ACTA UNIVERSITATIS LODZIENSIS
FOLIA PHILOSOPHICA. ETHICA – AESTHETICA – PRACTICA 38, 2021

https://doi.org/10.18778/0208-6107.38.06

C O P E
Member since 2019
JM14485

Marcin Saar
Independent researcher
fayv@wp.pl

RATIONALITY AS THE CONDITION OF INDIVIDUAL RIGHTS IN DAVID GAUTHIER'S *MORALS BY AGREEMENT*

Abstract
The topic of this paper is the foundation for individual rights proposed by David Gauthier in his seminal 1986 book *Morals by Agreement*, and particularly the role of conception of rationality in this foundation. The foundation of rights is a part of Gauthier's broader enterprise: to ground morals in rationality – more specifically, in the economic conception of rationality. Because of the importance of this conception for the whole of Gauthier's project, we reconstruct first the conception of rationality which can be found in decision theory and game theory, presenting simultaneously in a relatively non-technical way some basic concepts of the aforementioned disciplines. We proceed then to reconstruction of the foundation of rights itself – it turns on Gauthier's interpretation of the so-called "Lockean proviso." Lastly, we turn to the connection between rationality and foundation of rights. It is to be found in the narrow compliance – the disposition to enter only into cooperation which satisfies conditions of fairness set out in part by the Lockean proviso.

Keywords:
Rationality, utility, rights, Gauthier, choice

INTRODUCTION

Questions concerning rights possessed by individuals, their content and justification, are among the most vexing problems of modern and contemporary political philosophy. There are couple of ways one can treat the issue. One way consists of claiming rights to be natural, i.e., possessed in virtue of having some specified inherent features, such as being human being or being rational being. Alternatively, rights can be thought of as of purely artificial character, e.g., results of social contract, be it actual or hypothetical, explicit or tacit. Between these opposite approaches views can be found which do not fit neatly within any one of them – for example, rights can be treated as products of social evolution, in which case they are neither natural, nor purely artificial. In this paper we

discuss a particular justification of individual rights, one proposed by David Gauthier in his book *Morals by Agreement*. We will argue that Gauthier's account combines elements featured both by natural rights theories and social contract theories. This account is part of Gauthier's broader philosophical project, namely providing rational (in the sense of economic rationality) foundation for moral principles understood as impartial constraints on realization of one's own interests. In next two sections we reconstruct conception of rationality adopted by Gauthier. In section 3 we look at justification of individual rights itself. In the last section we reconstruct the relationship one complex of ideas has to the other.

GAUTHIER'S CONCEPTION OF RATIONALITY

Conception of rationality to which Gauthier subscribes can be summarized in the formulation "rationality is maximization of individual utility" – the action is rational provided that, in a given situation, its effect is to make certain quantity at least as high as it would be if some other possible action was performed. In decision-theoretical conceptual framework this quantity is called utility. It is a measure of preference, defined for a set of states of affairs obtainable as results of actions – state of affairs A has for a certain individual higher utility than does state of affairs B if A occupies higher place in the individual's preference ordering than B does (i.e., the individual prefers A more than B). Utility, being measure of preference, is dependent on preference, so this conception can be thought of as identifying rational choice with the choice that "realizes" given individual's preferences in the highest degree.

Before we discuss utility in some more detail, it is necessary to stress that Gauthier follows the conception which links rationality to realization of individual's **own** preferences. He rejects the view according to which a rationality of actions pursuing preferences is independent from the issue of who's preferences are in question. According to Gauthier, if X has certain preference ordering, the mere reality of that fact does not give Y any reason whatsoever to take it into account in his practical reasoning – it would be the case only provided that satisfying X's preferences had any place in Y's preference ordering (Gauthier, 1986, pp. 6–8). It is necessary, however, to point out that this conception of rationality does not by itself require individuals to be egoists. Although Gauthier does indeed assume individuals to take no interest in each other's interests, this assumption is not any simple consequence of adopting decision-theoretical conceptual framework.[1]

[1] For some critical discussions of Gauthier's theory of individuals' motivation, see: Morris, 1988; Thomas, 1988; P. Vallentyne, 1991.

The way utility is assigned to particular states of affairs is based on relations of preference these states of affairs enter into in particular choice situation. In connection with this, decision theory requires individual preferences to satisfy certain conditions of coherence which make it possible to define a measure representing these preferences.[2] First, any two states of affairs possible to obtain in a given situation have to be comparable to each other with respect to preference. In other words, for any pair of states of affair, the individual has to be able to tell which one of these states of affairs they prefer (alternatively, that they are indifferent with respect to them). Gauthier calls this condition the requirement of **completeness**. Second condition says that relations of preference be **transitive** – for any three states of affairs A, B and C such that A is more preferred than B, and B is more preferred than C, it has to be the case that A is more preferred than C (the same can be said for indifference relations).

Provided that individual's preferences satisfy two aforementioned requirements, all of states of affairs entering into preference relations can be ordered from the one preferred the most to the one preferred the least. And this is sufficient when we examine situations in which choice is made in conditions of certainty. One characteristic feature of this kind of situations is that any particular action (which is the object of the choice made) is correlated with only one outcome – the action entails particular outcome with probability of 1. This makes the choice of particular action tantamount to choice of particular outcome or state of affairs. In this kind of situations answering the question "what choice the individual should rationally make?" is simple: they should choose the action entailing the state of affairs most preferred by this individual. But what of situations when the choice is not made in conditions of certainty?

Besides choice in conditions of certainty, decision theory distinguishes choices made in conditions of risk and uncertainty. With respect to the first kind each possible action is correlated with more than one outcome, and the probability that given action will entail particular state of affairs is known. Let us illustrate this with the example of decision to flip a "good" coin: with this action there are two outcomes correlated: "heads" and "tails." In connection to this choice each of these outcomes is assigned probability of ½. Actions can be thus thought of as "lotteries" (probability distributions) with correlated outcomes as their "prizes." Concerning choices made in conditions of uncertainty, the situation is similar, with the difference being that probabilities can be only subjective – objective probabilities either are unknown, or it does not make sense to speak about them.

[2] Our discussion of these conditions follows Gauthier's own; see: Gauthier, 1986, pp. 38–46. For a more technical discussion, see Luce and Raiffa, 1957, ch. 2, esp. §§ 2.4 and 2.5.

In order to be able to choose rationally in situations in which we do not know with certainty what outcomes will be entailed by our actions, we need a measure of our preference such that not only it will represent our ordering of states of affairs in terms of their being more or less preferred, but also it will represent relative "strength" of these preferences: if we prefer A more than B, and B more than C, we want to be able to tell if our preferring A more than B is stronger or weaker than our preferring B more than C.

In order to define such a measure, further requirements for preference relations are introduced. The first is **monotonicity**. This condition says that for any two lotteries differing only in one prize (i.e., in one lottery A is one of the prizes obtainable, while in the other one B takes its place; all other prizes are the same), the individual has to prefer the lottery that gives them higher probability of getting the prize more preferred. As Gauthier points out, this condition excludes from consideration the attitude the individual can adopt towards the lottery itself, to the gambling *in se*. The last condition of discussed kind is **continuity**. It requires that for any states of affairs A, B and C (such that A is more preferred than B, and B more than C) there be a lottery with A and C as prizes such that the individual will be indifferent between this lottery and B.

Provided that given individual's preferences satisfy four discussed conditions, there is a possibility of defining a measure of them, which can be maximized even in situations in which certainty is not the case. In the first step utilities have to be assigned to every state of affairs which enters into given individual's preference relations. Utility is a measure, therefore the unit and the zero point can be assigned arbitrarily. For example, let us assign the "extreme" (in preference ordering) alternative states of affairs A and C utilities of 1 and 0, respectively. How is then assigning utilities to intermediate alternatives (in our example there is only one – B) to be done? By relating each of them to appropriate lottery with extreme alternatives as prizes. Let us imagine that the individual in question is given the following choice opportunity: they can either get B with certainty or participate in a lottery with A and C as prizes. In every lottery of this kind (and we can come up with infinitely large number of them) outcomes most and least preferred (our A and C) are assigned certain probabilities – p and (1 – p), respectively. We can surely imagine that if p is close to 1, our individual will prefer to take part in lottery. We can with equal ease imagine that if p is low almost down to 0, the individual will prefer to get the certain B. If we examine numerous possibilities of lotteries with varying values of p, we should be able to find the lottery with p such that our individual will be indifferent between that lottery and B. Let us say that it would be a lottery such that probability of getting A (or p) equals ⅔ (and probability of getting C – ⅓). In such a case, we can identify this lottery's value of p (i.e., ⅔) with B's utility. With respect to larger sets of outcomes entering into preference relations the procedure is essentially similar: we identify utility of each intermediate outcome with value of p of adequate lottery with extreme outcomes as prizes.

Having defined utility as appropriately refined measure, we have the tool which enables rational choice in conditions of risk and uncertainty. Knowing probabilities of each possible action entailing particular outcomes, we are able to determine for each of the several actions their **expected utility** – it is a sum of all of utilities of outcomes correlated with given action multiplied by probabilities with which these outcomes can be realized by performing that action. For example, if given action entails outcome utility of which is 1 with probability of 0,6 and outcome of utility of 0 with probability of 0,4, then we can determine the expected utility of this action as (0,6 * 1 + 0,4 * 0 =) 0,6. With the measure in hand we can define rational choice as the choice maximizing expected utility.

Let us illustrate all this with an example. Consider situation represented on the Figure 1.

		The World	
		W	X
P	A	3	2
	B	4	1

Figure 1.

The rows in Figure 1 represent actions P can perform, the columns – possible states of affairs which can exist in the world. Numbers placed in cells at the intersections of rows and columns represent utilities particular outcomes (dependent both on P's actions and on states of the world) possess for P. For example, situation resulting from P performing action B in state of the world X has utility of 1. Let us assume for a moment that P knows the probabilities both of world being in state W and it being in state X. Let us assume, moreover, that these probabilities are equal to 0,6 and 0,4, respectively. In this situation, we are able to calculate expected utilities of actions A and B – these are (0,6 * 3) + (0,4 * 2) = 2,6 for A and (0,6 * 4) + (0,4 * 1) = 2,8 for B. Therefore, P is required by criterion of expected utility to rationally choose action B. (If P does not know probabilities of particular states of the world being the case, situation is more complicated. In such a case, P can choose from one of available principles of choice. They can, e.g., determine for each of their action the worst possible outcome it can entail, and then choose the action the worst outcome of which is relatively the best.)

Before we can proceed, it is necessary to point out that, according to Gauthier, conditions imposed on preferences by decision theory are not the only requirements that should be satisfied. For the purposes of moral theory, Gauthier introduces additional conditions given individual's preferences have to satisfy in order for it to be possible to identify maximization of their measure with rationality. Generally speaking, these preferences have to be **considered**. Gauthier makes a couple of points to clarify. First of all, he rejects the view – often espoused by economists – that the only way to learn given individual's preferences is to observe the choices they actually make.[3] Besides the behavioural aspect of our preferences, Gauthier distinguishes the attitudinal aspect. Rationality requires congruity of these two aspects: preferences manifested in our choices have to agree with preferences expressed in our attitudes. The choices we make and the attitudes we declare have to confirm each other. If utility is to be the quantity maximization of which is identified with rationality, it has to be the measure both of preferences possessing of which we show in our actions and of preferences possessing of which we declare in our words (Gauthier, 1986, pp. 27–28).[4] In the second place, in order for given person's preferences to constitute the appropriate ground for evaluating their choices, the person has to have adequate **experience** concerning outcomes which are possible results of their actions. At last, preferences have to be **firm**, and not tentative – i.e., they should not be prone to change in the wake of examination of possible consequences of person's actions (Gauthier, 1986, pp. 30–32). What is important is that these conditions concern only the manner in which individual orders states of affairs according to their being more or less preferred, and not the states of affairs themselves.

Indeed, no one of the conditions on rational preferences Gauthier discusses pertain to "content" of these preferences, to what is preferred to what. These conditions pertain only to manner in which individual "has" their preferences, and to relations between these preferences. Conception adopted by Gauthier gives rationality only an instrumental role – the only task of reason is to find the most effective means to achieving given ends, it has (and can have) nothing to say about evaluating the ends themselves.

Utility as a measure of considered preference provides a norm or standard according to which the choices made are to be evaluated – an action is rational only in so far as it maximizes utility. Also, utility is identified in the economic conception of rationality with a measure of value for given individual. Value is thought to be subjective and relative. It is subjective, because – being the measure of preference – it depends on preference relations, a thus on the individual's affectations and attitudes towards particular states of affairs. Value is

[3] Baier (1988, pp. 27–29) rejects Gauthier's critique of economists' view of preference on grounds of it's not being sufficiently motivated.

[4] See: Baier (1988, pp. 30–34) for a critical discussion of the distinction between behavioural and attitudinal aspects of rationality.

relative, because it differs for different individuals – the fact that state of affairs A is higher in individual X's preference ordering does not automatically mean that it will be the case also for individual Y.[5]

STRATEGIC RATIONALITY

So far, we have discussed choices which are made in parametric contexts, i.e., in situations such that the individual makes their choice in immutable circumstances, the choice being the only variable. We should now say something about properties possessed by strategic choice, i.e., choice such that the individual is aware of presence of other choosing individuals, and of complications brought forth by their presence (Gauthier, 1986, p. 21).[6] Game theory is the discipline devoted to studying strategic rationality. A good point of departure for discussing some basic game-theoretical concepts is provided by recalling the situation represented in the Figure 1. It was an example of parametric choice in which individual P had to choose between one of two actions. The situation was complicated by the fact that the world around P could be in one of two states. Depending on whether P knew the probabilities correlated with states of the world or not, the choice was made in conditions of risk or uncertainty, respectively. We bring back this example now because choices of this kind are sometimes called games against nature. Figure 1 represents the situation as if the world was one of the players participating in the game, also choosing between two actions. It is a player of a very special kind, however, because it has no intention of obtaining any particular outcome in the game, and it is not a creature with full knowledge concerning other players' possible moves and preferences (and with capacity to use this knowledge).[7]

Let us now substitute a second player, Q, for the world in Figure 1, so we can see how situation presents itself in strategic contexts. Let us complicate the situation further by providing Q's utilities[8] (conventionally, utilities of "rows" are represented as first numbers in the pairs of numbers in the cells at the intersections of rows and columns).

[5] Compare somewhat more extensive discussion in Gauthier, 1986, pp. 46–59.

[6] Gauthier borrows the distinction between parametric and strategic rationality from Jon Elster.

[7] For a discussion of assumptions concerning knowledge available to players, see: Luce and Raiffa, 1957, pp. 49–50. Gauthier (1986, ch. 3) subjectifies these assumptions, because he speaks of players' predictions concerning other players.

[8] Doing so, we exclude from present consideration issues pertaining to zero-sum games (characteristic feature of which is that the players have strictly opposite preference orderings with respect to possible outcomes, so they can be represented using only utilities of one of the players). We do so because these issues have quite limited importance for Gauthier's project, and the present paper does not purport to be a systematic discussion of game theory (even a rudimentary one). For a discussion of zero-sum games, see: Luce and Raiffa, 1957, ch. 4.

		Q	
		W	X
P	A	3, 3	2, 1
	B	4, 2	1, 4

Figure 2.

What choices should P and Q rationally make? Let us consider how the situation looks from P's point of view: let us say that P, tempted by the perspective of gaining the highest utility, considers choosing the action B. But P, knowing that Q knows that P can be prone to choose B on this basis, predicts that Q may choose their action X. Therefore, P chooses A. However, Q – predicting P's reasoning – can choose W. P, knowing this, may return to considering choosing B, etc., etc. On the other hand, P could choose the action A on the basis that doing so, they can assure themselves utility of at least 2. Q, predicting this, would choose their action W. But P, knowing this, is once more tempted to choose B, and the whole circle begins again. Situation, when looked at from perspective of Q, presents itself analogously. And what of possibility that both, P and Q, choose actions that assure them the highest minimal utilities (A for P, W for Q)? Then we can expect the outcome giving each of the participants the utility of 3. But why P, anticipating Q to choose W, should refrain from choosing action B that gives them even higher utility? And so on.

So far, we have talked about actions as if each of them was an outcome of a separate choice. Strictly speaking, however, this is not the case. What is the proper object of choice is a **strategy**, i.e. probability distribution on possible actions (with each action being assigned the probability of at least zero, and all the assigned probabilities summing up to one). P and Q were choosing in the last paragraph among only **pure strategies**, i.e., strategies that assign probability of one to one of the possible actions. We have seen that there is at this level of analysis no way to unambiguously determine the pair of strategies P and Q should rationally adopt. Even the most promising pair (A, W) was not able to deliver: strategies constituting this pair were not – to use game-theoretical vocabulary – **in equilibrium**, i.e. they were not the best responses to each other: at least one of the players could benefit by unilaterally changing their strategy. In order to solve the problem of rational choice it is necessary to introduce the notion of **mixed strategies**, i.e., strategies such that at least two possible actions are assigned probabilities higher than zero. Consider following set of mixed strategies: let us assume that both P and Q adopt strategies assigning to each of their available actions the probability of ½ [we can refer to these strategies as (½A, ½B) and (½W, ½X), respectively]. Let us now see, how the situation looks

form P's point of view. If Q was to choose W, then P's expected utility is that of [(½ * 3) + (½ * 4) =] 3,5. On the other hand, if Q chooses X, P can expect utility of [(½ * 2) + (½ * 1) =] 1,5. P increases by adopting this mixed strategy the utility minima they can gain in each of the case of Q choosing one of their pure strategies (3,5 compared to 3 in the case of Q choosing W and 1,5 compared to 1 in the case of Q's X). Furthermore, P by adopting this strategy makes it impossible for Q to use knowledge of P's chosen strategy for their own gain. It is so, because when P adopts this strategy, Q can expect the utility of 2,5 irrespective of what pure strategy they adopt. The situation presents itself analogously from Q's perspective.

We can move quite comfortably from issues discussed so far to Gauthier's project itself by considering one of the best well-known and inspiring ideas from the field of game theory, namely **the prisoner's dilemma**. Let us imagine that X and Y commit some serious crime. They get caught and put in the cells isolated from each other. The prosecutor does not have enough evidence to get them behind bars for this serious crime, but they have just enough to get them sentenced for some minor offence. They present individually to each of the prisoners following offer: "if you confess to committing this serious crime, and your partner does not, you will get a year in prison, and they – 10 years. If they confess and you do not – well, you can tell yourself what is going to happen. If neither of you confesses, I will prosecute you for this less serious offence and both of you will get 2 years each. If you both confess, each of you goes to prison for 5 years." The situation is represented on Figure 3: strategies "to confess" and "not to confess" are signified by letters A and B, respectively, and numbers signifying years in prison are placed in the cells at the intersections of rows and columns[9] (of course, the goal of players in this particular game is to get maximally reduced number).

		Y	
		A	B
X	A	5, 5	1, 10
	B	10, 1	2, 2

Figure 3.

[9] Perhaps speaking of years in prison instead of utilities bears the certain risk of misunderstanding, it does, however, make easier to present and understand the problem itself, thus we chose this method of presentation. We can assume that utilities are inversely proportional or that utility chart can be obtained by adding a minus sign in front of each number.

What are we able to discern from Figure 3? Let us look at the situation from X's perspective. We can see that the strategy "to confess" guarantees that X will get possibly the lowest number of years in the case that game turns out badly for them. What is even more important is that this strategy gives better results than the strategy "not to confess" in every instance.[10] This circumstance constitutes good enough reason for X to choose this strategy. The situation is identical from Y's point of view, therefore we can expect the outcome which gives each of the players 5 years in prison. When we consider this outcome, we will see that strategies leading up to it are in equilibrium, i.e. each of them individually is the best (utility maximizing or, in this case, years-in-prison minimizing) response to the adversary's strategy. We see however that other outcome is possible that gives each of the participants only 2 years in prison. The outcome (5, 5) is **suboptimal** in the sense that the situation of each player could be better.[11] Prisoners could have got the outcome (2, 2), had neither of them confessed.

What the prisoner's dilemma[12] illustrates is a possibility of incompatibility of two desirable rationality properties: outcome which is in equilibrium is not necessarily optimal. In other words, actions which are individually rational may not be collectively rational. This conclusion is of great importance for the field of political and social inquiry: individuals unconstrained in their pursuit of their own ends, of their own interests, are able not only to worsen situation of other persons but may even bring about situations worse from their own standpoint than they could otherwise be. How to reconcile individual interests and the common good, how to obtain outcomes both in equilibrium and optimal, is one of the fundamental questions of moral, political, and social philosophy, and is one of the chief issues discussed by Gauthier in his *Morals by Agreement*.

INDIVIDUAL RIGHTS

In the type of situations which we have discussed so far the players choose their respective strategies and accept outcome which is the consequence of their acting on these strategies. Strategy choice was prior logically to outcome. There is, however, another way of solving the problem of rational choice: the individuals could choose their strategies **in order to** get some particular outcome. They could **agree** on what possible outcome they want to bring about, and then

[10] Speaking in game-theoretical terms, we can say that strategy "to confess" strongly **dominates** strategy "not to confess."

[11] Speaking more strictly, its suboptimality (or Pareto-suboptimality) consists of the availability of outcome bettering situations of at least one of the players without worsening the situation of any of the remaining players. The outcome (2, 2) is **optimal** because we are not able to find an outcome that would be superior to it (Straffin, 1993, p. 68).

[12] For a more extensive discussion of the prisoner's dilemma, see: Luce and Raiffa, 1957, pp. 94–102.

choose strategies which are necessary to realize that chosen outcome. For example, X and Y (see Figure 3) could agree on trying to obtain the outcome (2, 2) and on each of them choosing the strategy "not to confess." The outcome to be realized, taken as utility (or some equivalent, e.g., in money) distribution among the players, can be the object of a **social contract**, or – more specifically – of **rational bargain**. Instead of acting on the strategies chosen individually, the players can work out a joint strategy that assigns each of them appropriate actions. The idea of such a contract provides framework for our subsequent considerations. Unfortunately, given thematic constraints of this paper, this is the last thing we have to say about the contract itself.[13] Instead, we want to move to the topic of individual rights.

However, before we can discuss this issue, we must say something about the context in which individual rights appear in Gauthier's overall theory. This context can be presented by citing a story given by Gauthier (1986, pp. 190–192). Gauthier asks the reader to imagine a society consisting of masters and slaves. Perpetuation of slavery entails great financial expenses for the masters and is the cause of great suffering of the slaves. The masters realize that the situation for both sides is worse than it could be had slaves agreed to serve the masters voluntarily. Masters' situation would be better, because they could save the means they now use to assure the slaves' obedience and spend it in some other way. Slaves' situation would be better due to lack of physical violence and better conditions made possible by the savings mentioned. The masters decide therefore to abolish slavery, to enter into (rationally acceptable) agreement with the slaves, and to found the society on the principle of voluntary service provided by (former) slaves. Question: are they correct in predicting former slaves to keep their part of the bargain and to serve the former masters voluntarily?

According to Gauthier, they are not. The presence of coercion which ensured slaves' obedience was the only reason such a contract could have appeared to be rationally acceptable – given the slaves' actual situation, their situation would indeed better. But in the circumstances where the coercion is lacking, there is no reason at all to keep a bargain such that the results of past coercion are solidified.[14] The issue illustrated by this story is one of the so-called initial bargaining position – the situation which designates what the individuals "bring" to the bargaining table and what they are ensured to "get" after agreement is made. The social contract determines the distribution only of that part of utility (or its equivalent, e.g., in money) which results from social cooperation,

[13] See: Hardin, 1988 for a critical discussion of Gauthier's resolution of the problem of bargaining.
[14] Buchanan (1988, pp. 84–85) criticizes Gauthier's rejection of rationality of keeping this bargain. According to Buchanan, there is always a possibility of returning to the state of affairs which obtained before the agreement was made, so it is rational to keep this agreement (See similar point in Harman, 1988, p. 11).

therefore initial bargaining position does affect the individuals' positions in the society brought into existence by the contract. Gauthier argues that it would be irrational for individuals to cultivate in themselves a disposition[15] to keep the agreement which cements the results of past usage of coercion, because it would make other individuals more eager to use coercion and subsequently enter into bargain (1986, p. 195).[16] That is why Gauthier dismisses the Hobbesian state of nature (non-cooperative outcome in his terminology) as the appropriate starting point for bargaining, and endeavours to find such a point. This leads us to the idea of the so-called **Lockean proviso**.

John Locke proposed in his *Second Treatise of Government* (1988) a theory of individual property, its origin and justification. According to him, a just property title to an object which was previously unowned results from an act of acquisition which satisfies three conditions: (1) it involved "mixing one's labor" with the object (ownership of one's person, body and powers is assumed beforehand); (2) the acquired object was not spoiled, i.e. it was used (or at least the proprietor had the intention to use it); (3) there is "enough and as good" other goods left for others (Locke, 1988, §§ 27, 31) – it is the third of these conditions that was called the Lockean proviso. One of the issues pertaining to the proviso is the question of exactly how to understand it. Robert Nozick pointed out the serious problems resulting when we try to understand the proviso literally. According to Nozick, the role of the proviso is to assure that nobody's situation be worsened as a result of the acquisition done, therefore the proviso itself should be interpreted to be the prohibition of such a worsening (Nozick, 1974, pp. 173–176, 178–179). Gauthier follows Nozick, but he makes further modifications – in his view, the absolute prohibition of worsening another's situation is too severe because conditions can occur in which the only way to avoid worsening another's situation is to worsen one's own, and this cannot be required from rational persons taking no interest in one another's interest. The proviso constitutes in Gauthier's theory **the prohibition of bettering one's situation through interaction that worsens other's situation** (1986, pp. 203, 205).[17] We should add also that whether given person X's situation has been worsened or bettered by another person Y is to be ascertained by comparison of X's actual situation with the situation characterized by Y's absence (Gauthier, 1986, 204).

[15] It is worth noting that Gauthier speaks here of rationality as the feature of **dispositions** to choose, not of choice themselves. This reinterpretation of the notion of rationality is of much importance in Gauthier's considerations pertaining to issue of rationality of keeping one's agreement. We will return to this matter later.

[16] We will also return to this issue.

[17] Fishkin (pp. 46–54) argues that the Lockean proviso fails to fix an acceptable starting point for bargaining, because it doesn't adequately grasp the nature of coercion: it fails to recognize some blatant examples of coercion as such.

Not only does the proviso have in Gauthier's theory a different formulation, but it also occupies different place than it does in the theories by Locke and Nozick. In the last two theories, the proviso constitutes a special constraint imposed "from the outside" on rights of the specific kind, namely property rights, which possess a separate and prior justification. In Gauthier's theory the function of the proviso is to facilitate the introduction of both property **and** "personal" rights into the state of nature of otherwise Hobbesian character. The foundation of the specific sets of rights is laid down according to the following scheme: consider an action X done by a person A. Ask whether A violates the proviso by doing X. If they do not, ask whether some other person B violates the proviso if they intervene in A's performing X. If B does so, conclude that A has the right to perform X. Gauthier uses the proviso in this way to introduce rights to exclusive control over one's body, to gain from one's labour, and the right of ownership of external objects individual ownership of which is beneficial for everyone (Gauthier, 1986, pp. 208–217, 227).

THE LINK BETWEEN ECONOMIC RATIONALITY AND INDIVIDUAL RIGHTS

In order to appreciate the link in Gauthier's theory between the adopted conception of rationality and the acceptance of individual rights we must turn now to Gauthier's discussion of the issue of rationality of keeping the bargain one previously rationally entered into. Gauthier consider reasons individual interested in maximizing their own interest has to base their actions on previously agreed to joint strategy (where they know that other participants will keep their part of the agreement), instead of choosing their actions without regard for joint strategies (or rather with enough regard to profitably exploit the fact of others' compliance). We will not discuss this issue extensively, only the details which have a direct significance for topic that interests us. Gauthier distinguishes two dispositions to maximize utility which can be adopted by individual. **Straightforward maximization** is the first one. It is the disposition to choose the strategy which is the best (i.e., utility-maximizing) response to strategies chosen by others. Let us look once again at Figure 3. If X and Y agreed on joint strategy resulting in (2, 2), and X expects Y to follow then strategy "not to confess," then – supposing X to be a straightforward maximizer – X will choose the strategy "to confess," because this strategy guarantees the highest utility return (the lowest number of years in prison). The second disposition is called **constrained maximization**. Roughly speaking, it is the (conditional) disposition to enter into profitable cooperation. A constrained maximizer is ready to base their actions on agreed to joint strategy, provided that utility they can expect in the case of everyone following this strategy is higher than utility expected from

everyone acting on their individual strategies (Gauthier, 1986, p. 167 ff). For example, suppose that X and Y agreed to strive for outcome (2, 2). If X is a constrained maximizer, they will choose "not to confess." Gauthier argues on behalf of rationality of constrained maximization and examines its conditions. We will not discuss this problem in any detail, but the main idea is as follows: in some (reasonably probable) conditions it is more rational to be a constrained rather than a straightforward maximizer, because for a constrained maximizer some possibilities for gainful social cooperation are open which do not present themselves to straightforward maximizers (Gauthier, 1986, pp. 170–177).[18]

Before we go further, one issue is worth noting. The reader will notice that we have so far talked of rationality of individuals, actions, strategies, and outcomes. This manner of speaking is uncontroversial in the field of rational choice theory. However, Gauthier introduces a reinterpretation of the notion of rationality: rationality in his conception is a quality that can be predicated of **dispositions** to choose. Whether given disposition is rational or not is ascertained by considering if it would be chosen and adopted by an individual making a parametric choice (Gauthier, 1986, p. 170 ff).

The concept which properly links Gauthier's conception of rationality to his justification of individual rights is to be found by examining another set of dispositions, which we have now to introduce. This set is comprised of **broad** and **narrow compliance**. They both pertain to readiness to cooperate and keep one's part of the made bargain. Person who is broadly compliant is disposed to enter into cooperation if it is expected to bring out **any** benefit compared to the situation of no cooperation at all. A narrowly compliant person has higher demands: for them to want to enter into cooperation, the outcome of universal acting on joint strategy agreed upon must satisfy (or at least it must not be too far removed from satisfying) the conditions of optimality and impartiality imposed by principles of rational bargaining and by Gauthier's interpretation of the Lockean proviso (Gauthier, 1986, pp. 178–179). Gauthier presents two arguments on behalf of rationality of the narrow compliance. First, he points out that person who is broadly compliant would incentivise other individuals to exploit this person mercilessly, and then to strike a bargain which gives them minimal gain. The second argument is based on the assumption of the "equal rationality" of individuals. According to Gauthier, the assumption that individuals are equally rational means that for a disposition to be rational to adopt by anybody, it has to be rational to adopt by everybody. The broad compliance does not satisfy this condition: supposing one person to be broadly compliant, it would be rational for other persons to adopt the disposition of narrow compliance, so they can hammer out for themselves better provisions in the

[18] See a critical discussion of Gauthier's arguments in McClennen, 1988.

agreement they all enter into[19]. The situation is similar in the case of a disposition that is even more strict than narrow compliance with respect to conditions of entering into cooperation. In order for it to be rational for some person to adopt such – let us say – "more-than-narrow" compliance, it would have to be rational for some other persons to adopt broad (or "more-than-broad," perhaps) compliance. Only narrow compliance is the disposition which can be adopted rationally by everyone participating in the social game (Gauthier, 1986, pp. 226–227).

Narrow compliance is the link connecting rationality to individual rights. It is the disposition which it is rational for individual utility-maximizers to adopt. Simultaneously, due to the fact that narrowly compliant (and therefore rational in utility-maximizing sense) persons stipulate results of cooperation to satisfy certain conditions of impartiality, narrow compliance incorporates disposition to accept the Lockean proviso and individual rights engendered by it as the starting conditions of any social contract. Rationality turns out for Gauthier to be a source of rights, but it serves this function in a manner different than e.g., in Robert Nozick's theory of a more "Kantian" flavour. Rationality is not a feature which inherently commands some kind of respect for the being possessing it – respect, only adequate expression of which consists in honouring strict side-constraints imposed on actions of which this being is an object. As opposed to this "direct" approach, rights have in Gauthier's theory a "transcendental" meaning – they possess practical necessity because they are conditions of the social contract which possess this necessity on the basis of being expression of persons' practical rationality. They are natural in a sense because they are logically prior to society or state. In another sense, however, the rights are not natural: persons do not have them on the basis of any inherent features. Rather, individuals have rights only as potential cooperator or parts to an agreement.

BIBLIOGRAPHY

Baier, K. (1988). Rationality, Value, and Preference. *Social Philosophy and Policy*, 5 (2), pp. 17–45. https://doi.org/10.1017/S0265052500000042

Buchanan, J. M. (1988). The Gauthier Enterprise. *Social Philosophy and Policy*, 5, (2), pp. 75–94. https://doi.org/10.1017/S0265052500000078

Fishkin, J. S. (1988). Bargaining, Justice, and Justification: Towards Reconstruction. *Social Philosophy and Policy*, 5 (2), pp. 46–64. https://doi.org/10.1017/S0265052500000054

Gauthier, D. (1986). *Morals by Agreement*. Oxford: Clarendon Press.

Hardin, R. (1988). Bargaining for Justice. *Social Philosophy and Policy*, 5 (2), pp. 65–74. https://doi.org/10.1017/S0265052500000066

[19] See: Harman, 1998, p. 6 and further for a critical discussion of Gauthier's use of assumption of equal rationality.

Harman, G. (1988). Rationality in Agreement. A Commentary on Gauthier's 'Morals by Agreement'. *Social Philosophy and Policy*, 5 (2), pp. 1–16. https://doi.org/10.1017/S0265052500000030

Locke, J. (1988). *The Second Treatise of Government*. In: Locke, J., *Two Treatises of Government*. Edited by P. Laslett. Cambridge: Cambridge University Press, pp. 141–263.

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions. Introduction and Critical Survey*. New York: John Wiley and Sons, Inc.

McClennen, E. F. (1988). Constrained Maximization and Resolute Choice. *Social Philosophy and Policy*, 5 (2), pp. 95–118. https://doi.org/10.1017/S026505250000008X

Morris, C. W. (1988). The Relation between Self-Interest and Justice in Contractarian Ethics. *Social Philosophy and Policy*, 5 (2), pp. 119–153. https://doi.org/10.1017/S0265052500000091

Nozick, R. (1974). *Anarchy, State, and Utopia*. Cambridge: Basic Books.

Straffin, P. (1993). *Game Theory and Strategy*. Washington: The Mathematical Association of America, 1993.

Thomas, L. (1988). Rationality and Affectivity: The Metaphysics of the Moral Self. *Social Philosophy and Policy*, 5 (2), pp. 154–172. https://doi.org/10.1017/S0265052500000108

Vallentyne, P. (1991). *Contractarianism and the assumption of mutual unconcern*. In: Vallentyne P. (ed.), *Contractarianism and Rational Choice. Essays on David Gauthier's 'Morals by Agreement'*. New York: Cambridge University Press, 1991, pp. 71–75.