An Axiomatic Approach to Justice as Fairness

Takashi Suzuki[1]

[1]MeijiGakuin University

November 26, 2018

Abstract

Justice as Fairness by J. Rawls (1971 and 2001) will be reconsidered in a formal way. We reformulate his theory in an axiomatic manner and revise his original position. We propose a new concept of fundamental rights as membership of the original position, and on this basis, we justify Rawls's thesis of the priority of the first principle over the second. This would address the criticisms of Arrow (1973) and Hart (1973). This revision of the original position will enable us to deduce the two principles without reference to the primary goods or maximin principles. Therefore, the criticisms of Harsanyi (1975) and Sen (1980) do not apply to our theory. We do this by providing a rigorous definition of reflective equilibrium, which Rawls did not do, and we show that the two principles of justice are in reflective equilibrium but the libertarian principle of Nozick (1974) is not.

An Axiomatic Approach to Justice as Fairness

Abstract: *Justice as Fairness* by J. Rawls will be reconsidered in a formal way. We reformulate his theory in an axiomatic manner and revise his original position. We propose a new concept of fundamental rights as membership of the original position, and on this basis, we justify Rawls's thesis of the priority of the first principle over the second. This revision of the original position will enable us to deduce the two principles without reference to the primary goods or maximin principles. Therefore, the well- known criticisms of Harsanyi and Sen do not apply to our theory. We do this by providing a rigorous definition of reflective equilibrium, which Rawls did not do, and we show that the two principles of justice are in reflective equilibrium but the libertarian principles of Nozick is not.

Key Words: Justice as Fairness, 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 *A Theory of Justice* (hereafter *Theory*). This was similar to the contributions by Gödel (1931) in Mathematical Logic, Neumann and Morgenstern (1944) and Nash (1950) in Game Theory, Arrow (1963) in Social Choice Theory, and Arrow and Debreu (1954) in Market Equilibrium Theory, respectively. These theories share a common structure, that is, 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. Furthermore, we will see in what follows that most of the criticisms against *Theory* can be avoided. More specifically, Arrow (1973), Harsanyi (1975), Hart (1973), MacIntyre (1984), Sandel (1998), and Sen (1980), among others, criticized the basic and intermediate concepts of *Theory*, that is, Rawls's original position, the primary goods and maximin principles. The purpose of the present paper is to show that the two principles of justice can be rescued from these objections. We note that those authors cited did not criticize the principles themselves, unlike Nozick (1974) who proposed an alternative (libertarian) principle. We show in Section 6 that the libertarian principle is unacceptable as a principle of justice.

The paper is organized as follows. Section 2 presents general ideas and motivations; we propose two fundamental axioms and illustrate the logical structure of Justice as Fairness using an analogy from metamathematics. We also question Rawls's idea of fundamental rights as a primary good if it is not just a simple "metaphor" rather than a theoretical "concept." In Section 3, we revise the original position. We defend this revised original position against the criticisms of MacIntyre and Sandel. The basic idea is that the

fundamental rights stated in the first principle of justice are no longer considered to be a "good" but as "membership" by participants in the original position. The main results in this section can be summarized as two metatheorems. Theorem 1 states that this concept of rights as membership is not one of natural rights; Theorem 2 proves that there are no natural rights in Justice as Fairness. In Section 4, we prove that the difference principle will be selected over the utilitarian principle without any reference to the primary goods or maximin principles (Theorem 3). Therefore, the critiques of Arrow, Harsanyi, and Sen do not apply to our theory. In Section 5, we clarify the idea of reflective equilibrium, which was not perfectly clear in *Theory*, and provide a rigorous definition. Then, we prove the second fundamental metatheorem (Theorem 4), which validates that the original position with two axioms and two principles is in reflective equilibrium. In Section 6, we prove that the libertarian principle is not in equilibrium. Section 7 concludes.

2. Two Axioms and the Two Principles of Justice

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

From a careful reading of *Theory*, 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. (Rawls 1971a: 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.[2]

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: 101–102)

We note that Axiom 2 also represents reciprocity in a strong sense (natural talents as a common asset) and plays a dominant role in the deduction of the difference principle in our proof (Theorem 3 in Section 4). It is important to keep in mind that these axioms are postulated at the metalevel, which means that these are 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 reflective equilibrium (see Theorems 4 and 5 in Section 5).

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: 60)

The Second Principle. Social and economic inequalities are to be arranged so that they are both (a) attached to positions and offices open to all, and (b) to the greatest benefit of the least advantaged.[3] (ibid: 83)

Condition (b) of the second principle is the celebrated difference principle, which also expresses reciprocity.

> [T]he difference principle expresses a conception of reciprocity. It is a principle of mutual benefit. (ibid: 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.[4]

> In any proof of Gödel's Theorems 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: 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 that "there exists an undecidable proposition in *L* if *L* is ω-consistent." For Justice as Fairness, the original position corresponds to the formal system *L* of metamathematics. 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 metaethic[5] and the two principles are proved as metaethical theorems (Theorems 3 and 4). In the remainder of this section, we explain various criticisms raised so far and our own questions about Justice as Fairness. Our desire to address these questions motivates our revision of the original position.

First, we wonder that the concept of primary goods includes very different categories of "objects". On the one hand, they include wealth and income, which are definitely objects of economic theory. Moreover, they include (the social bases of) self-respect. On the other hand, primary goods include rights and liberties. Strictly speaking, the term should mean kinds of human "relationship" rather than kinds of "entity."[6] We suppose that when mentioning "obtaining" or "allocating" rights, liberties, and so on, the term is used in a metaphorical sense at best. We discuss this point further in the next section.

In contrast, Sen criticized the primary goods approach because "it takes little note of the diversity of human beings" (1980: 215). According to Sen, the appropriate measure for inequalities among citizens is not the primary goods but the "basic capabilities" of human beings. We claim that the concept of primary goods is at least conceptually ambiguous, and this ambiguity would have invited more serious criticisms, as shown below.

As is well known, Rawls concluded that people in the original position would choose the two principles rather than the utilitarian principle if their decision followed the maximin criterion. We note that primary goods were necessary for Rawls to allow them to "select" the two principles in that way, 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 metric or objective for their decisions that would correspond to utility or profit. We suspect that the "rational decision theory" is overkill here. This may be the reason why primary goods must bear such an artificial character, as described above. The use of the maximin criterion was also severely criticized by Harsanyi (1975) as placing almost irrational weight on the risks of the worst outcomes.

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 (1971a: 195–201). In the original position, there are 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?

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, by greater social and economic advantages. (ibid: 61)

Arrow (1973) and Hart (1973) questioned this thesis of the priority of the first principle over the second, or over the difference principle. Their objection also seems to arise from the

ambiguous concept of primary goods: if rights and liberties are items of primary goods, why must we give absolute priority to these specific items? Does this mean that people in the original position are assumed to have a lexicographic preference over them? If so, what is the justification for this assumption, because we can obviously point out several situations where other items of primary goods become more socially significant than rights and liberties? It seems difficult to answer those questions in a completely satisfactory manner as long as we maintain the idea of primary goods and include rights and liberties as their top priority items.[7]

The first principle regulates justice at the level of rights and liberties; the second at the level of economy and welfare. Rawls claims that the distinction between the two levels is absolute. The second principle (difference principle) would be considered superior to the utilitarian or libertarian principles. What about the first principle? In *Theory*, the alternatives presented to 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 average utility (with a certain social minimum, ibid: 124). In other words, the first principle overlaps in both alternatives, or as Rawls argues, it is not "chosen" by itself. Thus, why can we not include the first principle in the description of the original position, as it has already been agreed upon? Is this condition so stringent that it makes our original position too difficult to accept as a device to represent Justice as Fairness?

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 that *we* must accept it in advance (recall that *we* have already accepted the two axioms). We simply establish an ethically "thicker"[8] original position and observe its consequences. The justification of the first principle is finally confirmed by reflective 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 have 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 […]. 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: 585)

The basic idea of Rawls's statements and our own appears 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, 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 clarify the meaning of "an equal right to the most extensive basic liberty" stated in the first principle. We shall see that this revised original position avoids the criticisms of Arrow, Harsanyi, and Hart in a convincing way.

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 than 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: 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? 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. (2001: 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. (1971a: 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 postulate that they will accept the first principle as "obvious" at this stage. The second principle will be verified in Theorem 3, and the two principles are finally confirmed in Theorem 4, which asserts that the original position adopted the two principles is supported as a reflective equilibrium (see Definition 2 in Section 5).

We simply remark here on our supposition that the participants of our original position accepted Axiom 1 and the first principle simultaneously. This assumption can be justified by the fact that for a reasonable person, Axiom 1 and the first principle are closely related. Indeed, we will see in the proofs of Theorems 3 and 4 that the idea of reciprocity commonly included in the axiom and the principle connects them tightly. Hence, it seems difficult for at least reasonable persons to reject the first principle while accepting Axiom 1; that is, it seems natural that reasonable persons accept not only one of them, but always accept both of them.[11] We will also soon see why the priority of the first principle over the second is justified in original positions where Axiom 1 and the first principle are simultaneously held by the participants.

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. 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). (2001: 23)

Note that Definition 1 makes sense 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. We also note that this concept of rights 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 are no "natural rights" in this original position. Below, we elaborate on this point to enhance our own understanding of Justice as Fairness.

We can now support Rawls's thesis of the priority of the first principle against Arrow and Hart as follows. The rights and liberties are no longer items of primary goods. Hence, they are not "goods that are especially significant" among the other goods. Rather, as membership, the right is one of the most basic constituents of liberal societies, and the first principle frames it by determining its contents to the broadest extent; that is, it is the right to *liberties.* Taken together with Axiom 1, the first principle makes the participants recognize themselves as free and equal citizens in their society as a fair system of cooperation. This self-recognition would be fundamental for liberal societies to be well ordered. The distinction of the first and second principles is not a matter of weight given by the participants' preferences for liberty and inequality but that of their respective roles in the basic structure of societies. The priority of the first principle over the second is a consequence of these observations.

A precise and rigorous definition of natural rights is given by Hart (1955). The natural right (in the absence of certain special conditions that are consistent with the right being an equal right) means any adult human 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 (1955: 175). Hart called this right the natural right, because 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: 175)

The contents of the natural right (a) and (b) are distinctively different from those of the (liberal) right in the first principle. It is a comprehensive right that permits any action that does not coerce, restrict, or injure others. We could call this right the *libertarian right* (in a broad sense) and define the libertarian (first) principle.

The Libertarian Principle. Each person is to have an equal right that satisfies conditions (a) and (b).

We shall criticize the libertarian principle as a principle of natural right. The next metatheorem shows that the right stated by the first principle of Justice as Fairness is not the natural right.

Theorem 1. If the right defined as membership (a membership license) is the natural right characterized by conditions (1) and (2), 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 that 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[10] (empty or trivial). QED.

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 natural rights in Justice as Fairness. Indeed, we can prove the next theorem.

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

Proof. Suppose on the contrary that natural rights exist in Justice as Fairness. Then, they must exist in the original position; otherwise by condition (2), they would not exist in any subsequent stages of the four-stage sequence.[11]

Consider an original position in which there exists only one person. In such a society, he/she would be able to do everything he/she wanted; in other words, he/she has a "right" to do everything he/she wants to do. Obviously, this means that the concept of "rights"—whatever they are—loses its meaning. Indeed, right as a form of membership would make no sense; however, according to condition (1), natural rights claim that they maintain their meaning (otherwise they are not natural rights). This is a contradiction. Hence, there exists no such concept as natural rights in Justice as Fairness. QED.

Let us elaborate on the meaning and contents of the proof. There seems to be an obvious analogy between a moral agent endowed with a natural right and a consumer in economic theory endowed with characteristics such as a utility function or an initial endowment.

However, this analogy is rather superficial and restrictive. This is apparent if one realizes that markets with only one consumer make theoretical sense[12], while original positions with only one moral agent do not. 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 make no sense. On the other hand, the concepts in Justice as Fairness are constituted only by the relationships between moral agents. A single agent cannot form "relationships." In the original position with only one participant, 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 no meaning. In a society of only one person, what kind of "rights" could he/she have? This reiterates our point made in the previous section that a right is a relationship, not an entity.[13]

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 granted innately to moral agents are meaningless as *their* moral characteristics. It is now completely clear why there is no room for 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, 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.

One might note that in the original positions of the present paper, the concept of rights as membership might suggest that societies are considered as if they were voluntary association. Needless to say, any actual society is not an association or it is not voluntary. One does not choose a society in which one lives, one is simply 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 the communitarians, the indispensable and essential points for justice.[14] It should now be evident that these criticisms arise from confusion of reality and a device of representation. Rawls himself has repeatedly claimed this.[15] 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.

Because no natural right exists in Justice as Fairness by Theorem 2, neither does the libertarian right of Hart as a natural right exist. Consequently, the libertarian principle for a natural right would not hold as a principle of justice. However, we have not completed our criticism of the libertarian principle when the right stated in it was interpreted as membership of society as in Definition 1. We discuss this possibility in Section 5, where we criticize the (libertarian) entitlement principle of Nozick (1974) and show that the libertarian principle would not be selected in our original position even if its rights were taken as membership.

4. Deduction of the Difference Principle

In this section, we show that the second principle would be preferred to the utilitarian principle by people in the original position. The utilitarian principle is stated as a principle of restricted utility with social minimum (1971a: 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*. Therefore, we call it the Rawls's theorem. The following proof basically follows sections 34–39 of *Restatement*. Fortunately, the proof is much simpler than that of Rawls, thanks to our revised, or "thicker," original position. Further remarks on the theorem will follow the proof.

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

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 mutual advantage or reciprocity. Obviously, the difference principle is more consistent with this conception than the utilitarian principle, which merely demands that they 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 further benefits (they are already favored by their fortunate place in the distribution) provided that 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 regarding 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. QED.

As mentioned, because the *idea* of the proof of Theorem 3 is due to Rawls, we called it Rawls's theorem. However, Rawls himself could not set forth the statement of Theorem 3 as a *theorem*. When he "proved" that statement, invoking Axioms 1 and 2, he just kept them as his own philosophical belief in (his) metalevel. Therefore, the statement simply expressed *his own* view or essentially his *personal* opinion about what would happen in his original position. On the other hand, we have assumed Axioms 1 and 2 *in the original position*; hence, what we have proved is a theorem *about* the original position for which we inferred the decisions of the participants holding the axioms.

Note that Axioms 1 and 2 both express reciprocity, and the latter in particular in a very strong sense (natural talents are regarded as a common asset). As Rawls's original proof shows, the rational rather than the reasonable might be sufficient for the proof of Theorem 3, because it only shows inconsistency between the axioms and the utilitarian principle. As we see below, the reasonable is crucial for Theorems 4, because its proof is required to show the *stability* of society, which is more demanding than verifying the consistency.

5. Reflective Equilibrium and Stability of the Two Principles

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 reflective 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. […] 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 judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the contractual circumstances, at others withdrawing our judgments 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 judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium. (1971a: 20)

Then we ask: when does our original position "express reasonable conditions and yield principles that match our considered judgments, duly pruned and adjusted"? What do "considered judgments" mean? It is obvious that original positions should not be self-contradictory, nor should they contain incompatible conditions or assumptions from both logical and moral points of view. For instance, we cannot assume that the participants are mainly concerned with their own self-interest and at the same time are altruistic. However, the logical and moral consistency is just a necessary condition for "our considered judgments 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" (Rawls 2001: 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: 125)

When these conditions are fulfilled, we can sometimes 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* (or are *supported as*) *a reflective equilibrium* if and only if they are reasonably considered to be consistent and generate their own support.

To generate its own support, or achieve stability, society must have reasons to counterbalance the desire to violate its current terms of cooperation. Rawls mentioned three reasons for stability to be achieved in *his* original position after the two principles of justice are adopted.

> 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. 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: 125–126)

Obviously, these concepts come from Axiom 1 and the first principle. The difference principle also contains such an idea, so everyone has this reason to accept it. Note that Rawls had to assume ("we suppose …") the people's conception that they are "free and equal citizens engaged in mutually advantageous social cooperation." In *our* original position, Axiom 1 and the first principle are explicitly assumed to be accepted by the citizens. The second reason is as follows:

> 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. (ibid: 126)

Again, this proposition that 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).

We now state and prove the next metatheorem, and call it Rawls's second theorem.

Theorem 4. The original position with Axioms 1, 2, and the two principles is supported as a reflective equilibrium.

Proof. The consistency of the original position with Axioms 1, 2, and the two principles is apparent from the proof of Theorem 3, so we skip it. It remains for us to show that it is stable. As already noted by Rawls, Axiom 1 and the first principle provide the conception of citizens that they are free and equal citizens who are engaged in mutually advantageous social cooperation. Hence, they would understand that the difference principle contains such an appropriate idea of reciprocity required for a mutually advantageous social cooperation. Moreover, because they are reasonable and indeed accepting the difference principle, they can trust that their social class of origin, their native endowments, and good or ill fortune are taken advantage of only in ways that are to everyone's advantage, and that constant shifts in relative bargaining positions will not be exploited for ends motivated by self- or group interest. A society where all of those conditions are fulfilled is nothing but what generates its own support. QED.

We call the equilibrium stated in Theorem 4 the *Rawls equilibrium*. The remark after the proof of Theorem 3 also applies to Theorem 4. Rawls could not state the content of Theorem

4 as a theorem, since he did not assume Axioms 1 and 2 in his original position; he must invoke the axioms from "outside" of the original position. Hence, his "proof" is not for the original position but for his own "personal" opinion.

As stated, Theorem 4 requires a stronger assumption of the reasonable rather than the rational needed for Theorem 3, in which we only had to show the necessary condition for the reflective equilibrium, so it was enough to assume that the citizens were *rational*. In Theorem 4, we had also to show stability as well as consistency. To obtain the stronger conclusion of Theorem 4, the merely rational does not seem to suffice.

What the proof of Theorem 4 tells us is the following. For a well-ordered society to be stable rather than consistent, two things are required. One is that the citizens are reasonable. The other is that the truly just principles and institutions that make the citizens capable of revealing their reasonableness effectively are established. This observation would appear to be significant in the next section, which explains why the original position adopting the libertarian principle would not be stable.

6. Instability of Libertarian Principles

We finally turn to examine whether the libertarian principle can be endorsed as a principle of justice. It is stated as follows.

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 holding, is entitled to the holding. (3) No one is entitled to a holding except by (repeated) application of (1) and (2). (Nozick 1974: 150)

This inductive definition is completed by the next axiom.

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

According to Nozick (1974), the entitlement to natural assets (their native endowments, talents, etc.) in Axiom 3 must be taken as absolute so that it could be called a natural right. Of course, by Theorem 2, we must think of it as membership of society. However, the absolute and definitive character of the entitlement to natural assets (talents) in Axiom 3 would make it inconsistent with Axiom 2, which asserts that natural talents are a common asset. Note that the entitlement principle without Axiom 3 as the first step of inductive definition is meaningless. Therefore, we have no results of the Theorem 3 type to compare the difference and entitlement principles in an original position, because we cannot assume them simultaneously. Nevertheless, we can prove that the entitlement principle (and Axiom 3) will be rejected as a reflective equilibrium, a result corresponding to Theorem 4 for the difference principle that we support as the Rawls equilibrium. 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. The original position with Axiom 3 and the entitlement principle are not supported as a reflective equilibrium.

Proof. To examine the possibility of the entitlement principle with Axiom 3 to survive in an original position as a reflective equilibrium, it seems natural to start by considering the original position where Axioms 1 and 3 (instead of 2) are accepted, and the first and entitlement principles (instead of the difference principle) are agreed.

Step 1: We show that this original position is not consistent. Indeed, as the proof of Theorem 3 indicates, *reasonable* people who have accepted Axiom 1 and the first principle will recognize themselves as free and equal citizens in their society, which is a fair system of cooperation, and their recognition makes them consider that any principles of wealth distribution that apply to the basic structure of their society should contain mutual advantage or reciprocity. This contradicts the entitlement principle, which obviously does not contain any such ideas of reciprocity. Therefore, we conclude that this original position is inconsistent.

Step 2: The above conclusion obviously follows from the inconsistency between Axiom 1 with the first principle and Axiom 3 with the entitlement principle. In order for the Entitlement principle with Axiom 3 to survive, the compatible principle which protects the citizens' basic rights should not include the ideas of societies as a fair system of cooperation. We identify the libertarian principle of Section 2 as an alternative principle of the first principle. Hence, we consider an original position where Axiom 1 and the first principle are dropped, and replace them with the libertarian principle, and the entitlement principle with Axiom 3 are agreed (Axiom 1 is still maintained at the metalevel).

Step 3: Let us check whether this original position is stable or not. In this original position, there would exist the rights to natural assets and legitimate holdings, but the participants do not recognize themselves as free and equal citizens who are engaged in mutually advantageous social cooperation. As stated in Step 2, the libertarian principle does not yield such a self-recognition. Moreover, there does not exist Axiom 2 or the difference principle, there are no "effects of the educational role of a public political conception" or they have no reason to trust that the three contingencies are exploited only in ways that are to everyone's advantage, or that shifts in relative bargaining positions will not be exploited for ends motivated by self- or group interest, even if they are assumed to be reasonable. Reflecting on this situation from the viewpoint of Axiom 1 at the metalevel, we are uncertain whether political conceptions of justice in this society can generate their own support, or that the

institutions to which they lead must be self-enforcing, or we can have no confidence in the stability of this original position. Note that to conclude that a situation is not a reflective equilibrium, no definitive negative judgment is necessary. It suffices that positive judgments are uncertain (1971a: 20). Therefore, this situation is not in reflective equilibrium. QED.

Readers should notice that in the proof of Theorem 5, we actually executed the procedure of the reflective equilibrium described by Rawls. We can see that the assumption that the participants are reasonable played contrasting roles both in Theorems 4 and 5. In Theorem 4, we saw that the stability of the Rawls equilibrium was essentially brought about by the assumption of the reasonable. The two axioms and the two principles made it possible to become active. In the original position constructed in Step 2 of Theorem 5, there was no room for the reasonable to work effectively. We can even doubt whether the assumption of the participants' reasonableness is really consistent with the Entitlement principle, or rather that the rational is a consistent assumption for that principle. If the participants were assumed to be rational, the instability of the libertarian principles should have been more apparent in Step 3.

7. Conclusions

In this paper, we have revised the original position of Rawls and reformulated Justice as Fairness in such a way that the main results of *Theory* and *Restatement* can be obtained without any reference to primary goods or the maximin criterion. The essential point is the axiomatic approach in which a distinction between the meta and object levels of recognition is strictly maintained in the course of analysis. Exactly the same axioms or principles play different epistemological roles in philosophical discussions when they are placed at different levels of recognition. Consequently, we can use those axioms or principles *in the original position* to prove metatheorems, although it finally becomes apparent that they are the axioms and the principles of justice *of our own*. In Theorems 3, 4, and 5, the conclusions are not deduced from the parties' deliberations, such as maximizing the social primary goods according to the maximin criterion, but by the logical (philosophical) inference of *ourselves at the metalevel*. The proofs are essentially nothing but explications of philosophical implications of axioms, principles, and other assumptions (rationality, reasonableness, veil of ignorance, and so on) *executed by ourselves*. This is exactly the meaning of the *device of representation* in the axiomatic approach.

Although we have drastically revised Rawls's original position, we do not have to abandon it. Because it 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 own original positions, we can come back to Rawls's setting and reconsider them. Recall that this is part of the procedure of a reflective equilibrium.

The axiomatic approach also clarified some of the key concepts of Justice as Fairness that remained somewhat obscure in Rawls's discussions. First, we have provided a rigorous definition of the fundamental right stated in the first principle. We saw that the concept of right so defined as membership was the basis for the priority of the first principle over the difference principle. This definition was clearly obtained from our critique of the concept of primary goods. However, this does not mean that we should discard the idea of primary goods. We now comprehend 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 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 clear 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. Rawls constructed his theory as a form of game theory with rational players making their decisions under uncertainty. Primary goods and the maximin criterion were needed for this formulation. Although his theory gained a certain clarity and elegance from this formulation, it invited awkward but unnecessary questions such as validity of the maximin criterion, questions concerning the parties' "taking a chance" in the original position, and so on. The serious conceptual problem caused by social primary goods was already pointed out. The aim of the axiomatic approach is to avoid those difficulties.

Next, we rigorously defined the idea of a reflective equilibrium, which had previously been left ambiguous in Justice as Fairness. The elaborated formulation of the reflective equilibrium is especially important for the present paper, since it is the key for the derivation of the difference principle without primary goods or the maximin criterion. We proved that our original position with the two axioms and the two principles had support as a reflective equilibrium (Rawls's equilibrium), and any original positions where the two principles were replaced by the utilitarian or the libertarian principles could no longer be supported. In the course of the proof, we have recognized the crucial role of the reasonable rather than the rational. We need to recall that rationality and efficiency are originally the standards for people's behavior that are appreciated in economic markets rather than in political or social lives. Influences of economic theory in political philosophy seem to have been so overwhelming that the relevance of the rational for Justice as Fairness has been exaggerated. An important lesson of Theorems 4 and 5 is that those economic standards would be insufficient to support well-ordered societies.[16] Instead, those theorems tell us that such societies should be maintained and developed by reasonable people who appreciate reciprocity, and hence endorse the two principles of justice.

Foot Notes

1. Such as the original position in Justice as Fairness (see below), formal systems in the

proof of Gödel's theorem, (normal form) game models in Neumann–Morgenstern–Nash theories, market models in Arrow–Debreu theory, and so on.

2. See also Rawls (1971b).
3. The order of statements (a) and (b) is reversed relative to that of *Theory*. Our arrangement seems to be more convenient on account of the (lexicographic) order of the principles.
4. Rawls himself pointed out the similarity between metamathematics and moral philosophy:

> 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 to 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 logic and mathematics is profoundly altered by the discovery of logical systems 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 judgments can be found in no other ways. (1971a: 51–52)

5. The term "metaethic" is used differently from its traditional use, e.g., Hare (1952).
6. Habermas (1995) also pointed out this problem.
7. Rawls (1993) tried to justify the priority of the first principle in Lecture 8 of this book. His arguments were essentially based upon his elaborations on presupposed moral power of citizens in liberal societies. Unfortunately, his explanations generally do not seem to answer completely the criticisms of Arrow and Hart.
8. 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 (1971a: 396).

9. 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 reflective equilibrium in Theorem 4.
10. 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.

11. Once the parties in the original position have selected the two principles of justice, they move on to a constitutional convention and create their constitution subject to the principles that they have adopted. After that, they move forward to the legislative stage and the veil of ignorance is further lifted. The last stage is the application of laws to individual cases by judges and administrators. The participants are now citizens who obey the established rules of their society (1971a: Chapter IV, section 31).
12. Indeed, such a market model is the subject of optimal growth theory.
13. Sen stated:

    > [T]here is, in fact, "fetishism" in the Rawlsian framework. Rawls takes primary goods as the embodiment of advantage, rather than taking advantage to be a *relationship* between persons and goods. (1980: 216, italics by Sen)

14. MacIntyre:

    > If Rawls were to argue that anyone *behind the veil of ignorance*, who knew neither whether or how his needs would be met nor what his entitlements would be, ought rationally to prefer a principle that respects needs to one that respects entitlements, perhaps invoking principles of rational decision theory to do so, the immediate answer must be not only that *we* are *never* behind such a veil of ignorance, but also that this leaves unimpugned Nozick's premise about inalienable rights. (1984: 249, italics 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. (1998: 172)

15. 1971a: 12; 1993: 24–27, 35, 75; 2001: 17–18, 30, 80, 85–86; and many others.
16. Rawls wrote:

    > [W]e suppose that political and social cooperation would quickly break down if everyone, or even many people, always acted self- or group-interestedly in purely strategic, or game-theoretic fashion. (2001: 125)

References

Arrow, K. (1963). *Social Choice and Individual Values*, 2nd ed., Wiley, New York, USA.

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

Arrow, K., and Debreu, G. (1954). Existence of an Equilibrium for a Competitive Economy,

Econometrica **22**, 265–290.

Dworkin, R. (1977). *Taking Rights Seriously*, Harvard University Press, Cambridge, Massachusetts, USA.

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

Habermas, J. (1995). Reconciliation through the Public Use of Reason: John Rawls's Political Liberalism, *The Journal of Philosophy* XCII, 109-131.

Hare, R. (1952). *The Language of Morals*, Oxford University Press, Oxford, UK.

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.

Hart, H.L.A. (1973). Rawls on Liberty and Its Priority, *University of Chicago Law Review* **40**, 551–555.

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

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.

Nozick, R. (1974). *Anarchy, State and Utopia*, Basic Books, New York, NY, USA.

Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, USA.

Rawls, J. (1971a). *A Theory of Justice*, Harvard University Press; revised edition, 1999, Cambridge, MA, USA.

Rawls, J. (1971b). *Justice as Reciprocity* in *Collected Papers*, S. Freeman, ed., Harvard University Press, Cambridge, MA, USA, 1999, 190-224.

Rawls, J. (1993). *Political Liberalism*, Columbia University Press, New York, NY, USA.

Rawls, J. (2001). *Justice as Fairness: A Restatement*, Harvard University Press, Cambridge,

MA, USA.

Rosser, B. (1939). An Informal Exposition of Proofs of Gödel's Theorems and Church's Theorem, *The Journal of Symbolic Logic* **4**, 53–60.

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

Sen, A. (1980). Equality of What? in *The Tanner Lectures on Human Values*, vol. I., 195-220, Cambridge University Press, Cambridge, UK.