Œconomia
History, Methodology, Philosophy

15-4 | 2025
Varia

Fairness, Game Theory and Public Policy. The Case of Ken Binmore

Equité, théorie des jeux et politiques publiques. Le cas de Ken Binmore

Miłosz Ślepowroński

Electronic version
URL: https://journals.openedition.org/oeconomia/19389
DOI: 10.4000/15ea9
ISSN: 2269-8450

Publisher
Association Œconomia

Printed version
Date of publication: December 1, 2025
Number of pages: 603-631
ISSN: 2113-5207

Provided by Uniwersytet Jagielloński

Electronic reference
Miłosz Ślepowroński, "Fairness, Game Theory and Public Policy. The Case of Ken Binmore", *Œconomia* [Online], 15-4 | 2025, Online since 01 December 2025, connection on 09 March 2026. URL: http://journals.openedition.org/oeconomia/19389 ; DOI: https://doi.org/10.4000/15ea9

Fairness, Game Theory and Public Policy. The Case of Ken Binmore

Miłosz Ślepowroński*

The purpose of this article is to assess the normative force of the models of social contract formulated using evolutionary game theory, partially to find out how they can be more impactful outside of their current niche. As the clearest example of this approach, we focus on the theory developed by Ken Binmore. None of the other authors matched Binmore's scope and ambition. Binmore proposes a description of human evolution that leads to what can be called cautious egalitarianism. In the article, Binmore's theory is first reviewed, showing why we need a coordinating device for his "game of life" and how it evolved. Importantly, we show that Binmore does not only offer an abstract model or a description of how fairness norms work in human societies—his focus is not purely descriptive one. He explicitly writes about the construction of rational ethics and how to use it to guide public policy. Due to two types of problems, Binmore's theory can also be used to undermine the practical application of fairness norms, rather than promoting it. However, the solution to both types of problems lies in using Binmore's fairness along with other coordination mechanisms in a polycentric order that does not try to privilege a single one.

Keywords: social contract, fairness, egalitarianism, game theory

Equité, théorie des jeux et politiques publiques. Le cas de Ken Binmore

L'objectif de cet article est d'évaluer la force normative des modèles du contrat social formulés à l'aide de la théorie des jeux évolutifs, afin de déterminer comment ils peuvent avoir davantage d'impact au-delà de leur niche actuelle. À titre d'exemple, le plus clair de cette approche, nous nous concentrons sur la théorie développée par Ken Binmore. Aucun autre auteur ne propose une théorie d'ampleur et d'ambition comparable. Binmore propose une description de l'évolution humaine qui conduit à ce que l'on peut appeler un égalitarisme prudent. L'article commence par passer en revue la théorie de Binmore, en montrant pourquoi un mécanisme de

*Doctoral School in the Social Sciences, Jagiellonian University, Krakow, Poland. milosz.slepowronski@doctoral.uj.edu.pl
The author would like to thank the editors and anonymous reviewers as well as participants in the 2024 7th International Conference Economic Philosophy in Reims, and the members of The Polish Philosophy of Economics Network for their helpful remarks and suggestions.

coordination est nécessaire pour son « jeu de la vie » et comment celui-ci a évolué. Surtout, nous montrons que Binmore n'offre pas seulement un modèle abstrait ni une description du fonctionnement des normes d'équité dans les sociétés humaines ; sa perspective n'est pas purement descriptive. Il traite explicitement de la construction d'une éthique rationnelle et de la manière de l'employer pour orienter les politiques publiques. Toutefois, en raison de deux types de difficultés, la théorie de Binmore peut aussi servir à saper l'application pratique des normes d'équité, plutôt qu'à la promouvoir. La solution à ces deux difficultés consiste à mobiliser la conception de l'équité chez Binmore conjointement avec d'autres mécanismes de coordination au sein d'un ordre polycentrique qui ne cherche pas à en privilégier un seul.

Mots-clé : contrat social, équité, égalitarisme, théorie des jeux

JEL: C73, D63, D71, D02

The past thirty years have seen game theory applied by many philosophers and economists to explore the origins of justice and the social contract, by formalizing and naturalizing these ideas (Alexander, 2007; Skyrms, 2014; Vanderschraaf, 2018). This method is powerful because it introduces new tools to address the age-old questions of institutional design. Evolutionary game theory contributes in two key ways: by modeling the social contract as a theoretical normative framework and offering a descriptive approach that illustrates how certain norms or rules might have actually evolved. This article examines one of the most ambitious theories in this field, developed by Ken Binmore (1994; 1998; 2005), and argues that his normative goal of expanding the use of fairness may be undermined by his own theory.

Binmore argues that the Rawlsian concept of the original position is more than just a philosophical thought experiment; it reflects something fundamental about how fairness norms operate in reality. He sees it as a coordinating device for navigating the "game of life", helping us address essential problems of social cooperation. Moreover, Binmore believes that this mechanism is embedded in our genes, having enabled our ancestors to successfully coordinate their behavior over extended periods of time. Each aspect of this summary requires further exploration: Why do we need a coordinating device? What exactly does it entail? And, why does Binmore link it to genetics?

Section 1 of this article addresses these questions by outlining Binmore's theory to show how he builds his case for the normative role of fairness. Section 2 examines his practical aim of broadening the application of fairness. Finally, Section 3 highlights two issues in Binmore's theory: one internal, concerning the challenge of generalizing small-scale solutions (which Binmore himself warns against, but still attempts), and an external issue, the absence of comparisons with alternative solutions to coordination problems, which should be considered before applying his theory in public policy. I am not focusing on whether Binmore's descriptive theory is accurate, since there are already numerous critiques addressing specific issues in his theory (Gintis, 2006; Mackie, 2006; Schmidt-Petri, 2006; Skyrms, 2006). Assuming that these issues can be resolved, my goal is to examine whether Binmore's theory, if true, supports the justification for egalitarian policies that it aims to provide. My main argument is that despite its ambition and innovative approach, Binmore's theory ultimately fails to justify egalitarian solutions and can, in fact, be interpreted to support the opposite, undermining egalitarian policies. However, there is a solution to this problem, which is situating Binmore's fairness within broader, polycentric context.

1. Binmore's Theory

Ken Binmore proposes that fairness can be explained and naturalistically justified. He uses game theory to model the evolution of fairness and show that it exists universally and is an important coordination device for our daily problems. This part focuses on the basis of the descriptive part of his theory, as it is needed to critically discuss the requirements for the normative ideas and how he intends to use them.

1.1 Rawls Naturalized

Fairness takes a form similar to the Rawlsian original position:

> When two people use fairness to resolve an everyday coordination problem, I believe they are implicitly calculating the agreement that they would reach if they were to bargain on the assumption that their identities would be reassigned at random after negotiation was over (Binmore, 2005, 21).

Thus, fairness arises as the result of bargaining and is not used during it. To show an example, let us consider the following game, known as the Meeting Game. There are two players, Adam and Eve, who want

to meet in one of the ten places, and they get different payoffs from every one of the places. If they do not meet at all, Adam gets a payoff of 4 and Eve of 2.

Table 1. The Meeting Game

Location	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Adam's payoff	0	6	16	19	22	28	34	40	44	46
Eve's payoff	36	36	34	32	30	26	22	18	8	0

Source: Binmore (2005, 23).

The two most obvious solutions are 6 and 8, which can be named egalitarian and utilitarian, respectively. Binmore wants to focus our attention on number 7, selected by the Nash bargaining solution. Strategic bargaining among rational players, when the characteristics of the players and the nature of the bargaining problem are common knowledge, converges on that particular solution (Binmore, 2005, 25). It approximates the agreement that would be reached when two rational players with equal bargaining power negotiated face-to-face without appealing to any fairness norms.

But the question asked by Binmore is: What would Adam and Eve choose if they were to bargain in the original position, when the roles that they would end up with were unknown to them? To answer that, we need to consider Binmore's notions of the game of life and the game of morals. The game of life is a metaphor that describes all our daily coordination and competition problems, and its goal is to establish the rules of production and distribution of goods. It is repeated an infinite number of times, and its rules are the same as the rules of our natural world. They depend on technology, social structure, and the basic laws of physics and biology. They are binding to the same degree, which means that purely social rules are not included in them, which are binding only insofar as there are enforcement mechanisms present. Individuals cannot change the rules of the game of life.

The game of morals is played between the rounds of a game of life, and when playing it, the players have the possibility of an appeal to the

original position. If this happens, they negotiate behind the veil of ignorance, pretending not to know which person they would end up being (Binmore, 1998, 15). The object of their negotiation is the equilibrium in the game of life that should be operated. In this way, the game of morals becomes a coordination device for the game of life (Binmore, 1998, 42). This is despite the fact that no rules of the game of morals are truly binding. But why are we even considering this? Binmore says here something interesting for any game theoretical model: When modeling reality one must always consider how the game of morals would be played if its rule were binding, and that his work is akin to that of a novelist, populating his world with unreal people but then keeping to it and treating its made-up rules seriously (Binmore, 1998, 41). In fact, no rules, aside from those of a game of life, which include laws of physics and the like, are truly binding. Therefore, any "binding" social contract is interesting as long as playing by its rules is in the best interest of the parties involved. If someone respects the rules of a game of morals, it is because she wants to. Analyzing the game of morals and its rules is informative because its rules evolved as compatible with the game of life. No leaders or external coordination devices are required.

But the rules of the game of life and the game of morals do not answer quite a crucial question: Why should we bargain in the original position and not directly in this world? Why has such a peculiar apparatus evolved? Binmore here proposes that it works as an equilibrium selection device. Because Binmore represents our daily interactions as repeated coordination games, he has access to the Folk Theorem, which shows that any efficient outcome of a non-repeated game on which the players might agree on approximates an equilibrium outcome of the repeated version of this game (Binmore, 2005, 8). Binmore does not claim that any particular game must represent all social interaction. In fact, the specific game is not very important if we are analyzing long-term repetitions. He even claims that the question whether cooperation is possible is trivial, and the answer, for someone who actually cares about empirical facts, is obvious. The interesting question is what kind of cooperation is possible and which should be chosen. It looks like our first problem (how is cooperation possible) is solved if we focus on repeated games. However, introducing the Folk Theorem leads to additional problems. The space of possible stable cooperative arrangements becomes so large that even if we limit our choice to Pareto-efficient solutions, pure rationality cannot help us select a unique solution. We cannot count on social evolution to solve it for us either, as it may just

as well lead to inefficient equilibria. We can also be stuck in an old inefficient equilibrium when a new technology expands the space of possible social contracts. But even if evolutionary processes could be counted on, their work is too slow. Therefore, we should not expect to have a certain equilibrium ready and applicable for every situation. Instead, we need some kind of equilibrium selection device that could work in any situation. Binmore argues that fairness is the tool that solves all these problems, as it can get us quickly out of bad or inefficient equilibria (Binmore, 2005, 27).

Binmore proposes an evolutionary history of fairness as a tool of mutual insurance against misfortunes, such as hunting failures. First, it was used within families. Individuals behind an actual veil of uncertainty, where they did not know if they would be successful in hunting or foraging, insured each other against hunger. This contract required only imagining yourself in the future in two different situations. The second step involved the veil of ignorance, in which individuals did not know who they would turn out to be after the negotiations. It was then expanded to cover more distant relatives and then the non-kin. The reason is simple: fairness turned out to be effective (Binmore, 2005, 140). At the same time, the whole bargaining story told above is just a theoretical approximation. The real coordination operates much faster, without us, in fact, bargaining or simulating it and using the game of morals. We simply do not have time to bargain, but the mere possibility of it and its rare occurrences fixes its results in the currently operated social contract. Fairness evolved to solve problems that we face constantly, when there is no time to bargain over everything, and we use fairness as a shortcut to arrive at a workable solution. Additionally, it could have evolved even before we were able to use language (Binmore, 2005, 27). So, in fact, people do not really bargain but coordinate on a deal that they would agree on if they were bargaining behind the veil of ignorance. But for the solution to be meaningful, be it egalitarian, utilitarian, or Nash, players' rewards must be multiplied by social indices. Without any social index to rescale all rewards, it would not be possible to compare utilities.

To find what links different personal preferences, Binmore urges us to look at empathy, which is another facet of his general idea that game theory is a tool useful primarily to operationalize Hume's ideas (Binmore, 1994, 285). However, he does not want to abandon the model of self-interested persons as he carefully distinguishes empathy and sympathy. Sympathy is the ability to identify with another person so closely that one does not distinguish between one's own preferences

and that of the person with whom one sympathizes. Binmore follows Hume and Adam Smith in claiming that while altruism might have had some influence in the past, problems that can be solved using altruistic preferences are not important for modern society (Binmore, 1994, 56). Love is not the cement of society, but rather reciprocity. But while altruism, or sympathy, is not so important, empathy is said to lie at the very heart of human social organization: "without the ability to empathize, we would not be able to operate decentralized social contracts" (Binmore, 2005, 115). When we empathize with someone, we run our internal model of ourselves in someone else's situation, with the parameters of that person, to predict how we would feel under different situations. All "properly socialized" humans are said to have the ability to empathize (Binmore, 2005, 114).

Binmore claims that, in addition to empathy, people also have empathetic preferences. We express them while saying that we would rather be one person under one set of circumstances than another under a different set (Binmore, 1994, 59). Empathetic preferences must be distinguished from personal preferences since they exist precisely to take into account different tastes of different persons. When we empathize, we do not forget who we are and we retain our original preferences. We express empathetic preferences when we say that, for example, we would rather be Eve wearing a fig leaf (when we know that Eve is modest) than Adam eating an apple, knowing that Adam likes apples but does not mind being naked (Binmore, 2005, 113). If a person has a consistent set of empathetic preferences over a set of possible states, then that person can compare Adam's and Eve's utility on von Neumann and Morgenstern utility scales, which are derived from their personal preferences. In this way, one can determine the rate at which Adam's utilities are traded against Eve's and thus make comparisons of the utilities. But this is still an intrapersonal comparison, not an interpersonal one. In order to bridge inter- and intrapersonal preferences, Binmore claims that interpersonal comparisons of utility conducted by people with similar cultural background, social status, and personality are very similar, due to social evolution (Binmore, 1994, 62). Empathy formulated in this way began within the family, members of which are concerned about the well-being of another due to inclusive fitness. The device of the original position was postulated by Binmore to be useful in families, but then it was used by some human ancestors to coordinate on an equilibrium with a non-kin. This could have happened due to a mistake, but the result was that some part of the population started to coordinate with the non-kin in the

same way as they did with their kin—they "stumbled" upon the more successful equilibrium and stayed by it.

The important thing to mention here is that only ratios of social indices matter, not their absolute values. Any point in the meeting game can be achieved if we rescale the payoffs by some social indices, and they work differently if we use different solutions. If a utilitarian solution is selected, then it is better for a player to have a social index with a large value, if egalitarian, a small value is better. The details are not as important as the result that the evolution that changes social indices will change the outcomes.

1.2 The Deep Structure

The equilibrium selection mechanism is at work every time we see a coordination problem and automatically think of an equal split. According to Binmore, this is a clue that evolution endowed us with a preference for this particular mechanism. It is supposedly caused by a "deep structure" of human morality, which would unify at a deeper level all the different moralities observed. One proof of the existence of this deep structure is that all primitive societies that survived to modern times share a strikingly similar social contract. This is true even for groups living in very different environments, such as Greenland, the Kalahari Desert, or the Amazon rainforest. All of them are profoundly egalitarian, and relationships between people are based on sharing, without any long-term dependencies (Woodburn, 1982). All distinctions of wealth, power, and status are systematically eliminated. Some individuals strive for hierarchical power, but these societies counter these instincts by a "reverse dominance hierarchy" (Boehm, 1999). When an individual starts to accumulate power, he will be warned, mocked, ostracized, and finally punished, up to killing him. The fact that every hunter-gatherer society has a similar arrangement, despite very different environments and a lack of contact between them, is a clue that humans have a genetically determined propensity for this kind of social contract. Our hunter-gatherer ancestors probably shared this kind of arrangement, and due to the length during which it operated, it became written into our genes. As social contracts of human societies obviously differ, to reconcile this trivial observation with the postulated underlying unity of form, he proposes that what is different are standards of interpersonal comparisons of utility, which are needed as inputs in the original position. They are determined by specific cultures and change along with cultural evolution. So, every

culture and society has different standards of comparison, but they all share a deep structure.

Fairness is thus written into our genes, and it is supposed to work exactly like the Rawlsian original position. Binmore thinks that the original position is just another form of the old Golden Rule, do-as-you-would-be-done-by, but taking into consideration different tastes. Moral behavior started inside families, and sharing by using something similar to the original position allowed its users to fare better than others. Then, evolution, after teaching us to use the original position with the kin, "expanded the circle", and we started to coordinate with the non-kin. This was completely accidental, as we coordinated on equilibria in some games as if they were other games; we stumbled upon an equilibrium by pretending to be bound by rules of a more restrictive game, played with our family (Binmore, 2005, 171). Therefore, what was completely accidental and happened mainly for nonbiological reasons, after some time, became written into our genes and biologically cemented.

This shows that Binmore's original position is not the same as Rawlsian. In Rawls' theory, we must imagine ourselves behind the veil of ignorance, without any knowledge of our place in a society and our abilities. In Binmore's original position, we know who we are, just pretend that we do not know which specific person we are in some situation. Rawlsian theory leads to the usage of the difference principle and the selection of primary goods, whereas Binmore mainly concerns himself with the usage of the original position as a device for solving daily, mundane coordination problems, not for creating institutions. Rawlsian theory has a clear normative content, whereas in Binmore's it is somewhat ambiguous. Most importantly, unlike Rawls, Binmore's theory allows for very unequal outcomes, since all payoffs are calculated with the usage of social indices. The individuals behind Rawls' veil of ignorance are essentially the same, so assigning them unequal payoffs would be arbitrary. Binmore's theory takes into account the tastes of different people, but also their relative power, so interpersonal utility comparisons are needed as input. Its use is not justified normatively, as he claims that this is purely descriptive since this is what actually people do.

1.3 Binmore's Social Contract

The device of the original position resembles the Golden Rule: "Do as you would be done by". The Golden Rule is supposed to have arisen

in insurance contracts, made mostly within family, which were then extended to others. Generally, by coordinating on an equilibrium in a game of life, if two animals are able to monitor each other's behavior, they can achieve any result that could be achieved with a legally binding contract, that is, with external enforcement (Binmore, 2005, 139). Thus, mutual insurance contracts work like bargaining behind a veil of ignorance when we do not know whether we will be lucky or unlucky. To do that, we must imagine ourselves in the shoes of our future selves. Individuals capable of such an imagination were selected by evolution if mutual insurance was adaptive. Then, the neural wiring necessary to insurance and imagining in future selves evolved to operate the original position. But we should not expect such insurance contracts against hunger to be common among strangers. Their origins are in the family, since their members could trust each other to uphold the bargains and cooperate with them on a long-term basis. The next lucky step involved a mistake—misreading signals from the environment and taking some stranger for a kin. Thus, a behavior from one game was applied to another, rules of coordinating in an inner circle were applied to a wider circle. They were not always successful, but sometimes people "stumbled upon" an effective equilibrium.

In short, the evolutionary history of the original position can be described as follows. First, individuals behind a veil of uncertainty (where they did not know whether they will be successful in hunting or foraging) insured each other against hunger. This contract required only sympathy towards the self and imagining oneself in the future in two different situations. The second step involved the veil of ignorance, where the individuals did not know to whom they would turn after the negotiations and required empathy towards others. The final step involved either a utilitarian or egalitarian social contract, which depends on the existence of the external enforcement.

Binmore's idea makes sense in light of a U-shaped curve of human egalitarianism/despotism (Cushman et al., 2006). This theory refers to our knowledge that the closest relatives of humans, chimpanzees and gorillas, live in very hierarchical societies, apart from bonobos. It seems probable that some early human ancestors also formed groups with such a linear hierarchy. But, as evidenced by many anthropological studies, human hunter-gatherer groups formed very egalitarian groups which do not resemble anything that our primate cousins live in (Boehm, 1999). Obviously, contemporary societies are not as fiercely egalitarian as hunter-gatherers, even in countries which champion the ideas of equality the most. The change (or return) from egalitarianism

to different kinds of hierarchies seemed to coincide with the rise of agriculture. Small-scale societies of hunter-gatherers were organized by face-to-face sociality. This small and decentralized social contract was well suited to keep its egalitarian norms. But when the size of cooperating groups increased above the number that can be maintained by the human mind, the continuation of this early egalitarianism became complicated. Societies changed and saw the first hierarchical organizations at the same time as they became much more numerous than before. Once these hierarchical societies came into being, warfare between them created selection pressures for larger group size and better organization, which meant organized military. This led to a further centralization and enlargement of hierarchies, ultimately leading to leaders being sometimes recognized as gods or their kin. The U-curve view must be expanded since this extreme form of hierarchism has changed since then. Some lines of argument point to the role of great religions and philosophies in promoting egalitarianism and limiting the naked power with legitimate authority (Bellah, 2011).

Binmore's theory relies here on the claim that human ancestors have lived in strongly egalitarian societies for a long time, long enough for biological evolution to work and imprint on us the norms of these societies. Some genetic differences must have arisen, so evolution could suppress the instinct to dominate (Binmore, 2005, 134). The same cannot be said about societies living under nonegalitarian social contracts, be it feudal or democratic, as they are results of a cultural evolution. Therefore, Binmore concludes, present-day humans must share some propensities with hunter-gatherers. Thus, using some elements of our ancient social contract, ingrained in our minds, seems to be both effective and conductive to happiness.

Binmore's egalitarian solution is peculiar and is not what most philosophers would have called it. It does not specify any social institutions, rights, or liberties, does not describe any social system or practice in detail, and is limited to considerations about the abstract form of the distribution of goods, but nevertheless, the results would surprise any egalitarian or utilitarian. On the face of it, he claims that the system he discusses should lead to the maximization of the total utility. But this utility is not measured equally and added, but is rather rescaled by the social indices obtained due to the standards of interpersonal comparison of utilities. Therefore, gains are measured differently due to the cultural standards of a given society, which are in turn influenced by a power structure. Reasons for unequal social indices are thus distinctively nonutilitarian. A classical utilitarian would set social

indices equal for many different philosophical reasons, which are all discarded by Binmore in favor of social evolution (Schmidt-Petri, 2006). He also rejects Harsanyi's way of finding rational social indices, because they just do not exist, as every kind of social index is culturally determined. However, unsatisfactory from a utilitarian perspective, this is more in agreement with Binmore's overall approach, as he is not interested in philosophical reasons for what should be but what is, at least to some extent.

The first part of Rawlsian reasoning in the original position, specifying primary goods and basic liberties, is altogether discarded. We are left only with something similar to the difference principle, with unequal distributions of goods due to the role of different social indices and the state of nature being identified with the status quo (Binmore, 1994, 14), which makes the existing power structure important:

> Power is built into the rules of the game of life, which say who can do what, and when they can do it. Since the rules of a game determine what outcomes can be sustained as equilibria, it follows that power is reflected in the size and shape of the set of feasible social contracts (Binmore, 2005, 41, our emphasis).

This power should not be conflated with authority, as it is a property of the operated social contract. Inequalities in birth, naked force, bargaining power, intelligence, etc. are also not equalized here, which makes this egalitarianism highly unusual. Furthermore, Binmore's use of Rawls' idea might even be seen as contradictory to the reason behind that idea. Rawls' original position was constructed to remove, as much as possible, the influence of status when determining the principles of justice. In Binmore's version, interpersonal comparisons of utility are under the influence of power inequalities, and if people's empathy preferences are adapted to it, it undermines the original motivation (Qizilbash, 2009).

Moreover, as the payoffs are weighted by the social indices, the supposedly egalitarian outcome may have very little to do with equality of anything. Therefore, what many egalitarians would consider arbitrary from a moral point of view is actually allowed in Binmore's theory (Schmidt-Petri, 2006). Rawls argued that fairness in its basic form exists universally and in every society there must be a set of circumstances where fairness is applied. The differences between societies lie in where and how widely they apply fairness and also in how important it is compared with other values. Binmore agrees with this, but applies his fairness with utilities rescaled by social indices.

2. Binmore's Practical Goal

The description of the social contract and its source is not the only goal of Binmore's theory. While providing the naturalistic account of the evolution of fairness is in the center of Binmore's project, he also expressed a more normative aim. We find in his work a repeated call for a use of the original position on a larger scale than is already used:

> I would like to see the original position used as Rawls envisaged to focus the reforming zeal of those among us who think our children's children would lead more satisfying lives in a more fair society. But I differ from most people with similar aspirations in believing that its potential for solving social problems will evaporate if we fail to use it for large-scale purposes in the same way we currently use it when settling picayune fairness questions (Binmore, 2005, 21).

The call to "focus the reforming zeal" is not accidental: he believes that using fairness on a larger scale is also quite compatible with human nature—it is a very familiar tool, and adapting it on a wider scale is not difficult. He does not give any normative reason to do that, as he believes that there are no such things at all—every normative judgment is always just an example of preference. However, he promotes his view as feasible and optimal. He also frames it as an example of "rational ethics" (Binmore, 1998) when explaining that it requires interpersonal comparisons of utility.

Although Binmore frames his project primarily as a naturalistic explanation of fairness norms, as we can see, there is an undercurrent of prescriptive ambition in his writing. Some readers may regard this as a "slip" from description into normativity; but we might also see it as an interesting extension of his framework. For the purposes of this article, I bracket the question of whether this prescriptive strand is central to Binmore's theory. My concern is not to reconstruct his work as a wholly normative project, but to examine whether this ancillary proposal—using the original position more broadly—can withstand scrutiny, assuming his descriptive account is plausible when it comes to the general mechanism he describes. In other words, the focus here is on the potential validity and applicability of this proposal, regardless of whether Binmore intended it as the primary aim of his research.

But there is a serious problem with this argumentation—namely, that Binmore, while exhorting to use fairness on a wider scale, does not really specify what that means. Does it mean to adopt something similar to hunter-gatherer compulsory sharing? The social contract of hunter-gatherers is quite tight and probably absolutely unbearable for

most people living in modern societies. Is it some kind of egalitarianism, massive redistribution, taxing the rich to a larger extent? One would expect that after developing a quite complicated model of fairness, the author will use it to describe at least some social practices, features of social and economic system or practical consequences. One could also use this framework to discuss some aspects of modern society, e.g., social security or the public health system. Nothing of the sort happens, and we are left with the question: If the original position is the only game in town, how did we end up with all those large-scale solutions? If it is the only selection device that ever evolved, which however is not used to coordinate on a large scale, then how do we coordinate?

How does Binmore locate his egalitarian theory on the spectrum of more traditional moral and political theories? Certainly, he distances himself from what he sees as the political left and right. Every political theory must have two qualities: stability and efficiency. These qualities translate into questions: What social contracts are feasible? And, then, which are the most efficient? Binmore curiously accuses philosophies associated with left-wing positions of caring only about efficiency, and those associated with right-wing positions of limiting themselves to stability, while both areas are important. He also mocks philosophical theories of "The Good" or "The Right", claiming that he never saw their goodness or rightness, like Diogenes never saw Platonic tableness or cupness, and proposes that the good and the right are best understood as products of human evolution (Binmore, 2005, 93). These are necessary properties of a social contract, but they are not keys to understanding it. The actual source of our conceptions of the Good is said to lie in our understanding of how we choose the equilibrium coordination points, while what is Right is what sustains this equilibrium, i.e., social norms that keep it being an equilibrium. Therefore, saying what is good or right, or assigning blame or merit, is not helpful in looking for new equilibria. These ideas are only elements of our description of the existing equilibria. While he has no respect for the so-called theories of The Right and the Good, he sees himself, along with Aristotle and Hume, as a theorist of The Seemly (Binmore, 1998, 98). By that, he means the view that human morality is not something rooted in metaphysical concepts but is a way of balancing conflicting needs or a mutual advantage. Human conceptions of good are derived from observations of how societies in practice select equilibria, whereas conceptions of right are based on observing how to sustain an equilibrium (Binmore, 2005, 94).

3. Two Problems of Fairness

Binmore's theory suffers from a whole host of problems, such as the appropriateness of using the Folk Theorem and repeated games in this context (Gintis, 2006; Skyrms, 2006), how it approaches the empirical data on hunter-gatherers (Cosmides and Tooby, 2006; Mackie, 2006) or the validity of the concept of empathetic preferences. I assume that such problems can be resolved and follow Binmore in his assumptions for the sake of argument. Then, it seems like Binmore's theory can provide reasons to doubt that there is anything "special" about morality (Hédoin, 2018). However, we could go a step further and see that Binmore's theory actually shows why fairness norms are not just nothing special, but should actually be distrusted. There are two reasons for this in the following two parts. The first one is internal to the theory. Binmore warns many times about the danger of generalizing from small-scale solutions to the entire society. I argue that his theory of fairness is such a generalization. The second reason is external, as when assessing the usefulness of fairness as a coordinating device, we need to always ask the question "compared with what", and Binmore does not spend a lot of time considering competing coordination devices. I also argue that his theory does not necessarily lead to wider moral skepticism, just skepticism regarding rules of fairness.

3.1 Internal

Binmore's proposal might be considered puzzling, since when discussing what he calls "traditional moralists", he is fond of criticizing philosophers for not considering the feasibility of using their favored solutions on a large scale, and yet we do not find a proof that a large-scale usage of the original position is stable. He even says: "I have repeatedly emphasized that I think it highly misleading to generalize from the social arrangements of small close-knit groups to society as a whole" (Binmore, 1994, 287). We find numerous warnings against using norms and behavior patterns outside of the environments and situations for which they are adapted. Binmore repeats many times a criticism of behavioral laboratory studies when they purportedly show cooperation in one-off Prisoner's Dilemma or other games of this type: "My own view is that the result simply reflects our confusion over the concept of fairness when it is applied outside the domain for which it evolved" (Binmore, 1998, 36). The outside domain in this example is the laboratory, but we can also say that applying the same solution at different scales is changing domains, from picayune everyday

problems to large-scale redistribution. We can ask why then is it applied outside the domain, as such an application happens not only under the influence of egalitarian philosophies. We could explain, using Binmore's theory, that fairness evolved just as Binmore explains it, and it is used by people in other domains automatically. But they are not doing this as a result of careful consideration of what is needed to support modern institutions, but rather a natural affinity.

After all, the task of the naturalistic and descriptive theory of morality is not to create ethical principles on the basis of what behaviors are important to support institutions of a modern society. The fact that some moral behaviors would not support modern society does not imply that they do not exist or that we are not using them. After all, on several occasions, Binmore has illustrated the use of evolving norms that meet different needs over time compared to the needs that these same norms currently meet. If some of them, including sympathy, are ingrained by evolution in our minds and, because of that, are used or tried to be used, this cannot be overlooked by an empirical science of morality, even if some of those moral principles are not the right way to solve our modern problems. Binmore's theory is based on the claim that human ancestors have lived in strongly egalitarian societies for a very long time, enough for biological evolution to work and imprint on us the norms of these societies. Some genetic differences must have arisen, so evolution could suppress the instinct to dominate (Binmore, 2005, 134). Therefore, Binmore concludes, present-day humans must share some propensities with hunter-gatherers. Thus, using some elements of our ancient social contract, ingrained in our minds, seems to be both effective and conductive to happiness. However, even if this is true, so is having an abundance of resources at our disposal. And every social contract brings about different efficiencies in producing and distributing them. A social contract based on fairness might bring about lower efficiency, and thus we face a trade-off between happiness from plenty and happiness from fairness, and it seems impossible to be able to choose between them merely on the basis of his theory. Binmore strongly stresses the role of trade-offs, but in this he does not heed the warning himself. This leads to the comparison problem, addressed in the next section.

Nearly all of Binmore's examples involve "immediate return" societies, where members gain direct, short-term benefits from hunting or gathering, with food lasting only a few days, as opposed to "delayed return" societies, where food can be stored for longer periods and benefits from farming or herding are realized over the long term. Using the

social contract of an immediate-return society would most probably be highly ineffective for a delayed-return society that can lead to higher levels of material well-being. Moreover, virtually all modern hunter-gatherers live in environments that are either extreme or very problematic for agriculture. Due to that, their social contract is very rigid because attempts to experiment and change the way of doing things mostly end with starvation and death. Because ancestors of modern delayed-return societies lived in less harsh environments, they might have been less rigid; in fact, they surely were, since they did experiment and change their way of living into sedentary. Binmore acknowledges these problems, writing that it is risky to treat modern foragers as models of prehistoric societies (Binmore, 2005, 143). He also adds more reasons for this thinking: In the last ten thousand years, evolution, mainly cultural, has worked in both types of society and probably changed them. We also need to keep in mind that hunter-gatherers have actually quite diverse social structures and are more aptly characterized to occupy a spectrum of positions than to have one uniform type (Kelly, 2013). All these reasons should cast at least some doubts on the relevance of Binmore's social contract to our understanding of fairness.

He admits sometimes that fairness itself is a pattern that people have a problem applying beyond the realm where it evolved (Binmore, 1998, 36). This is in line with what modern evolutionary psychology calls the "mismatch theory"—that our moral psychology evolved for life in small groups but does not provide us with adequate incentives and restraints to be trusted in a modern society (Li et al., 2017). Equal sharing may well have been a good survival strategy in the past, and our genes may still predispose us to think that this is the case, but why should it be? Binmore, firstly an economist, himself knows it well, and even writes that "we are predisposed toward a kind of social behavior appropriate to a hunter-gatherer society, but these behavioral predispositions have been overwritten by cultural imperatives without which we couldn't been so productive" (Binmore, 2005, 137). Why then overturn these cultural imperatives? Binmore does not provide an answer to these questions, and he does not heed his own warning against applying a tool that evolved in a different environment.

We need to distinguish here between the well-known methodological issue of using simplified, small-world game-theoretic models to *explain* large-scale behavior, and the separate problem of treating such models as a *basis for normative prescriptions* in large and complex societies. The former is a standard and often defensible abstraction in the

social sciences: small-scale games can illuminate general strategic structures by bracketing away context-specific noise. As Robert Sugden puts it, economic models aren't simplified snapshots of reality but constructions of "credible counterfactual worlds," whose value lies in how reasonably they illuminate core strategic structures—not in literal realism (Sugden, 2001). Similarly, Don Ross—who champions highly idealized models in economic theory and cognitive science—argues that such abstractions are useful for uncovering generic mechanisms, provided we don't mistake them for ready-made policy tools (Ross, 2007). Treating models as a basis for prescription, however, assumes that the equilibria and norms identified in small, evolutionarily familiar settings can simply be scaled up as desirable policy goals in modern, heterogeneous societies—a move that, I argue, is far less secure, especially in light of Binmore's own repeated cautions about domain shifts. As Francesco Guala cautions in the methodology of experimental economics, the jump from laboratory models (or simple abstractions) to normative recommendations in complex, large-scale societies carries a heavy burden: structural differences and contextual contingencies may render such extrapolations unstable or unwarranted (Guala, 2005).

Here, we can see how Binmore's theory can be used for the opposite purpose: instead of bolstering Rawlsian or egalitarian thinking, it can undermine it. If we agree with Binmore that any normative proposal is strictly subjective and the role of the political or ethical theory is only to illuminate what are the evolutionary sources of some moral arrangements, the theory under consideration here can show how an egalitarian solution is not relevant for nonhunter-gatherer societies.

3.2 External

The external problem of Binmore's treatment of fairness can be summarized as "Compared with what?" To successfully argue for a wider usage of the original position, Binmore needs to not only prove that his coordination device did, in fact, evolve the way it did, but also show that no other coordination device did or is otherwise available—due to, for example, rational design. However, if there are other coordination devices, Binmore would need to propose a way to compare them. If we propose a change of our social contract (for example, when pressuring other people to abolish slavery), there is no objective code to which we may appeal, and Binmore claims that the choice is made according to taste. A wish to change a society is just that,

a wish. But this is also true, because when we use fairness and the original position to judge social contracts, we do not transcend any cultural prejudices. They do not give us any absolute scale of justice (Binmore, 1994, 72). He underlines that his theory does not rest on ethical presuppositions. What is ethical in theory lies in the use of the device of the original position. However, according to Binmore, there are no other coordination devices as successful and hardwired in our genes as the original position, but they are not even mentioned in the theory.

Among the alternative coordination mechanisms, a prime example is hierarchy and authority. Although Binmore acknowledges that most societies coordinate through dominance hierarchies, he dismisses this as mere "cultural evolution" without explaining why hierarchical coordination is so persistent across cultures and historical periods if fairness norms are truly "hardwired" in our genes (Dubreuil, 2010).). Christopher Boehm's work on "reverse dominance hierarchies" in hunter-gatherer societies actually suggests that egalitarianism requires constant vigilance against natural hierarchical tendencies—undermining rather than supporting Binmore's claim that fairness is our default coordination mode (Boehm, 1999). If fairness were genuinely our only coordination device, we would expect hierarchy to be unstable and require constant enforcement.

Market mechanisms present another coordination alternative that Binmore, despite being an economist, barely addresses. Hayek's theory of spontaneous order demonstrates how price signals coordinate complex economic behavior without requiring empathy, original position reasoning, or interpersonal utility comparisons (Hayek, 1945). Market coordination operates through what Hayek calls "*catallaxy*"—a system in which individuals pursuing their own ends inadvertently coordinate their behavior through price mechanisms. This coordination occurs without the cognitive demands of Binmore's empathetic preferences or the need to imagine oneself behind a veil of ignorance. The efficiency and scale of market coordination suggest that it may be a more robust solution to large-scale coordination problems than fairness norms.

Reputation and status competition offer yet another coordination mechanism grounded in evolutionary psychology but operating through different principles than Binmore's fairness norms. Costly signaling theory demonstrates how status competition coordinates behavior through prestige hierarchies rather than egalitarian sharing (Zahavi and Zahavi, 1997). Joseph Henrich's work on "prestige bias"

shows how humans naturally coordinate by copying high-status individuals, creating coordination cascades that require no fairness calculations or empathetic preferences (Henrich and Gil-White, 2001). This mechanism appears particularly relevant for understanding large-scale coordination, where direct reciprocity becomes impossible, but status signaling remains viable.

Similarly, the presence of a clear leader or a long-established convention can efficiently guide behavior and solve coordination dilemmas, such as which side of the road to drive on (Schelling, 1960). Binmore dismisses these external coordination devices in favor of an internal, genetically hardwired mechanism. However, the effectiveness of these alternative mechanisms in many real-world scenarios suggests that fairness is not the "only game in town" as his theory implies. Modern, large-scale societies have largely moved beyond the small, close-knit groups for which Binmore's fairness mechanism purportedly evolved. Instead of relying on decentralized fairness norms, they solve complex coordination problems through a vast array of formal institutions. For instance, legal systems, with their written laws and enforcement mechanisms, provide a binding framework for social contracts that goes far beyond what is possible in a face-to-face society. The market economy, as another key institution, uses price signals and property rights to coordinate the activities of millions of individuals, a system often criticized for its unequal outcomes, but nonetheless highly effective as a coordination device. These institutions operate on principles that may not align with Binmore's vision of egalitarianism and often represent a move away from the hunter-gatherer social contract. Therefore, for his theory to be truly impactful on public policy, it must justify why its "cautious egalitarianism" is preferable to the coordination achieved by these powerful, culturally evolved institutions.

Lastly, in domains such as international relations, corporate governance, or crisis response, we probably should not even start to apply Binmorean fairness. In international relations, attempts to apply Rawlsian fairness norms consistently lose out to realpolitik calculations—states coordinate through balance of power, threat credibility, and mutual deterrence rather than veil-of-ignorance reasoning (Butler, 2001). The persistence of international anarchy despite centuries of moral philosophy suggests that fairness is a weak coordination device on large scales. Corporate governance presents a similar pattern: firms that prioritize fairness over efficiency in decision-making tend to get outcompeted by those using hierarchical command structures and

market-based incentives (Skrastins and Vig, 2019). Even supposedly "fair" companies rely on authority gradients and profit maximization for actual coordination, with fairness relegated to HR rhetoric (O'Connor and Ihlen, 2018; Shamir, 2005). Crisis response offers perhaps the starkest counterexample: when coordination failures mean death, societies invariably abandon consensus-based fairness for command hierarchies (Chang, 2017). Disaster response or pandemic management all demonstrate that speed and clarity of authority trump fairness considerations when stakes are genuinely high (Farcas et al., 2020). If Binmore's original position were truly our primary coordination device, we would expect it to at least be present precisely when coordination matters most, but the opposite occurs.

For an empirical theory of morality, it might be strange not to consider other coordination devices. It is understandable if the goal was only to propose a theory of fairness as only a single part of justice. But above all, the sparse treatment of other solutions is understandable because Binmore does not face the problem of explaining some behavior with which his theory of fairness has problems, since, as critics pointed out, there is surprisingly little empirical data in this theory: "There are as few actual facts in *Natural Justice* as there are in Plato's *Republic* and far fewer arguments" (Mackie, 2006, 780). He does not face the problem of explaining some types of behavior that might depend on other coordination devices, so there is no need to analyze if there are other mechanisms and then compare them with fairness. If fairness is the only game in town, then Binmore's calls to use it on a large scale are perhaps reasonable. But it seems to be the only game in town only because of the goal of the theory, which is to build an account of fairness only, not to account for every coordination device. Using such a partial theory to guide public policy seems not to be justified if, by definition, we exclude every other consideration. If Binmore's theory ignores other devices, his normative suggestion is underdetermined: maybe fairness is the best large-scale coordination device, or maybe it's just the one he modelled. In the broader literature, fairness is only one among several plausible coordination devices humans employ, alongside reputation-based reciprocity, signaling systems, punitive institutions, and convention-formation through focal points (Ostrom, 1990; Schelling, 1960; Skyrms, 2010). Binmore's near-exclusive focus on fairness may be methodologically justified for a theory of fairness *per se*, but it leaves the normative argument for its large-scale adoption underdeveloped: without a comparative analysis, we cannot know whether fairness is uniquely suited to modern large-scale societies or

merely the most familiar. In policy terms, this is akin to recommending a single institutional arrangement, such as choosing a tax system, without benchmarking it against feasible alternatives—a move that risks importing unexamined cultural and historical biases into ostensibly naturalistic reasoning.

3.3 Local Justice in Polycentric Order

Rather than abandoning Binmore's insights entirely, we might reconceptualize his theory within a framework of "polycentric political orders" that acknowledges the proper scope and limitations of fairness coordination. If Binmore is correct that fairness norms evolved for small-scale, face-to-face interactions, then perhaps the appropriate response is not to scale up these mechanisms to encompass entire societies, but rather to design institutional arrangements that preserve spaces where such coordination can operate effectively while employing alternative mechanisms where fairness proves inadequate.

Gerald Gaus' *The Open Society and Its Complexities* suggests that complex societies require what he calls "institutional differentiation"—different coordination mechanisms operating at different scales and domains (Gaus, 2021). This approach would treat fairness not as the universal coordination device that Binmore sometimes presents, but as one mechanism in a portfolio whose optimal weights vary with ecological conditions, group size, and institutional complexity. David Thunder's analysis in *The Polycentric Republic* provides a framework for understanding how "partially autonomous local political units" can organize around local conceptions of fairness while leaving central authorities to coordinate through alternative mechanisms—markets, hierarchy, or technocratic expertise—in domains where fairness coordination proves ineffective (Thunder, 2025).

We would have a division of labor: local or municipal units deploy fairness to structure club goods and distributive practices (such as eligibility rules, "congruence" of contributions and benefits, or graduated sanctions), while higher-level centers handle tasks where fairness is neither necessary nor sufficient: scale-economy public goods, exit rights, baseline liberties, due process, and conflict adjudication. This matches well with Ostrom's design principles for commons governance, as we have clear boundaries, collective-choice arrangements, monitoring, graduated sanctions, low-cost conflict resolution, and nested enterprises, where fairness plays a role in reciprocity expectations but never substitutes for monitoring and enforcement (Ostrom,

1990). Fairness would then be a local heuristic inside an institutional scaffold, not the scaffold itself.

This polycentric approach solves both the internal and external problems identified earlier. The internal problem—Binmore's contradiction between warning against scaling up small-group solutions while doing exactly that—dissolves if we limit fairness coordination to appropriately scaled contexts. Local communities, professional associations, or other "human-scale" institutions could organize around fairness norms without requiring these mechanisms to coordinate national economic policy or international relations. The comparative case for scaling fairness must be made against other coordination devices such as prices (Hayek, 1945), hierarchy (Dubreuil, 2010) or ideology and religion (Harwick, 2020). The criteria would need to include (i) scalability and information processing, (ii) costs of enforcement and monitoring, (iii) robustness to heterogeneity, and (iv) distributional profile. Polycentricity does not pick a single winner; it permits domain allocation: markets coordinate production plans while fairness rules local distributive conflicts where empathy indices are shared enough to work. Without that comparative perspective, Binmore's large-scale fairness is underdetermined, but with it, his account works as a defensible local mechanism. Finally, the external problem—Binmore's failure to compare fairness with alternative coordination devices—becomes less damaging if we acknowledge that different mechanisms may be optimal for different coordination challenges rather than searching for a single universal solution.

Contemporary examples suggest that this division of coordination labor already operates informally. Local communities often coordinate around fairness intuitions for resource sharing, dispute resolution, and collective decision-making, while simultaneously relying on market mechanisms for economic exchange, hierarchical authority for emergency response, and bureaucratic procedures for large-scale administration (Ostrom, 1990; 2005). Rather than seeing this as a failure of fairness norms to achieve universal application, a polycentric perspective would treat it as evidence of adaptive institutional differentiation.

We need, however, to be cautious of two potential problems here. First, local tyranny and path dependence—polycentric orders need some level of constitutional meta-rules (baseline rights, exit/entry, transparency) so "local fairness" would not entrench caste-like suboptimal equilibria. Second, coordination externalities, when local fairness generates spillover effects (e.g., welfare magnets, zoning

exclusion), higher-level centers need harmonization tools (side-payments, shared fiscal bases) that are themselves not grounded in fairness. This illustrates the main point: fairness has a proper place alongside other devices; treating it as the unique, hardwired coordinator could be a category error.

Binmore's evolutionary account actually supports this more modest interpretation. If fairness evolved for specific ecological conditions, like small groups, repeated interactions, and limited resource accumulation, then we should expect it to work best under similar contemporary conditions. The mistake lies not in identifying fairness as an important coordination mechanism, but in treating it as the coordination mechanism that should dominate all others. A properly ecological understanding of coordination would predict exactly what we observe: fairness thriving in some contexts while being displaced by alternatives in others.

This reframing preserves what is valuable in Binmore's naturalistic approach—his demonstration that fairness norms have deep evolutionary roots and play important coordination roles—while avoiding the overreach that undermines his normative project. Instead of promoting fairness as a universal solution, we might focus on designing institutions that harness multiple coordination mechanisms appropriately, allowing fairness to operate where it works best while employing alternatives where it proves to be more effective. Such an approach would be both more empirically realistic and more likely to achieve the coordinated, stable societies that Binmore ultimately seeks to promote.

3.4 Fairness Skepticism and Moral Skepticism

An important problem for a naturalistic account of morality is that it can be used to actually undermine it. As Hédoin argues:

> 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 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, it is not clear why one should maintain that his beliefs about what is fair are justified. (Hédoin, 2018, 60)

This is because once people learn from Binmore's theory that their beliefs about what is fair are influenced by past power relations, they might stop wanting to follow the norms that are constituted by fairness. Hence, the theory that has the objective of justifying the usage of some coordination device actually undermines it.

However, we need to distinguish between broader moral skepticism and narrower "fairness skepticism". Because Binmore's theory is concerned with the role of fairness, it is difficult to assess whether it leads to a broader moral skepticism. At the same time, Binmore is not concerned with answering the challenge of a moral skeptic, he is, in fact, the one offering it:

> I have no absolute source of moral authority to which to appeal—but neither does anyone else. I know that my aspirations for what seems a better society are just accidents of my personal history and that of the culture in which I grew up. If my life had gone differently or if I had been brought up in another culture, I would have different aspirations. But I nevertheless have the aspirations that I have—and so does everyone else. The only difference between naturalists and traditionalists on this score is that naturalists do not try to force their aspirations on others by appealing to some invented source of absolute authority. The reality is that if enough people with similar aspirations—benign or otherwise—are sufficiently close to the levers of power, they shift the social contract because that is what they want to do (Binmore, 2009, 13).

Binmore relies exclusively on hypothetical imperatives: "if you want to achieve X, you should do Y as the best way to achieve it". This might not be what we want from moral theory if we want to produce a robust answer to a moral skeptic who requires a reason for the basic motivation of moral acts. However, despite what Binmore sometimes calls rational ethics and the need to produce it, his theory in fact assumes elementary moral motivation as being a part of human nature or just points out to the fact that immoral behavior is in fact punished and that should serve as a motivation (Binmore, 1998, 283). We can see how he is in line with moral antirealism and conventionalism, as he rejects the assumptions that moral claims should depend on justifiably true beliefs and that they should have an unconditional normative force (Hédoin, 2018, 66). We do not introduce additional moral realist criteria that Binmore's theory must address, but only show that even if we agree that all a naturalist has is his personal taste when deciding which social contract to choose, it does not mean that he is free from the need to confront his preferred solution with other coordination devices.

I need to note here that the criticism raised here does not prove that Binmore is incorrect in the descriptive part of his theory and in the

explanation of how fairness evolved. This argument rather adds to the doubts raised by Hédoin (2018) that Binmore, instead of convincing us of the merits of using his coordination device, shows rather why we are prone to think it is feasible and preferable, regardless of whether it is true. On the other hand, it does not necessarily mean that a naturalist should abandon this approach. The fact that a theory can lead to the destruction of intuitions it tries to explain does not mean that it is incorrect. It can show why we might distrust our own motivation to pursue some goal, but if we accept naturalistic presuppositions, then we can just accept this as a result and move on. Pursuing the naturalistic foundations of ethical theories can still be meaningful in terms of how they explain the sources of our behavior, even if they show that there is nothing distinctive about it.

Conclusion

This article identifies two key issues about the application of Ken Binmore's naturalistic theory of fairness. The first problem is internal: despite Binmore's repeated warnings against overgeneralizing from small-scale contexts, he himself extends these solutions to society as a whole. The second issue is that for Binmore to effectively promote the broader use of fairness, he would need to demonstrate either that fairness is the only viable solution in the game of life or that competing solutions are inferior. However, if there are other possible solutions, Binmore's theory simply explains why fairness seems like a natural choice—since we have used it for so long—but does not necessarily prove its effectiveness in modern societies. This raises doubts about whether fairness should be widely applied. Ironically, Binmore's theory could be interpreted as an argument against using fairness broadly, contrary to his original aim. However, I also argue that this naturalistic theory does not necessarily lead to broad moral skepticism, but rather to skepticism about the large-scale applicability of fairness. I also propose to resolve Binmore's contradictions through polycentric institutional design—fairness norms would coordinate at a local/small-group scales where they evolved to work, while markets, hierarchy, and other mechanisms handle large-scale coordination tasks where fairness proves inadequate. This division of labor preserves what's valuable in Binmore's evolutionary account without the overreach that undermines his normative project.

References

Alexander, J. McKenzie. 2007. *The Structural Evolution of Morality*. Cambridge: Cambridge University Press.

Bellah, Robert N. 2011. *Religion in Human Evolution: From the Paleolithic to the Axial Age*. Cambridge: Harvard University Press.

Binmore, Ken. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair*. Cambridge: MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract. Volume 2: Just Playing*. Cambridge: MIT Press.

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

Boehm, Christopher. 1999. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge: Harvard University Press.

Butler, Brian E. 2001. There Are Peoples and There Are Peoples: A Critique of Rawls' Law of Peoples. *Florida Philosophical Review*, *1*(2): 1-24.

Cosmides, Leda and John Tooby. 2006. Evolutionary Psychology, Moral Heuristics, and the Law. In Gerd Gigerenzer and Christoph Engel (eds), *Heuristics and the Law*. Cambridge: MIT Press, 175-205.

Cushman, Fiery, Liane Young, and Marc Hauser. 2006. The Psychology of Justice. *Analyse & Kritik*, *28*(1): 95-98.

Dubreuil, Benoît. 2010. *Human Evolution and the Origins of Hierarchies: The State of Nature.* Cambridge: Cambridge University Press.

Farcas, Andra, Justine Ko, Jennifer Chan, Sanjeev Malik, Lisa Nono, and George Chiampas. 2020. Use of Incident Command System for Disaster Preparedness: A Model for an Emergency Department COVID-19 Response. *Disaster Medicine and Public Health Preparedness*, *15*(3): e31-e36.

Gaus, Gerald. 2021. *The Open Society and Its Complexities*. Oxford: Oxford University Press.

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

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Harwick, Cameron. 2020. Inside and Outside Perspectives on Institutions: An Economic Theory of the Noble Lie. *Journal of Contextual Economics – Schmollers Jahrbuch*, *140*(1): 3-30.

Hayek, Friedrich. 1945. The Use of Knowledge in Society. *American Economic Review*, *35*(4): 519-530.

Hédoin, Cyril. 2018. Naturalism and Moral Conventionalism: A Critical Appraisal of Binmore's Account of Fairness. *Erasmus Journal for Philosophy and Economics*, *11*(1): 50-79.

Henrich, Joseph and Francisco J. Gil-White. 2001. The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the

Benefits of Cultural Transmission. *Evolution and Human Behavior*, *22*(3): 165-196.

Kelly, Robert L. 2013. *The Foraging Spectrum: Diversity in Hunter-Gatherer Lifeways*. Clinton Corners: Eliot Werner Publications.

Li, Norman P., Mark van Vugt, and Stephen M. Colarelli. 2017. The Evolutionary Mismatch Hypothesis: Implications for Psychological Science. *Current Directions in Psychological Science*, *27*(1): 38-44.

Mackie, Gerry. 2006. Ken Binmore, Natural Justice. *Ethics*, *116*(4): 776-780.

O'Connor, Amy and Øyvind Ihlen. 2018. Corporate Social Responsibility and Rhetoric. In Ihlen Øyvind and Robert L. Heath (eds), *The Handbook of Organizational Rhetoric and Communication*. New York: John Wiley & Sons, 401-415.

Ostrom, Elinor. 1990. *Governing the Commons. The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ostrom, Elinor. 2005. Policies That Crowd out Reciprocity and Collective Action. In Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr (eds), *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge: MIT Press, 253-275.

Qizilbash, Mozaffar. 2009. The Adaptation Problem, Evolution and Normative Economics. In Kaushik Basu and Ravi Kanbur (eds), *Arguments for a Better World. Essays in Honor of Amartya Sen. Volume I: Ethics, Welfare, and Measurement*. Oxford: Oxford University Press, 50-67.

Ross, Don. 2007. *Economic Theory and Cognitive Science: Microexplanation*. Cambridge: MIT Press.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge: Harvard University Press.

Schmidt-Petri, Christoph. 2006. Binmore's Egalitarianism. *Analyse & Kritik*, *28*(1): 89-94.

Shamir, Ronen. 2005. Mind the Gap: The Commodification of Corporate Social Responsibility. *Symbolic Interaction*, *28*(2): 229-253.

Skrastins, Janis and Vikrant Vig. 2019. How Organizational Hierarchy Affects Information Production. *The Review of Financial Studies*, *32*(2): 564-604.

Skyrms, Brian. 2006. Ken Binmore's Natural Justice. *Analyse & Kritik*, *28*(1): 99-101.

Skyrms, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.

Skyrms, Brian. 2014. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Sugden, Robert. 2001. Credible Worlds: The Status of Theoretical Models in Economics. *Journal of Economic Methodology*, *7*(1): 1-31.

Thunder, David. 2025. *The Polycentric Republic : A Theory of Civil Order for Free and Diverse Societies*. London: Routledge.
Vanderschraaf, Peter. 2018. *Strategic Justice : Convention and Problems of Balancing Divergent Interests*. Oxford: Oxford University Press.
Woodburn, James. 1982. Egalitarian Societies. *Man*, *17*(3): 431-451.
Zahavi, Amotz and Avishag Zahavi. 1997. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford: Oxford University Press.