to assume that they have an equal chance of being Adam and Eve once the veil is removed. As indicated in section 2, it follows that Oskar and John will bargain under symmetric empathetic preferences and will

[16] This issue cannot be dealt with in Binmore's framework since all his discussion is restricted to two-person bargaining games (though two-person games can be cooperative of course).

implement the egalitarian solution. Denote as U and V the unit of the empathetic scales that both Oskar and John use to value Adam's and Eve's payoffs respectively. As shown by Binmore (1998, Chapter 4), the value of U and V can be determined by choosing them such that the egalitarian solution *with Oskar's and John's empathetic utilities* correspond to the Nash solution *with Adam's and Eve's personal utilities*. Hence, we should have 63/U = 52/V, or U ≈ 6/5 V. Arbitrarily setting V = 1, we get $U \approx 6/5$. These values indicate how Oskar and John trade Adam's and Eve's personal utilities behind the veil of ignorance to reach an agreement. In this example, 6 units of Adam's personal utility are judged to be worth approximately 5 units of Eve's utility.

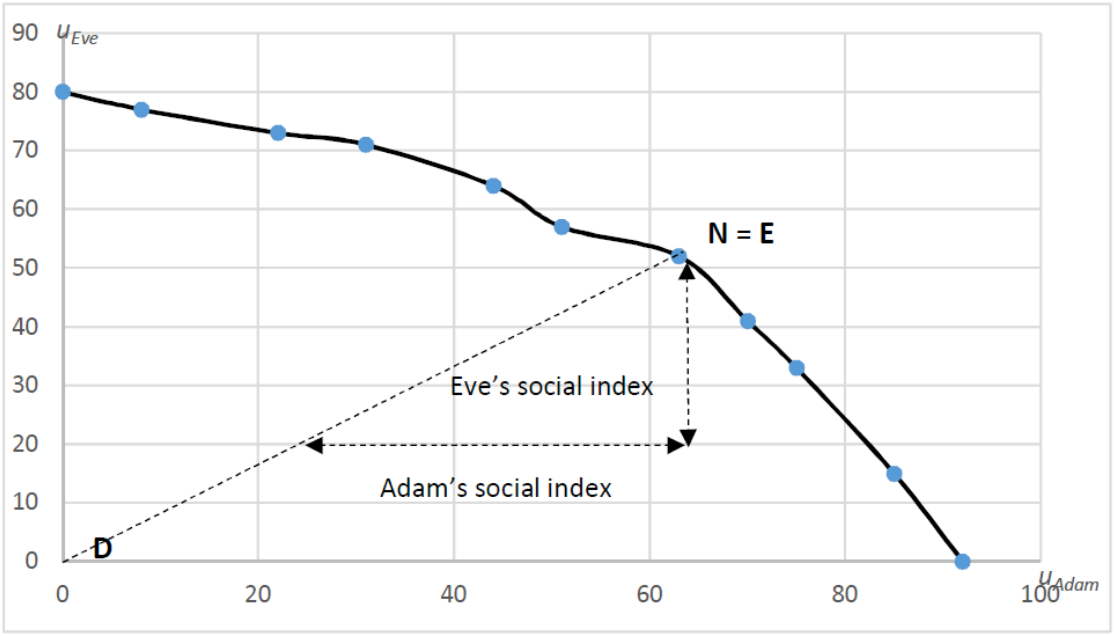

Figure 2

The empathetic preferences whose scales are determined by the variables U and V encapsulate the fairness judgments that Oskar and John use to solve coordination problems. This can be seen more clearly if we assume that the available quantity of the asset increases. As depicted in figure 3, this induces an expansion of the Pareto frontier and a modification of the Nash solution. The Nash solution now corresponds to the 50/50 bargain. However, in the short run, empathetic preferences remain unchanged by assumption. Oskar and John will thus continue to trade 6 units of Adam's personal utilities for 5 units of Eve's personal utilities. Using the OP device to coordinate, Oskar and John will implement the egalitarian solution for the *new* bargaining problem but

will use their *original* empathetic preferences. This leads to the coordination somewhere between the 60/40 and the 50/50 allocations. In the short run, the Nash and egalitarian solutions will thus no longer coincide, until cultural evolution induces a modification of empathetic preferences. Over the medium and long run, fairness norms are thus determined by natural and cultural evolution, especially bargaining power. But over the short run, they are used to coordinate in bargaining problems and do not reflect *current* bargaining power.

Binmore's fairness norms clearly have all the characteristics of moral conventions: they are grounded on past and current social practices, they could have had different content if past bargains had been different, and they are arbitrary in the sense that different empathetic preferences would also permit coordination on the Pareto frontier. Moreover, though this is not made explicit in Binmore's account, individuals bargaining under the veil of ignorance expect that the agreement corresponding to the egalitarian solution, given current empathetic preferences, will be reached.[17] I would argue, however, that it is insufficient to pass the critical test presented above. To see why, consider the reasoning of Oskar who ends up being Eve once the veil of ignorance is removed.

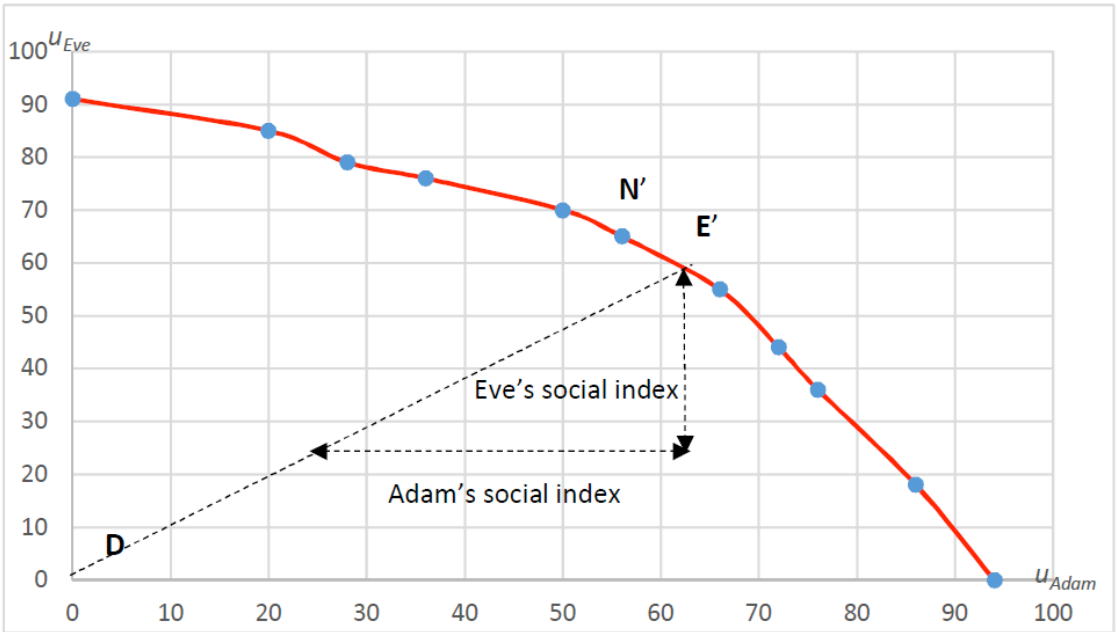

Figure 3

[17] This is true if we assume that players know their (identical) empathetic preferences and know (or at least strongly believe) that they are identical across the population.

Once the Pareto frontier has expanded, Oskar will obtain approximately 45 percent of the asset through playing the game of morals with John (whom we assume ends up being Adam once the veil is removed). Oskar is thus slightly disadvantaged, but *from his perspective*, the result is fair. This is due to the fact that the agreement is obtained by using Oskar's and John's empathetic preferences which they have inherited from past bargains. However, were Oskar to realize that his empathetic preferences are the result of bargains made in the past and whose outcomes have been determined by an *old* bargaining power structure that no longer prevails, there is no reason that to think that he would agree to an allocation that is worse for him than the one he would obtain using his *current* bargaining power. Indeed, actually, Oskar could use threats to implement the Nash solution and obtain half of the asset. John would, of course, disagree, arguing that by the prevailing standards the allocation is fair. But of course, this begs the question: Oskar would reply that what is fair is determined by bargaining relationships and that there is no reason to use *past* bargaining power rather than *current* bargaining power to allocate the asset. The point is that fairness norms play the role of coordinating devices if and only if individuals fail to reflect over the content and the origins of their empathetic preferences. Then, in this latter case, fairness norms are moral conventions that indeed play the role of Dennett's conversation stoppers. They put an end to the bargain and avoid costly negotiations.

Now, this may be an adequate account of how individuals actually solve coordination problems. It may be the case that in many situations, we play the game of morals almost automatically without reflecting on the content of our fairness judgments. The latter are just what they are, we expect others to make similar judgments, and we do not give more thought to this. However, this is not sufficient, because Binmore's account is explicitly about both descriptive *and* prescriptive ethics. While we may grant that this is an interesting account of individuals' actual reasoning in coordination problems, it is definitely not a convincing account of what makes morality special. Consider once again the above critical test. Undoubtedly, some persons in the population, even knowing theory *T*, would answer 'yes' to it. The reason for this is simply that it would be in their personal interest to continue to follow the set of norms *N* (it would probably be the case of John in our example). Cynics would concur: fairness norms are often nothing more than a '*cache-misère*' and advantaged people are well aware that fairness

is a convenient justification for the preservation of an unbalanced *status quo*. In some way, this is a vindication of Binmore's claim that fairness is ultimately grounded in power and nothing more. But this also shows that a pure behaviorist view of morality is ultimately untenable. In other words, the conventionalist cannot safely ignore the motivations and the reasons for action that are underlying the norm-abiding behavior, especially in the case of moral norms. This is well put by Philip Kitcher:

> … it's important to demonstrate that the forms of behavior that accord with our sense of justice and morality can originate and be maintained under natural selection. Yet we should also be aware that the demonstration doesn't necessarily account for the superstructure of concepts and principles in terms of which we appraise those forms of behaviour (Kitcher 1999, 222-3).

In Binmore's account, the morality of fairness norms is epiphenomenal since ultimately it reduces to (rather than merely supervenes on) power relations. Conventionalists like Binmore have to argue that there is nothing more to morality than 'conversation-stopping' devices. At the same time, if once they are aware of the genealogy of their fairness judgments, people continue to abide by them only because of necessity or personal interests, this should arguably matter to any account of morality. If there is nothing distinctive about morality, one may wonder why it is worth seeking to provide it with naturalistic foundations.

However, moral conventionalism is by no means condemned to failure. As I anticipated above, 'Moral Conventionalism with Public Justification' avoids almost all the difficulties discussed in this section. This form of moral conventionalism is grounded on its endorsement of what Gauss calls the "Public Justification Principle" (2013, 80):

> *The Public Justification Principle*: If a moral convention *C* in a community *G* is endorsable by all members of *G*, *C* is a genuinely moral convention.

A *genuinely* moral convention is a convention that has the moral authority that the moral skeptic claims a moral prescription must have. The *Public Justification Principle* thus implies that there are two kinds of moral conventions: those that are genuinely moral and those that are not. The latter are moral conventions that, though they exist in the relevant community *G*, do not impose any moral obligations on the

members of *G*. Whether a moral convention is genuine or not is left to the judgment of morally autonomous and competent agents. A morally autonomous and competent agent must determine for each existing moral convention if it provides justified desire-independent reasons for conforming *beyond* the desire-dependent reason constitutive of any convention. If this is the case, the agent will endorse it, i.e. the agent will follow it as long as a sufficient number of other individuals also follow it. If there are no such desire-independent reasons for following the convention, then one may be justified (though perhaps not required) in choosing not to follow the convention (for instance in the case one considers that there is an overriding desire-independent reason not to follow it). Given the fact that moral conventions solve coordination problems, there must exist, in a given community, a relative convergence over which conventions are judged to be genuinely moral. Too important a disagreement would entail that few, if any, moral conventions were consistently followed. Since a moral convention cannot exist if an insufficient number of individuals are ready to endorse it, the community would be deprived of any consistent and stable system of morality. Gauss (2013) suggests that such moral stability and consistency necessitate what Rawls (2005) has called 'public justification': there must be some public knowledge of which moral conventions are endorsable by all the members of the community. Conventional morality cannot exist without such public justification.

I think that *Moral Conventionalism with Public Justification* succeeds in passing the critical test. It also helps to make clear why Binmore's account fails: in Binmore's account, nothing indicates that empathetic preferences are publicly endorsable. This failure is due to the fact that what makes conventions genuinely moral is their ability to be endorsed for reasons that all members of the relevant community would accept after careful moral reflection.

6. CONCLUSION

This article has provided an examination of Binmore's game-theoretic account of fairness as an instance of moral conventionalism. I have suggested that Binmore's naturalization of the OP leads to a view that morality is conventional. In this sense, it seems to provide an answer to moral skepticism. However, in the specific case of Binmore's account of fairness, the moral conventionalist answer leads to a purely behaviorist view of morality and fairness. Moral motivations and reasons are then

completely ignored. There is, then, nothing special in morality. Still, I have suggested that other forms of moral conventionalism that emphasize the importance of reasons that establish moral conventions escape most of the difficulties of Binmore's account.

REFERENCES

Alexander, J. McKenzie. 2010. *The Structural Evolution of Morality*. Reissue. Cambridge, UK: Cambridge University Press.

Binmore, Ken. 1987. "Modeling Rational Players: Part I." *Economics and Philosophy* 3 (2): 179–214.

Binmore, Ken. 1994. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

Binmore, Ken. 1998. *Just Playing: Game Theory and the Social Contract.* Cambridge, MA: MIT Press.

Binmore, Ken. 2005. *Natural Justice*. New York, NY: Oxford University Press.

Bowles, Samuel, and Herbert Gintis. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton, NJ: Princeton University Press.

Dawkins, Richard. 1989. *The Selfish Gene*. Revised edition. New York, NY: Oxford University Press.

Dennett, Daniel C. 1995. *Darwin's Dangerous Idea: Evolution and the Meaning of Life*. New York, NY: Simon and Schuster.

Gauss, Gerald. 2013. "Why the Conventionalist Needs the Social Contract (and Vice Versa)." *Rationality, Markets and Morals* 4: 71-87.

Gintis, Herbert. 2006. "Behavioral Ethics Meets Natural Justice." *Politics, Philosophy & Economics* 5 (1): 5–32.

Haidt, J., S. H. Koller, and M. G. Dias. 1993. "Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog?" *Journal of Personality and Social* Psychology 65 (4): 613–28.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61, 434-435.

Kitcher, Philip. 1999. "Games Social Animals Play: Commentary on Brian Skyrms's Evolution of the Social Contract." *Philosophy and Phenomenological Research* 59 (1): 221–28.

Kumar, Victor. 2015. "Moral Judgment as a Natural Kind." *Philosophical Studies* 172 (11): 2887–2910.

Mackie, John Leslie. 1977. *Ethics: Inventing Right and Wrong*. New York, NY: Penguin.

Marmor, Andrei. 2009. *Social Conventions: From Language to Law*. Princeton, NJ: Princeton University Press.

Nucci, Larry P. 2001. *Education in the Moral Domain*. Cambridge, UK: Cambridge University Press.

Rawls, John. 2005. *Political Liberalism*. New York, NY: Columbia University Press.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50 (1): 97–109.

Sinnott-Armstrong, Walter. 2015. "Moral Skepticism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2015. Retrieved from <http://plato.stanford.edu/archives/fall2015/entries/skepticism-moral/>.

Skyrms, Brian. 1996. *Evolution of the Social Contract*. New York, NY: Cambridge University Press.
Skyrms, Brian. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge, UK: Cambridge University Press.
Smetana, Judith. 1993. "Understanding of Social Rules." In *The Child as Psychologist: An Introduction to the Development of Social Cognition* edited by Mark Bennett, 111-141. New York, NY: Harvester Wheatsheaf.
Southwood, Nicholas. 2011. "The Moral/Conventional Distinction." *Mind* 120 (479): 761–802.
Sugden, Robert. 2001. "Ken Binmore's Evolutionary Social Theory." *The Economic Journal* 111 (469): 213–243.
Sugden, Robert. 2005. *The Economics of Rights, Cooperation and Welfare*. 2nd ed. Palgrave Macmillan.
Sugden, Robert. 2008. "Is There a Distinction between Morality and Convention?" *Working Paper Series*, University of East Anglia, Centre for Behavioural and Experimental Social Science (CBESS) 8–1. Norwich, UK: School of Economics, University of East Anglia.
Turiel, Elliot. 1983. *The Development of Social Knowledge: Morality and Convention.* Cambridge, UK: Cambridge University Press.
Verbeek, Bruno. 2008. "Conventions and Moral Norms: The Legacy of Lewis." *Topoi* 27 (1–2): 73–86.
Young, H. Peyton. 2001. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions.* Princeton, NJ: Princeton University Press.

Cyril Hédoin is full professor of economics at the University of Reims Champagne-Ardenne (France). His academic work essentially belongs to the philosophy of economics and to institutional economics. He has recently published papers in *Economics and Philosophy*, *the Journal of Economic Methodology* and the *Journal of Institutional Economics*.
Contact e-mail: <cyril.hedoin@univ-reims.fr>