RAWLSIAN "JUSTICE" AND THE EVOLUTIONARY THEORY OF GAMES: CULTURAL EVOLUTION AND THE ORIGINS OF THE NATURAL MAXIMIN RULE

Mantas Radžvilas
MSc in Philosophy of the Social Sciences
E-mail: mantas.radzvilas@alumni.lse.ac.uk

This paper is dedicated to the analysis of the maximin principle, which is one of the key theoretical concepts of John Rawls's theory of justice, and the problem that this principle creates for any attempt to provide a naturalistic interpretation of Rawls's concept of fairness . Analysis shows that maximin principle is, in fact, incompatible with the Bayesian decision theory. This paper is intended to show that recent breakthroughs in evolutionary game theory could help to reconcile the maximin principle with a certain naturally emerging and evolutionary stable pattern of human social behaviour.

Keywords: *Maximin principle, Nash bargaining game, Evolutionary Game Theory, Evolutionary Stable Strategy, learning rules.*

There is an almost unanimous agreement that John Rawls was an intellectual giant who deserves a place among the greatest political philosophers of the 20th century. However, the situation becomes less clear when an attempt is made to specify the reasons why Rawls deserves such a place in the history of political philosophy. In fact, it is not difficult to distinguish two large groups holding radically different interpretations of Rawls's contribution to political philosophy.

For the first group, which consists mainly of philosophers working in the analytical tradition, Rawls is the leading figure of political philosophy, whose theory of justice is the first *real political theory,* primarily because it is not based on the highly speculative metaphysical arguments that were, and still are, widely used in political philosophy to justify various concepts of justice. What is more, this group holds the view that the strength of Rawls's theory lies in its attempt to *reconcile rationality and morality*. Rawls attempted to show that *under certain circumstances*, it is rational for an individual to choose certain concept of fairness as a foundational principal of the social contract (Rawls 1971: 17–22).

Although, as it will be shown later, Rawls's argument is faulty, it nevertheless deserves a place in the history of political philosophy. Rawls gave an original and innovative answer to the question about the relation between rationality and morality while at the same time escaping the Kantian style arguments, often, at least implicitly, based on a questionable assumption that a truly rational individual will necessarily

observe one categorical imperative, even regardless of the consequences that such behaviour would bring about (Binmore 2005: 38).

The second group consists primarily of philosophers working in the continental tradition of political philosophy, especially those who use metaphysical arguments to defend certain concepts of justice. This group supports the view that Rawls *probably* deserves his place in the history of political philosophy for the sophistication of his arguments, as well as for his contribution to bringing the Kantian philosophy back into the contemporary political debate. However, this group seems to believe that Rawls's theory of justice is merely a *sophisticated intellectual exercise* that has no direct relevance for our everyday problems, primarily because of the *highly abstract and academic nature of arguments that support the foundational components of Rawls's theory* (a characteristic comment that exemplifies such interpretation of Rawls's theory can be found in Ankersmit 1996: 6). According to this view, Rawls's key ideas, such as his *original position* concept, are theoretical constructs that serve well as thought experiments supporting his argument, but have little to do with the *real world* situation, where justice hardly gets along with rationality.

I believe that the second view is a serious misinterpretation of Rawls's theory. The main cause of such misinterpretation is the crucial misunderstanding of the actual strengths of Rawls's theory of justice. Its actual strength lies in the fact that this theory *captures our intuitions about the nature of fairness*. In fact, there are serious reasons to believe that Rawls was able to capture, albeit in a *stylized form of the original position*, some *deep structure of human fairness norms*. Although Rawls uses the *original position* as a *hypothetical standpoint* from which to make judgments about the fundamental principles of a "fair" social contract, it nevertheless grasps some important *patterns of human social behaviour* that could be considered as *representing* our understanding of "fairness" (Binmore 2005: 15–16, 150–151).

However, it must be admitted that critics are right in claiming that some of Rawls's arguments are *merely hypothetical*. In order to reveal the real potential of Rawls's theory, it is necessary to show that his ideas about "fairness" correspond to the *actual patterns of human behaviour in real-world situations.*

Although I do not consider myself an orthodox naturalist, I do believe that social norms are products of the *cultural evolution* of a particular society. If Rawls managed to capture the deep structure of our "fairness" norms, it should be possible to provide a *naturalistic interpretation of his theory of justice.*

I do believe that current developments in *evolutionary game theory* and its increasingly wide application in philosophy and social sciences can bring us a step forward towards the naturalistic interpretation of Rawls's theory of justice.

In this paper I look into the most problematic element of Rawls's theory, the so-called *maximin* principle, which creates significant problems for any attempt to provide a naturalistic account of Rawls's theory.

In the first part of this paper, I clarify the nature of the problems that are faced by the *maximin* principle. In the following part of the paper I look into the contemporary attempts to explain certain stable patterns of "fair" human behaviour in evolutionary game-theoretic terms. I believe that they provide the necessary key for the naturalistic *reinterpretation* of the *maximin* principle. Finally, I indicate some of the most serious problems with the evolutionary reinterpretation of the *maximin* principle and suggest *some viable solutions to these problems.*

1. *Maximin* as a Violation of the Orthodox Principles of Rational Decisions

In order to demonstrate the problem of the *maximin* principle, I use a traditional method of demonstration based on a simple version of the *Nash bargaining game* (Figure 1).

The game is simple: Player X and Player Y have to divide 10 dollars between them. Each player must write a certain sum of money on a piece of paper and hand it to a referee. There is no communication allowed between players, so both of them have to make independent decisions. What is more, the relation among the players is *symmetrical. This means that neither Player X nor Player Y has any prior claim to the 10 dollar sum and that neither of them has any special needs that should be taken into consideration when playing this game*. It is *common knowledge* that *if* the sum of both claims is equal or less than 10 dollars, *then* each player gets the sum that he/she has written on the paper. *If* the sum *exceeds* the 10 dollar limit, *then* the referee takes the money and both players get nothing (another popular version of the Nash bargaining game is the *divide-the-cake game*. Its description can be found in Skyrms's 1996: 3–4).

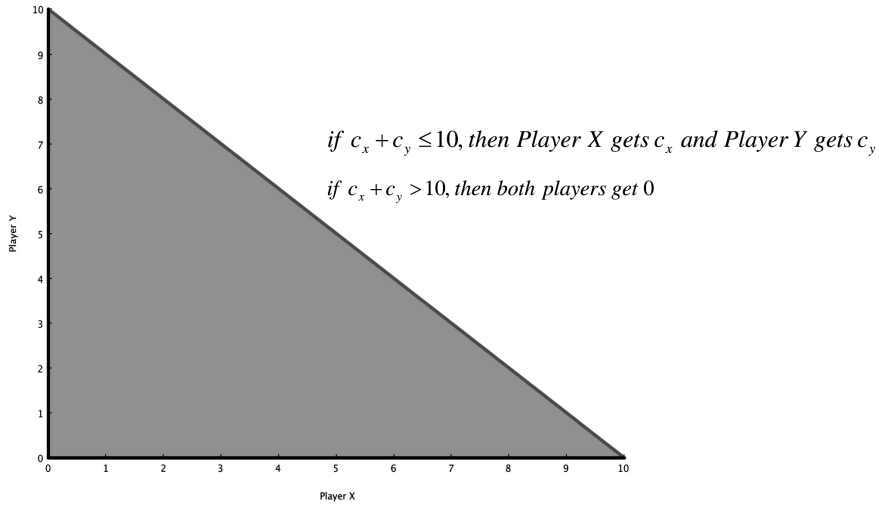

Figure 1: Nash bargaining game

In other words, Player X makes a claim for a certain sum of money c_x. At the same time Player Y must make his/her claim c_y. *Each player's choice is based on his/her beliefs about the decision made by the other player*.

At this point it should be clear that every possible money claim could be interpreted *as strategy that Player X and/or Player Y uses to play this game.*

Let I be the set of players of this game. For each player $i \in I$, let S_i be a *set of pure strategies* for this game. Let $s_i \in S_i$ denote a *pure strategy* for player i. I will use s_i, s_{-i} to denote a combination of strategies of the game. Finally, the finite set of real numbers $P_i(s_i, s_{-i})$ denotes the *payoff* for every player $i \in I$ who is using strategy s_i (a detailed formalization of the elements of Noncooperative Game Theory can be found in Weibull 1997: 1–31).

Using this notation, we can summarize the game in the following form:

$$P_i(s_i, s_{-i}) = \begin{cases} s_i \ \ if \ \ s_i + s_{-i} \leq 10 \\ 0 \quad otherwise \end{cases}$$

This game has one crucially important feature: every strategy pair s_i, s_{-i} satisfying the condition $s_i + s_{-i} = 10$ *is a strict Nash equilibrium.* Strict Nash equilibrium means that *neither player benefits by changing his/her strategy if the other player does not change his/her strategy as well*. Because such strategy pair is a *strict* equilibrium, it constitutes a *Pareto-efficient* strategy pair (Alexander 2007: 148–155). In other words, any *unilateral deviation* from strict Nash equilibrium would bring payoff that *would definitely be worse than the equilibrium payoff.*

This Nash bargaining game is a perfect example of the so-called *equilibrium selection problem*: the game has an *infinite number of equilibria* (Skyrms 1996: 4–6). For example, if $c_x^* = 4$ *and* $c_y^* = 6$[1], any unilateral deviation made by Player X would make him/her strictly worse-off: if player X decides to deviate and claims more than 4 dollars, then *both players receive nothing*. If player X decides to claim less than 4 dollars, then *Player X gets a smaller payoff than his/her equilibrium payoff.*

The problem of this game is to find a principle that would justify the selection of one particular money division rule among the other possible division rules. *In other words, the aim of this game is to find an appropriate decision rule that would help players to select one particular Nash equilibrium in a game that has an infinite number of Nash equilibria.* Of course, it could be claimed that *each player should claim half of the total sum* because such division would be "fair". However, it turns out to be difficult to provide a *philosophical justification for such division*.

Rawls attempted to solve this bargaining problem using the *veil of ignorance principle*. This principle can be easily summarized as follows: Player X and Player Y have to divide 10 dollars between two *hypothetical individuals* – Individual A and Individual B. After both players make their decisions, a referee *tosses a fair coin.* This flip of a coin determines which one of the players will take the position of Individual A and which one will step into the position

[1] The asterisk symbol (*) indicates that strategies constitute one of the Nash equilibria of this game.

of Individual B (Binmore 2005: 129–145). What principles of reasoning should be used in such situation?

Rawls thought that under such conditions *both players should use the maximin decision criterion to select their game strategies*. According to this principle, *each player's decision* should be based on the assumption that a toss of a fair coin will put the player into the shoes of the individual *who has been given the smallest sum of money*. Such method of reasoning would lead to the conclusion that the *optimal decision for each player* is to *choose the strategy that maximises his/her minimum payoff*. According to the maximin decision criterion, *both players should demand 5 dollars* (Skyrms 1996: 5–7).

Unfortunately, this decision criterion has some serious shortcomings. First of all, Harsanyi (1973: 37–40) is right to claim that Rawls's maximin principle is not based on the principles of modern Bayesian decision theory. According to Harsanyi, the maximin principle only states that every strategy must be evaluated in terms of the worst possible outcome. *However, it does not take into account the fact that some of the worst possible outcomes could have nearly insignificant probability.* Even if the probability of the worst possible outcome *were extremely low*, Rawls's maximin criterion would nevertheless *rule such strategy out*. According to Harsanyi, such decision criterion leads to highly unacceptable practical decisions. For example, this rule would imply that no one would ever be able to fly to another country in order to collect the biggest lottery prize ever (no matter how large the prize would be) because there always is an infinitesimally small probability that the plane will crash (of course, it is assumed that death is the worst possible outcome).

Harsanyi has also shown that a maximin decision criterion inevitably requires giving absolute priority to the interests of the worst-off individuals, even in cases where such policy would seriously damage the interests of other individuals (Harsanyi 1973: 40–43). Therefore, it seems that at least in some real-world cases Rawls's maximin principle would actually lead society to very 'unfair' policy decisions.

What is more, as Skyrms notes, it makes no sense for both players to assume that for some reason *fortune will work against both of them*. That would undermine the very idea that one of the players of this game must be the lucky one (Skyrms 1996: 6–7). In fact, this argument against the maximin principle is extremely compelling. Rawls assumes that the veil of ignorance prevents contracting parties from knowing their place in society, class position, social status, their natural assets, abilities, intelligence, strength and other things that give them individual advantage over other individuals. What is more, individuals behind the veil of ignorance do not know the particular circumstances of their own society (e.g. economic, political or cultural situation) and have no information as to which generation they belong (Rawls 1971: 136–138). However, it is also obvious that individuals negotiating behind the veil of ignorance *do not have any reason to believe that the probability of the worst possible outcome is more likely than the probability of any other possible outcome*. Rawls's

maximin criterion is not Bayesian, but it is based on an *implicit assumption* that the worst possible outcome is *always more likely than any other possible outcome*.

Moreover, since this decision criterion leads to very unreasonable outcomes in some simple every-day situations, it is hard to imagine how such decision criterion could have *naturally* become the primary criterion for decision-making in risky or uncertain situations. In fact, it seems that Rawls's maximin principle would actually have negative effect on the fitness of individuals who would base their survival strategies on such restrictive decision criterion.

One possible improvement over the maximin criterion is the *minimax regret* (or *minimax risk*) criterion, which was suggested by Savage in 1951. According to Savage's principle, every decision problem with utility entries should be associated with a new table with regret (risk) payoffs, where regret is defined as the difference between the actual payoff and the payoff that would have been obtained if a different course of action had been chosen. The minimax regret criterion means than it is rational to choose the act that minimizes the maximum risk index for each course of action (Luce and Raiffa 1957: 280–282).

Unfortunately, this decision criterion also has some serious shortcomings. According to Chernoff (A detailed account of Chernoff's objections to the minimax regret principle can be found in Luce and Raiffa 1957: 280–282), no one has succeeded in demonstrating that differences in utility can be used to measure regret. What is more, Chernoff has shown that it is very easy to construct examples where this decision criterion, just as the previously discussed maximin criterion, rules out good strategies because it allows a very small advantage of the strategy in one possible state of the world to outweigh a considerable advantage in another possible state of the world. Finally, the most serious criticism raised against the minimax risk criterion is that it violates the *independence of irrelevant alternatives* principle. As Chernoff has shown, in some cases the minimax risk criterion will select strategy s_3 among the available strategies s_1, s_2, s_3 and s_4. However, when for some reason strategy s_4 is made unavailable, the minimax risk criterion will select s_2 among the remaining strategies s_1, s_2 and s_3. The biggest problem with this shortcoming of the minimax risk criterion is that it cannot be easily eliminated. As Luce and Raiffa showed, the obvious way to cope with the problem of non-independence of irrelevant alternatives is to make paired comparisons among strategies instead of evaluating all available strategies simultaneously. Unfortunately, they also showed that in some cases this modification leads to the violation of the *transitivity principle*, which is one of the fundamental principles of decision theory (Luce and Raiffa 1957: 281–282).

Finally, it is again hard to believe that the minimax regret criterion could have been *naturally* selected as a standard response to risky or uncertain situations. Subjective probabilities play a crucial role in the every-day decision-making process, so it seems that any adequate assessment of regret should involve the assessment of probabilities of various possible states of the world. This criterion, on the other hand,

is too restrictive and inefficient because it is not dependent on the probabilities of the various possible outcomes. It is also based on an implicit and highly problematic assumption that the worst possible outcome is always more likely than any other possible outcome. This assumption is not suitable for Rawls's social choice theory because players *have no information that could justify such probability distribution.*

According to the *orthodox Bayesian decision theory*, every rational player should choose the course of action that maximises his/her *expected* payoff. In the case of Nash bargaining game, Player X and Player Y have a probability of $\frac{1}{2}$ to end up as Individual A or Individual B. Their *expected utility* (*EU*) *function* can be expressed in the following form:

$$EU = \frac{1}{2}u(c_A) + \frac{1}{2}u(c_B)$$

In this function, $u(c_A)$ and $u(c_B)$ denote the *utility payoffs* that each player receives from the *shares of money* attributed to Individual A and Individual B (an axiomatic treatment of utility can be found in Luce and Raiffa 1957: 12–38).

O*rthodox decision theory implies that rational individuals should be indifferent between lotteries with the same expected utility* (Binmore 2009: 50). On the other hand, orthodox interpretation of this bargaining game also *leads us back to the original equilibrium selection problem* that we have already discussed in detail.

Consider the following situation: Players have to make a selection between *two possible* options O_1 and O_2:

O_1: Individual A gets 4 dollars; Individual B gets 6 dollars.

O_2: Each Individual gets 5 dollars.

According to the orthodox decision theory, there is no rational reason for Player X to prefer O_2 *because both options have the same expected payoff* (*EU*). Player X should *be indifferent* between O_1 and O_2:

$$EU_{O_1} = \frac{1}{2} \times u(4) + \frac{1}{2} \times u(6)$$

$$EU_{O_2} = \frac{1}{2} \times u(5) + \frac{1}{2} \times u(5)$$

$$EU_{O_1} = EU_{O_2}$$

On the other hand, such result seems to contradict our intuitions about the nature of "fair" behaviour. Bayesians are probably right to claim that individuals behind the veil of ignorance have no reason to believe that the worst possible outcome is more likely than any other possible outcome. However, people still perceive equal division rule as a "fair" division rule in their every day decision-making process. This is especially visible in cases where people have to interact with complete strangers. Unfortunately, even if we accept the idea that Rawls's *maximin* principle somehow reflects the *actual human behaviour*, we still need to provide a well-founded explanation of why such behaviour is selected among *other equally rational options.*

2. "Fair" Division Rule as an Evolutionary Stable Model of Social Behaviour

An alternative explanation of the emergence of "fairness" is based on the idea that "fair" division rule is a social norm, which, when

interpreted in evolutionary game-theoretic terms, *is an Evolutionary Stable Strategy (or ESS) that has the potential to eliminate all other strategies during the process of evolution of human interactions.*

According to Alexander, since human beings are not perfectly rational, any project that attempts to explain and predict individual choice by positing human beings as perfectly rational agents appears to be misguided. In reality, human beings are boundedly rational and often rely on less-than-perfect calculations derived from heuristics. Alexander defines heuristics as *"'common knowledge' acquired by participation in the common culture"*. Heuristics *"encapsulates this common knowledge in comprehensible and readily apprehended forms, which can then be applied in contexts different from the one in which it was originally acquired"* (Alexander 2007: 5).

It seems that "fair" division rule is nothing more than a part of the heuristic toolbox that we use in our culture. *In other words, "fair" division rule can be interpreted as a social coordination norm that evolved as a response to repeated bargaining interactions among boundedly rational individuals.*

This interpretation avoids a lot of problems related to the maximin decision criterion because a heuristic is not a universally applicable decision-making principle. As Alexander notes, "[E]*ach heuristic works well for a certain class of problems whose structure satisfies certain necessary conditions required for the reliable functioning of the heuristic"* (Alexander 2007: 7). Since heuristic can be perceived as an evolved trial-and-error style response to a certain type of decision problem, it is not necessary to find an abstract rational principle that would justify the applicability of one or another universal decision criterion to a particular type of decision problem.

Consider a two-player game G played by the players who are *randomly selected* from a population I of individuals. There are n possible *pure strategies* denoted $(s_1, ..., s_n) \in S_i$. According to Weibull, since every player of the game has a *finite set* of pure strategies S_i, it is possible to represent any *mixed strategy* x_i of player i as a vector x_i in m_i-dimensional Euclidean space R^{mi}, where its h^{th} coordinate $x_{ih} \in R$ is the *probability assigned by some mixed strategy* x_i *to player i's pure strategy* h (Weibull 1997: 2). The vector $x_i \in R^{mi}$ belongs to the unit simplex Δ_i that has dimension $m_i - 1$. It is then possible to study a projection of the simplex $\Delta_i \subset R^{mi}$ to $m_i - 1$-dimensional Euclidean space. Each vertex e_i^h of Δ_i then represents *one of the pure strategies of the game* while every mixed strategy of the game $x_i \in \Delta_i$ can be expressed in the following form:

$$x_i = \sum_{h=1}^{m_i} x_{ih} e_i^h \quad \text{(Weibull 1997: 3)}$$

A *mixed-strategy profile* is a vector $x = (x_1, ..., x_n)$, where each component $x_i \in \Delta_i$ is a mixed strategy of player $i \in I$. In other words, a mixed strategy profile is a *point in the mixed strategy space* Θ. The expected payoff $u_i(x)$ to player i that

is associated with a *mixed strategy profile* $x \in \Theta$ can be expressed in the following form:

$$u_i(x) = \sum_{s \in S} x(s) P_i(s),$$

where P_i(s)denotes player i's payoff associated with *pure strategy profile* $s \in S$ (Cressman 2003: 19–20 and Weibull 1997: 1–31).

Let us again consider a population *I* of individuals, who are all programmed to play one particular strategy $x \in \Delta$ where Δ denotes a *mixed strategy space*. Suppose that a certain small group of *"mutants"* appears in this population and they all play the same "mutant" *strategy* $y \in \Delta$. The payoff to strategy $x \in \Delta$, when played against $y \in \Delta$, will be denoted as $u(x, y)$. The payoff to strategies $x \in \Delta$ and $y \in \Delta$ *when each strategy is used against itself* will be denoted as $u(x, x)$ and $u(y, y)$ respectively. The proportion of the "mutants" in the population is ε, where $\varepsilon \in (0, 1)$. Two individuals are *randomly selected from the population I for each round* of the game. The probability for every player to face an opponent playing "mutant" strategy $y \in \Delta$ is ε, while the probability that opponent will play the "normal" strategy $x \in \Delta$ is $1 - \varepsilon$.

According to the formal definition of *Evolutionary Stable Strategy*, evolutionary forces will *select against* the "mutant" strategy $y \in \Delta$ *if and only if its payoff is lower than the payoff of the "normal" strategy*. What is more, "normal" strategy $x \in \Delta$ will be evolutionary stable only in case strategy *x* is a *better reply to strategy y than strategy y is to itself*. In such case, the set of evolutionary stable strategies Δ^{ESS} must meet two criteria:

1. $u(y, x) \leq u(x, x) \quad \forall y$
2. $u(y, x) = u(x, x) \Rightarrow u(y, y) < u(x, y) \quad \forall y \neq x$

(Weibull 1997: 35–68)

To put it simply, *Evolutionary Stable Strategy (or ESS)* is a strategy such that *population of individuals adopting it cannot be invaded, in evolutionary time, by "mutants" adopting different strategies* (Maynard Smith 1982: 54–67). Such "mutants" using *any other possible strategy* will eventually be *taken over by the population using the ESS*. Such "takeover" could mean various things but in this context it simply means the *disappearance of "mutants" from the population*.

In a Nash bargaining game, the *"fair" division rule has the properties of the Evolutionary Stable Strategy*. This can be easily shown by comparing the performance of the strategy "Demand 5 dollars" with all other strategies that could *possibly* be used in this game.

It is convenient to use the same notation as before, so the two players will be denoted as Player X and Player Y. The Nash bargaining game is played *repeatedly* by the pairs of individuals, who are *randomly selected* from a population of individuals for every new round of the game. Each of the randomly selected individuals takes the position of either player X or Player Y. Every individual in a population is *programmed* to play one particular strategy of the game. *The dollar payoff that every individual gets after the pairwise interaction with other player corresponds to the number of players who will adopt his/her strategy in the next round of the game* (almost similar

descriptions of the evolutionary Nash bargaining game has been used by Skyrms 1996: 9–11 and Alexander 2007: 148–155). Thus, according to this logic, *a strategy that yields higher payoff to the player will increase the proportion of individuals in the population who will be using this strategy when selected to play the game.*

Let us suppose that *every randomly selected individual* who is playing the strategy "Demand 5 Dollars" *always takes the position of Player X* while "mutants" using other possible strategies *always take the position of Player Y*. In such case every "mutant" using the strategy c_y, which satisfies the condition that $c_y > 5$, will get *nothing in every interaction with "fair" player because* $c_x + c_y > 10$. If "mutant" strategy c_y is such that $c_y < 5$, then *it will always yields a lower payoff than the "fair" strategy*. What is more, *the strategy "Demand 5 Dollars" is a better response to itself than any other strategy.* If "mutants" play *against each other* using the same strategy, any strategy that satisfies the condition $c_y > 5$ will leave them with *nothing* while the strategy that satisfies the condition $c_y < 5$ will leave them with *lower payoff* than the strategy "Demand 5 Dollars" (a somewhat less technical description of the evolutionary stability of "fair" strategy can be found in Skyrms, 1996: 9–11). This means that "Demand 5 Dollars" is a strategy that *satisfies the evolutionary stability criteria*. If this strategy takes over the population, other possible strategies *cannot invade it.*

The fact that "fair" division rule has such evolutionary stability properties can provide a firm ground for a highly plausible explanation of the emergence of "fair" behaviour. Rawls attempted to explain the stability of this pattern of behaviour by inventing the problematic *maximin* principle. In the case of evolutionary explanation, we can explain the emergence of "fair" behaviour without any assumptions that are obviously incompatible with the orthodox decision theory.

On the other hand, we still need to answer the question *how likely is it* that "fair" division rule would take over the population. Evolutionary stability criteria help to identify those strategies that are *resistant to invasion* by other strategies. However, these criteria are not sufficient to determine the probability that evolutionary stable strategy will eliminate all other competing strategies and become the dominating strategy of the population.

In order to test the plausibility of the scenario in question, it is now possible to use certain computer-based simulations of evolution. There are various more or less complex ways to model the evolutionary dynamics, but I will focus on a relatively simple model of the *replicator dynamics*, which was used by Skyrms in his attempt to provide the evolutionary explanation of the emergence and stability of the "fair" division rule and many other social norms. I believe that this model is sufficient for my purposes.

In the simple version of the replicator dynamics, it is assumed that each individual in the population is *programmed* to play *one particular pure strategy i* from the *set of pure strategies available* to him/her. If the population is at the state $x \in \Delta$, the expected payoff of any pure strategy *i* at a

random match is $u(e^i, x)$ while the payoff of any individual drawn at random from such population, or the *population average payoff*, is $u(x, x)$ (Weibull 1997: 71–72).

In the Nash demand game, the dollar payoff must be interpreted as *individual fitness*, measured as *the number of offspring per time unit*. In other words, the amount of dollars that each individual earns during his/her pairwise interaction with other individual in the game corresponds to the number of offspring who will inherit his/her strategy. In this case, *the higher the payoff that strategy i earns, the higher the number of individuals who will be using this strategy in the future, thus increasing the share of population using the strategy i*. The *share of population* programmed to pure strategy *i* at a time *t* will be denoted as $x_i(t)$.

It will be assumed that the death rate $\delta \geq 0$ is equal for all individuals in the population. Then the *instantaneous rate of change for* x_i *is*

$$x_i\left[u(e^i,x) - u(x,x)\right]$$

(a detailed mathematical account of the replicator dynamics can be found in Weibull 1997: 71–119).

The most important question with regard to the replicator dynamics of the Nash bargaining game is whether the *probability* that the strategy "Demand 5 Dollars" will take over the population is *sensitive to the different possible initial conditions of the population state*. In other words, it is necessary to consider the possibility that the actual success of the "fair" strategy is *largely determined by the properties of other strategies that are present in the population and the outcomes of their interactions*.

It is reasonable to begin by considering a case where the population is still in the *"inefficient"* state. In other words, there is no specific *"harmony"* between the strategies competing with the "fair" division strategy. Such a situation is perfectly plausible because of two reasons:

1. In the model of the replicator dynamics, it is assumed that individuals are programmed to play one particular strategy. It does not mean that strategies are somehow *adjusted* to each other.
2. In the case of *cultural evolution*, individuals *still cannot be perfectly rational*. Bounded rationality can be realistically modeled as a *situation where players do not know the actual population state and cannot identify individuals playing particular strategies*. As Skyrms notes, every individual in this population game is under some kind of *Darwinian veil of ignorance*, because he/she cannot tell in advance, which strategy the *other player* will use in any of the random pairwise interactions that he/she might be selected to participate in (Skyrms 1996: 9–11).

In the first simulation, it will be assumed that there are three strategies represented in the population:

s_1: "Demand 5 Dollars", or *D*5 ("fair" players)

s_2: "Demand 2 Dollars", or *D*2 ("risk averse" players)

s_3: "Demand 6 Dollars", or *D*6 ("risk seeking" players)

All players of this game have the same

payoff matrix, which can be expressed in the following form:

	*D*5	*D*2	*D*6
*D*5	5	5	0
*D*2	2	2	2
*D*6	0	6	0

All three strategies are equally represented in the population (*the shares of population playing different strategies are equal*). However, the individuals programmed to play different strategies are *dispersed randomly* in the population. The dynamics is displayed as a two-dimensional *dynamical picture of the population* (Figure 2), where *each vertex of the triangle represents one pure strategy represented in the population*[2]:

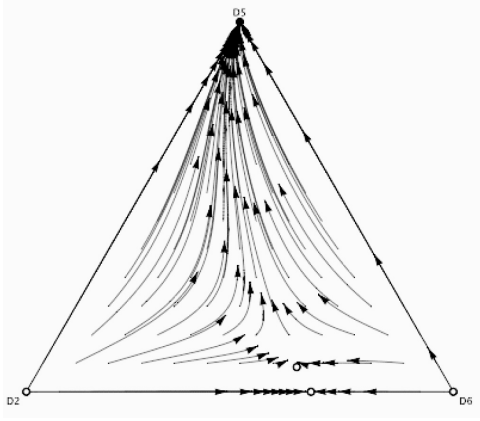

Figure 2: The evolution of "inefficient" population

The picture shows that the strategy "Demand 5 Dollars" *eventually takes over the population*. The shares of population using the strategies "Demand 6 Dollars" and "Demand 2 Dollars" quickly start to vanish, while the share of individuals playing the "fair" strategy increases rapidly, thus allowing the strategy "Demand 5 Dollars" to become the *dominating strategy in the population*.

Simulations show *essentially the same results under a substantial number of various "inefficient" initial conditions and different combinations of possible strategies*. Such results from various types of simulations should be considered as evidence that supports the evolutionary explanation of the origin of "fairness". It shows that "fair" division rule is an *Evolutionary Stable Strategy* that *can* become the dominating strategy under various different initial population states.

Of course, it must be admitted that such explanation is a serious detour from Rawls's original project. First of all, *it does not show that "fair" division rule is somehow more compatible with rational behaviour than any other rational division rule*. It only shows that such strategy is *very robust against other strategies* and has the potential to eliminate other strategies in the *evolutionary competition*. In this sense, such explanation leads to the abandonment of Rawls's original intention to reconcile rationality and morality. What is more, it is not really possible to claim that we could base our rational decisions about the best possible welfare policy on arguments drawn directly from the evolutionary explanation of "fairness" *because it does not capture the whole complexity of the macro-level public policy problems.* For example, it does not give an adequate answer to the question *what does it mean to say that individuals are equal*. In evolutionary explanation of "fairness", it is simply assumed that

[2] All diagrams were generated using *Dynamo* workbooks (v.0.2.5).

individuals are exact copies of each other. However, real world welfare problems are primarily related with the fact that *human beings are very different and thus have very different needs* (Dasgupta 2009: 580–606).

I believe that such real-world problems guarantee a solid place for traditional political philosophy and welfare economics because these disciplines try to provide an answer to such context-sensitive questions. Nevertheless, evolutionary explanation of "fairness" is very important for deepening our *understanding of the emergence of the stable patterns of social behaviour.*

One of the seemingly strong arguments against the evolutionary interpretation of the maximin principle is that Rawls was concerned with *one-shot decision problem* while evolutionary account deals with *iterated games*. This argument is invalid for several reasons. First of all, Rawls's original position is a merely hypothetical situation. If we want to explain the natural emergence of "fairness", we must eliminate all hypothetical arguments from his theory and consider realistic evolutionary scenarios. Secondly, if "fair" division rule is a *learned social coordination norm*, it means that *boundedly rational individuals would probably use this norm in any single shot bargaining game*. The positive experience gained from repeated micro-level interactions using the "fair" division norm would only lead to *even more automated response to various bargaining problems*. It is very unlikely that individuals would change their behaviour when faced with a single-shot decision problem. In fact, some experiments have shown that human beings who were selected to play repeated games in the laboratory environment still followed the social norms for a considerable period of time. Eventually they began to adjust their strategies to the laboratory game they were actually playing. *However, this adjustment was gradual and took a considerable amount of time* (for example, experiments with Nash bargaining game by Binmore et al. 1993: 67–101).

Unfortunately, there are several real problems with the evolutionary explanation of "fairness". The most serious shortcomings and some solutions to these problems will now be discussed in detail.

3. Some problems with the Evolutionary Explanation of Fairness

There are two major criticisms raised against the evolutionary explanation of the emergence and persistence of the "fair" division rule that require particular attention.

3. 1. The Problem of the Polymorphic Population States

This is probably the most serious shortcoming of the evolutionary explanation of "fairness". *Polymorphic population states* emerge under certain initial population conditions, when there is a specific *"efficient harmony"* between the strategies competing with the "fair" division rule.

Consider a population with three strategies:

s_1: "Demand 5 Dollars", or *D*5
s_2: "Demand 3 Dollars", or *D*3
s_3: "Demand 7 Dollars", or *D*7

The payoff matrix looks like this:

	*D*5	*D*3	*D*7
*D*5	5	5	0
*D*3	3	3	3
*D*7	0	7	0

Simulation provides the following results (Figure 3):

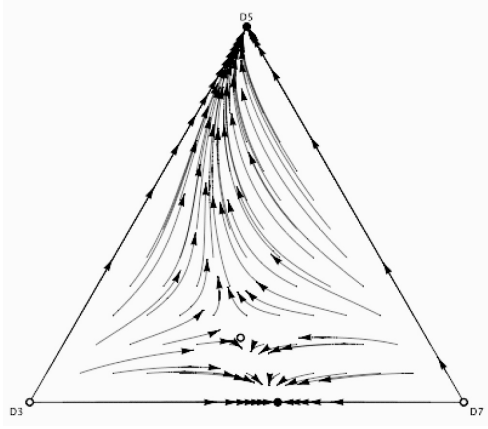

Figure 3: Polymorphic population state

This dynamical picture of the population shows the emergence of the so-called *polymorphic population state*. As the picture shows, if certain proportions of the population adopt the strategies that are competing with the "Demand 5 Dollars" strategy, there is a possibility that these strategies will form a *stable polymorphic state preventing the "fair" strategy from taking over the population*. A black dot between the vertices representing strategies "Demand 3 Dollars" and "Demand 7 Dollars" represents this *stable polymorphic state*. The *white dot* in the interior of the dynamical picture shows the *unstable polymorphic state*, which *should not be considered as a serious obstacle* for the emergence of "fair" population. However, simulation shows a relatively large *basin of attraction of the stable polymorphic state*. The size of this basin of attraction can be used to evaluate the probability that the population will turn into the "polymorphic trap". Skyrms claims that if the size of the basin of attraction of a polymorphic state is relatively small compared to that of a "fair" division rule, the "Demand 5 Dollars" strategy should take over the population *from a larger set of initial conditions than the polymorphic state.* However, he also admits that in some cases the basin of attraction of the polymorphic state is of *considerable size*, so it is not impossible that the population in question will get into a "polymorphic trap" and the "fair" division strategy *will not be able to take over the population* (Skyrms 1996: 11–16).

One of the possible solutions to this problem is to abandon the random interactions assumption by introducing into the model some *positive correlation* between *individuals playing the same strategy*. Positive correlation means that individuals with the same strategy *tend to interact more frequently* with each other than with individuals using other strategies. According to Skyrms, even the slightest correlation between individuals with the same strategy works favorably for the "fair" division rule – the *basins of attraction of polymorphic states tend to decrease* (Skyrms 1996: 11–16).

On the other hand, such a solution is highly problematic. First of all, such positive correlation *must be justified*. This idea comes from the argument used in evolutionary biology: because populations do not tend to disperse too widely, any given individual is more likely to interact

with those individuals who share more of his/her genes than does a randomly selected member of the population at large. However, there are serious reasons to believe that complex patterns of social behaviour *are not genetically transmitted from one generation to another.* What is more, if we seriously consider the *cultural evolution* model, we cannot neglect the fact that human beings are *at least boundedly rational*. It then becomes necessary to consider the possibility that individuals *are able to learn how to behave in a particular context*. For example, individuals using the strategy "Demand 7 Dollars" should be able to learn to avoid interactions with individuals using the same strategy as he/she does because in such interactions he/she gets nothing. However, D'Arms argues that in such case *correlation assumption is just as plausible as the anti-correlation assumption* (D'Arms 1996: 621–627). In fact, certain simulations show that anti-correlation is favorable to polymorphic states. Even in case when both positive and negative correlation is present in the population, the "fair" division rule "Demand 5 Dollars" will not do well (Alexander 2007: 160).

On the other hand, it seems that this problem is not as serious as it looks. It is true that positive correlation seems to be a rather weak argument for cultural evolution, even though *intuitively* it seems that the positive correlation assumption is *more realistic* than the anti-correlation assumption because anti-correlation is a much more sophisticated process which requires relatively high intelligence necessary for such "strategic" thinking. However, in case of simple replicator dynamics, we do not need to search for complex justifications of such intuitions. It is better instead to focus on the fact that *the model of replicator dynamics is based on assumptions borrowed from biology and captures only the very basic elements of the complex process* of *cultural evolution*. For example, since prehistoric communities were relatively small, it is reasonable to believe that human beings with even very limited cognitive abilities could have developed an ability to *identify the inequalities caused by the "unfair" distributions of goods. High intelligence is not necessary in simple comparative welfare evaluations, especially in cases when the basket of goods is very basic*. What is more, in case of some *serious exogenous shock*, such as serious food shortage caused by sudden and/or uncontrollable changes in the natural environment of the population, *individuals getting smaller payoffs could be motivated to challenge the existing "unfair" polymorphic state by changing their strategy.* Finally, the most elementary, albeit not the most sophisticated, "Darwinian" argument is that *individuals getting smaller payoff are less fit and could simply die-out in case of serious exogenous shock*. Since prehistoric societies had only a very basic basket of "survival goods", it is *reasonable to believe that unequal division of such goods would also mean unequal fitness*. On the other hand, Skyrms showed that polymorphic state is eliminated *only in case* the share of population getting small payoff virtually ceases to exist (Skyrms 1996: 11–16). Such scenario seems to be merely hypothetical.

I believe that players' ability to change their strategies during the game according to some kind of simple evolutionary

learning rule plays a very important role in cultural evolution. In the case of simple replicator dynamics, this element of cultural development is virtually eliminated. *If, on the other hand, this element is reintroduced, it could solve the problem of polymorphic states*. In fact, Smead claims that *polymorphic states* create ideal conditions for the evolution of *learning individuals* (Smead 2009: 11–16).

If we allow the possibility that individuals who are getting the lower payoff are able and motivated to develop their cognitive skills and adjust their behaviour in response to stimuli coming from the existing population state according to some evolutionary learning rule, the polymorphic state can become unstable (a detailed account of evolutionary learning rules can be found in Smead 2009). It seems that even very primitive and, from the perspective of perfect rationality, "inefficient" trial-and-error type *randomization of strategy* may lead to the elimination of the stable polymorphic state. One of the examples of such randomization is represented in the following dynamical picture (Figure 4). In this simulation, the strategy "Demand 3 Dollars" is replaced by the *randomized strategy D3L5*: "play 'Demand 5 Dollars' with probability $\frac{1}{3}$ and 'Demand 3 Dollars' with probability $\frac{2}{3}$". In other words, this model is based on assumption that these players quickly learn to imitate the strategy "Demand 5 Dollars" that they encounter in their pairwise interactions with other players and try to use this new strategy in *some* of their future interactions with other players. The *stable polymorphic state (black dot)* is gone (although the *unstable* one still persists) and population moves towards "fair" game:

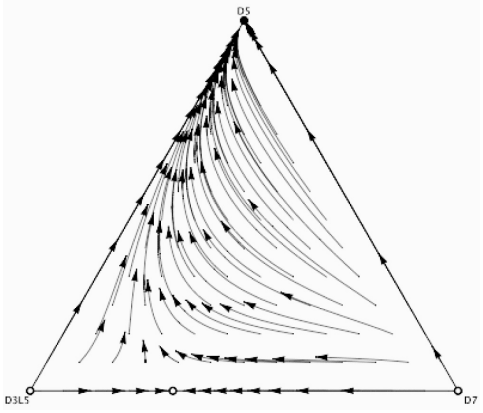

Figure 4: Evolution with one randomized strategy

Unfortunately, it seems that this scenario requires that all individuals using the strategy "Demand 3 Dollars" would adopt the strategy D3L5 relatively quickly. Therefore, such scenario requires a *realistic evolutionary model* of the pre-historic learning process that could explain how a share of population learns to randomize its strategy. Any attempt to develop such model inevitably involves some guesswork. However, since we know that pre-historic human societies were small and compact communities, such a rapid learning process seems to be plausible. What is more, complex agent-based models of evolution also show that if the "fair" division strategy is present in the population from the beginning of simulation, "fairness" takes over the population *significantly more often* than the "polymorphic traps" (Alexander 2007: 148–198).

4.2. The Superstructure of "Fairness"

Another criticism of the evolutionary explanation of "fairness" is that although evolutionary account can explain the *emergence and stability* of a certain stable pattern of behaviour, it does not explain the *superstructure of concepts and principles in terms of which we appraise such form of behaviour.* For example, it does not explain why there are certain negative emotional reactions and subtle informal sanctions that act as punishment for "unfair" behaviour (Kitcher 1999: 221–224).

Evolutionary explanation does not account for the superstructure of "fairness". On the other hand, *it does not mean that evolutionary explanation of "fair" behaviour cannot be combined with some realistic account that explains the origin and functions of the superstructure of "fairness"*.

Probably the best possible explanation of the superstructure of "fairness" is related with the fact that "justice" is a product of *cultural evolution*. According to Bicchieri, the *normative force* of social norms is created by a *set of interrelated expectations and beliefs that are common knowledge for members of society*. In other words, each individual in a particular society is aware that a certain social norm exists and that other members of society expect him/her to conform to this norm. What is more, in many cases it is common knowledge that transgression of a certain norm will cause certain informal sanctions. This knowledge might be the reason why individuals are willing to conform to existing norms even in cases when the chances of someone punishing the transgressor are negligible (Bicchieri 2006: 1–46).

It seems that these complex sets of beliefs about the existence of social norms (especially when such beliefs involve some supernatural arguments justifying such norms) and the *punishment of transgressors* help to *speed up the replication process*. What is more, sanctions and, especially, beliefs about existing sanctions also help to make the social norms stable because potential transgressor loses the motivation to "experiment" with his/her behaviour in social interactions.

Conclusion

Analysis shows that Rawls's *maximin* principle often leads to highly unacceptable practical decisions and is incompatible with the orthodox principles of modern decision theory. In order to solve this problem, a radical new approach is required that could explain certain *stable patterns of behaviour* that are *hypothetically predicted* by the *maximin* principle but without violating the orthodox conception of rational choice.

One way to achieve this goal is to reinterpret this problem in evolutionary game-theoretic terms and show that such pattern of behaviour is *evolutionary stable* and has the potential to *become the dominant model of behaviour in a particular society*. However, evolutionary explanation of "fairness" has certain shortcomings, such as the problem of polymorphisms and serious inadequacy in explaining the superstructure of "fairness".

Nevertheless, it seems that such approach is *a viable alternative to traditional "rational" arguments for "fair" behaviour*.

As I have shown, some important theoretical additions to the original evolutionary model give promising results. It is not unreasonable to believe that some highly sophisticated evolutionary model will eventually succeed in thoroughly explaining the "strange" persistence of our notion of "justice" without any controversial ideas that cannot be reconciled with the realistic notion of bounded rationality.

REFERENCES

Alexander, J. M., 2007. *The Structural Evolution of Morality*. New York: Cambridge University Press.

Ankersmit, F. R., 1996. *Aesthetic Politics. Philosophy Beyond Fact and Value.* Stanford: Stanford University Press.

D'Arms, J., 1996. Sex, Fairness and the Theory of Games. *The Journal of Philosophy,* Vol. 93, No. 12 (Dec., 1996), p. 615–627.

Bicchieri, C., 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms.* New York: Cambridge University Press.

Binmore, K., 2009. *Rational Decisions*. Princeton and Oxford: Princeton University Press.

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

Binmore, K., Swierzbinski, J., Hsu, S. and Proulx, C., 1993. Focal Points and Bargaining. In: Ken Binmore. *Does Game Theory Work? The Bargaining Challenge*, p. 67–101.

Cressman, R., 2003. *Evolutionary Dynamics and Extensive Form Games*. Cambridge, Massachusetts and London: The MIT Press.

Dasgupta, P., 2009. Facts and Values in Modern Economics. In: H. Kincaid, D. Ross, eds. *The Oxford Handbook of Philosophy of Economics*. New York: Oxford University Press, p. 580–640.

Harsanyi, J. C., 1973. Can the Maximin Principle Serve as a Basis of Morality? A Critique of John Rawls's Theory. In: *Essays on Ethics, Social Behaviour, and Scientific Explanation*, p. 37–63.

Kitcher, P. 1999. Games Social Animals Play: Commentary on Brian Skyrms's Evolution of Social Contract. *Philosophy and Phenomenological Research*, Vol. LIX, No. 1, March 1999.

Luce, R. D., Raiffa, H., 1957. *Games and Decisions: An Introduction and Critical Survey*. New York: Dover Publications Inc.

Rawls, J., 1971. *A Theory of Justice*. Cambridge, Massachussets: The Belknap Press of Harvard University Press.

Sandholm, W. H.; Dokumaci, E.; Franchetti F., 2010. *Dynamo: Diagrams for Evolutionary Game Dynamics*, version 0.2.5.

Skyrms, B., 1996. *Evolution of Social Contract*. Cambridge: Cambridge University Press.

Smead, R., 2009. *Social Interaction and the Evolution of Learning Rules*. Available at: <https://webfiles.uci.edu/rsmead/Papers/smead_learningrules.pdf> [Accessed 21 January 2011].

Maynard Smith, J., 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Weibull, J. W., 1997. *Evolutionary Game Theory*. Cambridge, Massachussets and London: The MIT Press.

ROLSIŠKASIS „TEISINGUMAS“ IR EVOLIUCINĖ LOŠIMŲ TEORIJA: KULTŪRINĖ EVOLIUCIJA IR NATŪRALIOS MAKSIMALAUS MINIMUMO TAISYKLĖS KILMĖ

Mantas Radžvilas

S a n t r a u k a

Straipsnyje nauju aspektu analizuojama kertinė Johno Rawlso teisingumo teorijos koncepcija – vadinamasis maksimalaus minimumo principas. Aptariami sunkumai, kuriuos kelia šis principas, kai mėginama pateikti natūralistinę J. Rawlso teisingumo teorijos interpretaciją. Atlikta analizė atskleidžia, kad maksimalaus minimumo principas, kurį Rawlsas laikė racionalia ir „teisinga“ elgesio strategija nežinojimo uždangos situacijoje, iš tiesų nesuderinamas su modernios, tikimybių analize grindžiamos racionalių sprendimų teorijos principais. Straipsnyje nagrinėjamos galimybės išspręsti šią problemą pasitelkiant naujausius evoliucinės lošimų teorijos laimėjimus. Jie atveria galimybę suderinti maksimalaus minimumo principą su tam tikru natūraliai kultūrinės evoliucijos būdu įsitvirtinančiu žmogaus socialinio elgesio modeliu.

Pagrindiniai žodžiai: maksimalus minimumas, Nasho derybų lošimas, evoliucinė lošimų teorija, evoliuciškai stabili strategija, mokymosi taisyklės.

Įteikta 2011 02 2