An Axiomatic Approach to Justice as Fairness

Takashi Suzuki*

July 31, 2017

Abstract

Justice as Fairness of J. Rawls (1971 and 2001) will be reconsidered from the formal point of view. We reformulate the theory in an axiomatic manner and revise the original position. The revised original position will enable us to deduce the two principles without reference to the primary goods or maximin principle. Therefore, the critiques of Arrow (1973), Harsanyi (1975) and Sen (1980) do not apply to our theory. In the course of revision, we will obtain a new concept of general and basic (human) rights which is distinct from that of the natural rights. Moreover, once Justice as Fairness is formulated axiomatically, we can give a precise meaning to reflexive equilibrium which was not given by Rawls and show that two principles of justice are in reflexive equilibrium but the libetarian principle of Nozick (1974) is not.

Key Words: Justice as Fairness, Axiomatic Approach, Basic Human Rights, Difference Principle, Utilitarian Principle, Libertarian Principle.

1 Introduction

There is a general consensus among moral and political philosophers that J. Rawls laid a theoretical framework (Justice as Fairness) for the philosophical area of social justice in his *A Theory of Justice* (hereafter *Theory*). This was similar to the contribution of Gödel (1931), Neumann-Morgenstern (1944), Nash (1950), Arrow (1963) and Arrow-Debreu (1954) in Mathematical Logic, Game Theory, Social Choice Theory and Market Equilibrium Theory, respectively. These theories share a common structure, that is to

*Department of Economics, MeijiGakuinn University

say some "models (devices of representation)"[1] are established at the formal level of theories, and theoretical results derived from them are "interpreted at" or "applied to" the metalevel: namely, the real world. When this common structure is appropriately recognized, our understanding of Justice as Fairness is more transparent, so most of the criticisms of Rawls can be avoided.

More specifically, Arrow (1973), Harsanyi (1975), MacIntyre (1984), Sandel (1998), and Sen (1980), among others, criticized the basic and intermediate concepts of *Theory*: Rawls's original position, the primary goods and the maximin principle. They did not criticize the principles themselves, but Nozick (1974) did. Indeed, he proposed an alternative (libertarian) principle which we can not accept as a principle of justice. The purpose of the present paper is to show that the two principles of justice can be rescued from these objections.

After presenting general ideas and motivations in Section 2, we revise the original position and defend it against MacIntyre's and Sandel's criticisms in Section 3. Our revision is based upon the idea that people taking the original position now hold ethically "thicker" moral characteristics than in Rawls's original formulation[2]. Our original position will enable us to deduce the two principles without reference to the primary goods or maximin principle. Therefore, the critiques of Arrow, Harsanyi, and Sen do not apply to our theory. Moreover, from this revised interpretation of the original position, we obtain in Section 3 a new concept of the (general) rights to liberties that is distinct from (traditional) concepts of natural rights (see Hart (1955)), and may include human rights. In Section 4, we prove that the difference principle will be selected over the utilitarian principle. In Section 5, we propose a precise definition of reflexive equilibrium and show that the original position with two axioms and two principles is in reflexive equilibrium, but that with the libertarian principles is not. Section 6 concludes.

[1]The original position in Justice as Fairness (see below), a formal system (e.g., Peano arithmetic) in the proof of Gödel's theorem, a (normal form) game model in Neumann–Morgenstern–Nash theories, a society model in Arrow theory and a market model in Arrow–Debreu theory.

[2]Recall that Rawls called his theory "thin".

> Since these assumptions [assumptions about the player motives in the original position] must not jeopardize the prior place of the concept of right, the theory of the good used in arguing for the principles of justice is restricted to the bare essentials. This account of the good I call thin theory: its purpose is to secure the premises about primary goods required to arrive at the principles of justice (*Theory*, p.396).

2 Two Axioms and the Two Principles of Justice

Below, "society" usually refers to actual (our own) society, but it sometimes means the totality of people in the original position. The meaning will be specified in each context. If necessary, the former is called the actual society.

When one reads *Theory* carefully, one recognizes that it has fundamental postulates, which we refer to as axioms. The first is:

Axiom 1. A society is a cooperative venture for mutual advantage (*Theory*, p. 4).

A view of (actual) societies represented by Axiom 1 is the basis of Justice as Fairness. We note that it expresses the idea of mutual advantage. This is nothing but *reciprocity*, and we will see that reciprocity plays a key role in our theory as a whole[3]. The second axiom is less obvious:

Axiom 2. No one deserves greater natural capacity, nor merits a more favorable starting place in society. The distribution of natural talents should be regarded as a common asset (*ibid.*, pp. 101–102).

We notice that Axiom 2 also represents the reciprocity in a strong sense (natural talents as a common asset) and plays a dominant role to deduce the difference principle in our proof (Section 4). It is important to keep in mind that these axioms are postulated at the metalevel, which means that this is the axioms for ourselves (including Rawls). We accept both axioms as *our* truth, and they require no further justification (hence, *axioms*). Different axioms would yield different theories. However, axioms may be supported (not proved) or rejected according to reflexive equilibrium (see Theorems 4 and 5 in Section 4).

The goal of Justice as Fairness is, of course, described by the two principles of justice.

The First Principle. Each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others (*ibid.*, p. 60).

The Second Principle. Social and economic inequalities are to be arranged so that they are both (a) attached to positions and offices

[3]Although we cannot discuss this in sufficient depth in the present paper, this is also the case for Rawls's original arguments in *Theory*. He stressed reciprocity as a foundation of Justice as Fairness in his 1971 paper "Justice as Reciprocity" in *Collected Papers*, S. Freeman, ed., Harvard University Press, 1999.

> open to all, and (b) to the greatest benefit of the least advantaged[4] (*ibid.*, p. 83).

The condition (b) of the second principle is the celebrated difference principle which also expresses the reciprocity.

> . . . [T]he difference principle expresses a conception of reciprocity. It is a principle of mutual benefit (*ibid.* p.102).

A (successful) theory of justice is nothing but a whole body of arguments from which the two principles from the two axioms may be deduced in the most continuous and smooth manner possible. Rawls's fundamental idea on achieving this is well known. First, he sets the original position whereby free and rational persons (moral agents) face a set of alternative principles. They choose a (set of) principle(s) behind a veil of ignorance in order to maximize the (index of) primary goods, which include basic rights, liberties, wealth and self-respect. Rawls concludes that they choose the two principles rather than the utilitarian principle under the assumption that their decision follows the maximin criterion.

The analogy between Rawls's logic and that of metamathematics is very impressive.

> In any proof of Gödel's theorem or Church's theorem, two logics [languages] are concerned. One serves as the "logic of ordinary discourse" in which the proof is carried out, and the other is a formal logic L, about which the theorem is proved (Rosser (1939, p.53)).

A proposition of metamathematics, such as Gödel's theorem, is a formal result that is proved in the formal system L by means of ordinary language, and the "meaning" of the proposition interpreted by ordinary language is a statement of the incompleteness theorem, which says "there exists an undecidable proposition in L if L is ω-consistent[5]." For Justice as Fairness, the

[4]The order of statements (a) and (b) is reversed to that of *Theory*. Our arrangement seems to be more convenient on account of the (lexicographic) order of the principles.

[5]The idea of proof (an application of the liar paradox) is simple enough and well known. For expository simplicity, we outline the proof which is due to Gödel himself for a system L that proves only true propositions (this condition is called "correctness," which is stronger than "ω consistency"), and the metamathematical concepts and statements expressed in ordinary language are placed within quotation marks. Gödel assigns a (prime) number to each symbol of L: "1" to $\mathbf{0}$, "3" to f (successor function), "5" to $\sim$ (not), and so on. Then, using the recursive functions, metamathematical concepts such as "formulae," "propositions," "proofs," and "provable formulae" are all represented as numbers in L.

original position corresponds to the formal system L of metamathematics[6]. Rawls proves a proposition (in L) that people of the original position (not ourselves) will select two principles as the best (most desirable) ones. We interpret this proposition to mean "the two principles are just." Justice as Fairness is a meta-ethic and the two principles are proved as meta-ethical theorems (Theorems 3 and 4 in Section 4). In the rest of this section, we explain our own questions about Justice as Fairness. Answering these questions motivates our revision of the original position.

First, we wonder whether the concept of primary goods is legitimate, because they include very different categories of objects. On the one hand, they include rights and liberties. Strictly speaking, the term should mean kinds of human "relationships" rather than "properties." We discuss this point further in the next section. On the other hand, they include wealth and income, which are definitely objects of economic theory. Moreover, they include self-respect. It is difficult (in particular for economists) to understand any "goods" that encompass such a vast variety of items. When

---

There exist many countable formulae for a natural number x, $\phi_1(x)$, $\phi_2(x)\ldots$ in L. The crucial step of the proof is that in L, we can construct the formula $NP(x)$, which means "$\phi_x(x)$ is not provable." From $NP(x)$, we obtain the "undecidable proposition (sentence)" as follows. Because $NP(x)$ also appears in the above list, we have for some k, $NP(x) = \phi_k(x)$. Then, the formula $G \equiv \phi_k(k)$ means "$\phi_k(k)$ is not provable," or equivalently, G expresses the meaning "G is not provable" (the existence of G is a formal result obtained in L). We can show that G is the desired "undecidable proposition." Indeed, suppose that "G is provable in L." Because "L is correct," "G is true." Hence, "G is not provable," which is a contradiction. Next, suppose that "$\sim G$ (not G) is provable." Then, "$\sim G$ is true," which implies that "G is provable." Then "G and $\sim G$ are both provable in L," which is a contradiction. Therefore, "G and $\sim G$ are both unprovable (G is undecidable)." Q.E.D.

[6]Rawls himself pointed out the similarity between metamathematics and moral philosophy. He stated:

> Note, for example, the extraordinary deepening of our understanding of the meaning and justification of statements in logic and mathematics made possible by developments since Frege and Cantor. A knowledge of the fundamental structure of logic and set theory and their relation to mathematics has transformed the philosophy of these subjects in a way that conceptual analysis and linguistic investigations never could. One has only observe the effect of the division of theories into those which are decidable and complete, undecidable yet complete, and neither complete nor decidable. The problem of meaning and truth in logical system illustrating these concepts. Once the substantive content of moral conceptions is better understood, a similar transformation may occur. It is possible that convincing answers to questions of the meaning and justification of moral judgements can be found in no other ways (*Theory*, pp.51-2).

one mentions "obtaining" or "allocating" rights, liberties, and so on, we suppose that the term is used in a metaphorical sense at best. We claim that the concept of primary goods is at least conceptually ambiguous, although not necessarily wrong.

Recall Rawls's emphasis that the first principle is more fundamental than the second one.

> These principles [the two principles] are to be arranged in a serial order, with the first principle prior to the second. This ordering means that a departure from the institutions of equal liberty required by the first principle cannot be justified by, or compensated for, greater social and economic advantages (*ibid.* p. 61).

The first principle regulates justice at the level of rights and liberties, the second principle at the level of economy and welfare. Rawls claims that the distinction between the two levels is absolute. Now the second principle (difference principle) would be considered superior to the utilitarian or libertarian principles. What about the first principle? The alternatives presented for the participants of the original position are essentially two packets of principles, one is the combination of the first and the second principles, the other is that of the first principle and the principle of the average utility (with a certain social minimum, *ibid.*, p.124). In other words, the first principle overlaps in both alternatives, or it is not "chosen" by itself in Rawls's argument. Thus, why can we not include the first principle in the description of the original position, as they have already agreed with it? Is this condition so stringent that it makes our original position too difficult to accept as a device of representation for Justice as Fairness?

We notice that the primary goods were necessary for Rawls to allow people in the original position to "select" the first principle in particular, because behind a veil of ignorance they have no objectives that are available from a standard type of choice problem, such as utility or profit in microeconomics. Then, something is needed as a theoretical "measure" or "criterion" for their decisions that would correspond to utility or profit. As we have pointed out, the "choice" of the first principle has almost no substantial meaning in Rawls's proof. The original position of Rawls seems to be "too thin," and evidently the "rational decision theory" is overkill here. This seems to be the reason why primary goods must bear such an artificial character as above.

Next, we ask the meaning of "an equal right to the most extensive basic liberty" stated in the first principle. Any rights in general appear in reality,

and their content becomes clear when they have been written into various laws or constitutions. In Justice as Fairness, constitutions and laws are settled at the second and third stages of the four-stage sequence, respectively (*Theory*, pp. 195–201). In the original position, there exist no such laws or rules. If rights existed there, they would be rights in a very general and abstract sense. Nowadays (apart from the 17–18th-century advocates of natural rights), no one but Dworkin (1977) would believe in such an abstract right. Must we assume that people in the original position are Dworkinian jurisprudential philosophers?

In the next section, we define an original position in which people are assumed to have already accepted the two axioms and the first principle. Note that the assumption of *people's* acceptance of the first principle does not mean for *us* to take it in advance (recall that *we* accepted the two *axioms*). We simply establish an ethically "thicker" original position and observe its consequences. The justification of the first principle is finally confirmed by reflexive equilibrium (Theorem 4). The crucial character of Rawls's arguments has not been lost at all. Indeed, Rawls himself indicated the possibility of an original position that included some ethical content. He even suggested the possibility that people in the original position accepted the second (difference) principle.

> Occasionally, we have touched upon some possible ethical variations of the original situation. ... [T]hey may be said to accept a principle of reciprocity requiring that distributive arrangements always lie on the upward sloping portion of the contribution curve. ... There is no a priori reason for thinking that these variations must be less convincing, or the moral constraint they express less widely shared. Moreover, we have seen that the possibilities just mentioned appear to confirm the difference principle, lending further support to it (*ibid.* p. 585).

The basic idea of Rawls's statements above and our own seem to be the same. The point is that we can incorporate reciprocity into the original position. We do so by postulating that the two axioms and the first principles are accepted. As we have seen above, these propositions include reciprocity. In our original position, reciprocity as the basis of Justice as Fairness is manifest. A precious by-product is that we can make clear the meaning of "an equal right to the most extensive basic liberty" stated in the first principle.

3 The Original Position and Basic Rights

We should start from Axiom 1, which states that a (well-ordered) society is a cooperative venture for mutual advantage. Rawls wrote:

> Yet one basic characteristic of human beings is that no one person can do everything that he might do; nor a fortiori can he do everything that any other person can do. The potentialities of each individual are greater those he can hope to realize, and they fall far short of the powers among men generally. Thus everyone must select which of his abilities and possible interests he wishes to encourage; he must plan their training and exercise, and schedule their pursuit in an orderly way. Different persons with similar or complementary capacities may cooperate, so to speak, in realizing their common or matching nature (*ibid.* p. 523).

The purpose of a society is to advance the plans and hopes of each of its members. What else could be more natural and rational for a society and its members than mutual cooperation among them? Reciprocity is fundamentally based upon the rationality of human beings rather than so-called higher-order morality such as benevolence or altruism. We assume that people in the original position have accepted Axioms 1 and 2. It should be emphasized that this is an assumption *of the original position* (the formal system) which is different from the assumption that these axioms are true *for us* at the metalevel (reality).

Rawls distinguishes the reasonable from the rational.

> As applied to the simplest case, namely to persons engaged in cooperation and situated as equals in relevant respects (or symmetrically, for short), reasonable persons are ready to propose, or to acknowledge when proposed by others, the principles needed to specify what can be seen by all as fair terms of cooperation. Reasonable persons also understand that they are to honor these principles, even at the expense of their own interests as circumstances may require, provided others likewise may be expected to honor them (*Restatement*, pp. 6–7).

A reasonable person is rational, but not vice versa. We assume that the people in the original position are reasonable persons. This assumption is stronger than the corresponding assumption of Rawls that people are simply supposed to be rational. Our assumption, however, is consistent with the

two axioms that are already assumed to hold in the original position, and it does not require higher-order morality. Moreover, the assumption of the veil of ignorance also has to be maintained:

> ... [T]he parties are situated behind a veil of ignorance. They do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations (*Theory*, pp. 136–137).

Finally, in this situation, we assume that people have agreed unanimously with the first principle. This completes the description of our original position.

Now let us compare it with that of Rawls. He postulates the two axioms at the metalevel, and invoking the primary goods and the maximin principle he "proves" that rational people behind the veil of ignorance in the original position will choose the two principles as "best." We postulate the two axioms at the metalevel and assume that reasonable people behind the veil of ignorance in the original position hold to the axioms. We also "prove" that they will accept the first principle as "obvious" (the second principle will be verified in the next section). For the reasons described in the previous section, we claim that the "proof" is self-evident[7]. We simply remark here that for a reasonable person, Axiom 1 and the first principle are closely related. It seems difficult for a person to reject the first principle while accepting Axiom 1, because the former is a(n almost) logical consequence of the latter. In other words, Axiom 1 is the reason for the first principle. Hence, it seems unreasonable for a person to accept only one of them[8].

We now turn to the interpretation of "an equal right to the most extensive basic liberty." What does this "right" mean? We know that all members of society hope to achieve their life plans successfully, and that society will honor every success, taking each one as a contribution. Therefore, we give the next definition.

Definition 1. The "right" stated in the first principle is a *membership license* authorized by society.

This right as a membership license entitles and qualifies people to pursue their plans freely if (and only if) they are compatible with those of others.

[7] If required, we could provide Rawls's argument.

[8] Here you could ask: "If reasonable persons should accept both Axiom 1 and the first principle, why do *I* not do it, because *I am* reasonable and have accepted Axiom 1?" If this question occurs to you, our original position has been successfully formulated. Obviously *your* confidence in the first principle came from your consideration of *this* original position, and it is part of the reflexive equilibrium in Theorem 4.

By "authorized by society" we mean mutual agreements and respect among people. We claim that this is what the first principle states. Rawls wrote:

> ... [T]hey [i.e., people in the original position] regard themselves as self-authenticating sources of valid claims. That is, they regard themselves as being entitled to make claims on their institutions so as to advance their conceptions of the good (provided these conceptions fall within the range permitted by the public conception of justice) (*Restatement*, p. 23).

Note that this definition is possible because the first principle is agreed by people, not selected from alternatives. In our original position, rights and the (first) principle become effective at the same time. This "right" is meaningless for any individual who is isolated from society. In other words, there are no "proper and inherent" rights in such a situation—there exist no "natural rights" in this original position. Below, we elaborate this point further to enhance our own understanding of Justice as Fairness.

A precise and rigorous definition of natural rights has been given by H.L.A. Hart (*ibid.*). According to Hart, the natural right means (in the absence of certain special conditions that are consistent with the right being an equal right) any adult human being capable of choice (a) has the right to forbearance on the part of all others from the use of coercion or restraint against him, save to hinder coercion or restraint, and (b) is at liberty to do (or is under no obligation to abstain from) any action that is not one coercing or restricting or designed to injure other persons (*ibid.*, p. 175), and it is characterized as follows.

> (1) This right is one which all men have if they are capable of choice; they have it *qua* men and not only if they are members of some society or stand in some special relation to each other. (2) This right is not created or conferred by men's voluntary action (*ibid.*, p. 175).

Probably the content of natural rights (a) and (b) are not exactly the same as those of the (equal) right in the first principle. We suppose at least that Rawls would not reject Hart's natural right in his intention of the first principle, and show that in Justice as Fairness no natural right characterized by conditions (1) and (2) exists; hence, our right is not the natural right. First, we prove the next metatheorem.

Theorem 1. If the right defined as a (membership) license is the natural right, then the first principle is vacuous.

Proof. Suppose that our right is the natural right. Then by condition (2) it would not exist in any stage (of the four-stage sequence) after the original position; hence, it must exist in the original position. Obviously it cannot be derived from any of the propositions (the two axioms and the first principle) postulated there, so we must *assume* its existence. However, the assumption that the natural right exists in the original position implies that the first principle holds (recall that the natural right means the right to do *any action* which is not one coercing or restricting or designed to injure other persons). In other words, *we* must assume *at the metalevel* that the first principle is true. (This is completely different from our actual assumption that *people in the original position* accept the first principle as true.) Now it is clear that if our right were the natural right, then the first principle is vacuous (empty or trivial)[9]. Q.E.D.

Theorem 1 shows that original positions that include natural rights are "too thick." However, it does not necessarily imply nonexistence of natural rights. Furthermore, the task of deriving the second principle is still left to the original position. Hence, one could say that this weak (too thick) original position could make sense if the second principle is proven to be true. Therefore, we continue to investigate the meaning of the natural rights in Justice as Fairness.

There seems to be an obvious analogy between a natural right endowed with a moral agent and a characteristic such as a utility function endowed with a consumer in microeconomics. However, this analogy is rather superficial and restrictive. This will be apparent if one realizes that markets with only one consumer make theoretical sense[10], while original positions with only one moral agent do not[11]. In such an original position, he/she could choose whatever he/she wanted, and the first, utilitarian, and libertarian principles would be reduced to the same principle, which means that there would be no problems of justice. Moreover, the concept of rights would have

[9] Any mathematical theorem is provable if you assume that it is true (proof: obvious from the assumption), but of course this "theorem" makes no sense.

[10] Indeed, such a market model is the subject of optimal growth theory.

[11] The reason is that theoretical concepts in microeconomic theory are constituted by the relationships between economic agents and commodities. For instance, the utility functions specify the agent to whom utility belongs, and are defined on the consumption set (the domain of the utility function), which is a subset of the commodity space. Markets with only consumers (and no commodities) or one with only commodities (and no consumers) would be nonsense! On the other hand, the concepts in Justice as Fairness are constituted only by the relationships among moral agents. A single agent cannot form "relationships."

no meaning. In a society of only one person, what kind of "rights" could he/she have? This is exactly our point when we asserted in the previous section that a right is a relationship, not a property.

For moral agents to be well defined, however, their theoretical description must be complete, or it must be complete even if the agents are isolated from society and placed in a situation where rights play no role. Therefore, natural rights endowed with moral agents are meaningless as *their* moral characteristics. It is now clear that there is no room for the natural rights in Justice as Fairness. Indeed, the foregoing discussion has proved:

Theorem 2. There exist no natural rights in Justice as Fairness.

We have assumed that people in the original position are reasonable and agree with the first principle; that is, they know that everyone, including themselves, in society accepts the first principle. Probably the only moral characteristics that can be meaningfully assumed are these kinds of intellectual properties and knowledge. In fact, we can imagine a person with some knowledge and intelligence living alone, but not a person living alone with any meaningful rights. That is to say, for Justice as Fairness, the concept of rights must be constructed and explained within the theory, not postulated and given from outside of the theory.

Note that in both the original positions of Rawls and the present paper, society is considered a voluntary association. It goes without saying that an actual society is not an association, and it is not voluntary. One does not choose a society in which one lives, one is just born there. One lives there for a lifetime unless unusual situations such as emigration or exile occur. We must keep in mind that the original position is not a description of reality but a device of representation. Although this is an obvious theoretical fact, it is worth emphasizing.

The "communitarian" criticism aimed at Rawls is well known. The essential point of the criticism seems to be that people in the original position are too abstract, and Justice as Fairness fails to capture the moral personality of each individual. Consequently, the justice it describes is "prior to" or "independent of" value and desert, which are, according to them, the indispensable and essential points for justice[12].

[12]MacIntyre:

> If Rawls were to argue that anyone *behind the veil of ignorance*, who knew neither whether and how his needs would be met nor what his entitlements would be, ought rationally to prefer a principle which respects needs to one which respects entitlements, invoking perhaps principles of rational decision theory to do so, the immediate answer must be not only that *we* are *never*

It should be now evident to us that these criticisms arise from confusion of reality and a device of representation. Rawls himself has claimed this again and again[13]. Once Justice as Fairness is formulated axiomatically, then its theoretical character and the status of the original position become perfectly clear. We can now confirm that Rawls's claim is valid.

4 Deduction of the Difference Principle

In this section, we show that the second principle would be preferred to the utilitarian or the libertarian principles by people in the original position. The utilitarian principle is stated as a principle of restricted utility with social minimum (*Theory*, p. 124).

Principle of Restricted Utility. The basic social institution should be organized so that average utility is maximized under the constraint that a certain social minimum is maintained.

Recall that in our original position, the first principle (an equal right to the most extensive basic liberty) has been already agreed. Hence, the social institution stated in the utilitarian principle has to be consistent with the first principle. We also assume that condition (a) in the second principle (positions and offices are open to all) is understood under the utilitarian principle. The next celebrated (meta)theorem, proved by Rawls, was a fundamental result of *Theory* and *Restatement*. Indeed, the following proof basically follows sections 34—39 of *Restatement*. Fortunately, it is much simpler than that of Rawls, thanks to our "thicker" original position.

Theorem 3. (Rawls) People in the original position choose the second principle over the principle of restricted utility.

---

behind such a veil of ignorance, but also that this leaves unimpugned Nozick's premise about inalienable rights (*ibid.*, p. 249, italic by MacIntyre).

Sandel:

> But as our discussion of agency and reflection suggests, we are neither as transparent to ourselves nor as opaque to others as Rawls's moral epistemology requires. If our agency is to consist in something more than the exercise in "efficient administration" which Rawls's account implies, we must be capable of a deeper introduction than a "direct self-knowledge" of our immediate wants and desires allows (*ibid.*, p. 172).

[13] *Theory*, p. 12; *Political Liberalism*, pp. 24–27, p. 35, p. 75; *Restatement*, pp. 17–18, p. 30, p. 80, pp. 85–86; and many others.

Proof. First we recall that the participants in the original position accept Axiom 1 and the first principle. This implies that they recognize themselves as free and equal citizens in their society as a fair system of cooperation. Given that they are reasonable persons, their recognition contains the mutual advantage or reciprocity. Obviously the difference principle is more consistent with this conception than the utilitarian principle which merely orders to maximize the sum (average) of their utilities. Moreover, the difference principle expresses the idea that the better endowed (who have a place in the distribution of natural endowments they do not morally deserve) are encouraged to seek still further benefits (they already favored by their fortunate place in the distribution) provided they train their endowments and use them in ways that contribute to the good of all, and in particular to the good of the least endowed (who have a less fortunate place in the distribution, a place they also do not morally deserve). This idea of reciprocity is embodied in Axiom 2 regrading the distribution of native endowments as a common asset which is also already accepted by the citizens. Given these conceptions of reciprocity, the participants would be inconsistent with their axioms if they selected the utilitarian principle over the difference principle. Q.E.D.

Note that the essential point is that we have assumed Axioms 1 and 2 *in the original position*, both of which express reciprocity and the latter in particular in a very strong sense (natural talents regarded as a common asset). On the other hand, as Rawls's original proof shows, the rational rather than the reasonable might be sufficient for the proof of Theorem 3, since it only shows consisitency between Axioms and the utilitarian principle. As we see below, the reasonable is crucial for Theorems 4 and 5, since their proofs require to show *stability* of society (see the next section) which is more demanding than the consistency.

5 Reflexive Equilibrium and Stability of Societies

The axiomatic approach clarifies and enhances the idea of reflective equilibrium. We begin our elaboration by recalling Rawls's own explanations for the notion of reflexive equilibrium:

> In searching for the most favored description [of the original position] we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield

> a significant set of principles. If not, we look for further premises equally reasonable. But if so, and these match our considered convictions of justice, then so far well and good. But presumably there will be discrepancies. In this case we have a choice. We can either modify the account of the initial situation [original position] or we can revise our existing judgements, for even the judgements we can provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the contractual circumstances, at others withdrawing our judgements and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgements duly pruned and adjusted. This state of affairs I refer to as reflexive equilibrium (*Theory*, p.20).

Then we ask: when we can say that our original position "expresses reasonable conditions and yields principles which match our considered judgements duly pruned and adjusted"? It is obvious that original positions should not be self-contradictory, or it should not contain incompatible conditions or assumptions from both logical and moral point of views. For instance, we cannot assume that the participants mainly concern their self-interests and at the same time they are altruistic. But the logical and moral consistency is just a necessary condition for "our considered convictions of justice". It is not sufficient. We also require that "a political conception of justice must generate its own support and the institutions to which it leads must be self-enforcing, at least under reasonably favorable conditions (*Restatement*, p. 125)."

> This means that those who grow up in such a well-ordered society develop ways of thought and judgment, as well as dispositions and sentiments, that lead them to support the political concept for its own sake: its ideals and principles are seen to specify good reasons [for compliance]. Citizens accept existing institutions as just, and usually have no desire either to violate or to renegotiate the terms of social cooperation, given their present and prospective social position (*ibid.*).

When these conditions are fulfilled, we can say that the society described by the original position is *stable*. Given these observations, we obtain the next definition.

Definition 2. An original position and the principles derived from it are

> said to be in *reflexive equilibrium* if and only if they are reasonably considered to be consistent and stable.

Sometimes we will also say that an original position is *supported as* a reflexive equilibrium.

In order to maintain the stability, society must have reasons to counterbalance the desire to violate the current terms of cooperation. Rawls mentioned three reasons.

> First, there is the effect of the educational role of a public political conception. Thus we suppose all members of society to view themselves as free and equal citizens who, in and through the basic structure of their institutions, are engaged in mutually advantageous social cooperation [cf. Axiom 1 and the first principle]. Given this conception of themselves, they think that the principle of distribution that applies to that structure should contain an appropriate idea of reciprocity (*ibid.*, pp. 125–126).

Obviously, the difference principle contains such an idea, so everyone has this reason to accept it. Notice that Rawls had to assume the people's conception that they are "as free and equal citizens engaged in mutually advantageous social cooperation". In our original position, this assumption is fulfilled by Axiom 1 and the first principle. The second reason is:

> We also suppose that in addition to the reason which all have, the more advantaged have a second reason, because they are mindful of the deeper idea of reciprocity implicit in the difference principle when it is applied to the basic structure: namely, that it tends to ensure that the three contingencies [their social class of origin, their native endowments, the good or ill fortune] are taken advantage of only in ways that are to everyone's advantage [cf. Axiom 2] (*ibid.*, p. 126).

Again this proposition which Rawls had to "suppose" is guaranteed in our original position by Axiom 2. The third reason is that the difference principle encourages mutual trust and cooperative virtues, because it will make people understand that the three contingencies tend to be dealt with only in ways that advance the general good, and that constant shifts in relative bargaining positions will not be exploited for ends motivated by self- or group interest (*ibid.*, p. 126) which is obviously satisfied by Axiom 2 in our original position.

Note that the assumption of the reasonable has been implicitly used in these arguments, which show that the original position with the two axioms

and two principles would make a well-ordered society stable. Therefore, we have also proved the next metatheorem because of Rawls.

Theorem 4. (Rawls) The original position with Axioms 1 and 2 and the two principles is supported as a reflexive equilibrium.

We call the equilibrium stated in Theorem 4 *the Rawls equilibrium.*

Next, we turn to the comparison with the libertarian principle. It is stated as follows (*Anarchy, State and Utopia, p. 150*).

Entitlement Principle. (1) A person who acquires a holding in accordance with the principle of justice in acquisition is entitled to that holding. (2) A person who acquires a holding in accordance with the principle of justice in transfer, from someone else entitled to the holdings, is entitled to the holding. (3) No one is entitled to a holding except by (repeated) application of (1) and (2).

This inductive definition is completed by the next axiom.

Axiom 3. A person is entitled to his/her own natural assets.

The entitlement to natural assets (their native endowments, talents, etc.) in Axiom 3 is absolute and could be called a natural right. The entitlement principle without Axiom 3 (the first step of induction) is meaningless. Therefore, we have no results of the Theorem 3 type to compare the difference and the entitlement principles in an original position, because Axioms 2 and 3 are obviously inconsistent; hence, we cannot assume them simultaneously.

Nevertheless, we can prove that the entitlement principle (and Axiom 3) will be rejected as a reflexive equilibrium, a result corresponding to Theorem 4. In so doing, we have to suspend Axiom 2 at the metalevel, because it would not be fair to use it to reject the contradicting proposition (in the formal system). However, we keep Axiom 1 at the metalevel, and people in the original position are still assumed to be reasonable.

Theorem 5. Justice as Fairness with Axiom 1 rejects the original position with Axiom 3 and the libertarian principle as a reflexive equilibrium if people in the original position are reasonable.

Proof. Suppose that Axioms 1 and 3 (instead of 2) are accepted, and the first and entitlement principles (instead of the second principle) are agreed in an original position. We show that this original position is not in equilibrium. Indeed, as the proof of Theorem 4 shows, *reasonable* people who have accepted Axiom 1 and the first principle will

"think that the principle of distribution that applies to that structure should contain an appropriate idea of reciprocity." (*Restatement*, p.126. See the first reason of Rawls for resisting the violation of the current terms of cooperation.) This contradicts the entitlement principle, which obviously has no reciprocity. Hence, we can have no confidence in this original position. Note that to conclude that a situation is not a reflexive equilibrium, no definitive negative judgment is necessary. It suffices that positive judgments are uncertain (*Theory*, p .20). Therefore, this situation is not an equilibrium.

This conclusion obviously follows from the inconsistency between Axiom 1 and the first principle, which includes reciprocity, and Axiom 3 with the entitlement principle, which does not. Hence, the only possibility for the libertarian principle to survive seems to be an original position where only Axiom 3 and the entitlement principle are agreed (Axiom 1 is still maintained at the metalevel). Because people are reasonable, it is not necessary to consider original positions where there remains only one of Axiom 1 and the first principle (see a remark made above in our definition of "rights for liberty" in Section 3). Let us consider this possibility.

In this original position, there would exist (natural) rights to natural assets and legitimate holdings, but no rights to the most extensive basic liberty. Reflecting this situation from the viewpoint of Axiom 1 (at the metalevel), we are not certain that "political conceptions of justice in this society can generate their own support and the institutions to which they lead must be self-enforcing." In other words, we cannot be sure that this society can be stably sustained. Hence, again this original position is not an equilibrium. Q.E.D.

In retrospect, from the proofs of Theorems 4 and 5 we recognize the stability of the Rawls equilibrium.

6 Concluding Remarks

In the present paper, we have shown that the original position can be reformulated so that the main results of *Theory* and *Restatement* can be obtained without the primary goods or the maximin criterion. This does not mean that we should discard Rawls's original position or the primary goods. Because the former is probably the "ethically thinnest" original position, it is valuable as a touchstone from which we can estimate the balance and the strength of the theory. When we have doubts regarding the theoretical

settings of our original positions, we can come back to Rawls's setting and reconsider them.

Regarding the latter (primary goods), because it is now apparent that they are inessential for deduction of the results, there would be no further problems in using them as a theoretical tool, with an understanding of their theoretical restrictions and possible problems. This situation is the same as that for utility functions in microeconomics. When they were proposed as cardinal utilities, they were criticized for their economic meanings and the possibility of interpersonal comparison. Then they were replaced by indifference curves or ordinal utilities, and recently by preference relations. Now that it has become apparent that cardinal utilities are inessential for deduction of results in modern economic theory, no one doubts the value of utility functions as a theoretical tool.

References

Arrow, K., 1963, *Social Choice and Individual Values*, 2nd ed., Wiley.

Arrow, K., 1973, "Some Ordinalist–Utilitarian Notes on Rawls's Theory of Justice," *The Journal of Philosophy* **70**, 245–263.

Arrow, K., and G. Debreu, 1954, "Existence of an Equilibrium for a Competitive Economy," *Econometrica* **22**, 265–290.

Dworkin, R., 1977, *Taking Rights Seriously*, Harvard University Press.

Gödel, K., 1931, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," *Monatshefte für Mathematik und Physik* **38**, 173–198.

Harsanyi, J., 1975, "Can the Maximin Principles Serve as a Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review* **64**, 594–606.

Hart, H.L.A., 1955, "Are There Any Natural Rights?" *The Philosophical Review* **64**, 175–191.

MacIntyre, A., 1981, *After Virtue* (2nd edition, 1984), University of Notre Dame Press.

Nash, J.F., 1950, "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences of the U.S.A.* **36**, 48–49.

Neumann, J.v., and O. Morgenstern, 1944, *Theory of Games and Economic Behavior*, Princeton University Press.

Nozick, R., 1974, *Anarchy, State and Utopia*, Basic Books, New York.

Rawls, J., 1971, *A Theory of Justice*, Harvard University Press; revised edition, 1999 (quoted as *Theory* in the text).

Rawls, J., 1993, *Political Liberalism*, Columbia University Press.

Rawls, J, 2001, *Justice as Fairness: A Restatement*, Harvard University Press (quoted as *Restatement* in the text).

Rosser, B., 1939, "An Informal Exposition of Proofs of Gödel's Theorems and Church's Theorem," *The Journal of Symbolic Logic* **4** no. 2, 53–60.

Sandel, M., 1998, *Liberalism and the Limits of Justice 2nd edition*, Cambridge University Press.

Sen, A., 1980, "Equality of What?" *The Tanner Lectures on Human Values* vol. I, Cambridge University Press.